Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Ciudad de México

School of Engineering and Sciences

Analysis and Use of Textual Definitions through a Transformer Neural Network Model and Natural Language Processing

A thesis presented by

Germán Eduardo Baltazar Reyes

Submitted to the
School of Engineering and Sciences
in partial fulfillment of the requirements for the degree of

Ph. D.

in

Engineering Sciences

Ciudad de México, México, December, 2021

Dedication

Acknowledgements

In the first place, I would like to thank my parents Guadalupe Reyes Vadillo and Germán Eduardo Baltazar Robles. They taught me the love of learning and the discipline and constancy required for achieving this gigantic goal. Thanks to you, I'm becoming the man that I am today.

I also want to thank my wife, Martha Angélica Peimbert Oceguera, for accompanying me during this whole process. It was not an easy process, but I grew and persevered thanks to your love and support.

I want to express my deepest gratitude to my advisors, Edgar Omar López Caudana and Pedro Ponce Cruz, for giving me this incredible opportunity, for your patience, and for constantly believing in me to reach this milestone in my life.

I also want to thank David Christopher Balderas Silva and Luis Miguel Ibarra Moyers for their dedication and support in developing this research. All your comments, questions, and feedback allowed me to reach this point. Now I can finally say that I've finished.

I would also like to thank Tecnológico de Monterrey for the tuition support, as well as the opportunity to become a professor during these nine semesters. Finally, but not least, I thank CONACyT for their support on living expenses throughout this whole academic process.

Analysis and Use of Textual Definitions through a Transformer Neural Network Model and Natural Language Processing

by

Germán Eduardo Baltazar Reyes

Abstract

There is currently an information overload problem, where data is excessive, disorganized, and presented statically. These three problems are deeply related to the vocabulary used in each document since the usefulness of a document is directly related to the number of understood vocabulary.

At the same time, there are multiple Machine Learning algorithms and applications that analyze the structure of written information. However, most implementations are focused on the bigger picture of text analysis, which is to understand the structure and use of complete sentences and how to create new documents as long as the originals. This problem directly affects the static presentation of data. For these past reasons, this proposal intends to evaluate the semantical similitude between a complete phrase or sentence and a single keyword, following the structure of a regular dictionary, where a descriptive sentence explains and shares the exact meaning of a single word.

This model uses a GPT-2 Transformer neural network to interpret a descriptive input phrase and generate a new phrase that intends to speak about the same abstract concept, similar to a particular keyword. The validation of the generated text is in charge of a Universal Sentence Encoder network, which was finetuned for properly relating the semantical similitude between the total sum of words of a sentence and its corresponding keyword.

The results demonstrated that the proposal could generate new phrases that resemble the general context of the descriptive input sentence and the ground truth keyword. At the same time, the validation of the generated text was able to assign a higher similarity score between these phrase-word pairs. Nevertheless, this process also showed that it is still needed deeper analysis to ponderate and separate the context of different pairs of textual inputs.

In general, this proposal marks a new area of study for analyzing the abstract relationship of meaning between sentences and particular words and how a series of ordered vocables can be detected as similar to a single term, marking a different direction of text analysis than the one currently proposed and researched in most of the Natural Langauge Processing community.

Acronyms

AI	Artificial Intelligence
BoW	Bag of Words
BERT	Bidirectional Encoder Representations from Transformers
DAN	Deep Average Network
DL	Deep Learning
DNN	Deep Neural Network
EHR	Extended Hyperbolic Representation
xLSA	Extended Latent Semantic Analysis
FFN	Feed Forward Network
GAN	Generative Adversarial Network
GPT-2	Generative Pre-Training Model-2
GloVe	Global Vectors
IBE	Inference to the Best Explanation
LSA	Latent Semantic Analysis
lr	learning rate
LSTM	Long-Short Term Memory
ML	Machine Learning
MLM	Masked Language Model
MSE	Mean Squared Error
NLP	Natural Language Processing
NSP	Next Sentence Prediction
NaN	Not a Number
POS	Parts of Speech
PMI	Pointwise Mutual Information
RNN	Recurrent Neural Network
RC-S	Relation by Contrast-Synonymy
STS	Semantic Textual Similarity

SNLI Stanford Natural Language Inference

SoA State-of-the-art

USE Universal Sentence Encoder

List of Figures

List of Tables

Contents

Chapter 1

Introduction

We currently live in an environment where it is crucial to find, analyze, and use reliable information daily [91]. At the same time, there have been multiple technical advances that allow us to have access to more affluent, complex information sources in different formats and types [9]. One of the most noticeable sources of information has been the Internet, which after the Web 2.0 model, allowed any person to publish, share, and comment on new content. However, the amount of information available to everyone has surpassed the human capacity to analyze and use it completely.

According to [89], Google's search index has increased from approximately 45 million web pages in 2017 to more than 60 million in 2019. Regarding the scientific world, in 2014, Google Scholar indexed almost 160 million documents, while the CrossRef database registered a total of 97 million DOIs. In 2018, Web of Science included 70 million article records in one month, and 33,100 active scholarly peer-reviewed English-language journals published more than 3 million articles [48].

This phenomenon is called information overload, referring to the difficulty a person faces when deciding in the presence of excessive information [96], the overabundance of relevant information that exceeds the human processing capability [68, 91], or the burden of equally excessive unsolicited information [33]. Thus, it can be said that information overload is a problem related to the management of data, information, and knowledge [91].

Information overload is mainly caused by the existence of multiple sources of information, the overabundance of information, difficulty in managing it, the vast amount of irrelevant data, and the scarcity of time on the part of information users to analyze and understand all the received data [44]. All these problems can be reduced to three particular issues:

1. Information is **excessive**. Nobody can analyze every existent document.
2. Information is presented in a **disorganized** way. There is no proper way of knowing in which order information should be evaluated when searching for different documents.
3. Information is **static**. One document of reference could be completely understandable for a person, but at the same time, confusing and considered not useful for another one looking for the same information.

It is then necessary to develop techniques and resources aimed at maximizing information handling efficiency [10], transforming the way it is presented to the user according to their particular needs [11, 22, 38]. It is then necessary that text and data mining techniques modify the way information is given to people [48] to achieve such goals.

One of the main reasons written information resources are not considered helpful for the user is that the reference was not comprehended enough. Some studies have observed that this problem is related to reader vocabulary. If the reader has enough vocabulary, it is easier for a better text comprehension [45]. At the same time, it has been noticed that it is necessary to understand around 98% of the document vocabulary to gain unassisted comprehension of the document [45]. In the case of academic research documents, the 2,000 most common English words constitute 76.1% of the document's vocabulary. The remaining 23.9% is divided into low-frequency words and proper names (13.9%) and academic vocabulary (10%) [101]. Following these studies, improving the reader's comprehension level is needed without new documents explaining those terms. Such unknown words should be described in the same document to prevent more disorganized and excessive information.

Evaluating what constitutes an explanation is necessary before determining how explanations should be developed and presented to the user. An explanation is divided into two subjects: an illocutionary and a perlocutionary act. The first one is denoted as an explaining act that represents the action of uttering words with a particular intention [6]. The second subject is a perlocutionary act that represents the effect of the person understanding of the explaining act [6]. It can be observed that what we usually call an explanation is only an illocutionary act. Thus, one can develop an explanatory sentence, but the final user is the one in charge of deciding if the idea was explained or not.

When dealing with an unknown term, usually the question to answer would be ”what does that word means?”, in which case the answer, or explaining act, would be to define the word. In order to define or explain the meaning of a particular word, different semantic relationships are used to structure the conceptual information of any word [75]. By generating multiple inferences and arbitrary representations of concepts, a person can generate new knowledge and build new concepts as well [75]. The relationships required for doing these tasks are considered as *ad hoc* inferences, which mean that a person does not assign meaning for a concept until he mentally relates the previous conceptions he has about a specific word or term [8].

It is logical to assume that an unknown term remains that way because the person who finds such a concept cannot generate the corresponding relations that allow him to assimilate the concept to a mental inference that makes sense. Also, by maintaining unknown words in the references or documents, the information remains unclear, keeping the information static for the user. Due to these reasons, it is necessary to modify the way concepts of unknown words or terms are shown to the user, using different words and terms that englobes a general concept, facilitating the generation of mental inferences to give sense to the particular unknown term.

1.1 Problem Definition

Information access nowadays presents three main problems. First of all, it is excessive, making it impossible to evaluate all the information available regarding any topic. Second, the way is given to people is disorganized, making it more difficult to analyze which references are helpful and not. Finally, information is static, meaning that while one person could understand one reference, it can also be entirely confusing for others. Regarding the latter, it has been observed that most of the documents or references are considered confusing to the user due to not knowing most of the used vocabulary because the user cannot generate inferences about what such a term means.

At the same time, current Machine Learning algorithms are focused on evaluating the semantical similitude between two long texts or phrases. However, there is no official benchmark or methodology for evaluating the semantical context between a complete sentence and a single word. The lack of such a validation method makes it difficult to evaluate if a given sentence's general context or meaning remains the same when interpreting or explaining a particular concept or idea.

1.2 Justification

Current text generation algorithms are focused on copying the writing style of human beings to create conversational platforms capable of talking and communicating in a given scenario. However, it is still necessary to evaluate how well these algorithms can analyze the general contextual meaning, rather than just selecting one word after another to create a readable document.

The ability to analyze a given text and interpret it into a new understandable phrase or concept may reduce the staticity of textual information, making information on every topic more accessible to every person. Two significant examples could be the educational and medical fields; in the first case, the opportunity to rephrase an unknown term would allow any student to understand new topics better without having to study or relearn past, complete topics. Regarding the medical field, this proposal could interpret medical, specific terms to make them more understandable to a general patient or medic of different specialties or even interpret personal descriptions of a patient when explaining their symptoms during a consult.

Those examples are just a few implementations of how interpreting written information and presenting it to a person makes it easier to understand general knowledge, reducing the difficulties of understanding different topics, making access to information more user-friendly, and helping with eliminating the Information Overload problem.

1.3 Objectives

This research's main, general objective is to create a model that interprets how a word in a written document is explained, based on the semantic relations between the main word and the different phrases that could define or explain the same term. In order to achieve this goal, the following particular objectives are pursued:

- The proposed method can relate a complete sentence with a single word, evaluating its semantical similitude regardless of the size of the original descriptive phrase.
- The proposed model has an accuracy of above 50% for interpreting a descriptive sentence and generating a new phrase similar to the original input and its corresponding keyword.
- Implement a method for getting a sentence embedding considering its semantical and syntactical structure, obtaining similarity scores above 70% in a given dataset.
- The validation of the generated text should correlate higher than 60% when evaluating the semantical similarity between the generated text and the human perspective.

1.4 Hypothesis

First, the algorithm will evaluate how semantic relationships between words determine their corresponding level of similarity. Once the similarities between two words can be measured, the next step will be to evaluate how the union of different terms generates phrases with a similar meaning to a single word. The structure of such phrases and their relation to specific words will generate different word combinations to explain a particular term. Those ways will train an ANN model capable of determining the best phrases for explaining a given word.

This configuration should be capable of producing multiple explanations for an unknown term coming from a written document. Once the explanation is produced, it is still necessary to validate the semantical similitude between the generated text and the original concept being tried to explain. A second Neural Network model is required to evaluate the similitude score between a given text and a single key concept.

Once these two models are combined, it will be possible to create a platform capable of receiving a written phrase or sentence and interpreting it into a new sentence to be more understandable to the user than the original text.

1.5 Methodology

This Ph. D. study is focused first on analyzing the different ways written information is examined and used to modify the way it is presented. Once the general characteristics were comprehended, a new model for generating multiple explanations for any word was implemented to compare their performance against other state-of-the-art methods.

The approach of this research is centered on written text generation based on the semantic and contextual relationship between words to make it easier to understand their meaning. Once the new explanation model is compared, the new model for dynamic explanations generation will be validated to observe if the generated text does not lose context when using new vocabulary to interpret and explain a given phrase.

Chapter 2

Natural Language Processing

The development of algorithms and methods related to the computational analysis of words is directly derived from one of the main applications of Deep Learning (DL): Natural Language Processing (NLP). This chapter will describe the objectives and uses of NLP and its State-of-the-art (SoA) algorithms used to analyze semantic similitude between words.

2.1 Stages of Artificial Intelligence

Artificial Intelligence (AI) refers to the study of making machines perform as well as, or even better, as a human being in particular tasks, being such activities are currently done by humans or that they represent currently unsolved problems that nobody is capable of doing nowadays [84]. During the first stages of AI, the performed tasks were focused on solving problems that required particular features that human beings do not possess [84], like generating complicated calculations, or storing vast amounts of information and retrieving it as fast as possible. However, with the technological advancements, the AI tasks began to focus on assimilating human behavior [84], such as evaluating visual or written perception, analysis of human language, or even optimizing human processes in the financial or industrial field.

As tasks became more complex in the type of actions required to solve a problem, a particular question began to arise: are machines capable of developing intelligence to adapt their behavior and solve a given problem? During the 50s, Alan Turing developed an evaluation method to analyze that. As was further known, the Turing test tried to evaluate if a computer could think on its own. A person had to interact with two entities, A and B, from which one was a person and the other a machine. The subject needed to determine which was the human entity; in the case, he selected the machine, it would be concluded that the latter was capable of thinking [98]. As the problems given to machines became more complex and challenging, a more particular implementation of AI has been developed, as seen in Figure 2.1.

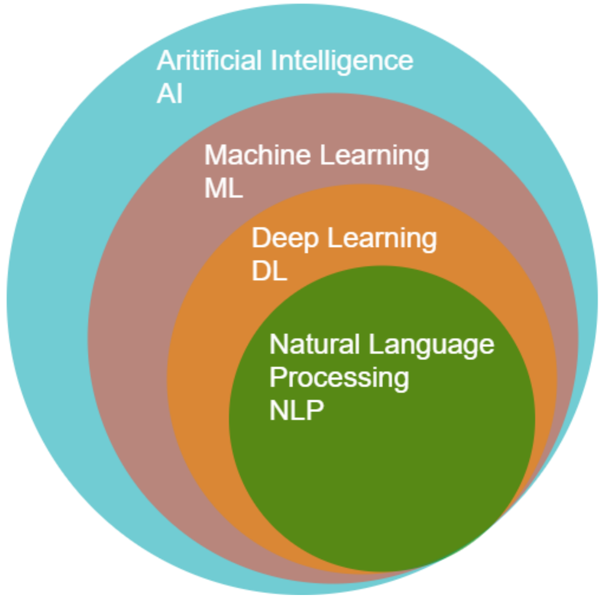


Figure 2.1: Artificial Intelligence composition.

2.1.1 Learning Models in Machine Learning

The concept of Machine Learning comes from the Turing test. It is referred to as the analysis of a compound of data, also known as a dataset, allowing the development of a statistical model that intends to solve a particular problem [18]. Machine Learning (ML) focuses then on training a model with data and answers in order to generate the rules that allow the machine to solve the problem on its own [35]. The training of ML models requires enough relevant data to generate a statistically robust model that is capable of evaluating new data correctly while updating the way it evaluates new information.

Depending on the nature of the dataset, there are four types of learning for the model to represent the rules that could solve the given problem. The first one, and the most common one, is supervised learning. The structure of the dataset for this type consists on labeled examples $\{(x_i, y_i)\}_{i=1}^N$, where x_i represents the input data, and y_i its corresponding label. The main objective of this type of learning is to evaluate an input x_i and output a deduction of which label y_i corresponds to [18]. It is usually used for sequence generation, syntax tree prediction, object detection, or image segmentation [35].

The second type is called unsupervised learning. As the name suggests, it is the opposite of supervised learning. In this case, the dataset only has the input data $\{x_i\}_{i=1}^N$, and its goal is to transform the input into a different vector representation for analyzing data [18]. Typically, unsupervised learning is the first approach for understanding the nature of the dataset, using clustering or dimensionality reduction techniques [35].

The third type of learning is a hybrid between the other two previously mentioned. Self-supervised or semi-supervised learning datasets contain both labeled and unlabeled data for training the model. While its objective is the same as supervised learning, self-supervised learning has more unlabeled data to train [18] to use heuristic models to generate the data labeling on its accord [35].

Finally, the fourth type is known as reinforcement learning. It is the most recent way of learning that has been developed so far. The intention is to create an environment where the machine (denoted as agent) interacts with the dataset to learn policies that takes the feature vector x_i of the environment and outputs an optimal action through a reward-punishment evaluation [18, 35].

Depending on the given task, the learning approach of the ML model will determine an easier and proper way of dealing with the data and solving the problem.

2.1.2 Specialization of Machine Learning

ML generates different data representations to solve the given task. However, as the complexity of the problem increases, a single representation is not enough for treating the given data. It is then necessary to evaluate successive layers of data representations in order to analyze how data behaves, obtaining more complex and detailed representations from the previous ones [35]. This process is called DL.

The implementation of DL has performed considerably well regarding different tasks that were considered achievable only by humans [35]: image classification, speech recognition, machine translation, improved information search and classification, autonomous driving of vehicles, among others. The rest of this chapter will focus on one of these areas, specifically the one that studies how humans use language and how a machine could implement a similar behavior. This area of DL is known as Natural Language Processing.

2.2 Principles of Natural Language Processing

NLP is a research field of ML focused on understanding, interpreting, and implementing human language and its structure to make machines capable of communicating as similar as a human being [51]. The term of Natural Language refers to any language that has been used daily by people, evolving without requiring a proper construction [51]. The analysis of text has multiple research fields, as shown in Table 2.1; each of them contributes to a particular procedure in the communicative process. Even though these categories may sound like easy procedures, it is pertinent to notice that this thought is typically produced because we, as human beings, generate each process unconsciously every time we speak with someone or transmit a message. However, the theory and structure behind each category involve a more profound and complex methodology that involves the management of intangible data. For the rest of this document, we will focus on the analysis of semantical and syntactical analysis of text to better understand the discourse structure of a complete phrase and how it is related to the user's pragmatics.

Table 2.1: Categories of Language analysis [51]

Morphology	**Field of study**
Lexical	Segmentation of text into words
Syntax	Grammatical rules for text generation
Semantics	Meaning of words in a given context
Discourse	Meaning among different sentences
Pragmatics	Meaning through speaker intent
Acoustics	Representation of sound
Phonetics	Mapping sound into speech
Phonemics	Mapping speech into a language
Prosodics	Stress, pitch, tone, and rhythm when communicating

Table 2.2: Stop words examples in English language [14]

i	me	he
it	a	the
as	into	no
so	by	or

For computers to understand and work with language and text documents, it is necessary to follow specific preprocessing procedures that reduce every document, paragraph, sentence, word, or letter into modular information for better processing. The first step is the *tokenization* of long phrases, which consists of segmenting text into relevant units (usually words) [51]. After separating each word, they are evaluated individually to obtain its nuclear representation. *Stemming* is the process in which particular ending characters of a word are eliminated in order to get a general word representation[51]; this process is usually practical with verbs since it is easier to analyze a global representation of an action rather than a particular conjugation. However, this process not always gets a nuclear word and focuses on just eliminating ending characters of a word.

The process of *lemmatization* is usually implemented with nouns (although not exclusively). In this case, the ending part of a word is eliminated to get n accurate, modular term [51], unlike the stemming procedure that only cuts certain parts. A final lexical preprocessing is the localization and elimination of stop words. These terms are tokens commonly used in a given language, having a poor or almost no particular meaning or value for the general context of a sentence. Bird et al. [14] shows some of the most common stop words used in the English language; some examples are shown in Table 2.2. Figure 2.2 shows an example of every preprocessing procedure.

The syntactical structure of words is directly related to particular POS that generalize and classify every term into a grammatical group. Table 2.3 shows the most common POS tags for NLP analysis. These categories are used to identify the grammatical relationships between a sentence's words and analyze particular blocks inside the phrase [26]. Figure 2.2 also shows an example of dependency grammar, where the root word is linked to their dependant

Table 2.3: General POS in English language [51]

Label	**POS**
N	Noun
V	Verb
A	Article
ADJ	Adjective
ADV	Adverb
AUX	Auxiliary
DET	Determiner
P	Preposition
CON	Conjunction
PRO	Pronoun
INT	Interjection

terms by an arrow.

2.2.1 Applications of NLP

The robustness of a NLP task drastically varies depending on the level of specialization for the representation of words, the level of semantical, lexical, or syntactical analysis, or even the level of specialization of the used vocabulary. Zhai and Sean [110] classify different implementations based on the level of text processing and *easiness*. This last concept of easiness refers to the complexity of the language rules required for analyzing a given document, being text classification or word retrieval the more direct implementations, and question answering and dialogue generation the most complex activities. At the same time, that easiness factor is deeply correlated to the fact that most NLP applications are based on co-occurrence frequencies, which means that the effectiveness of the model is deeply based on the vocabulary (both extension and complexity) used for the model's training [19].

According to Cambria and White [19], every NLP application can be divided into three practical categories: syntactic evaluation, semantic evaluation, and pragmatic (or narrative) evaluation. The first group, as the name suggests, is focused on the grammatical structure of a text, and its primary purpose is to retrieve and extract information, categorize and model different topics based on the vocabulary, among others [19]. Some of the most representative syntactical contributions in the SoA are the Penn Treebank, a corpus of more than four million words of the English language categorized by their corresponding POS [66], and PageRank, a Google ranking algorithm for web pages based on their content [77]. Other syntactical implementations are the first language models developed by Manning and Schütze [65] and Hofmann [41]; these models used ML models to evaluate the valence of each word in a dictionary, as well as attributing a co-occurrence value based on the probability of each word being present in a given text.

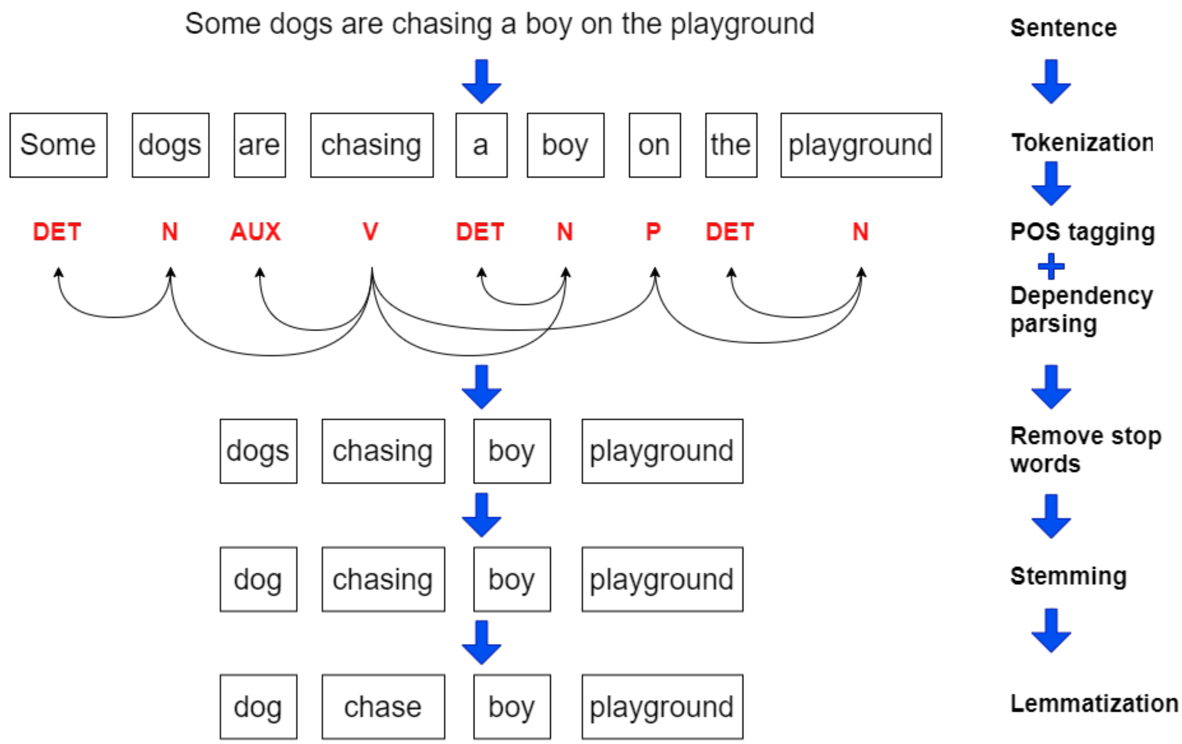


Figure 2.2: Basic preprocessing of a sentence

The semantic analysis of a text is based on evaluating the meaning of each word in a given context. This particular field of study assigns a numerical representation to every corpus word based on different probabilistic or hierarchical relations. Some of the most relevant implementations in this area are LSA [42] and the WordNet database [70]. Both of these implementations will be furtherly described in this document.

Finally, the pragmatic analysis consists of implementing both a syntactical and semantic analysis to create a narrative behavior [19]. Some SoA models were proposed by Asher and Lascarides [5], who established one of the first discourse structures, while Bex et al. [12] developed specific hierarchies for argument reasoning in a conversational model. These models allowed the generation of multiple algorithms capable of analyzing vast extensions of text to imitate the writing style of famous authors and creating new works [106], or even translate phrases into different languages with grammatical accuracy [7] and summarizing long documents through semantic graphs [52, 53] or neural network models [90].

The rest of this document will focus on the different databases, algorithms, and implementations used for semantical and syntactical analysis in sentences to relate a semantical similitude between those phrases and a particular, single word.

2.3 NLP Databases for Text Analysis

As mentioned before, the analysis of NLP involves representing and evaluating textual data in a machine. In order to do that, it is necessary to establish some structures and relationships that allow the algorithm to accurately detect each token or concept for further semantic or syntactical implementations. Just like with any other ML database, the compendium of textual data depends on a particular application, dividing the information and its corresponding ground truth values according to a specific purpose. The following databases are some of the SoA corpora for analyzing the semantical structure of sentences and words.

2.3.1 Thesaurus Databases

Some databases of semantical analysis are frequently used, even in research outside the NLP field; these compounds of words are based on the relationship of how a small phrase can describe a single word or concept. These databases are the standard and traditional dictionaries and thesaurus (a document that groups words based on similar meaning [31]). Two of the most accessible databases online are the Oxford English Dictionary [31] and the Thesaurus (as redundant as it sounds) compendium [1]. The first database is in charge of assigning a descriptive phrase to every English-language term to establish a semantical similitude between a single token and a complete sentence. The thesaurus database indicates the synonymity relationships between different tokens.

Even though these databases were not designed for a particular NLP analysis, they do offer a complete interpretation of the English language, which allows offering an initial structure of the relationships between multiple words. Also, these documents are the main compendia of words that evaluate the relationship of text following a word-phrase structure. Other SoA sets evaluate the semantical similitude of text, but only by comparing single words with other tokens (following a one-on-one analysis) or the whole context of complete phrases that combine the individual similarities of their forming tokens to get a general phrase representation and similitude.

2.3.2 WordNet Database

This database is a thesaurus-like, lexical compound of nouns, verbs, adjectives, and adverbs based on a semantical hierarchy. In this structure, every word is represented in a synonymy set (better known as *synset*) that includes the lemma (nuclear word), POS tag, and a definition number [70]. This structure allows differentiating between *bank.n.01* that talks about a slope of land, the *bank.n.02* that is a financial institute, and the *bank.v.01* that involves the action of tipping laterally.

Besides the assignment of multiple synsets to every token, each of them is classified in a hierarchical semantical structure based on synonymy and other semantical relations, such as hyponymy and meronymy (these concepts are furtherly described in Chapter 3). This hierarchy, shown in Figure 2.3, follows a tree design where the father node corresponds to the term that englobes its corresponding children [70]. For the sake of representation, this database

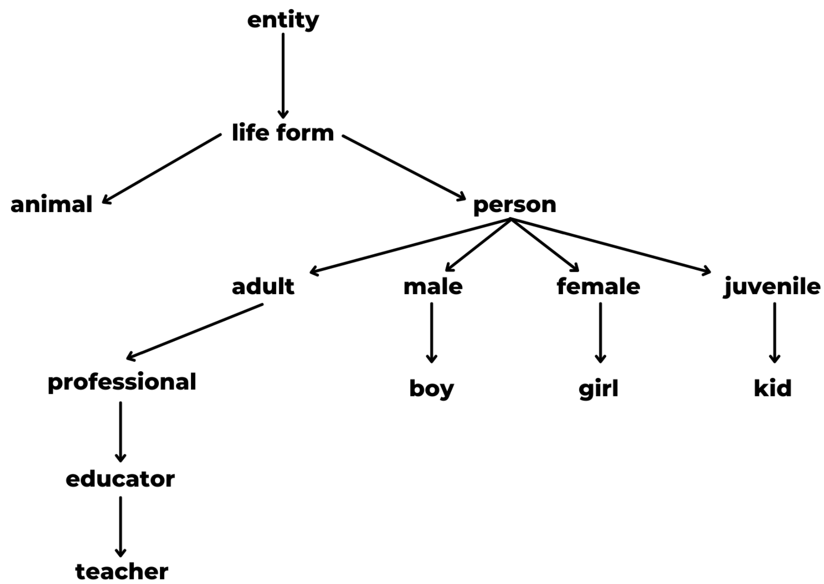


Figure 2.3: Wordnet hierarchical structure

Table 2.4: Synset distribution in the WordNet database [70]

POS	Unique definitions	Synsets	Total word-sense pairs
Noun	117,798	82,115	146,312
Verb	11,529	13,767	25,047
Adjective	21,479	18,156	30,002
Adverb	4,481	3,621	5,580

uses as root term the concept of *entity*; the rest of vocables are children nodes of this term and branch into multiple contexts.

At the same time, this database includes a brief definition for every synset, giving a descriptive text for each lemma based on its context. This pair of concept-definition generates a total of 147,278 different synsets distributed into each POS tag as shown in Table 2.4. Even though this database has a representative amount of words for analysis, one main disadvantage is that the data is presented textually, lacking a numerical vectorization or embedding that helps position each token into a point in space that allows a NLP algorithm to evaluate the semantical structure of its vocables. Fortunately, there are multiple methods in the SoA that assign these vectors based on supervised learning or knowledge-based representations.

2.3.3 Flickr Corpus

Commonly, every descriptive narration or text we use is based on visual information. These descriptions are also known as *visual denotations* [32, 72], and are the general structure of the Flickr Corpus [109]. This database has 31,783 images of multiple situations; each of these pictures with five descriptive sentences (158,915 in total) that talk about the general context

Figure 2.4: Examples of the Flickr database [109]

observed on the image. Figure 2.4 shows an example of the structure of this database.

Each of the five selected sentences for describing a picture was preprocessed according to the methodology described by Hodosh et al. [40] and the WordNet hypernym structure to create generic phrases that could be used for describing more than one image at the same time. Each denotation s was evaluated to determine how much information was covered from a different denotation s'. This process, known as Pointwise Mutual Information (PMI), measures the probability of how many of the N images were described by each denotation, as described in equation (2.1), where $P(s) = \frac{|s|}{N}$ and $P(s, s') = \frac{|s \cap s'|}{N}$ [24].

$$PMI(s, s') = \frac{\log\left(\frac{P(s,s')}{P(s)P(s')}\right)}{-\log(P(s, s'))} \tag{2.1}$$

This database assumes that the descriptive phrases give the same context when evaluating the same image; however, no numerical score or ground truth value determines the degree of similitude between each sentence. Equation (2.1) shows the level of information covered between two phrases, but not the similitude with the input image. This lack of numerical data makes this database a more qualitative compendium of phrases based on the abstract way the human brain correlates denotations.

2.3.4 SNLI Corpus

The SNLI database follows the idea that two sentences can be related to each other based on three different inferences: a direct, positive relationship, better known as an *entailment*, the opposite meaning, or *contradiction*, and *neutral* relationships with independent meanings [64]. It consists of 570,152 pairs of descriptive sentences (some rescued from the Flickr Corpus [109]) directly classified by human subjects into the three mentioned inferences [16]. Table 2.5 shows an example of each of the inferences used in the database.

Table 2.5: Examples of text inferences in the SNLI database [16]

Sentence 1	Sentence 2	Inference
A soccer game with multiple males playing.	Some men are playing a sport.	Entailment
A man inspects the uniform of a figure in some East Asian country.	The man is sleeping	Contradiction
A smiling costumed woman is holding an umbrella.	A happy woman in a fairy costume holds an umbrella.	Neutral

This database was designed solely for classification purposes, demonstrating its usefulness for inference classification with dependency tree models [100], Deep Neural Network (DNN) models like one-layered architectures [15] or even RNNs [39].

Even though this database provides an extensive amount of sentences for classification, there is no evaluation of the similarity between each pair. At the same time, the entailment and contradiction, which can be assumed as qualitative ways of showing similarity, do not determine in which sense the phrases may be identical or different in meaning. Again, this type of analysis and observations in text data makes it difficult to determine the grade of similitude between a complete phrase and its relationship with a single token. The entailment comparison can only assign a class thanks to the union of words into a complete phrase; the general context disappears whenever a word is eliminated.

2.3.5 STS Benchmark

The STS benchmark is one of the best SoA databases for actually evaluating the semantical similitude between pairs of phrases. Unlike the previous databases, this compendium of phrases generates a ground truth value that measures different graduations of meaning overlap between each phrase [20]. In order to produce such methods, each pair was evaluated by a test population that scored with a 5-Likert metric, as described by Agirre et al. [4]. These scores are described in Table 2.6.

The numerical score of each pair of sentences allows measuring semantical similarity as the cosine similarity between the sentence embeddings (each of these procedures is described in the following subsection). A total of 8,628 pairs were divided into a train, dev, and test set [20] for the tuning of multiple NLP algorithms like a conversational prediction model [107] or the use of multiple RNNs [94].

Even though STS works with complete phrases, like the rest of the mentioned databases, the numerical representation allows for better classification and relationship tasks, allowing a direct evaluation of how the use of different vocabulary can refer to the same general context by combining multiple databases of semantical and lexical information. For example, the method of Wu et al. [104] combines the STS ground truth values with the cosine distance between the sentence embeddings, both generated with the WordNet corpus for evaluating the

Table 2.6: Similarity scores for the STS similarity classification [4, 20]

Score	**Similarity description**
0	The two sentences are completely dissimilar
1	The two sentences are not equivalent, but are on the same topic
2	The two sentences are not equivalent, but share some details
3	The two sentences are roughly equivalent, but some important information differs/missing
4	The two sentences are mostly equivalent, but some unimportant details differ
5	The two sentences are completely equivalent, as they mean the same thing

hierarchical structure of the sentence's tokens.

2.4 Word Representation Methods

Sentences and other word groups are analyzed and comprehended based on the principle of compositionality, which establishes that the meaning of a phrase is the result of the collective meaning of the sub-phrases or words that compose it [36]. This principle can be applied to every sub-phrase until reaching word-level meaning. Even then, the meaning of a single word could be determined by the composition of every single letter that compounds it.

Fig. 2.5 shows an example of the principle of compositionality, where the meaning of phrase A is obtained by the combined meaning of words/sub phrases B, C, and D. This observation suggests that semantic rules can work as mappings from local phrase structure trees to determine its meaning, since we, as human beings, are capable of creating a mental process that determines that the combined meaning of B, C, and D is equal, or at least similar, to A. In other words, the representation (furtherly referred to as embedding) e of a particular term t, is the result of the total sum of every ith word embedding w_i that is present on the evaluated sentence, or document D, as shown in equation (2.2).

$$t_e = \sum_{i=1} w_i \Leftrightarrow w_i \in D \tag{2.2}$$

An embedding will be the numerical, vectorial representation of each word w inside a given compendium of vocables (better known as dictionary or vocabulary). These vectors represent the particular point in multidimensional space to distinguish them from other terms or concepts. One of the first, more common ways of representing a word embedding was the

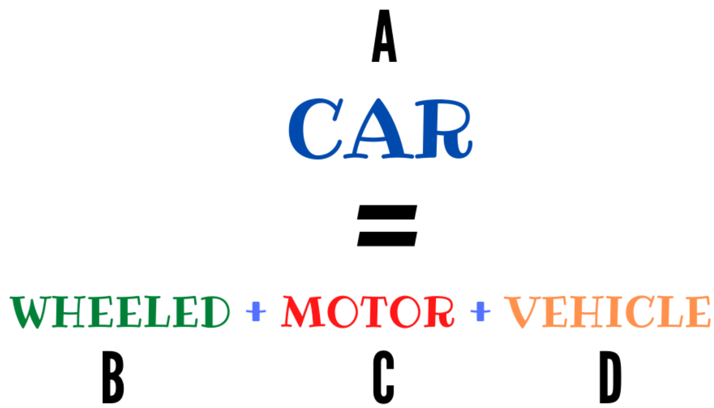


Figure 2.5: Example of the principle of compositionality.

one-hot vector, shown in Figure 2.6. This vector positions each word of the dictionary with a value of 1 for the space located and 0 where not. Even though this approach was common during the first implementations of NLP, one disadvantage is that the vector length augments with an additional word, making it impractical for long documents and extensive vocabulary usage.

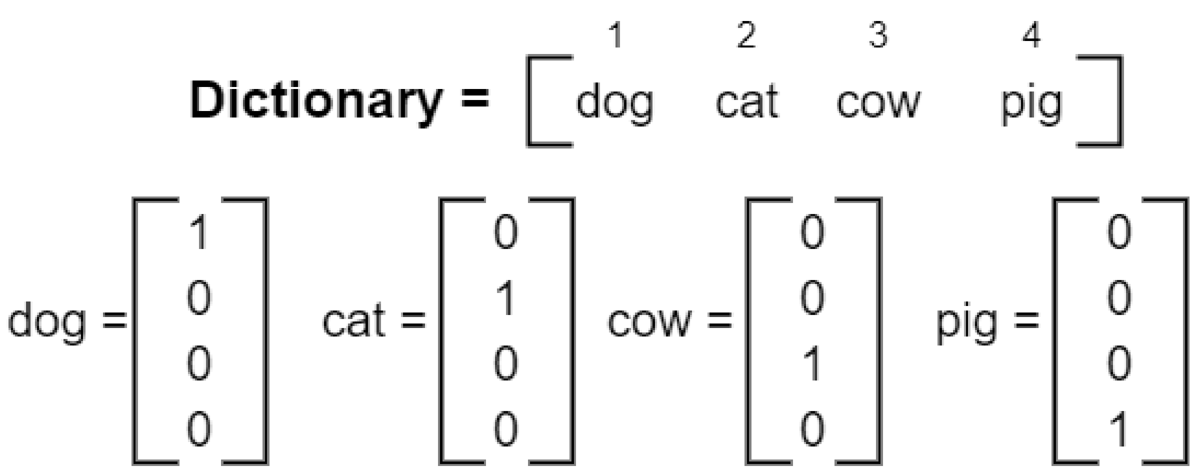


Figure 2.6: Example of one-hot embeddings.

Current SoA implementations have evaluated different ways of getting word embeddings for better representations, including not only the position of each word but also semantical or syntactical information as well. The following subsections describe the two main branches for word embedding generation.

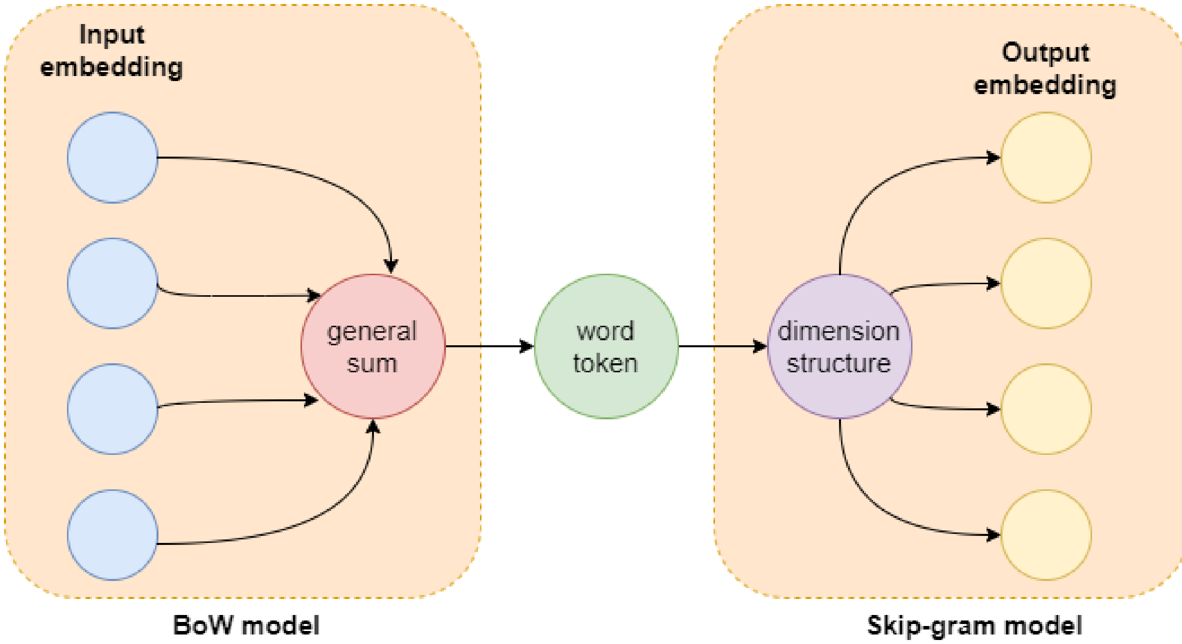


Figure 2.7: General structure of the BoW and skip-gram model

2.4.1 Corpus-based Representations

One of the most common methods for getting a word embedding is evaluating the similarity of vocabulary in a particular dictionary, or corpus [92]. Under these circumstances, the representation of every word is based on the co-occurrence factor: the fact that two or more documents occur in the same document often [65].

One SoA implementations for knowledge-based representations is the combined use of a BoW and a skip-gram architecture, shown in Figure 2.7. This architecture trains to DNN models to analyze the context in a sentence and then determine the numerical representation of the following term [69]. Following this line of thought, the BoW is trained with multiple n-gram representations of words to be capable of assigning a particular word to the input embedding. The skip-gram architecture uses the opposite architecture: by receiving a particular word, the neural network can assign an embedding based on the training context. Even though this implementation is commonly used among regular-sized corpus, the embedding training may generate representation problems for the overspecialization on particular topics, meaning that the model requires finetuning for any additional topic.

Another common algorithm for corpus-based embeddings is LSA. This algorithm uses co-occurrence between multiple documents and phrases to evaluate how many times a word appears in the same passage, summing its presence into a numerical vector that indicates all the different mentions through the corpus documents. This vector, in the end, represents a probability of how commonly this word or term may be present in other documents [56, 102]. Figure 2.8 shows the general structure of how LSA operates.

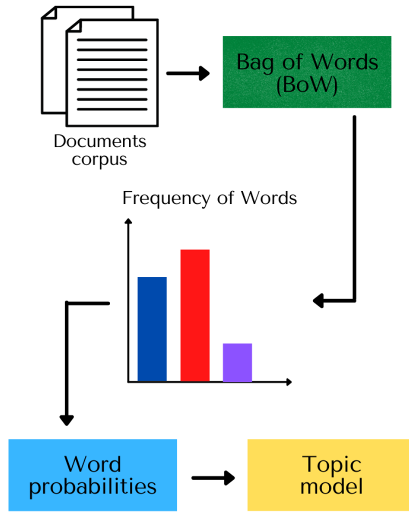


Figure 2.8: LSA general structure

This embedding results from a co-occurrence matrix between m phrases and n documents creates an $m \times n$ table that measures the number of times each phrase is present in every document. It is essential to mention that the n documents used for this matrix are regularly related to the same topic to achieve better results. After that, each term is evaluated through the weight function (2.3) to discard low-frequency terms inside the corpus, where $p_{mj} = \frac{tf_{mj}}{gf_{mj}}$, being tf_{mj} the frequency of each term appearing in a single document j, and gf_{mj} the occurrence of the same term along the whole corpus of n documents [57].

$$w = 1 + \sum_{n} \frac{p_{mj} \log_2(p_{mj})}{\log_2(n)} \tag{2.3}$$

One of the most recent enhancements to this algorithm is Extended Latent Semantic Analysis (xLSA), which additions a syntactical analysis to the word representation of LSA by evaluating how every POS affects different parts of a sentence, modifying the dependency between each part, improving the word embedding [92]. This model assigns a POS tag to every token of the sentence and divides it into three different blocks: the subject of the phrase, the main verb, and the subject's object. Each block is then evaluated individually, getting the LSA embedding for each token; the sentence's embedding will be the result of the mean score between the total embedding sum of each block.

According to Mohler and Rada [71], corpus-based representation of words has the best correlation scores between embeddings since each term is evaluated in a compendium of documents that talk about similar topics. However, even though this process generates more specified embeddings, it is crucial to notice that these representations are only effective when working with the same text topic. Under different circumstances where a new topic is introduced, this method cannot get accurate similarity results due to the new document's lack of term frequency.

2.4.2 Knowledge-based Representations

A more general embedding of words is implemented with the knowledge-based representation, where a semantical hierarchy is designed and implemented to evaluate the position and similitude between concepts. Since every word is represented in a general context, the similitude accuracy of this implementation is lower than the corpus-based [71]. However, it is capable of analyzing multiple contexts and topics without needing to reevaluate the weighting function of each term, like the case of LSA and xLSA.

Regularly these semantical networks or hierarchies are directly designed and evaluated by people, like the WordNet database [70], rather than measuring the appearance probability of corpus-based representations. Unfortunately, this hierarchy is mainly designed with textual data, making it impossible to solely evaluate the relationship of words by detecting their semantical relationships. It is then necessary to complement this textual data with a numerical representation for its use in NLP.

A relatively new method for getting word embeddings is presented by De Sa et al., and Nickel and Kiela [76, 87]. These authors analyze that since a hierarchical structure, like knowledge-based representations, follows a tree structure rather than a linear behavior, word embeddings should follow that tree structure instead of trying to represent each vector in a linear, Euclidean space [87].

They propose to compute word embeddings in hyperbolic space, specifically on a $\mathbb{H}_2$ Poincaré ball model, like the one shown in Figure 2.9, since it represents and respects better the hierarchical relationship between terms when determining their position in space. In this image, each node represents a particular word linked with others by its hierarchical structure. This implementation uses equation (2.4) to measure the distance between words, distributing them in the hyperbolic space and respecting their hierarchy position [76, 87]. The primary node of the tree structure is located at the origin O, while the remaining vocables are distributed in the same way as the semantical hierarchy.

$$d_H(x,y) = \operatorname{arccosh}\left(1 + 2\frac{||x-y||^2}{(1-||x||^2)(1-||y||^2)}\right) \tag{2.4}$$

The hierarchical structure is measured by a distortion D. For each word embedding n, its distortion will depend on the distance between its extremes U and V and their corresponding embeddings $f(u)$ and $f(v)$, respectively [87]. The closer the distortion is to zero, the word embedding follows a Euclidean behavior, while distortion of $D = 1$ allocates the embedding in hyperbolic space. Equation (2.5) shows this behavior.

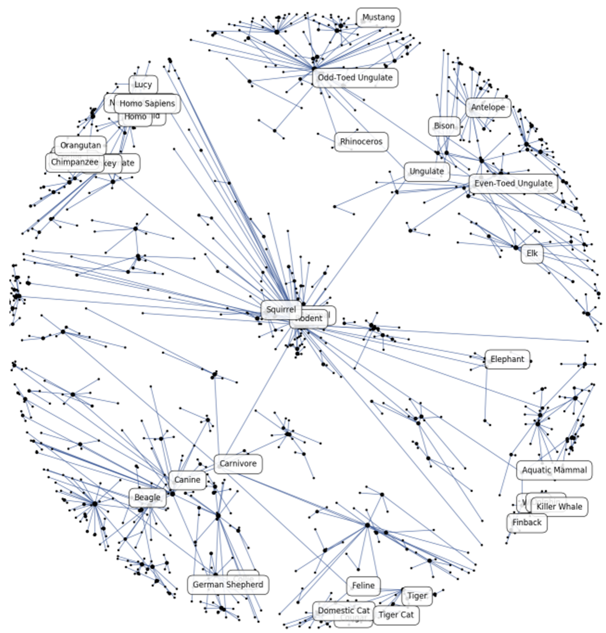


Figure 2.9: Representation of words in a Poincaré hyperbolic space [76]

$$D(f) = \frac{1}{\binom{n}{2}} \left(\sum \frac{|d_V(f(u), f(v)) - d_U(u,v)|}{d_U(u,v)} \right) \tag{2.5}$$

By maintaining the hierarchical structure of words, this implementation can improve the accuracy of knowledge-based representations. These authors [76, 87] have distributed a hyperbolic representation database for the WordNet corpus. This database will be the one furtherly used for the text generation block of this dissertation's proposal.

2.5 NLP Algorithms

The union of word embeddings with ML algorithms are based on recursive models, not because the input data of text is a signal that varies through time, but because there is a direct relationship between consecutive tokens. The most common algorithm for this type of process is known as the *autoencoder* model. This architecture is an unsupervised neural network that implements dimensionality reduction of the multiple input words in order to create a corresponding word or sentence embedding [51]; this part of the model is better known as an *encoder*. At the same time, the generated embedding is used for the autoencoder to map an approximate (or the same) text representation; this block is better known as a *decoder*. The optimization of this method looks to minimize the Mean Squared Error (MSE) between the input text x and the newly generated text $\hat{x}$, as shown in equation (2.6), where the generated text is the result of the decoded embedding z, $\hat{x} = Dec(z)$, and the latter being the result of encoding x, $z = En(x)$ [51]. Figure 2.7 resembles the general structure of this architecture, where the BoW simulates the encoder and the skip-gram the decoder.

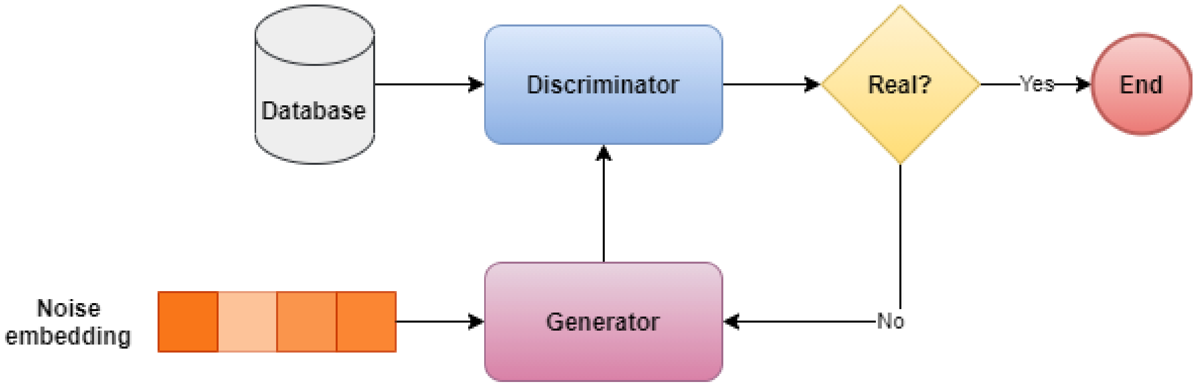


Figure 2.10: GAN architecture [51]

$$\mathcal{L}(x, \hat{x}) = ||x - \hat{x}||^2 \tag{2.6}$$

A more recent and complex model for text analysis is the GAN, which uses two separate neural network models to separate the *generation* of text based on a noise embedding and a *discriminator* that discerns from the generated text and a ground truth text [37]. The model's training consists of using noise embeddings as input for the generator to create a ground truth text; simultaneously, the discriminator is trained with the generated text and the actual ground truth value. The discriminator is in charge of classifying the generated text as a fake representation or a ground truth text. The intention is that the generator must ”fool” the discriminator, while the latter recognizes the difference in texts; it is said that the model finishes training when the discriminator is no longer capable of discerning between the two text inputs. Figure 2.10 shows a general diagram of this procedure. Equation (2.7) describes the objective function of the model between the generator G and the discriminator D, where $\mathbb{P}_r$ and $\mathbb{P}_g$ are the real and generated data distributions, respectively; also, it is considered that the generated text is the result of the generator's noise embedding z: $\tilde{x} = G(z)$ [51].

The following architectures were designed with the general structure of these two models, obtaining remarkable results for both the generation and analysis of text data.

$$\min_G \max_D \mathbb{E}_{\hat{x}\sim\mathbb{P}_r}[\log(D(x))] + \mathbb{E}_{\hat{x}\sim\mathbb{P}_g}[\log(1 - D(\hat{x}))] \tag{2.7}$$

2.5.1 Long-Short Term Memory Neural Networks

One of the first models for NLP analysis were RNNs, an architecture capable of analyzing sequential data x (like time-depending functions or text inputs). The difference with regular neural network models is that for every member t of the sequence, $x^{<t>}$, a particular layer is assigned to evaluate and generate its corresponding output $y_{<t>}$; at the same time, each layer has a unique activation function based on the value of the previous layer $a^{<t-1>}$, as shown in Figure 2.11. Besides, this architecture also separates the network's weight matrix into three different ones: W_{ax} for the input, W_{aa} for the activation, and W_{ay} for the output of every layer

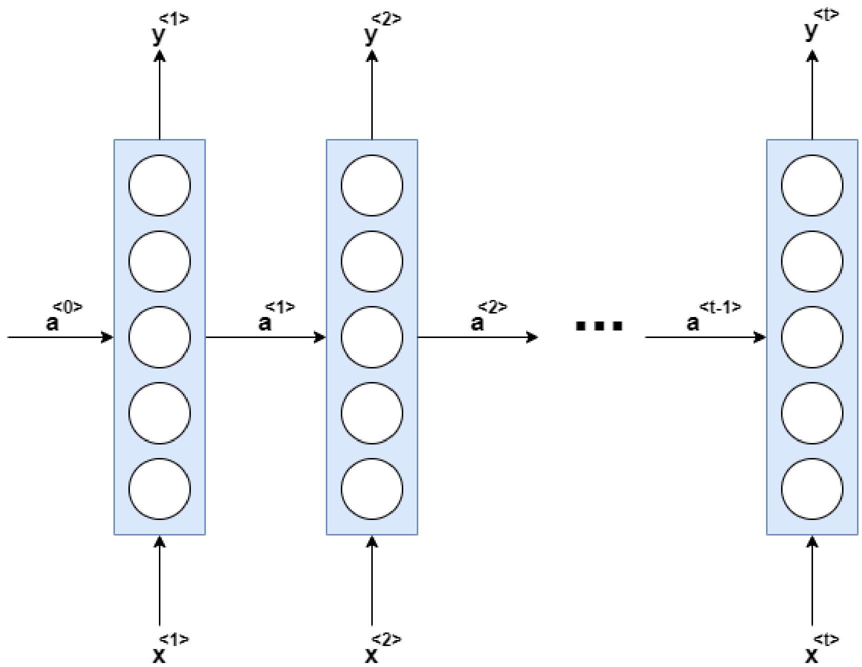


Figure 2.11: RNN architecture

[93]; the forward-propagation of the model is described then by equations (2.8) and (2.9), where σ usually represents the sigmoid function; in the case of the first layer, a^0 will be represented as a zero-vector.

$$a^{<t>} = \tanh(W_{aa}a^{<t-1>} + W_{ax}x^t + b_a) \tag{2.8}$$
$$y^t = \sigma(W_{ya}a^{<t>} + b_y) \tag{2.9}$$

This architecture has a good generation accuracy when dealing with short text inputs thanks to the architecture mentioned before (better known as a *short-term memory* architecture). Unfortunately, when working with long sentences and a more significant amount of vocabulary, RNNs begin to lose the dependency between the first sentence's token and the one generated at the end; this problem is also known as *vanishing gradient*. Besides, the network's gradients begin to increase during backpropagation when working with long inputs, causing a numerical overflow on the operations of the model; this is known as a *exploding gradient* problem [39].

The model of the LSTM neural network intends to reduce these problems by implementing additional hidden units inside the RNN 's architecture, as shown in Figure 2.12. This unit consists of a *long-term memory* cell c that helps "remembering" the structure of further past tokens. This cell is capable of determining when to update, forget, and output the remembered token with their respective gates Γ_u, $Gamma_f$, and $Gamma_o$ [23, 39]. These gates are

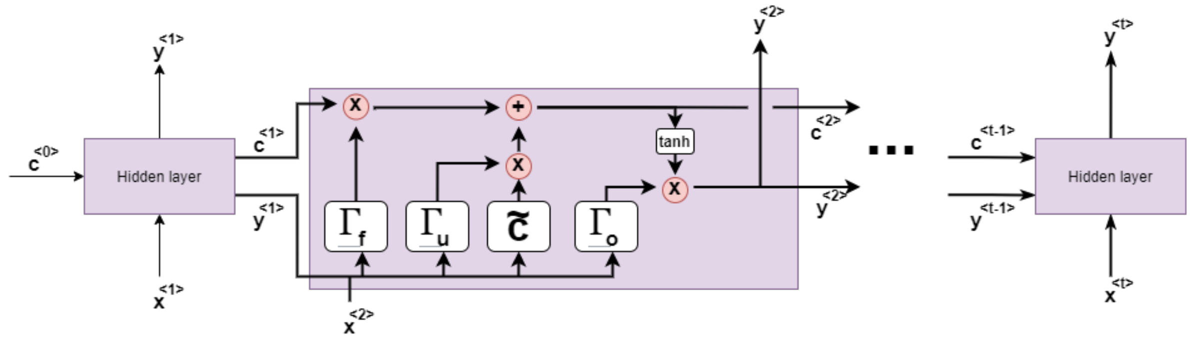


Figure 2.12: LSTM architecture

described by equations (2.10), (2.11), and (2.12), respectively. At the same time, the over-writing of the cell memory, $\tilde{c}^{<t>}$ will follow a $\tanh$ function to determine when to update the function and when to forget it, as shown in equation (2.13). Finally, the resulting value of the memory cell, described by equation (2.14), will help to generate the new text token based on its activation function, represented by equation (2.15) [23, 39].

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u) \tag{2.10}$$

$$\Gamma_f = \sigma(W_f[c^{<t-1>}, x^{<t>}] + b_f) \tag{2.11}$$

$$\Gamma_o = \sigma(W_o[c^{<t-1>}, x^{<t>}] + b_o) \tag{2.12}$$

$$\tilde{c}^t = \tanh(W_c[a^{<t-1>}, x^t] + b_c) \tag{2.13}$$

$$c^t = \Gamma_u \tilde{c}^{<t>} + \Gamma_f c^{<t-1>} \tag{2.14}$$

$$a^{<t>} = \Gamma_o \tanh c^{<t>} \tag{2.15}$$

The use of LSTM networks has been common practice in NLP during the last decade, particularly with the combination of two networks, simulating the behavior of the encoder-decoder model described before. This particular model is referred to as *bidirectional LSTM*, an architecture that evaluates the input sentence's tokens through the encoder LSTM and creates a general sentence embedding at the output of it. That embedding is then used for the decoder LSTM to interpret those numerical values into new tokens, producing a new phrase. A further description of how bidirectional LSTM are used in textual analysis is shown in [46].

2.5.2 Attention Model

Even though RNN diminish the vanishing gradient problem, its architecture still presents computational problems when evaluating long phrases. The *attention* model is a novel SoA implementation based on the psychological concept of the same name: given a particular sentence of length S, the attention model maps smaller parts of it, blurring and ignoring the rest of the sentence during the learning phase of the network [51]. This implementation resembles the encoder-decoder model of a bidirectional LSTM, but implements an additional attention layer between the two architectures, as shown in Figure 2.13a. Using the global (total) or local (portion of the sentence) tokens of the encoder's hidden layers $h_x^{<t>}$ and the decoder's output of the given token h^t, the attention layer generates a context vector c_t (Figure 2.13b) that helps with the output of the token $y^{<t>}$ [7, 63].

The model's novelty is that the attention model ponderates the weight of $y^{<t>}$ based on more than one word of the original input text at the same time, unlike the LSTM model that only evaluates the direct past token. The attention model is trained to predict the position of the generating token $p^{<t>}$ with equation (2.16), where W_p and v_p are the parameters to be learned [63]. Additionally, the position is aligned in a Gaussian distribution by equation (2.17), where D is an empyrical window for value selection and score$(h^{<t>}, h_x^{<t>}) = v_a^T \tanh() W_a [h^{<t>}; h_x^{<t>}]$[63].

$$p^t = S\dot{\sigma}(v_p^T \tanh(W_p h_t)) \tag{2.16}$$

$$a^{<t>}(s) = \frac{\exp(score(h^{<t>}, h_x^{<t>}))}{\sum \exp(score(h^{<t>}, h_{x'}^{<t>})} \exp\left(-\frac{(s - p^{<t>})^2}{2\left(\frac{D}{2}\right)^2}\right) \tag{2.17}$$

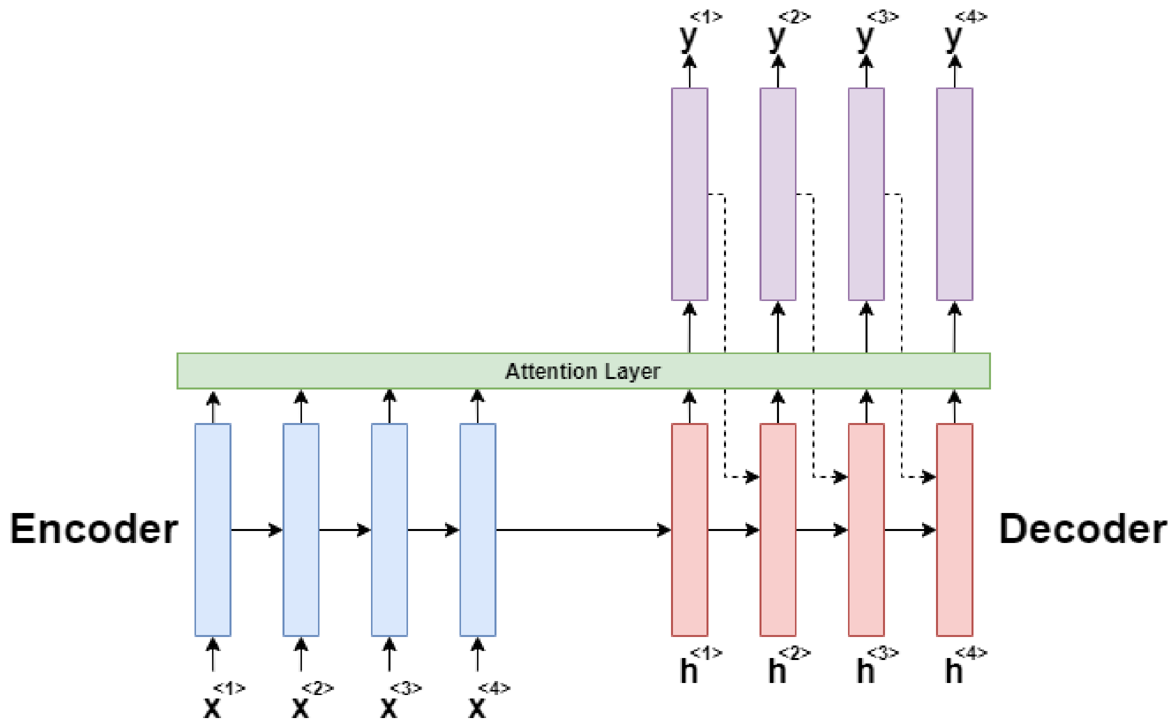
In general, the advantage of the attention model is its capacity for evaluating the overall context of the input sentence and how it interprets a particular relationship between tokens between the input and generated phrases. Figure 2.14 shows an example of the attention matrix of a sentence translation according to the original tests of Luong et al. [63].

2.5.3 Transformer Neural Network

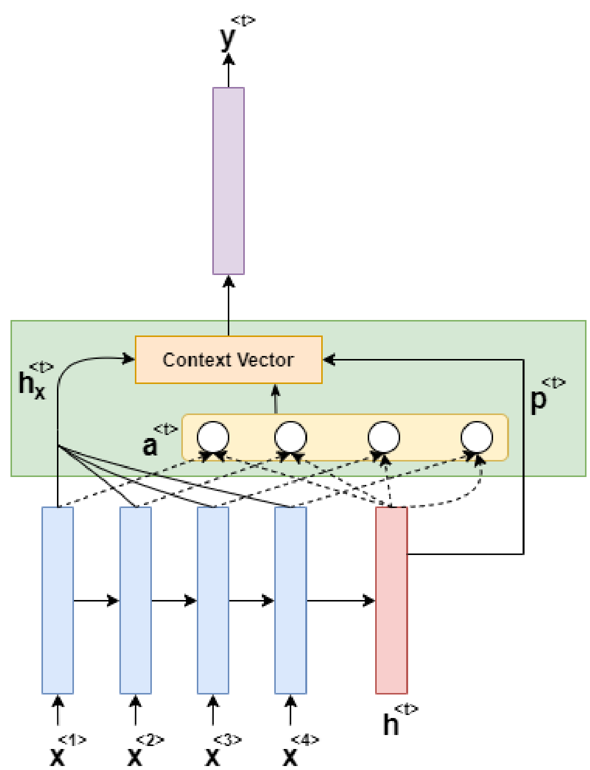
The Transformer architecture proposed by Vaswani et al. [99] improves the training time and generation accuracy of any RNN model, making it the current prime selection for NLP analysis. This analysis implements SoA results by implementing parallel DNNs for evaluating each token of the input sentence individually. Figure 2.15 shows the complete structure of this architecture.

Each token embedding is ponderated by a positional encoding, shown in equations (2.18) and (2.19), where each token vector is modified depending on the position pos in every dimension i [99]. Besides that, each embedding is evaluated to create a *query*, *key*, and *value* matrix (Q, K, and V, respectively) that multiply each token's embedding in order to get an attention score that evaluates the relationship between each word in the sentence. This score is obtained by multiplying the query and key values of each word. At the same time, this multiplication is implemented in a softmax function and ponderated by the value matrix to eliminate the vanishing gradient problem. Equation (2.20) shows the expression of the attention score, where $\sqrt{d_k}$ is the length of the key vector [51, 99]. The evaluation of the attention score is also known as *multi-head attention*. The Feed Forward Network (FFN) that is applied in the encoder after obtaining the attention score through equation (2.21) consists on twoReLU functions on traditional fully-connected neural networks alongside every N layer (usually $N = 6$) [99]. This function evaluates each input token positionally.

The decoder part of the Transformer consists first of using multi-head attention to mask



(a)



(b)

Figure 2.13: General description of the (a)attention model architecture and the (b)context vector generation

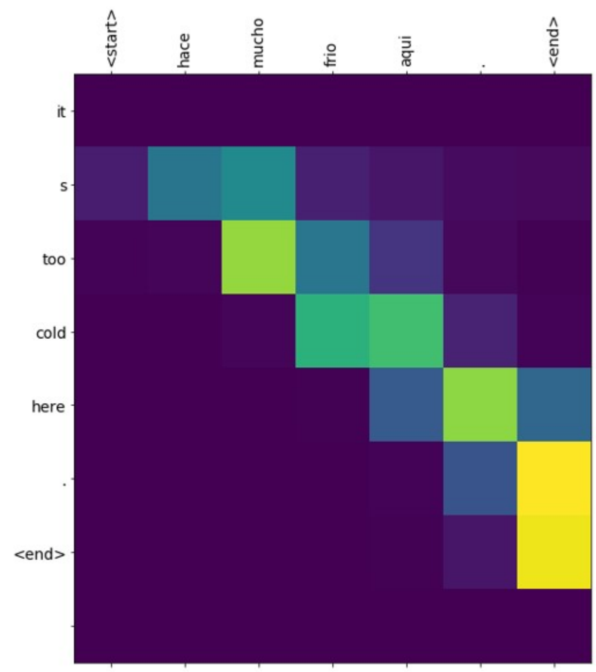


Figure 2.14: Attention matrix of sentence translation

the position of the decoders data, evaluating only the current and previous tokens of the current generated word, and ignoring the ones that come after that. The second multi-head attention block evaluates the relationship between the desired generated token and the encoder's attention score. Finally, a layer normalization is applied to get the final token embedding [99].

The network separates and evaluates each token of the input sentence individually and in parallel, allowing a faster text generation. Different SoA modifications have been designed during the last years, improving the speed and accuracy of the Transformer model. However, it is still important to remember that every model described so far (and the following ones) was designed to analyze and generate long texts. This dissertation proposal will implement the following algorithms but modify them to create an analysis between a complete sentence and a single token.

$$PE_{pos,2i} = \sin \frac{pos}{10000^{2i/d_{model}}} \tag{2.18}$$

$$PE_{pos,2i+1} = \cos \frac{pos}{10000^{2i/d_{model}}} \tag{2.19}$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK_T}{\sqrt{d_k}}\right) V \tag{2.20}$$

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{2.21}$$

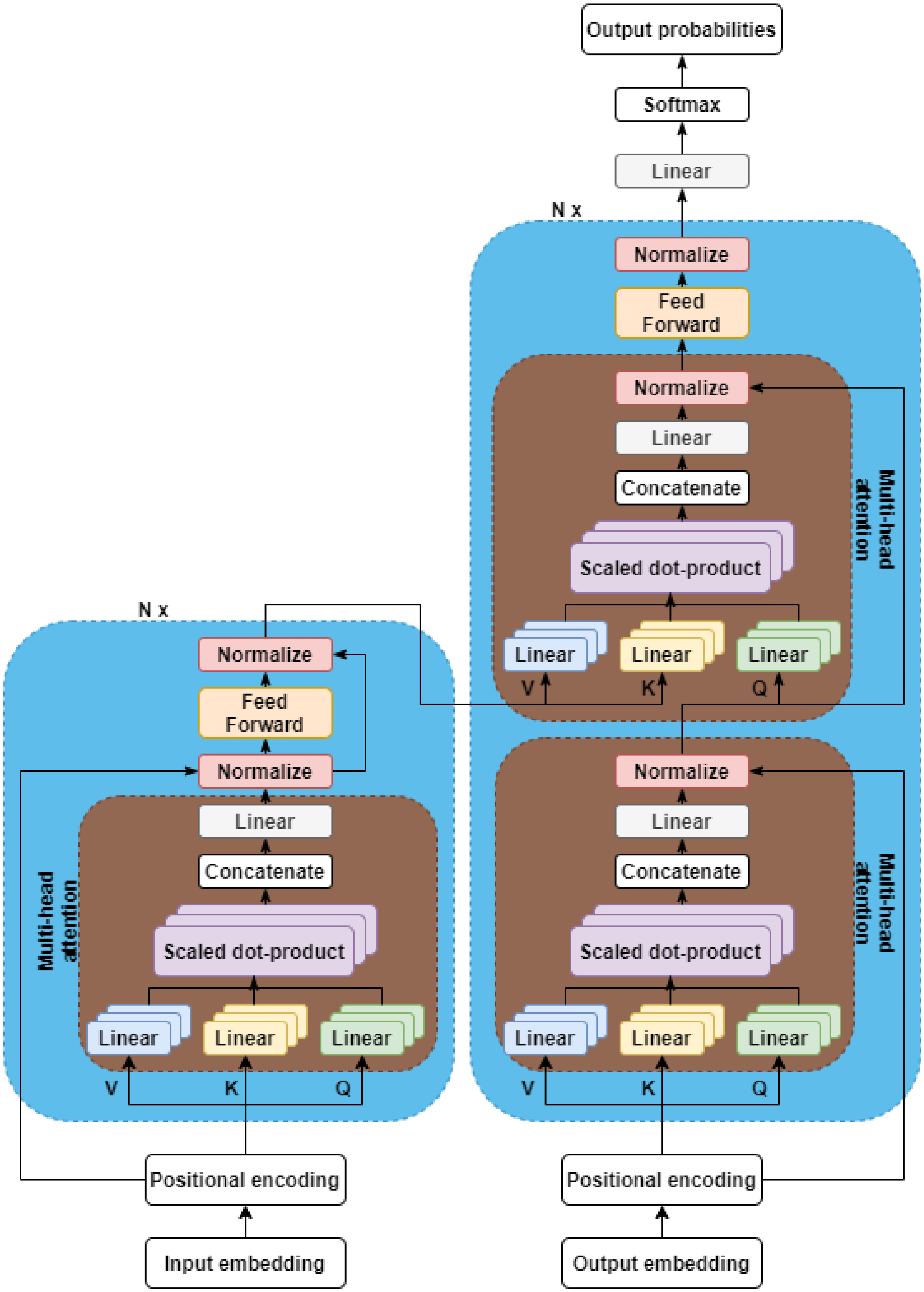


Figure 2.15: Transformer layer architecture

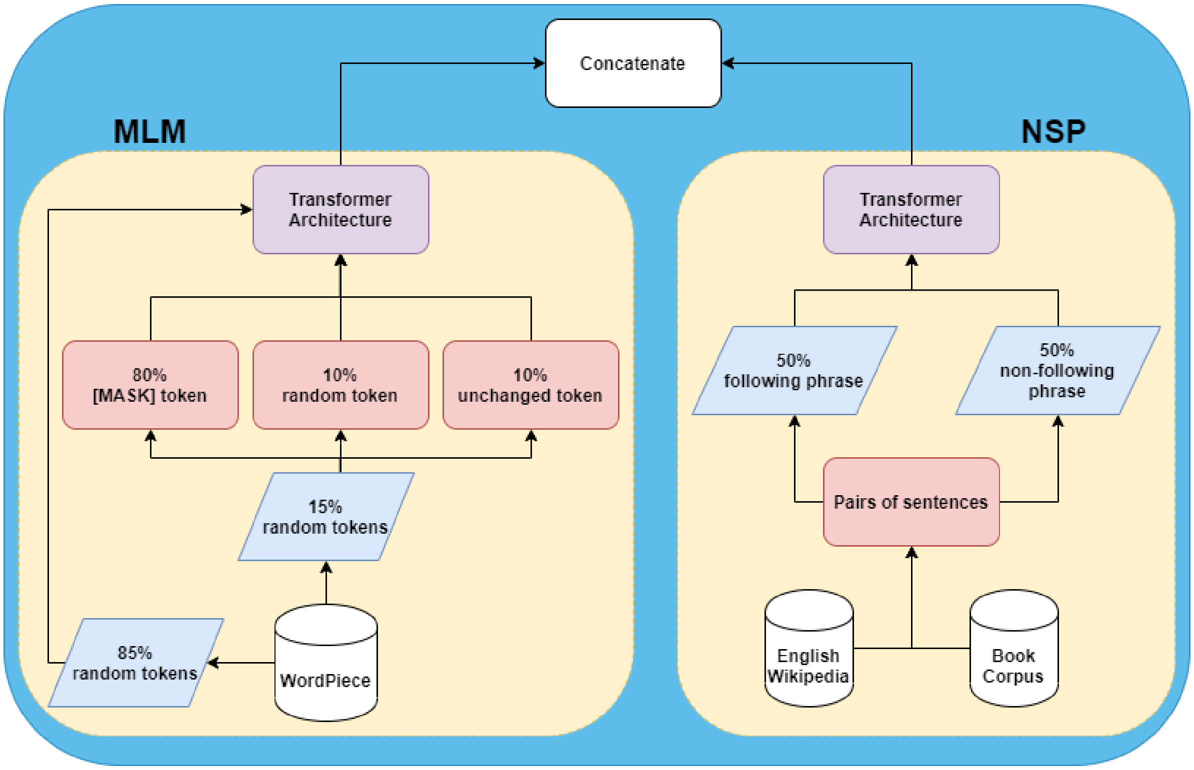


Figure 2.16: BERT architecture

2.5.4 BERT

The BERT network is one of the most common algorithms for text generation. Its architecture is entirely based on the Transformer model of Vaswani et al. [99]. The difference between each algorithm is based on the model's training and how the word embeddings are masked for the next token generation. This model is first pre-trained on two tasks at the same time, as shown in Figure 2.16.

The first task is for the word embedding of tokens and is better known as Masked Language Model (MLM). For this process, the BERT model separates the WordPairs database [105] in a 75-15% distribution. The 15% selected words were modified as follows: 80% of the token's embeddings were changed to a *[MASK]* token to be analyzed by the architecture's output; the other 20% were changed into random values or remained the same (10% for each of them) [29]. The second task is known as Next Sentence Prediction (NSP); consists of using pairs of sentences from the BooksCorpus [111] and English Wikipedia databases to determine which sentence followed the other. In this case, both sentences were joined into a single vector and sentence embedding, separated by a *[SEP]* token to indicate each phrase. 50% of the training set were from consecutive sentences and the other 50% of non-consecutive ones [29].

Once the model finished both tasks, the resulting word embeddings were concatenated

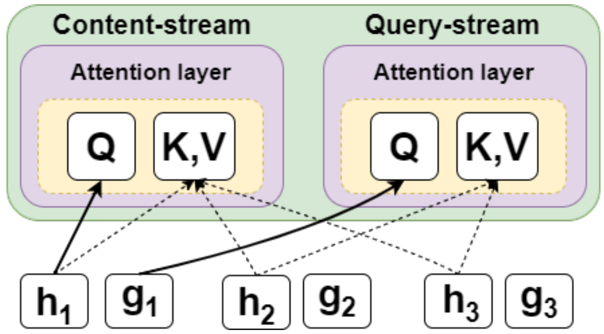


Figure 2.17: XLNet architecture

into a single word vector per generated token. The model can be furtherly finetuned into a particular case of study; however, the architecture remains the same, and the output data depends on the given task. BERT is one of the most used algorithms for text analysis in NLP and has been modified into multiple sub-models for a more specific task. For example, Liu et al. [61] modified the architectures pretraining to improve the overall text generation accuracy; Lan et al. and Sanh et al. created smaller, cheaper versions of BERT to reduce the computing time and cost [55, 88]; there are even BERT models finetuned for a specific language, like French [58, 67].

One of the main problems of this model is that the generation of text is based on a *bidirectional mask* procedure: the embedding of the current token depends on both the words before and after it on the sentence. In other words, the context of a particular word depends not only on previous vocabulary (which seems logical) but also on the future tokens that come after the word that is not generated yet. This masking procedure may generate some accuracy problems when dealing with large corpus or the use of specified vocabulary.

2.5.5 XLNet

The XLNet is a bidirectional Transformer neural network based on the one-directional Transformer XL [27]. Its purpose is to evaluate the encoder's input sentence h separately from the context of every generated token g, as shown in Figure 2.17. This procedure is known as *masked two-stream attention* and consists of using all the possible permutations between the generating token and its predecessors in order to improve the context's log-likelihood [108]. This evaluation is expressed in equation (2.22), where Z_T is the set of all possible permutations of length T, z_t is t-th element and $z_{<t}$ the first $t-1$ elements of the permutation.

$$\max_{\Theta} \mathbb{E}_{z \sim Z_T} \left[\sum_{t=1}^{T} \log p_\Theta(x_{z_t} | x_{z<t}) \right] \tag{2.22}$$

Suppose that given an input sentence x of 3 tokens, the generation of the second token is in process. The XLNet model evaluates every single permutation in the sentence, using as reference the second token $t = 2$. Only the tokens before reference one are considered for the

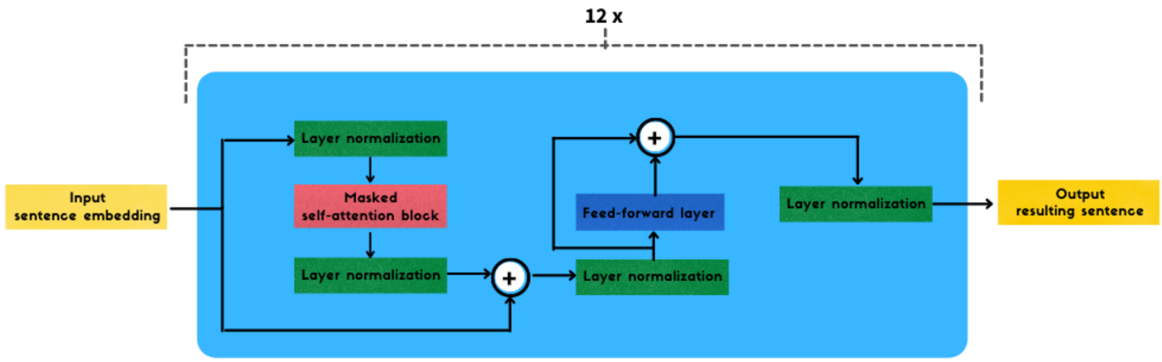


Figure 2.18: GPT-2 architecture

context evaluation in every permutation, omitting the reference itself. So, given the possible permutations $1-2-3$, $2-1-3$, $3-1-2$, $1-3-2$, $2-3-1$, and $3-2-1$, only the context of the tokens $[1, 3-1, 1-3, 3]$ are evaluated for the context generation of $t = 2$. The same process is repeated for every other token.

2.5.6 GPT-2

The Generative Pre-Trained model is another SoA modification of the original Transformer architecture. Like the XLNet, this architecture works with masked-self attention layers to evaluate only the tokens before the currently generated word. However, this architecture follows a more straightforward structure of the Transformer model, designed by Liu et al. [60], who decided to eliminate the encoder part of the network, join the two input sentences into a single embedding (similar to the BERT network), and use the decoder part of the Transformer. This simpler model is known as the *Transformer-Decoder*. The general architecture of the network is shown in Figure 2.18.

This new structure allows a better handle of long phrases, reducing the computation and hyperparameters required for that purpose [82]. Even when the first version of this model was outperformed by the BERT and XLNet networks, the second version, furtherly known as GPT-2, achieved better results on multiple NLP tasks than those two other models, such as language translation, language modeling, summarization, and question answering. The second version applied the layer normalization before each sub-block (unlike the original Transformer that applies it after), adding a second normalization block after calculating the attention score. Besides that, the number of data used for training was considerably augmented, creating a generalized dataset capable of evaluating multiple topics and contexts without further finetuning [83]. Table 2.7 shows the difference in each version's size.

Last year, the third version of this architecture was published (GPT-3). This new model uses the same architecture as the GPT-2, but with a much more considerable amount of parameters (see Table 2.7). At the same time, this new architecture evaluates the finetuning process

Table 2.7: GPT size comparison [17, 83]

Version	Parameters	Units per layer
GPT	117M	768
GPT-2	1542M	1600
GPT-3	175B	12288

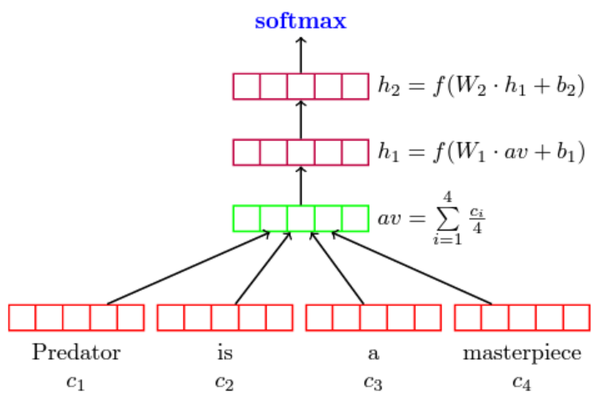


Figure 2.19: DAN architecture [47]

in four different ways [17]. The first model is the traditional finetuning, where a large, new dataset is given to the model, and the general weights are updated to reach a specific task. The second process is known as *few-shot*, where a small part of the new dataset is given to the network, but it is not allowed to update any hyperparameter. The third one is called *one-shot* and follows the same logic as before, but only one single example is allowed in the finetuning process. Finally, the *zero-shot* gives no additional data for the new task. Even though it may seem that the omission of specific data for new tasks may affect the model's accuracy, this new proposal demonstrates the opposite, showing results much better or similar than current versions of BERT and other NLP architectures.

2.5.7 Universal Sentence Encoders

The past architectures were designed to analyze the semantical and syntactical structure of the input text to generate new phrases. The USE model follows a different approach, where only the semantical structure is analyzed, maintaining the same accuracy as the rest of the models.

This architecture is based on the DAN proposed by Mohit et al. [47], where the word embeddings of the whole sentence are connected into an averaging layer to then evaluate the overall sentence's mean embedding through the different layers of the network. At the same time, this model implements a novel dropout algorithm that, instead of dropping some of the network's neurons, randomly drops a token's embedding. Figure 2.19 shows the structure of the DAN network.

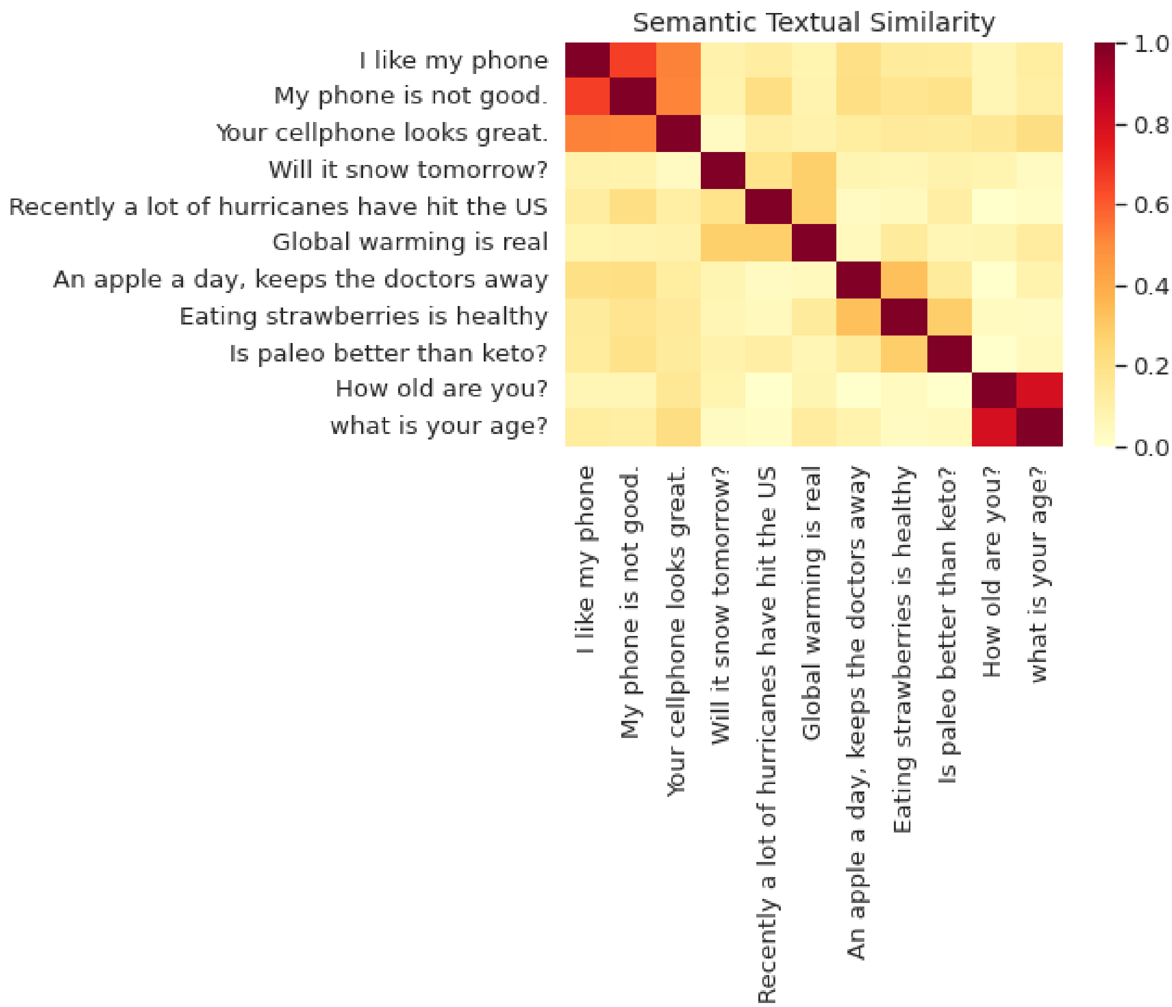


Figure 2.20: USE confusion matrix for phrase semantical relationship

The USE architecture uses then the DAN to average the individual and bi-grams of the sentence's word embeddings, resulting in a 512-dimensional vector that represents the whole phrase [21]. The high accuracy of this model resides on the finetuning on the SNLI corpus [16, 25] and the STS benchmark [20]. The semantical evaluation of this model is visualized through a confusion matrix, like the one shown in Figure 2.20. In this image, the semantical relationship between different sentences is visualized, demonstrating how the usage of the specialized corpus allows for indirect topic classification.

2.6 Metrics for Semantic Similarity

Semantic similarity refers to evaluating how identical the meaning of a word or phrase is with a different text input. Most of the SoA use the word embeddings of their vocabulary to evaluate the distance between two different concepts in a multidimensional space, assuming that the closer the distance, the more similar in meaning they are. This assumption is based

Table 2.8: Theme words for token similarity [85]

Group A		Group B	
asylum	gem	automobile	midday
autograph	glass	bird	monk
boy	graveyard	cemetery	pillow
brother	grin	forest	rooster
car	mound	fruit	sage
coast	noon	hill	serf
cock	oracle	implement	shore
cord	slave	jewel	signature
crane	tool	journey	smile
cushion	voyage	lad	stove
food	wizard	madhouse	string
furnace	woodland	magician	tumbler

on the advantages of word representation, which assigns a numerical value and space for each word in a particular environment. The following subsections will talk about two of the referring similitude metrics that will be implemented in this work: one talks about the first approach that was made to get a similitude metric of words based on human perception, while the other explains one of the most common SoA methods for measuring the distance between word embeddings.

2.6.1 Human Contextual Similarity

Rubenstein and Goodenough [85] were some of the first researchers that evaluated how the common context of a word A may be evaluated to get a degree of similarity with the context of a word B. To create this analysis, they follow the thought of Joos, that the similarity of words is based on *"the set of conditional probabilities of its occurrence in context with other words"* [50].

Their experiment consisted of evaluating 48 colloquial words (listed on Table 2.8), separated into two blocks: A and B. These terms were randomly paired to create a total of 33 pairs of words. A testing population had to determine the similarity level of the corresponding pair using a 4-Likert score, where 4 represented completely identical meanings and 0 different meanings. The evaluated results showed a correlation value of $r = 0.95$ among the population and a positive context overlap [85]. Table 2.9 shows the general similarity scores obtained from this experiment.

More recent implementations are the ones developed with the datasets described before. Two of the first compendium for phrasal semantic similarity were the Flickr and SNLI corpus [66, 109]. The first database consisted of images described by multiple phrases, assuming that each was semantically identical since they shared the same context. The SNLI corpus worked with the same sentences but divided into pairs of words (without the image input) and assigned them a similitude category: entailment, neutral, and contradiction. It is essential to

Table 2.9: Mean similarity scores between theme words [85]

Word A	Word B	Score	Word A	Word B	Score
cord	smile	0.02	hill	woodland	1.48
rooster	voyage	0.04	car	journey	1.55
noon	string	0.04	cemetery	mound	1.69
fruit	furnace	0.05	glass	jewel	1.78
autograph	shore	0.06	magician	oracle	1.82
automobile	wizard	0.11	crane	implement	2.37
mound	stove	0.11	brother	lad	2.41
grin	implement	0.18	sage	wizard	2.46
asylum	fruit	0.19	oracle	sage	2.61
asylum	monk	0.39	bird	crane	2.63
graveyard	madhouse	0.42	bird	cock	2.63
glass	magician	0.44	food	fruit	2.69
boy	rooster	0.44	brother	monk	2.74
cushion	jewel	0.45	asylum	madhouse	3.04
monk	slave	0.57	furnace	stove	3.11
asylum	cemetery	0.79	magician	wizard	3.21
coast	forest	0.85	hill	mound	3.29
grin	lad	0.88	cord	string	3.41
shore	woodland	0.90	glass	tumbler	3.45
monk	oracle	0.91	grin	smile	3.46
boy	sage	0.96	serf	slave	3.46
automobile	cushion	0.97	journey	voyage	3.58
mound	shore	0.97	autograph	signature	3.59
lad	wizard	0.99	coast	shore	3.60
forest	graveyard	1.00	forest	woodland	3.65
food	rooster	1.09	implement	tool	3.66
cemetery	woodland	1.18	cock	rooster	3.68
shore	voyage	1.22	boy	lad	3.82
bird	woodland	1.24	cushion	pillow	3.84
coast	hill	1.26	cemetery	graveyard	3.88
furnace	implement	1.37	automobile	car	3.92
crane	rooster	1.41	midday	noon	3.94
			gem	jewel	3.94

notice that these two databases worked with solely qualitative data without assigning a numerical score of similitude. This type of data show problems when determining a specific level of similitude between texts since there is no way of knowing which word or concept is more similar to other and why.

The most recent corpus for semantical similarity evaluation is the STS benchmark [20]. This compendium of words followed the same structure of evaluating pairs of sentences and assigning them a numerical value of similitude based on a 5-Likert score. Those scores were determined by a human population, getting the mean score of each pair and assigning it as the ground truth value. This database is one of the most complete models in the current SoA, even being the selected training set for the USE algorithm [21].

The evaluation of semantical similarity has made significant progress in obtaining data for evaluating different documents. However, except for Rubenstein and Goodenough's first implementation, every database is focused on analyzing the structure and similarity between complete sentences since their structures and methodologies require the total sum of each sentence token (word) to get a similarity score, as described by equation (2.2). When evaluating this formula, it is evident that each sentence's length is a critical point for getting the final embedding, and hence similarity score. Even though these databases apport in the development of NLP, they are ignoring the fact that a single sentence can also be explained with a single word. So, following the way of thought of Joos [50] and the experiments of Rubenstein and Goodenough with pairs of words, it is possible to detect a general similitude score between the general context of a complete sentence and a particular token or term. However, it is necessary to evaluate how a given population analyzes such pairs to get a benchmark value for comparing future algorithms.

2.6.2 Co-occurrence Vectors

One of the first algorithms that created a model for representing a word in a multidimensional space is GloVe, designed by Pennington et al. [80]. This algorithm uses a weighted least-squares method for training the co-occurrence of words in a vast corpus. They use the co-occurrence matrix of LSA to determine the probability of a word k appearing after a word i and before a word j, as shown in equation (2.23). Another advantage of this implementation in comparison with LSA is that it manipulates the weighting function to create a more significant embedding for common words and practically ignores rare tokens that are seldom used. The resulting embedding of words creates a more specified semantical space for locating different concepts. At the same time, each dimension in this space refers to a particular context, making it possible to use traditional arithmetic operations to determine new terms based on the context of others. One example is shown in Figure 2.21, where $(king-man)+woman = queen$.

$$F(w_i, w_j, w_k) = \frac{P_{ik}}{P_{jk}} \tag{2.23}$$

Since every word in a corpus is located in a multidimensional space, their corresponding

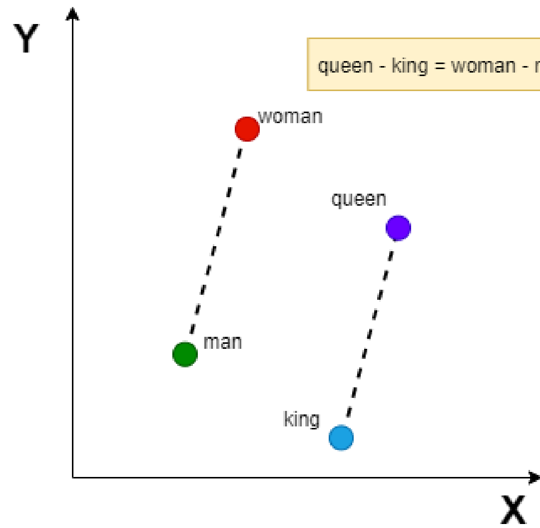


Figure 2.21: Word representation and selection with GloVe [80]

embeddings are numerical vectors that locate the exact location of each term in space. Following the precepts of vectorial spaces, the distance between two points indicates how similar they are, being a distance of zero the same point in space, hence, the same concept. The most practical way of determining that distance is through cosine similarity, shown in equation (2.24). This formula measures the similarity between two vectors and their inner product to determine the cosine angle [28]. The angle indicates the percentage of similarity depending on how close those vectors are, as shown in Figure 2.22.

$$similarity = \cos\theta = \frac{A \cdot B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{2.24}$$

Nowadays, SoA algorithms are robust enough to produce their word embeddings based on a vast training set, such as the Transformer algorithms mentioned before. In their case, the initial word embedding starts with a random value and is updated through supervised learning to get closer to their specific probability of occurrence after a word or phrase. However, the relationship of vectors is still evaluated through cosine similarity due to the practicality of linear algebra.

There are other novel implementations for evaluating the semantical similitude between embeddings, such as the one presented by Pawar and Mago [79]. This approach uses the WordNet database's semantical and hierarchical representation to determine the similarity between two words. Another implementation is by representing the GloVe embeddings on hyperbolic space, as shown in [95]. In this case, the Euclidean space of GloVe is moved into a Poincaré disk model, following the structure mentioned in [76, 87]. However, by evaluating the similarity inside a hyperbolic or hierarchical space, the model's accuracy decreases the longer the distance between two concepts. In that case, the most practical representation would be to use the hyperbolic representation for jus word selection and then return the

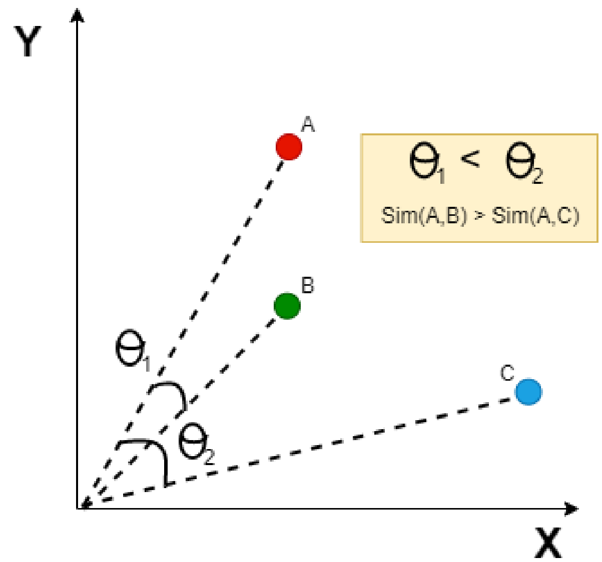


Figure 2.22: Representation of the cosine distance between words

embedding into a Euclidean space to evaluate through cosine similarity.

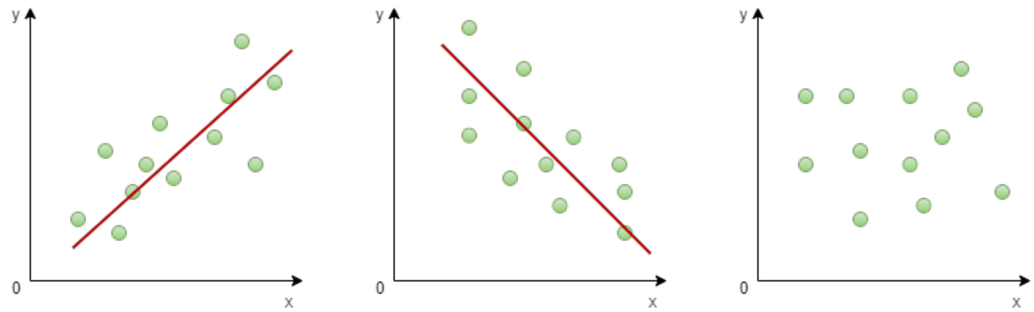
2.6.3 Pearson Correlation and P-score

Even though there are particular metrics for measuring the accuracy of a given method, it is still necessary to evaluate and compare every algorithm and model with other SoA implementations. In the NLP field (and generally in almost every ML research), one of the most influential and standard metrics for comparing different algorithms is Pearson's correlation r. The correlation refers to how strong is the relationship between two variables x and y is.

So, given n samples of data $(x_1, y_1 \ldots, (x_n, y_n))$, the correlation r of these samples would be defined as shown in equation (2.25), where the difference between each variable and its mean value determine a positive or negative relationship [30].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.25}$$

The relationship of these two variables could have three tendencies, as shown in Figure 2.23. Considering that equation (2.25) has a range of $-1 \leq r \leq 1$, variables x and y will have a positive relationship when the lineal relationship between them follows a positive grade, as shown in Figure 2.23a. In this situation, it can be said that x and y have a similar behavior under the same circumstances. When the relationship has a descending relationship, like in Figure 2.23b, both variables have the exact opposite relationship. On the other hand, a correlation value of zero, like in Figure 2.23c, demonstrates that those variables have nothing in



(a) (b) (c)

Figure 2.23: Correlation of two variables x and y with (a) positive, (b)negative, and (c)zero relationship

common, and each of them acts separately [30].

Besides the correlation metric, it is also necessary to evaluate how probable it is to get the opposite behavior under the same circumstances. This metric is known as the *p-value* and consists of evaluating the expected value of a test $E(Z)$, as described in equation (2.26). The difference between the alternative value of the sample $\hat{p}$ and the null value of the same test p_0 follow a Normal distribution that determines the probability of getting opposite results [30].

$$E(Z) = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \tag{2.26}$$

The combination of these two metrics demonstrates the feasibility of the results of a given method compared to other SoA implementations. The closer the correlation to 1, and the lower the *p-value*, the results of the proposed method will be more believable.

Chapter 3

Explanation Structure

3.1 What is an Explanation?

Explanations are phrases used to generate an understanding of a particular theme and are used regularly in the scientific and daily lives of everyone. The main objective of explanations is to accommodate novel information within previous beliefs to generalize [62] an idea and create an understanding to facilitate category learning [73] and foster conceptual coherence [74, 78].

Two components build an explanation: a description of the phenomenon that is being explained, or *explanandum*, and the logical relations made to make sense of the explained concept, also known as *explanans* [43]. In fewer words, an explanation consists of an illocutionary act [6] (a description of what is being explained), and a perlocutionary act [6] (how that explanation is produced), respectively. To make sure that the effect of explaining is helpful, it is necessary a strong relation between the *explanandum* and the *explanans*. In order for the final user to understand an explanatory action, certain conditions must be met [3]:

- The user needs to connect the parts of the explanatory act with others that the user previously knew.
- The explaining act needs to be coherent with the user's beliefs.
- The user needs to know that the explaining act answers a particular question.
- The level of puzzlement of the user is reduced after given the explaining act.

However, it is subjective to explain something [43]. So the subject must generate multiple inferences based on the given evidence and generate a conclusion of accepting or not the explaining act. This principle is known as Inference to the Best Explanation (IBE) [59].

This principle evaluates three conceptions of understanding for elaborating an effective explaining act [59]. First, it considers the premise that it is necessary more than knowledge for understanding something; in other words, just describing the phenomenon is not enough for a person to understand. The second criteria declare that an explanatory act is enough by itself, assuring that the answer to any question should only provide enough information without worrying about further questions that may arise after the explanation. Finally, and the main reason why an explanation is considered subjective is that the *explanandum* evokes our reasoning for believing that the explanation is correct or not.

IBE parts from the idea that the explanation act derives from giving information about the causes [59] implied to the *explanandum*. This premise is based on the idea that causes make the difference between the phenomenon occurring or not [59]. At the same time, IBE uses the concept of self-evidencing, assuring that the observations given to support the hypothesis of the explanation are valid because the hypothesis support such given observations. So, in conclusion, it is by evaluating different hypotheses and evidence that people can determine which of them are accepted as correct.

Until now, it has been mentioned what constitutes an explanation in general, where the *explanans* is a task solely for the person that receives the explanation since he is the one that generates the relations that prove correct or not. Hence, it is the responsibility of the one generating the *explananandum* to produce the best hypothesis to ensure a proper *explanans*. For this reason, it is necessary to evaluate how the explaining act should be developed.

3.2 The Explanatory Act

An explanatory act could be described, according to [3], as the action where a subject S explains a question q by uttering a sentence u, only if the following condition is met [2]:

Theorem 3.2.1 *S utters u with the intention that his utterance renders q understandable.*

At the same time, it needs to be considered that any explanatory act should only be done when the formed hypothesis answers the main question Q of the explanation [2]:

Theorem 3.2.2 *S believes that u expresses a correct answer to the main question Q, which presupposes q.*

Another aspect related to the previous theorem evaluates that the explanatory act needs to give a sufficient hypothesis to prove it correct or not, without making allegiance to explaining an utterance u to generate a secondary utterance u' that answers Q [2]:

Theorem 3.2.3 *S utters u with the intention that his utterance renders q understandable solely by producing the recognition that u expresses a correct answer to Q.*

Finally, the explanatory act should not be considered a mere sentence but instead a proposition p. So, it is not important which sentence is used for the explaining act, as long as the preposition correctly answers to the main question [2]:

Theorem 3.2.4 *If p is the explanation of q given by S, then it is a correct explanation if and only if p is a correct answer to Q.*

As observed, every theorem is included by its predecessor. The final premise concludes that it is not important how it is intended to explain something, as long as the general concept or idea, the *explanandum* per se, is formulated correctly to allow the person receiving the explanation to produce a proper *explanans* and understand the hypothesis.

3.3 Explaining Unknown Words

When dealing with an unknown term, two options could be the cause of the confusion. First, the word could be entirely new for the user, so a definition should be given to him to make him understand what the document is talking about. On the other hand, the definition or idea may be somewhat understood, but it is still necessary to explain the definition of the concept. According to the Cambridge Dictionary [81], a definition is a ”statement that explains the meaning of a word or phrase,” while an explanation consists of ”the details or reasons that someone gives to make something clear or easy to understand”.

In both cases, when dealing with an unknown word, the main idea is to utter a phrase that contains a similar meaning to the term but using a different vocabulary. In other words, it is necessary to express a synonymy of the concept differently from the one currently presented to the user. The following sections will talk about the structure of the conceptual knowledge used for creating definitions and meaning and the three primary paradigmatic relationships that form those meanings.

3.3.1 Conceptual Relationships between Words

According to [75], the meaning of a word is dependent on all the words that are related to it. However, such meanings are not formed by those words but among their corresponding concepts instead. It is the responsibility of the person to transform inferences into conceptual knowledge. The conceptual knowledge of a person is generated through the application of arbitrary representations and rules of previously known concepts to generate new ones [75]. Such representations are based on mental inferences produced by *ad hoc* relationships that the person creates to conceptualize the characteristics of any word concept [8].

Nevertheless, it is still necessary for the person to have enough information for producing those inferences since ambiguous words cannot be disambiguated without knowledge of the extensions of other words [75]. That information focuses on the meaning postulates of the words used to define the main concept to create the relationships between the extensions of each word [34]. Even though the mental inferences are made between the abstract meaning of each word, the meaning components of each term are highly related to several semantic characteristics [49].

Conceptual information of words is highly related to paradigmatic semantic relations, such as phonetic, morphological, and morpho-syntactic relationships [75]. Those relations can be classified into three main groups, also illustrated in Figure 3.1:

- Synonymy. Evaluates the knowledge about the similarity between words.
- Hyponymy - Hypernymy. Hierarchical relationship between a general word and its nearer, more particular words.
- Meronymy - Holonymy. Hierarchical relationship between a whole word and the ones mentioning its corresponding parts.

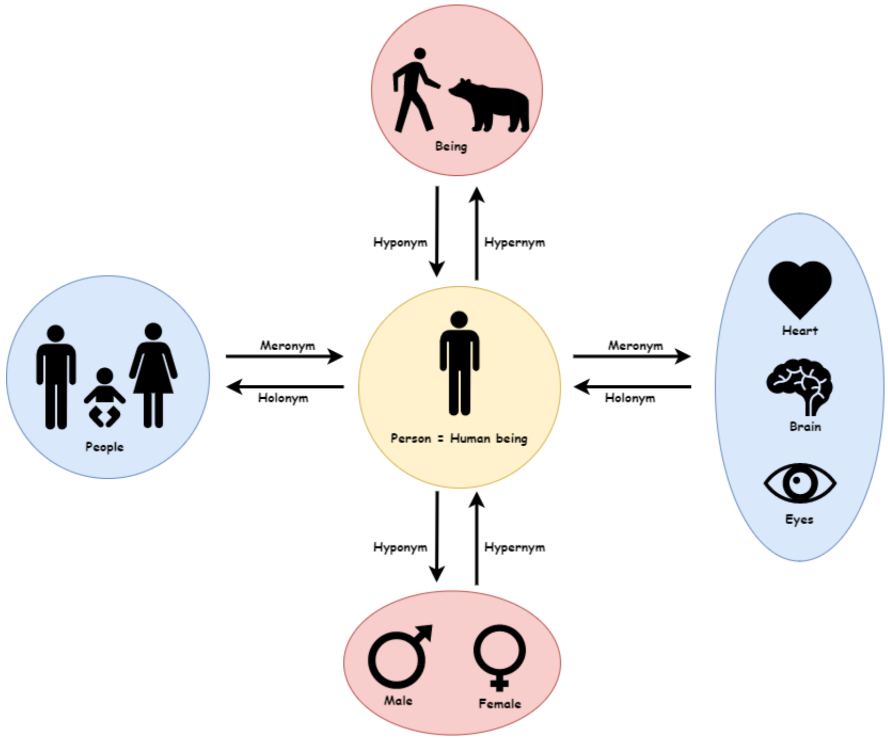


Figure 3.1: Paradigmatic semantic relations of the word *person*.

3.3.2 Synonymy of Concepts

Synonymy between words is the most important factor when generating definitions and explanations of unknown terms. The reason behind this statement is that synonymy exists when two concepts share a similar meaning. Murphy [75] denotes this phenomenon as Relation by Contrast-Synonymy (RC-S), or "A synonym set that includes only words or concepts that have all the same contextually relevant properties, but differ in form." The similarity between words occurs when they share more common attributes, or at least they match closely enough to be considered similar based on the context implied. Furthermore, the similitude given by the person will depend on the level of specificity of the relevant properties that are mentioned [75].

Considering that the synonymy of words depends only on how specific the used terms are, it is possible to assume that the similarity is bounded between a pair of words and between a word and particular sentences. Given an expression E, there exists a conceptual structure (CS) that expresses the message of E, a semantic form ($SF(E)$) that expresses the contribution of E to explaining more complex expressions, and a conceptual context Ct that represents the overall representation of CS under particular circumstances [13]. There are four cases of synonymy between two expressions A and B, according to the studies shown in [13] and considering those previously mentioned structures.

The first case represents the ideal case of synonymy, where two expressions SF are completely identical in every context. This situation is complicated to achieve, especially when A and B are complex expressions. However, [13] uses a simpler example to help visualize this situation:

1. (a) enter the garden
   (b) go into the garden

In both sentences, the words *enter* and *go into* share similar lexical entries, denoting the same intentions on both sentences. Hence, this case of synonymy can be expressed as:

Theorem 3.3.1 *$SF(A)$ is identical to $SF(B)$.*

The second case is a more general expression of similitude between $SF(A)$ and $SF(B)$, where both expressions are different, but their SF are equivalent. Considering the sentences

2. (a) John is not tall enough (to be in the team)
   (b) John is too short (to be in the team)

In this situation, both sentences express the same thought, but the contradictory vocabulary between them modifies their respective SF. However, considering that both sentences are using as context the inclusion of John into the team, the meaning of both sentences is similar enough to be considered synonyms. Hence, the second case of synonymy can be expressed as:

Theorem 3.3.2 *$SF(A)$ is equivalent to $SF(B)$.*

A more abstract type of synonym occurs when there are two expressions completely different, but utter the same condition or criteria. For example, consider the sentences:

3. (a) Draw a square around the box!
   (b) Draw a box around the square!

In this situation, the SF of both *box* and *square* is completely different. However, the overall expressions express the same idea, making their corresponding CS identical. This reasoning is expressed as:

Theorem 3.3.3 *$CS(A)$ is identical to $CS(B)$ relative to Ct.*

Finally, the fourth type of synonymy is highly related to the ad hoc relationships mentioned by [8] and the fact that every person is capable of generating proper inferences based on their particular knowledge. Given the next sentences as an example:

4. (a) We just talked about the president
   (b) We just talked about John's brother

By evaluating those two expressions, it could be assumed that they do not share a similar SF or CS, making them completely unrelated. Nevertheless, if by some chance a person posses the information that the president is also John's brother, then both sentences become synonyms because they share the same conceptualization. Hence, the fourth type of synonymy can be expressed as follows:

Theorem 3.3.4 *A and B are identical in reference (or denotation).*

Considering that the explanations and definitions of unknown words depend on every person's conceptual knowledge and inference processes, it is more likely to generate better explanations of unknown terms based on the fourth type of synonymy. In other words, it is necessary for the expressions used to explain or define an unknown term to be identical to the conceptualization of every person's conceptual knowledge. At the same time, it is complicated to achieve sentences containing the same SF and CS, so the only way to achieve synonymy between expressions is by using contextually equivalent terms. Those equivalences are divided into hyponymy and meronymy relationships.

3.3.3 Hyponymy and Meronymy Relationships

The hyponymy-hyperonymy relationship is defined as an *IS-A* [86] or *IS-A-MEMBER-OF* [54] function. As the name suggests, this relation refers to a hierarchical accommodation between concepts, where the more specific term is a member of the more general one (hyperonym) and vice versa (hyponym). Observing Figure 3.1, it can also be observed that it is an asymmetric relationship [75], where given two words, p and q, respectively:

$$p \text{ is a hyperonym of } q \longleftrightarrow q \text{ is a hyponym of } p \tag{3.1}$$

As well as the hyponymy-hyperonymy relationship, the meronymy-holonymy function follows a hierarchical, asymmetric relation that follows (3.2); however, this function follows an *IS-A-PART-OF* or *HAS-A* relation [103]. In this case, the general concept has a series of particular terms (meronym), while the particular ones are part of something more complex (holonym).

$$p \text{ is a meronym of } q \longleftrightarrow q \text{ is a holonym of } p \tag{3.2}$$

Using these two semantic functions makes it possible to talk about every aspect related to an ambiguous term or concept. The advantage of this approach is that every word has a direct hyponym or meronym, while the second word has direct functions of them both. By implying that a term A has a direct hyponym B, which also has a direct hyponym C, it is accurate to say that A is related to C as an indirect hyponym well. These concatenations of different terms allow finding deeper relationships between concepts, forming a complex linguistic net structure that unites every term.

All of these concepts have been studied to determine how explanations linguistically operate when analyzing natural language. However, this same study and analysis allowed the development of a research area in ML that tries to analyze, comprehend, and apply the use of human linguistic behavior using computational analysis.

Chapter 4

Proposal’s Methodology

The use of NLP allows evaluating the written structure of a human being quantitative, making it possible to create new documents, paraphrase them, copy the writing style of a given author, or even evaluate how similar grammatically or semantically two sentences are. However, current SoA algorithms are focused on the production of whole sentences and text documents. It is still necessary to evaluate how a series of phrases and words can explain one key concept. Hence, this thesis proposal intends to use current NLP algorithms, such as Transformer Neural Network models, to evaluate the similitude between a descriptive phrase and its key concept.

The novelty of this approach resides in the actual evaluation of a single term against a complete phrase. SoA validation methods for text similarity (both knowledge and corpus-based) have effectively proven how similar the general context of two phrases are; unfortunately, all these methods require the comparison of every single word in the phrase actually to produce a similarity value. The current problem is that the general sum of the phrase’s words is vital to create a closer similitude value with the other sentence. Hence, when evaluating a single word’s general meaning or context, the results lack understanding of how a given sentence refers to a single term.

Figure 4.1 shows the general behavior of the thesis proposal. It parts from the widely used dictionary model, where a particular key concept and its related descriptive phrase share the same meaning. So, given any text document, the idea is to separate the phrase intended to interpret and the key concept that defines it. These two textual data will be used in two stages: a text generation module and a validation.

The generation module will input the descriptive phrase, while the key concept will function as the ground-truth value. Through the use of a GPT-2 Transformer neural network model, the data will finetune the network to determine that they share the same context. The model’s output will consist of a new phrase that interprets the context of the descriptive one with the usage of new vocabulary to talk about the same context of the ground-truth value.

The validation module will work with both the original key concept and the GPT-2 ’s generated phrase. This block will determine how similar the general context of the produced text and the ground-truth value are. This block needs to be finetuned to evaluate the similitude

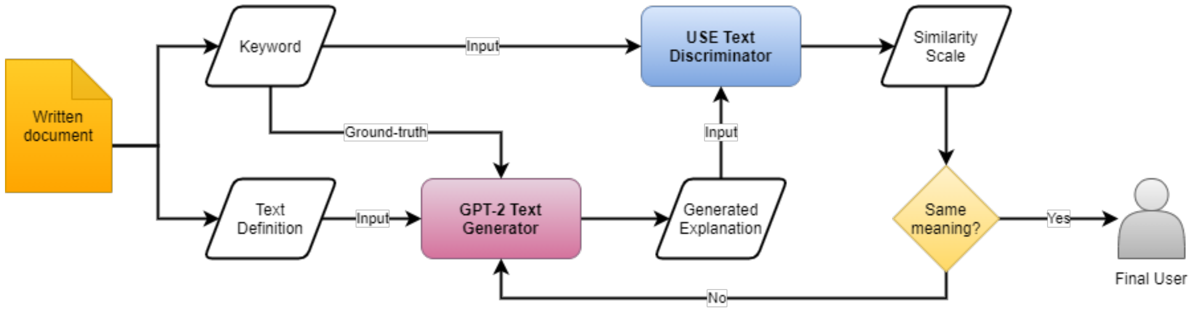


Figure 4.1: Thesis proposal general structure.

between a single context and a complete sentence like the generation module. An acuse model will be re-trained to evaluate this new similitude metric. As a result, the USE model will determine the semantical similitude between the two input texts, and if the score reaches a given similitude benchmark, the newly generated text can be given to the user in order to describe the unknown concept. If the benchmark is not reached, it will be necessary to create a new sentence in the generation block, repeating the validation process until a successful phrase is created.

The following sections will describe how each module was designed, including the training and testing data management, the finetuning process, and the different validation tests used.

4.1 Text Generation Procedure

This block is in charge of evaluating a descriptive sentence as input and producing a new phrase or selecting a particular word (in the best case scenario) that reassembles the general context of the input and the ground-truth concept being described. So, for example, if the model received the definition of ”a building where people can live,” the network should be able to determine that the text is referring to ”a house.” It was necessary to select a database with enough definition-concept pairs and obtain a way of numerically representing the relationship between both parts to ensure the neural network produces such results. Figure 4.2 shows the general structure of the generator block.

4.1.1 Dataset Preprocessing

The database was divided into two sections: one for the general training of the model and a second one for the testing phase of the proposal. The training dataset was built with the WordNet [70] database, selecting 206,978 random pairs of definition-concept texts. Those pairs were preprocessed by converting each letter into its lower case version. Since the idea was for the network to analyze the different textual ideas presented in each definition, the punctuation signs such as points, commas, among others, remained in the input data. Each letter was converted into its lower case version for the ground-truth (the word or concept of the given

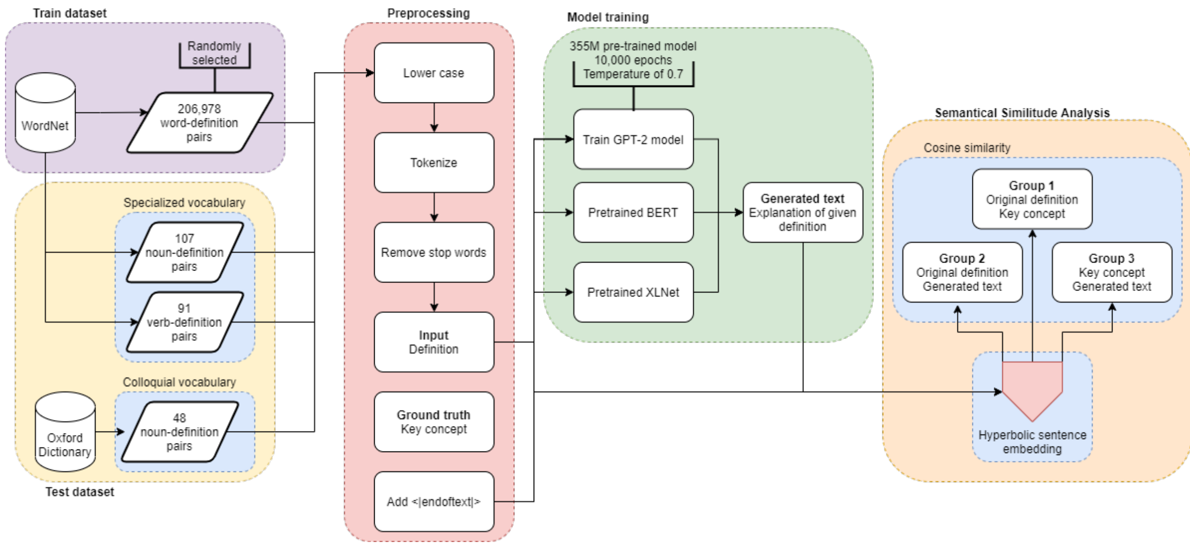


Figure 4.2: Generator block structure.

definition); additionally, to specify the ending of the ground-truth value, an <|endoftext|> token was used. It is essential to mention that this final token was only used for the specifications of the neural network training model, and the final result considered from the network consisted of the text generated before it. An example of how the resulting training data was implemented is shown in Table 4.1.

The testing dataset of the model consisted of a total of 246 definition-concept pairs, divided into three blocks: 48 of the pair texts were selected from the synonymy study of Rubenstein and Goodenough [85]; even though the test only evaluated the similitude between single words, the defining phrase of each term was obtained from Oxford dictionary [31]. In this way, it was possible to apply the same evaluation logic but compare a single word with its definition. These pairs were selected to evaluate the network's performance using more colloquial and straightforward use of words; for the rest of this paper, these pairs are referred to as the colloquial language dataset.

The other 198 pairs came from the WordNet database since the vocabulary used for the definitions was more specialized than those in the colloquial dataset. From these pairs, 107 of them defined nouns only, while the other 91 pairs consisted of verb-only pairs. For the rest of the document, these pairs of words are referred to as the specialized dataset. Table 4.2 shows an example of each of the three different datasets used to test the model. The test dataset is fully shown in Table A.1 of Appendix A.

4.1.2 Extended Hyperbolyc Representation

As mentioned in Chapter 2, when implementing a NLP algorithm, it is necessary to give a numerical representation to every word of the vocabulary. The method can only assimilate the

Table 4.1: Examples of the training dataset

Input text	Ground-truth text
existing in abundance	abounding $< \vert endoftext \vert >$
existing in abundance	galore $< \vert endoftext \vert >$
not recognized	unrecognized $< \vert endoftext \vert >$
lacking activity; lying idle or unused	inactive $< \vert endoftext \vert >$

Table 4.2: Examples of the testing dataset

	Input Text	Ground-truth text
Colloquial dataset	the land next to the sea; seashore	coast $< \vert endoftext \vert >$
Noun specialized dataset	the smallest administrative district of several European countries	commune $< \vert endoftext \vert >$
Verb specialized dataset	make an effort or attempt	try $< \vert endoftext \vert >$

relationship between each token and create the corresponding sentence, based on the numerical probability of which word should follow in the phrase. There are two ways of representing a word in a multidimensional space, each of them with its advantages and disadvantages: corpus-based and knowledge-based.

For this proposal, a new method called EHR is presented. The objective of this procedure is to combine the advantages of the two existing representation methods: the syntactical analysis of xLSA [92] based on the corpus-based representation, and the use of hyperbolic word embeddings [76] for representing on a general context (knowledge-based) each vocable of the used datasets. Figure 4.3 show the general diagram of how these two implementations were combined in order to get the general sentence embedding of the generated text.

The first step would be to preprocess the given sentence by tokenizing it and converting every word into its lower-case version. The following step would be to remove every punctuation sign and stop word since these characters do not prove much information for word selection.

The algorithm searched each token inside two datasets depending on the POS it belonged: noun or verb. Each token was searched first on the noun dataset to retrieve its corresponding 10-dimensional embedding; if it was not found on the noun dataset, the token was searched on the verb dataset. In the given case that the token was still not found, a second preprocessing procedure was carried.

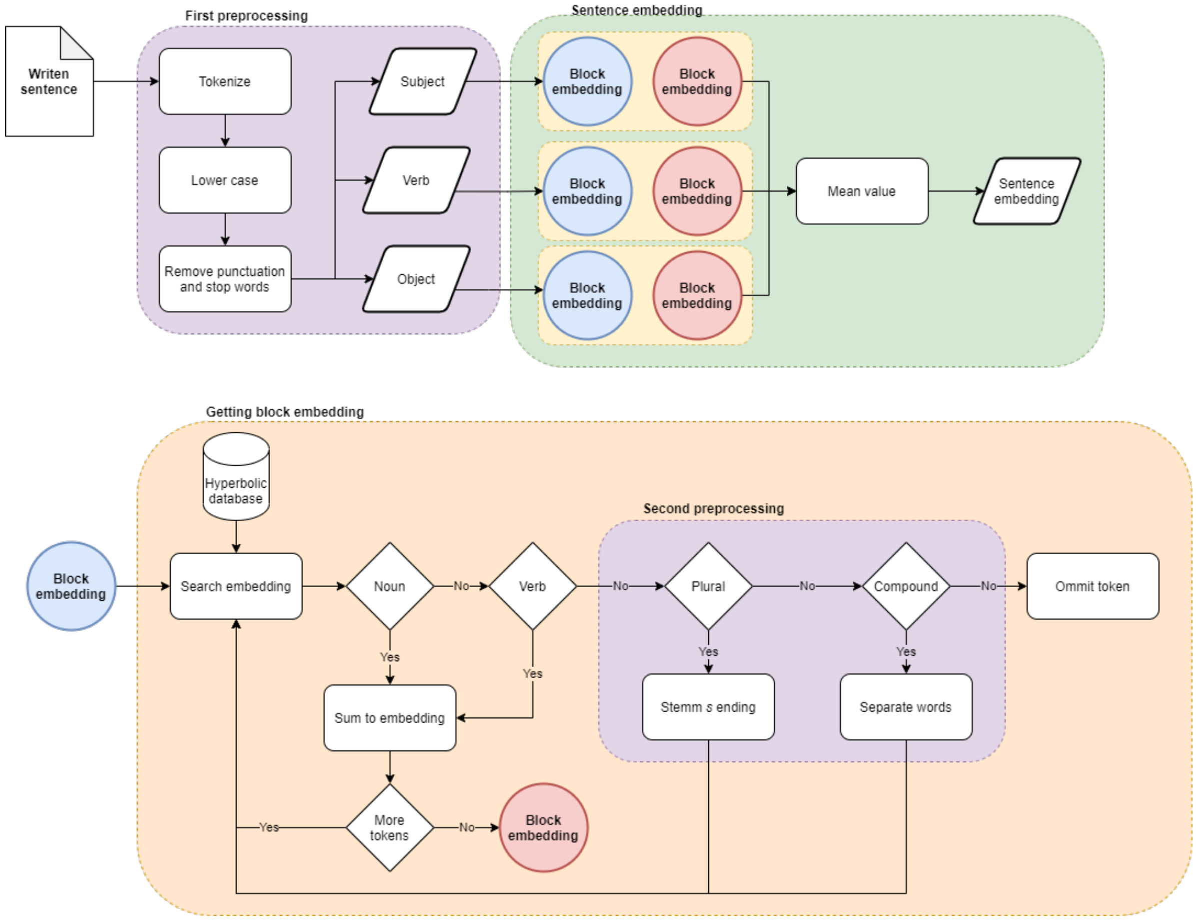


Figure 4.3: Methodology for the hyperbolic representation of a sentence

The token was evaluated to analyze if it ended on *s*, assuming it was a plural token. Under these circumstances, the token was stemmed from that character and returned to the embedding search. When the word was not detected as a plural term, the algorithm evaluated if it was a compound word (multiple terms in the same token) by checking for underscores and separating them; each new token was individually searched on the datasets. Finally, the token was ignored from the embedding search if none of the second preprocessing blocks helped assign an embedding.

Once every token in the sentence had its corresponding embedding, the sentence was divided into three groups: subject, verb, and object, as shown in Figure4.4; the subject block had all the vocabulary that appeared before the main verb of the sentence, and the object the ones that appeared after. Each block had its embedding from the total sum of each token representation, as described in equation (2.2). The general sentence embedding was the mean value of the three blocks.

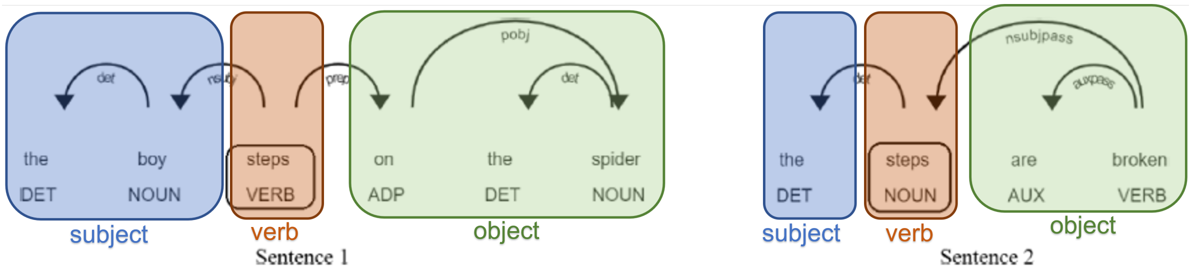


Figure 4.4: Sentence analysis with EHR

4.1.3 Model Training

The GPT-2 model has already demonstrated its capabilities for general text generation [83]. Also, it is quite advantageous to use the available pre-trained models as a starting point for this proposal. This model can analyze and generate long texts while following the grammatical structure required for a coherent document. Hence, the best approach would be to finetune this model to learn the grammatical structure of the dictionary model described before. In this way, it is possible to indicate that the descriptive sentence is the same as the ground-truth keyword.

There is currently four different sizes of the GPT-2 model. Each of them corresponds to the number of parameters of the overall architecture. These sizes are $117M$, $345M$, $762M$, and $1542M$ parameters; the bigger the number of parameters, the better the accuracy. However, it is also necessary for a more powerful computation for the finetuning process. Due to the particular equipment available, this proposal used the $345M$ model.

The general hyperparameters used for this model are shown in Table 4.3. The model was trained for 10,000 epochs with a learning rate of $1e-4$. The model was saved every 500 epochs and after finishing training. Regularly there are two forms for evaluating the training of a neural network: by the output's accuracy and the general loss function of the model. The selection of the accuracy method would imply that the model's purpose is to generate precisely the ground-truth concept, which could suggest that the model only tries to "memorize" the relationship between a phrase and its key concept. On the other hand, the evaluation through the loss function allows us to detect if the generated texts are similar to the ground-truth but not the same vocable. This other approach is preferred since the general objective of the model is to create multiple texts that interpret the original descriptive text. By doing so, it is possible to create multiple combinations of vocabulary and sentences that try to explain the same key concept in different ways.

For the current generation of the text, it was necessary to input the original descriptive sentence followed by two dots. This structure allowed the GPT-2 model to analyze the complete original sentence, to then create a different interpretation of it. The generated text was limited to a total length of 10 words since the objective is to create the smallest text possible. The model's results were created by selecting different vocabulary, controlled by a temperature hyperparameter; the bigger the temperature (maximum 1.0), the more random the

Table 4.3: GPT-2 hyperparameters

Hyperparameter	**Value**
Model size	345M
Learning rate	1e-4
Epochs	10,000
Output length	10
Temperature	0.7
top_k	40
top_p	0.9

vocabulary was selected. For this purpose, the standard value of 0.7 was used. The selection between the different guesses or sentences the model could create was achieved by selecting the best sentence that reached a cumulative probability of $top\_p = 0.9$ of being similar to the original text and was among the $top\_k = 40$ best-generated options. Finally, just like the model's training data, the generated sentence was finished once the $<|$endoftext$|>$ token was used.

4.1.4 Model Evaluation

The measurement of the semantical similitude between the generated text of the model and the ground truth value was made in three stages. The first step consisted of preprocessing both the original definition and the generated text of the model; it was not necessary for the ground-truth text since it was just one word for every case. The preprocessing consisted of tokenizing the texts and removing punctuation signs and stop words. Since the model was trained to work with lower case words, it was unnecessary to repeat this process.

Once the texts were preprocessed, every text (including the ground-truth words) was interpreted into its respective word embedding using the 10-dimensional hyperbolic knowledge-based dataset of [76]. The resulting text embedding t_e of each text consisted of the sum of the embedding of the i words w_i that it had, as described by equation (2.2). When the text had only one word, the value of t_e would be the same embedding as the one in the dataset. In the remote case that a word w_i had no hyperbolic embedding representation in the dataset D, the word was omitted.

The embeddings of each representation were used to measure the cosine similarity between each text, generating three different values: one comparing the ground-truth text with the original definition, a second value for comparing the semantical similitude between the generated text of the model with the original definition, and a third one comparing the semantical similitude between the ground-truth and the generated texts. The cosine similitude was obtained by comparing the two embeddings A and B through the ten dimensions of each embedding, as described in equation (2.24), where two words are semantically identical if the distance (in this case, the cosine similarity) between each word embedding is zero, and completely different if the value reaches a value of one.

Table 4.4: BERT and XL-Net architectures

	Cased	Layers	Hidden neurons
BERT	Uncased	24	1024
XL-Net	Cased	12	768

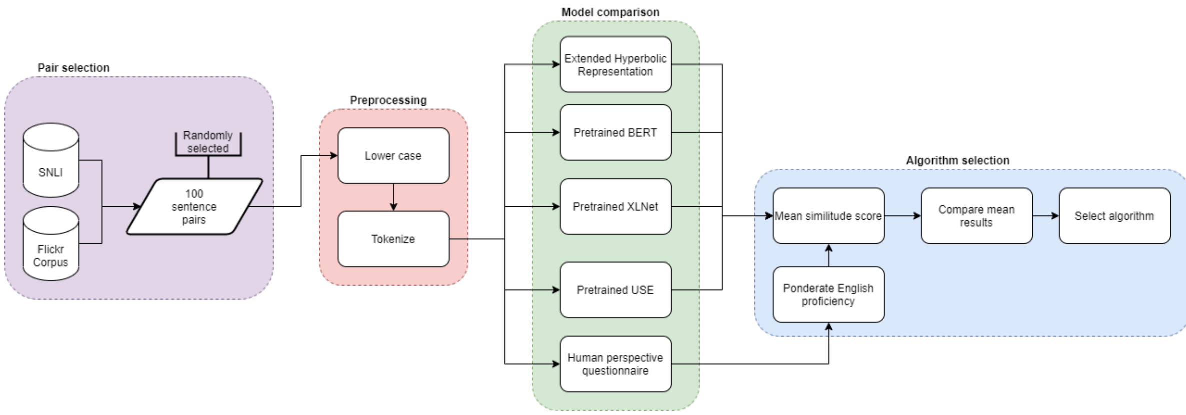


Figure 4.5: Similarity evaluation structure.

The performance of the model was compared with two other SoA algorithms currently used for text generation: BERT [29, 97], and XL-Net [108]. For these two algorithms, the pre-trained models were used during the comparison; the general structure of each model is shown in Table 4.4. According to its training, the dataset usage was the same as before; only the embedding generation changed from each model.

4.2 Selection of Algorithm for Similarity Evaluation

Just knowing the cosine similarity between the generated texts of each method is not enough to determine the algorithm's accuracy compared to the human perspective. There is no use if the resulting phrase has a cosine similitude of 0.0 if it changes the general context of the key concept wholly or it is the same keyword. The first case would suggest that the model is not capable of evaluating the similitude between the given texts (resulting in an undertraining problem), while the second one would result in the complete opposite: an overtraining that only memorizes the relationship between phrases and key concepts, without being capable of interpreting and changing the vocabulary at will.

The algorithm selection for the validation block came from the comparison of four different algorithms and the general human perspective resulting from a given questionnaire. Figure 4.5 describes the complete process for this purpose.

4.2.1 Dataset Preprocessing

Two different databases were used during this phase for the evaluation of the algorithms: the SNLI [16] and the Flickr [109] corpus. These databases were selected since the two of them evaluate the similitude of multiple short sentences. Even though none of them has a quantitative value for determining such similitude, they have separate ways to assume so. The SNLI corpus uses the two sentences as input, and a category as a ground-truth value between them, classifying them into entailment, contradiction, and neutral sentences; considering the research purposes of this proposal, only the entailment pair of sentences were used in this phase. The Flickr corpus does not have a categorical classification of its sentences, but each group of sentences shares a common image that is explained by each phrase; considering that every group sentence talks about the same image, it is assumed for this proposal that each of them share the same general context.

The data from the two repositories were combined, and from them, a total of 100 pairs of sentences were randomly selected. In the case of the sentences from the Flickr corpus, only the sentences were selected; no context image was used for the rest of the evaluation. Since the objective of the similitude evaluation was to observe the overall similitude of a complete phrase, the data preprocessing just consisted of turning every letter into its lower case and tokenizing every sentence. The similitude metrics of the implemented algorithms required no finalization token, so the $<|$endoftext$|>$ character was omitted as well.

At the same time, a 10-Likert questionnaire was made to evaluate the similitude of the sentences regarding the human perspective. In this way, every person could give a numerical value of similitude between each pair of sentences; a 10 score represented that the two sentences had the same general context or meaning, while a 0 score represented no relationship in their meanings. In addition to the sentence evaluation, each surveyed person indicated their English proficiency before completing the test with a 5-Likert scale (5 being wholly proficient and 0 not proficient at all). Depending on their proficiency level x, each subject's scores s were ponderated according to equation (4.1). If English was the native language of a subject, a proficiency score of 5 was implemented. The complete list of sentence pairs is shown in Table A.2 of Appendix A.

$$similitude_{human} = (0.2x)(0.1s) \tag{4.1}$$

4.2.2 Model Training

Besides the questionnaire, a total of four different algorithms were used to evaluate the semantical similitude of the 100 pairs of sentences: BERT [29, 97], XL-Net [108], USE [21], and EHR. The first two algorithms followed the same structure mentioned in Table 4.4; the USE model was implemented with the structure of the pretrained model mentioned in [21], while the overall structure of EHR is mentioned in detail in Chapter 2.

The similitude validation of EHR was made through the use of hyperbolic representations, as well as applying equation (2.2). For the case of the other three algorithms, each of them produced its word and sentence embeddings. At the end of each text generation, every algorithm was measured using the cosine similitude of equation (2.24).

4.2.3 Model Evaluation

The main objective of this block was to determine which algorithm had the best results compared with the human perspective. Once every model generated its corresponding results, an average similitude score was obtained from every one of them. It is essential to mention that in this case, the highest mean similitude score was not necessarily the best outcome, which seems to be a common practice when comparing algorithms of NLP analysis. The mean human similitude score would function as a reference point to determine which algorithm is capable of evaluating sentence similarity correctly and that their results were closer or similar to the ones of the human perspective.

It is necessary to evaluate the human perspective since the proposal's final user is a person; hence, it is logical to assume that the data processing must follow their perspectives to generate comprehensive results. The algorithm that showed the best correlation with the questionnaire results will be used in the validation block of the general algorithm.

4.3 Validation of Text Similarity

Even though cosine similitude was used to compare the results of the different methods, it is still impossible to determine the correlation between their corresponding accuracies and the human perspective. The lack of a benchmark is another current problem inside the acstoa since the current methodologies are only capable of evaluating the general context between long texts or sentences, but not between a single concept and its corresponding descriptive phrase. Because of this problem, it is necessary to compare different algorithms to analyze which correlates better with the human perspective.

For this reason, the validation block intends to evaluate the semantical similitude between a descriptive phrase and a particular term. In this way, it is possible to evaluate if the generative block creates understandable phrases without modifying the original context of the key concept being interpreted. Figure 4.6 shows the design description of this current block.

4.3.1 Dataset Preprocessing

For evaluating the semantical similitude between texts, one of the best databases is the SNLI corpus [20]. This compendium of phrases follows a similar structure to the previous section's questionnaire: a series of sentence pairs were given to a particular population, and they had to evaluate how similar their meaning was between each other. The database shows the two evaluated sentences and a 5-Likert numerical similitude score, 5 for expressing the same context or meaning, and 0 that there is no relation between the sentences. For the best performance of this proposal, only the sentences with a score bigger or equal to 4.5 were selected from the train, dev, and test datasets. This selection resulted in a total of 789 pairs of sentences similar

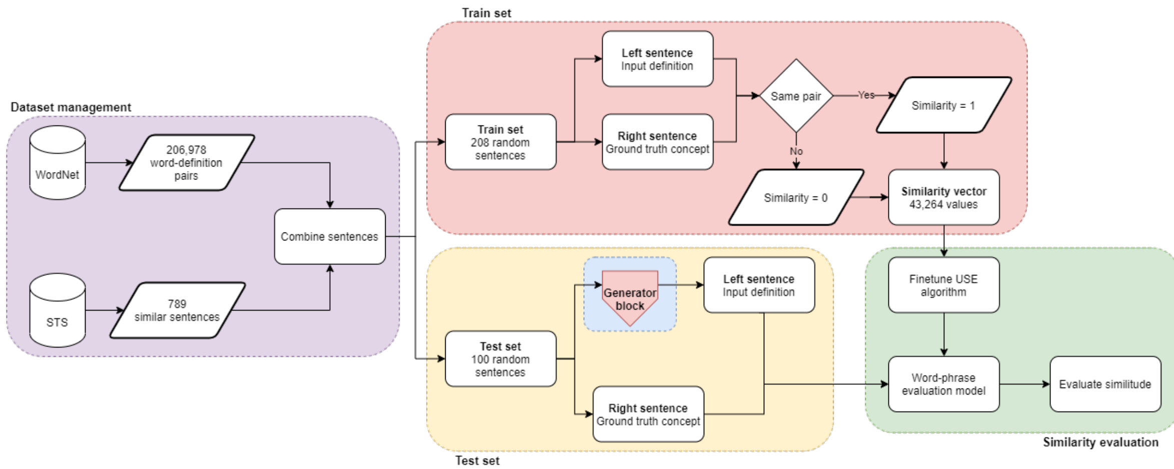


Figure 4.6: Validation block structure.

in context.

As mentioned before, one of the problems this dataset has (as well as others in the SoA) is that they evaluate complete sentences. Since the proposal's objective is to evaluate the similarity between the descriptive phrase and a unique word, it is necessary to use a database to determine such relationships. In this case, the WordNet database achieves the goal since their text relationships are the key concept-descriptive text pair. The 206,978 random pairs of definition-concept texts used for the Generator block training were also applied in this phase.

Both compendia of sentences were grouped into a single dataset of 207,768 sentence pairs for the train and test datasets division. For computational reasons, a random selection of 208 pairs of sentences was used for the model's finetuning; Table A.3 of Appendix A show the selected texts. Since the model had to evaluate two sentences, each pair was divided into two blocks: a left sentence for the original input definition and a right sentence for the ground truth concept. Besides this separation, the model required a specification of how the similarity among every sentence was. Hence, every sentence had a similarity score of 1 for the same pair and a score of 0 for different pairs. In the end, the training set consisted of 43,264 pairs of sentences with their corresponding similarity score.

The test set consisted of another 100 random pairs of sentences from the general set. For this case, it was not necessary to assign a similarity score since that would be the model's result. Instead of that, the input right sentence belonged to the ground truth concept, while the left input sentence was generated by the GPT-2 model, using the original descriptive sentence as reference. Table A.4 of Appendix A show the complete list of the test set.

4.3.2 Model Training

The selected algorithm for the validation block was the USE model since it showed the best results in the previous similarity evaluation. One of the main advantages of this algorithm is that it does not require a particular preprocessing of the data since it is capable of evaluating the use of lower and upper case and punctuation signs.

The model received the training dataset and was finetuned through an Adagrad optimizer for a total of 50 epochs. Since two sentences were used as input of the model, a regularization parameter was needed for both of them, being 0.0 for the first sentence and 0.01 for the second one to prioritize the second sentence (considered the ground truth). The model training was early stopped if the loss function reached a score below 0.05 before passing through all the epochs. Table 4.5 show the general hyperparameters used for the finetuning of the model.

Once the training was completed, the test dataset created a confusion matrix and observed the differences before and after finetuning.

Table 4.5: USE training hyperparameters

Hyperparameter	**Value**
Model's version	4
Optimizer	Adagrad
Adagrad learning rate (lr)	0.1
Sentence 1 regularization strength	0.0
Sentence 2 regularization strength	0.01
Classes	256,797,824
Adam lr	0.001
Loss function limit	0.05

4.3.3 Model Evaluation

Once the model was completed, the test dataset was used in the Generator block to create new interpreted texts of one of the original descriptive sentences. Once the texts were created, a similitude comparison was made with the USE algorithm to measure the semantical similitude between the ground truth concept and the generated text. Those results were compared with the model's evaluation before finetuning to evaluate the changes in the similarity evaluation between the same texts.

Chapter 5

Results

5.1 Text Generation Procedure

After inserting the training dataset into the GPT-2 model, the loss function of the neural network was measured every ten epochs, obtaining the behavior shown in Figure 5.1. As it can be observed, the model shows, at first sight, a decreasing average loss, which could mean that the network is being capable of detecting the semantical relationship between the given definitions and the resulting concept. However, it is essential to remember that the GPT-2 model and most neural networks used for text generation mainly focus on copying a writing style rather than evaluating the semantic relationship between words.

This behavior can be observed in Figure 5.1 at the right side of the graph, where the epoch's loss function achieves almost a 0.20 value, even though the average loss keeps decreasing. This fact shows that even though the network tries to follow the same structure and determine the word, this task can be more complicated by just following a writing style.

After the model's training, the colloquial and specialized datasets evaluated the neural network results. Some examples of the given results are shown in Table 5.1. The cosine similarity between the generated texts and the ground-truth words was made to evaluate if the model can determine a particular context-based on any definition. Table 5.2 shows the average values of the cosine similitudes between the given definitions, the ground-truth concepts, and the generated words. Three similitude evaluations were made: Group 1 evaluated the ground-truth concept and the input definition, Group 2 compared the model's generated text with the definition, and Group 3 compared the ground-truth concept with the definition of the generated text. The complete generated texts and their corresponding similarity distance are shown in Tables A.5 and A.6 of Appendix A, respectively.

Depending on the vocabulary used in each sentence, some sentences included words that had no hyperbolic representation in the database. In this case, a value of zero was given to every dimension. The other possible situation was that the generator produced no sentence or vocabulary at all; under this situation, a value of Not a Number (NaN) was implemented for the embedding, and a $<|\text{None}|>$ token for the sentence; Those phrases were omitted from the mean analysis. The rest of the sentences are presented without stop words to specify the evaluated words in each phrase.

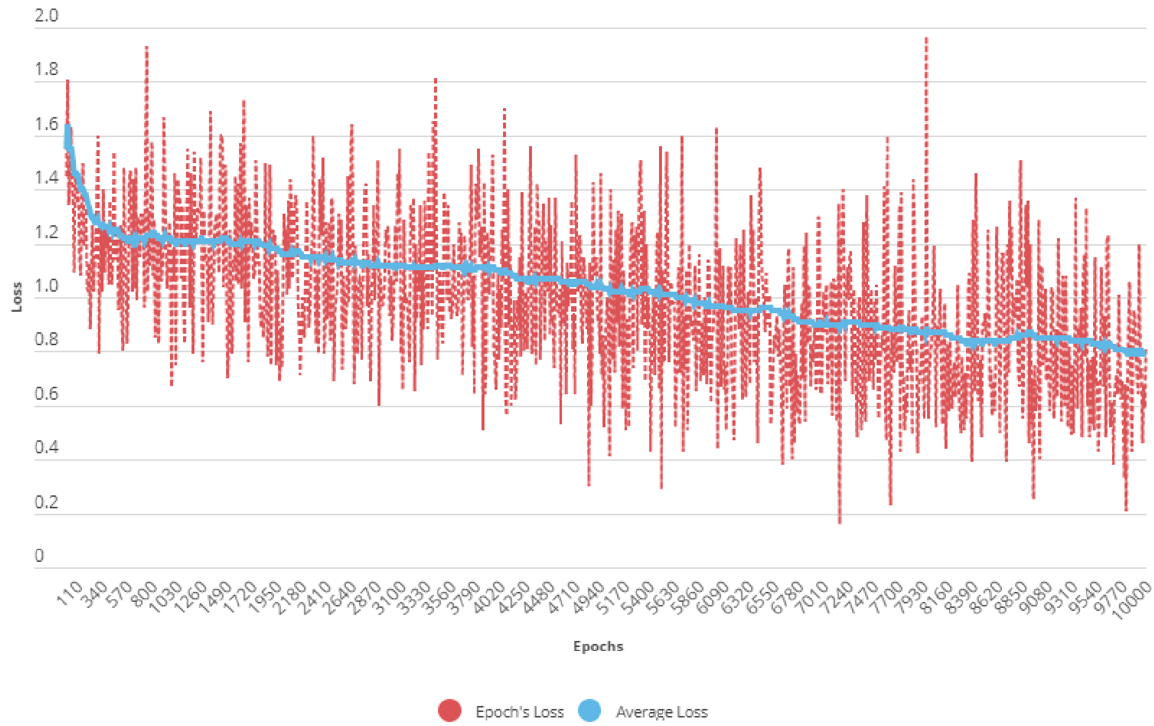


Figure 5.1: GPT-2 loss function

The results from Table 5.1 show that the generated texts of the model are capable of generating results similar to the ones of the ground-truth. For example, with the definition of the colloquial dataset, even though the generated text was larger than the answer, it can be observed that the context is similar between the two answers. Similar behavior is observed with the example of the verb specialized dataset. In that case, the general context between the generated and ground-truth answers is very similar, although the generated word was not the expected. Finally, different behavior is seen with the noun specialized dataset. In this case, the model only tried to imitate the definition structure, using parts of the original text instead of another vocabulary.

The results shown in Table 5.2 demonstrate that colloquial text benefits the selection of vocabulary that better describes the given definition. However, it appears that the semantical similitude between the generated answer and the ground-truth text is lesser than the similitude with the definition text. When analyzing the use of more specialized text, the definition of nouns and the generated texts is more similar with the distance between the generated text and the ground-truth concept, rather than the similitude between the original text pair (definition and ground-truth).

Analyzing the GPT-2 model results with the BERT and XL-Net results, it can be observed that there is no standard value for the similarity of texts. Particularly with the similarity results of Group 1, it is observed how every model determines a different similitude between the main word or concept and the original definition. Nevertheless, it can also be observed

Table 5.1: Examples of generated texts with the testing dataset

Definition	**Generated answer**	**Ground-truth**	**Dataset**
(especially formerly) an institution for the maintenance and care of the mentally ill, orphans, or other persons requiring specialized assistance	a hospital where psychiatric services are available	asylum	Colloquial
a peddler who shouts to advertise the goods he sells	a peddler who shouts	crier	Specialized (noun)
find unexpectedlyd	discover	fall_upon	Specialized (verb)

Table 5.2: Average distance between generated and ground-truth texts

		GPT-2		**BERT**		**XL-Net**	
		Distance	*Std. Dev.*	*Distance*	*Std. Dev.*	*Distance*	*Std. Dev.*
Group 1	**Colloquial**	0.5137	0.4627	0.3545	0.1560	0.8972	0.0989
	Noun	0.7365	0.3125	0.4982	0.1238	0.9363	0.0824
	Verb	-0.0134	0.4245	0.4987	0.1490	0.9280	0.0456
Group 2	**Colloquial**	0.6579	0.3780	0.6394	0.1602	0.9708	0.0171
	Noun	0.5463	0.3996	0.6697	0.1463	0.9725	0.0214
	Verb	0.5611	0.4133	0.6209	0.1554	0.9617	0.0292
Group 3	**Colloquial**	0.5136	0.4605	0.5457	0.0850	0.9072	0.0913
	Noun	0.5009	0.4308	0.5645	0.1115	0.9433	0.0812
	Verb	-0.1225	0.3682	0.5851	0.1171	0.9280	0.0452

how the evaluations of the colloquial vocabulary of Group 2 and the colloquial and specialized nouns of Group 3 are more similar between the GPT-2 model and BERT. In this way, it is possible to say that there is little consensus between these two models when evaluating the similitude between the generated text and a ground truth phrase.

Different behavior is observed with the definition of verbs, where the average semantical similitude is closer between the original text pair and the distance between the ground-truth and the generated text. However, the semantical similitude between the generated text and the definition is bigger.

These three different behaviors suggest that the sole analysis of the semantical relationship between the given definition and the concept text (generated or ground-truth) is insufficient to determine if the model can generate texts that genuinely evaluate the similitude meaning of the given definition. These results express that the model can generate texts similar in meaning, but only because it tries to imitate the writing structure of the training dataset. It is necessary to evaluate if the analysis of the documents' syntactical structure influences the generation of better answers.

5.2 Selection of Algorithm for Similarity Evaluation

The SoA has multiple algorithms and architectures that claim to produce accurate texts. However, it is crucial to determine the application and implementation of the text documents to determine how genuinely effective an algorithm is. Since most algorithms are designed to evaluate only the overall structure of a sentence and that the produced text is somehow readable, it is necessary to compare each model with the human perspective to determine which algorithm can maintain the general context of the original text.

A population of 23 people evaluated the 100 pairs of sentences and determined the corresponding similitude score of each. The graph in Figure 5.2 shows the distribution of the subject's English proficiency. Among the population, 26.1% were native speakers, represented by a proficiency level of 5 in the graph. Even though most of the subjects were not native English speakers, most of them (91.3%) were affluent on the language.

Every sentence was evaluated as well with four different algorithms and procedures: the use of the standard model of USE, the uncased model of BERT, and cased version of XL-Net (described in Table 4.4), and the result of evaluating the knowledge-based representation of words in hyperbolic space (EHR). Since the first three models were already pre-trained for this analysis, it was only necessary to introduce both sentences and evaluate their similarity scores through cosine distance.

The implementation of EHR required a more manual approach. Following the description of Figure 4.3, every sentence was transformed into lower case and tokenized to eliminate every stop word and punctuation sign. Then every sentence was divided into its corresponding subject, verb, and object blocks to evaluate each block's embedding, later joining them with

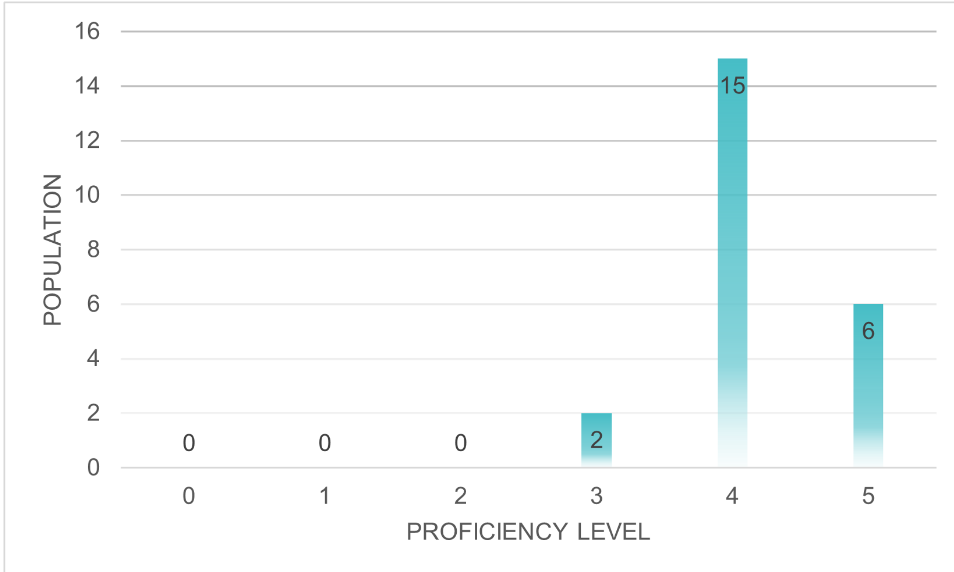


Figure 5.2: Surveyed population's English proficiency

the mean score between them. As well as the other implemented models, cosine similarity was used to determine the difference in similitude between each pair of sentences.

An example of the first five pairs of sentences and the overall mean score of each algorithm is shown in Table 5.3. The mean scores show that even though two of the SoA algorithms: BERT and XL-Net, as well as the proposed EHR, claim a similarity score above 80%, those scores do not represent the human perspective of the test population. At the same time, the Pearson correlation (r) and its corresponding p-value demonstrate how the implementation of the USE algorithm is the one that best follows the human perspective. Table A.7 of Appendix A shows the individual results of each pair of the questionnaire that were used with every algorithm to generate these statistic measurements.

The best performance is also observed in Figure 5.3, where even though every algorithm follows a positive correlation, the USE implementation follows a more distributed interpretation of the sentence's general context when comparing them with the survey results.

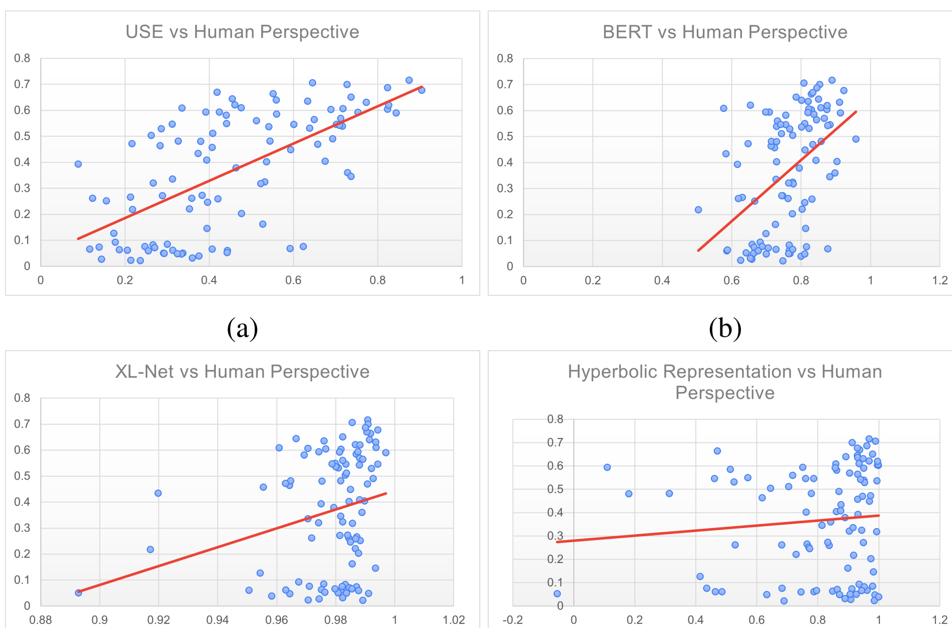
5.3 Validation of Text Similarity

The model finetuning used a big lr considering that the training dataset was not big enough for evaluating more minor compensations on the network's parameters. Nevertheless, the lr of 0.1 was enough to reach a loss value of 0.0447 on the first epoch, early-stopping the training of the model. However, the finetuned model showed a relevant modification on the similarity scores of the model.

The test set was first used to generate new interpreted texts from the original input sentences (left sentences) through the GPT-2 text generator. Table 5.4 show some of the generated

Table 5.3: General results of the similarity scores test

	USE	**BERT**	**XL-Net**	**Hyperbolic**	**Survey**
Pair 1	0.20507446	0.765544128	0.981895491	0.787412789	0.06173913
Pair 2	0.5315156	0.775600547	0.982348109	0.943932376	0.324347826
Pair 3	0.8228531	0.873916928	0.976600763	0.995200041	0.604347826
Pair 4	0.6453475	0.808043558	0.985431424	0.988872588	0.706086957
Pair 5	0.3005308	0.658887544	0.975924086	0.979368666	0.085217391
Mean	0.454266365	0.763986481	0.978878184	0.812471648	0.367678261
Std. Dev.	0.202946565	0.090598291	0.015706973	0.206246681	0.228288228
Pearson r	0.637847375	0.466795134	0.249547033	0.09725548	
p-value	9.53828E-13	9.80837E-07	0.01228525	0.335742403	



(a) (b) (c) (d)

Figure 5.3: Correlation graphs of the similarity survey

Table 5.4: Generated phrases for the finetuned USE testing

Pair	Input phrase	Generated phrase	Ground truth
1	the cardinal number that is the sum of three and one	three_quarters	quatern
2	the fifth sign of the zodiac; the sun is in this sign from about July 23 to August 22	the fifth sign of the zodiac; the sun	Leo
3	German composer of many operas; collaborated with librettist Hugo von Hoffmannsthal to produce several operas (1864-1949)	composer_of_songs	Strauss
4	capable of being diagnosed	describable	diagnosable
5	having or deserving or conferring glory	praiseworthy	glorious

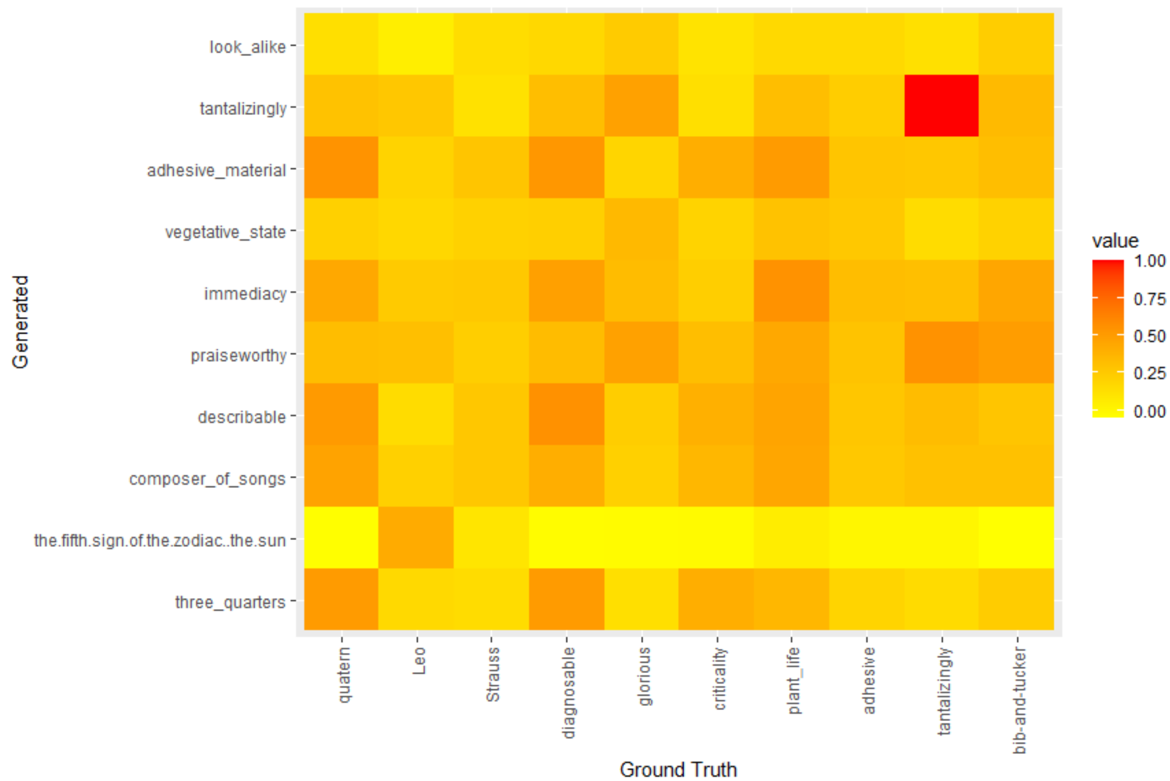
texts, while Table A.8 of Appendix A show the complete list of resulting sentences. It is possible to observe how the generator model could produce a one-word sentence most of the time while assumingly maintaining the general context of the ground truth concept.

In order to validate the context of the generated texts, these sentences were used among the test set ground truth concepts in the finetuned model. After evaluating their corresponding similarity through cosine similarity, a heat map was obtained between each pair of sentences to evaluate the similarity in context. The produced heatmaps were obtained with the original and the finetuned USE model to observe the changes in context interpretation.

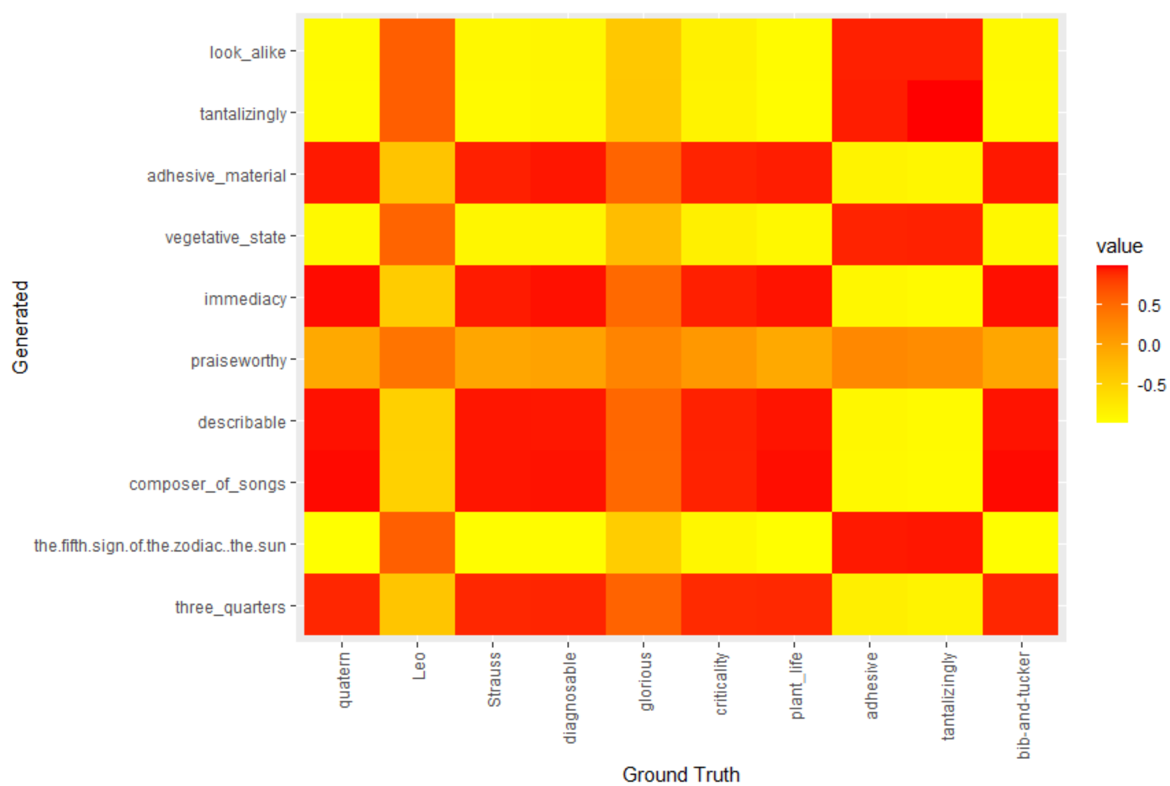
Figure 5.4 shows an example of the first ten sentences of the test set, before and after the model's training. It can be readily appreciated that before finetuning the USE algorithm (Figure 5.4a), the model was not capable of determining the similarity between short phrases and a single concept. This act demonstrates the widespread problem with SoA algorithms when comparing concepts: they are designed to compare complete sentences between each other due to the evaluation of the total sum of each word's representation; however, when evaluating multiple words against a single concept, the total sum overrides the similarity score, assuming that they are talking about entirely different concepts.

After finetuning the model (Figure 5.4b), it is observable how the similarity score of each pair of sentences improves, closing the gap between the generated phrase and its corresponding ground truth concept. Unfortunately, a similar effect appears between concepts of different pairs of sentences indicating that the model cannot separate the overall context between a large group of sentences and topics. The complete heatmap comparison of the test set is shown from Figure A.1 to A.10 on Appendix A.

Even though the model cannot accurately distinguish between multiple contexts in a given dataset, it can still evaluate the relationship between the generated phrase and the original keyword, improving the semantical similitude evaluation between them. After finetuning, an increase of 21.5167% in the similarity score was achieved. This percentage demonstrates the model's capabilities for evaluating the similitude in meaning between a complete phrase and a single token. Table A.9 on Appendix A shows the difference of the similarity scores before and after finetuning the model.



(a)



(b)

Figure 5.4: Similarity heatmap example (a) before and (b) after finetuning

Chapter 6

Discussion and Future Implementations

The results obtained from this proposal show that the analysis of a single concept or word compared with a complete sentence is quite challenging. Considering that most of the SoA models are focused on working with complete sentences (or even documents) instead of single vocabulary shows the difficulties for analyzing the abstract context behind the idea of a word or phrase.

The generation of text based on a single concept, following a traditional dictionary-description style, may sound relatively straightforward. However, it is still necessary to remember that any ML model begins generating results from scratch without previous knowledge. When training a model, this lack of information makes it challenging to analyze a complete sentence's general structure and meaning based on a subject, verb, and object and transform it into a single word (or a smaller phrase) that uses different vocabulary without changing the general context.

The results obtained from this proposal's text generator demonstrate that difficulty. The representation of a whole idea by using a single word is an abstract process that a ML algorithm is not capable of doing. For this reason, it is necessary to evaluate how training the model with different vocabulary, both colloquial and specialized, allows the algorithm somehow to find a relationship between a series of words and its abstract definition or concept. One of the first experiments of this type was the one made by Rubenstein and Goodenough [85], who evaluated how human beings made a mental process to assign an abstract idea for a particular word and then determine how similar it was in meaning with another vocable. The next step in that experiment should be to evaluate the same abstract concept of a word but with a longer phrase that defines it.

Most validation metrics in NLP evaluate similitude by observing if the two compared sentences share the same vocabulary. Nevertheless, it is a known fact that a human being does not explain a particular concept by using the same explained word. Following this reasoning, the training and testing datasets used different phrases to define a particular word. Even though most of the used keywords (mainly from the colloquial dataset) were similar, each term had its proper definition and description. This implementation allowed the GPT-2 model to analyze not only the relationship between a complete sentence with its corresponding keyword but also how different phrases could use similar vocabulary or context to explain the same (or a similar) concept.

The generated texts of Table A.5 show how the model tries to interpret the given concept by summarizing the input sentence. Pair 2 of the test dataset is an example of this; even though the expected word or concept for *a person's signature* was an *autograph*, the model only simplified the same sentence, denoting a *persons signature*. The same phenomenon occurs with Pair 75; it was intended to describe a *pipe* by describing it as a *long tube made of metal*, but the model only reduced the phrase as a *long tube made*. Multiple phrases copied the original sentence, and the determined text length was not enough for the model to interpret and generate a complete sentence. A similar effect happened with the pairs that had no generated text, like Pair 109.

The fact that some sentences were very similar to the original inputs does not mean that the GPT-2 model was not capable of finding a relationship between concepts and different vocabulary. The model could interpret the original descriptive text and transform it into a keyword synonym, particularly with the verb dataset. One example can be observed with Pair 69, where *gain knowledge or skills* was transformed into *know*, trying to explain the concept of *learn*. Another situation can be seen with Pair 129, where the original concept of *warn* was defined as an *advise or counsel in terms of someone's behavior* and interpreted as *express opinion hope*. Even though the resulting phrase was not closer in meaning with the keyword's general context, it tries to emphasize the idea of saying something with a particular purpose. It can be said then that the model can evaluate a given input and create a different, short sentence that tries to imitate the general context of the keyword while ensuring the use of different vocabulary.

It is also important to mention that giving a numerical value to the semantical similarity of two phrases or words is not always easy. Table 5.2 shows how the comparison between the original inputs and the generated text create different similitude scores without reaching a particular benchmark. However, that does not mean that the word *dog* is not similar to a *domestic canine* or an *animal that barks*. The mean similarity scores demonstrate that the evaluation of similitude is based mostly on the vocabulary *per se* and not on the whole idea the phrase tries to give. It is also important to notice how the similarity scores between different algorithms vary, making it even more difficult to evaluate and determine a benchmark that indicates the correct grade of similitude between two concepts.

On the other hand, it is interesting to notice how the mean similitude scores are similar between the USE and BERT algorithms when evaluating the original input sentence with the generated one of colloquial vocabulary (Group 2 of evaluation) and the original keyword with the generated text (Group 3 of evaluation) of the colloquial and noun datasets. That means that at least by two different algorithms, it was possible to evaluate practically the same way the relationship between a generated text and one part of the original phrases. It is curious how the semantical similitude between the original description and its keyword showed lower

similarity scores, considering that these phrases were used to train, interpret, and select different vocabulary to create new descriptive phrases.
The first stage of this dissertation's proposal demonstrates that even though multiple SoA models claim that their results are superior to others for evaluating and generating text, it is still necessary to compare those results with a human being since they are considered to be the final user of every implementation. Their abstract, personal perspectives are the ones that decide if a phrase makes sense with another one or not.

The implementation of the questionnaire included recollecting different postures and opinions when evaluating the overall meaning of two sentences. Much likely to the experiment of Rubenstein and Goodenough [85], but comparing complete phrases instead of single words. Unfortunately, the surveyed population was not big enough to produce a statistically substantial comparison. This problem was due to the size of the questionnaire, making it difficult and tedious to answer completely. Nonetheless, it was necessary to evaluate a considerable number of sentences to create a robust mean evaluation, asserting that the given results were significant. Another advantage obtained from the survey results is that most of the population had a higher English proficiency, obtaining closer results to the original answers after pondering each subject's answers.

For the evaluation of the SoA algorithms, it is known that one of the most common implementations for sentence analysis and word representations (and hence concept similarity evaluation) is LSA since it has word embeddings based on their statistical presence inside a document. At the same time, recent implementations [92] show how separating the sentence into three syntactical blocks improved the representation of words. This dissertation proposed then that by combining this idea with the general representation of a knowledge-based word representation [76], it was possible to create a model capable of representing a word (and therefore a sentence) without regarding the topic or particular topic that was being described. This thought was then defined as EHR, and compared with different SoA Transformer models.

Unfortunately, the proposal of EHR proved unuseful for evaluating textual context similarity based on the human perspective. The results can be observed in Figure 5.3a, where most of the given definitions were marked as outlines of the general correlation line. At the same time, the mean similarity scores of Table 5.3 demonstrate that the proposal was the one farthest from the human way of thinking for similarity evaluation. These implications suggest that the model of EHR is not adequate for general context evaluations and that the syntactical structure of the sentence needs to properly evaluate the topic that is being treated in order to determine if some texts are similar in meaning efficiently.

Besides the undesired results of the EHR proposal, the similarity evaluation test allowed to select USE as the best algorithm that correctly follows the human perspective when evaluating the semantical structure of text. This result shows how the Transformer architecture is one of the best options for evaluating abstract relationships between documents and phrases. On the other hand, it is essential to mention that even though this algorithm had the best results for evaluating, it does not mean that the GPT-2 generator block lacks in its capabilities for producing and interpreting text. This particular experiment shows how the architecture of different NLP algorithms allow the combination of multiple blocks and stages to create a complex model for a complete analysis and text production. In this case, the parallel architecture of the GPT-2 network and its masking procedure allow it to interpret better a given sentence and create a new one similar in context. On the other hand, the combination of the Transformer model with a DAN network and the vast topics and documents used to train the USE algorithm allows it to evaluate better the relationship between two sentences. Following this idea, the combination of these two algorithms resembles the structure of a LSTM neural network for the encoder-decoder task or even a GAN with its generator and discriminator structure.

The finetuning of the USE algorithm allows us then to implement the relationship between a single keyword and its descriptive sentence. As mentioned before, the advantages of this neural network architecture permit a similarity context evaluation closer to the one of the human perspective. Considering this fact, it is only necessary to finetune it to reduce the problem of evaluating a complete sentence with a single word, as shown in Figure 5.4a. Even though the pair of sentences refer to the same concept, the similarity score indicates almost the opposite.

The finetuning of the model used a relatively big lr and a smaller training set mainly for the given limited computational capabilities during the development of this dissertation. However, the model was finetuned during the first epoch (of 50) with this learning rate. After changing the value to minor lr, the model reduced their loss function at the limit of generating a NaN value, indicating that the model was not capable of continuing with the evaluation. However, by indicating the relationship of similitude between similar sentences (with a value of 1) and different pairs (assigning a value of 0), the model could increment the semantical similitude between the general concepts of the ground truth keyword and the generated descriptive sentence.

Besides the fact that the model could now evaluate the similitude between these text inputs, another interesting result occurred. Observing the behavior of the result's heatmaps, it is appreciated that the semantical similitude between each corresponding pair is augmented as well as between different sentences. One example can be observed in Figure A.1, where there was practically no semantical similarity between any phrase before finetuning. Following this logic, the USE algorithm specified for Pair 7 that the key concept of *(botany) a living organism lacking the power of locomotion* could not refer to a *vegetative_state*, but only to a *plant_life*. It is correct to assume that by specifying in parenthesis the concept of botany, the input phrase was referring to a *plant*. However, most of the used vocabulary refers to a living organism incapable of moving, which alludes to the generated phrase. Another example can be seen in Figure A.2, where Pair 12 talks about a *legal holiday in the United States*, and the model before finetuning indicates that the answer *Fourth_of_July* is not similar to *July_4*. A similar case occurs in Figure A.4, where the input of Pair 36: *express willingness to have in one's home or environs* is similar to the keyword *receive*, but not the generated answer *adopt*.

One problem after the finetuning of the model is that multiple generated sentences reflect a high similarity score between different sentences apart from their corresponding pair.

For example, the heatmap of Figure A.3 after finetuning indicates that the generated text of Pair 29: *Pisa*, is similar to the keywords of Pairs 22, 23, 25 and 27 (*Mbeya*, *heterozygosity*, *genus_Sphaeralcea*, and *appendicularia*, respectively). Only in the first case, both pairs of sentences talk about a particular region in a country. However, the rest of the sentences use complex vocabulary to talk about particular classes of plants or animals. In this case, the heatmap shows how specialized vocabulary directly affects the analysis of the phrase's context. This phenomenon is linked with the general results of Table 5.2, where the use of specialized vocabulary directly affects the general semantic similarity score when evaluating the input descriptive sentence and its corresponding keyword.

The preliminary results of this new evaluation and implementation for analysis of phrase-word text documents begin to show promising aspects for the interpretation and understanding of the abstract meaning of a particular concept. Even though the results still demonstrate problems when comparing contexts of different sentences, it is necessary to mention that, as mentioned in [92], the knowledge-representation of words presents an accuracy problem when evaluating different phrases. This problem is due to the general representation used for every word embedding, which directly affects the evaluation of particular topics and contexts in a broad vocabulary.

However, it is still essential to notice that current SoA implementations are unable to create such evaluations. This fact proves the novelty of this proposal, not only for the structure of the implemented algorithms for text generation but also for the designed methodology for analyzing an abstract idea, such as the general meaning of a phrase or a word, and using it to determine how similar a phrase (with multiple words) is regarding a single term. Hence, this methodology marks the beginning of a new branch in NLP. Whereas most implementations and models are being used to create and analyze long documents of text and complete phrases, this proposal analyzes the relationship between a modular concept and a complete sentence, sharing the same abstract idea of meaning.

The results of this dissertation make room for multiple cases of study and evaluation metrics. In the first place, it is still necessary to determine at which point a given vocabulary is too specialized to influence the generation of new descriptive texts negatively. Another aspect to consider for future implementations is that a phrase's syntactical structure directly affects the similarity evaluation. Hence, it is essential to evaluate how the same generation and validation methodology works when using a corpus-based representation (more specialized than the implemented knowledge-based representation).

Finally, it would seem attractive to analyze the direct usefulness of the given model when interacting directly with a human being. In this scenario, the idea would be to select multiple phrases regarding a particular topic and use those descriptive sentences to interpret and generate new ones intended to explain to the user what the main topic is. As mentioned before, this case study will have to consider how technical those analyzed concepts will be, making use at first of a more colloquial vocabulary and context, to then try to explain a more advanced topic with its vocabulary.

Chapter 7

Conclusions

The world currently lives with an information overload problem, where data is excessive for analyzing and evaluating every existing reference regarding a particular topic. Also, due to the multiple platforms, journals, and available sites for getting information, data is disorganized, making it more challenging to determine which reference should be analyzed before another properly, or even if it is valuable or not. Finally, information is static, meaning that the way data is presented to the user could be helpful for someone but at the same time confusing or disposable to a different person that looks for the same information. These three problems are deeply related to the vocabulary used in each document; since the usefulness of a document is directly related to the number of understood vocabulary, the excessive generation of data and its static problem could be solved by dynamically modifying a given redaction to make it more comprehensive to each user.

At the same time, there are multiple algorithms and applications of NLP involving the analysis of different documents of text to paraphrase, resume, or even detect a particular writing style to create new content and information. However, even though the advancements achieved by the SoA are remarkable, it seems that most of the current implementations are focused on the bigger picture of text analysis, which is to understand the complete structure of how human beings communicate when most of the times are better to analyze and use lesser vocabulary than a complete document. This current SoA problem directly affects the static presentation of data. For these past reasons, this dissertation proposes a new structure and methodology for analyzing the general context of a given sentence and how similar it is with the abstract definition of a particular keyword.

This proposal intends to evaluate the semantical similitude between a complete phrase or sentence, and a single keyword, following the structure of a regular dictionary, where a descriptive sentence explains and shares the exact meaning of a single word. With this purpose in mind, the implemented architecture was divided into a text generator and text validation block. The generation of text was implemented by using a GPT-2 transformer model to evaluate a given definition text and generate a smaller text corresponding to the concept word. The difference between this implementation and the other uses of the SoA is that in this situation, the neural network model needs to learn the semantical relationship between the words of the definition to select the correct phrase that best describes it adequately. This model used a hyperbolic knowledge-based representation of words to evaluate the semantical relationship between words and compared the semantical similitude between the given definition, the ground-truth concept, and the generated text of the model.

The results demonstrated that even though the GPT-2 model selects semantically similar words, there are different behaviors when using colloquial or more specialized vocabulary. These results do not determine if the neural network is analyzing the abstract meaning and relationship between words, or it is only trying to imitate the writing structure of the training dataset. Implementing the validation block is necessary to determine the similitude between a given definition and the concept text.

As human beings, we have the innate ability to process information and analyze it to comprehend new or unknown concepts or terms. However, elaborating the same process through supervised models to replicate human thinking is not trivial. Since the final intention of this dissertation is for a human user to understand a given phrase better, the model in charge of determining the validity of the generated text must follow the human perspective for semantic evaluation. However, multiple NLP algorithms claim that their accuracies for semantic evaluation are superior to others, without specifying how similar their results are in comparison with the opinion of a human user. The validation algorithm was then selected based on the comparison of different NLP models and a proposed combination of word representation: EHR. This last algorithm combined the syntactical analysis of a corpus-based representation with the general knowledge-based representation of a hyperbolic structure of the WordNet database. Every algorithm was evaluated with a questionnaire where a given population determined the semantical similarity of 100 pairs of sentences.

The obtained results demonstrated that the USE algorithm was the model with better correlation and reproducibility in terms of human perspective. Even though the proposed EHR algorithm showed relatively good similarity scores, its relationship with the human way of evaluating was the farthest among the implemented algorithms.

As previously mentioned, one of the problems of current NLP algorithms is that they are designed for evaluating two pairs of long text inputs. Since the objective of this proposal is to evaluate the similarity between a complete sentence with a single concept, it was necessary to finetune the validation network to analyze such relationships properly. The results obtained from this new training allowed the model to assign a higher similarity score between these phrase-word pairs, increasing the similarity score validation of 21.5167% after finetuning the proposed model. Nevertheless, this finetuning process also showed that it is not enough to indicate when two sentences are similar and when they are not, since multiple sentences began to indicate higher similarity scores between different pairs of text (mainly when an advanced vocabulary was used).

This final result shows that another evaluation metric is still needed to differentiate the context between text pairs. This is an expected result from a knowledge-based implementation since the whole methodology followed the structure of a general word representation, reducing the accuracy for defining particular contexts or topics between different inputs. On the other hand, it is also vital to notice that there is no current evaluation of the general context between a single word and a complete sentence, proving the novelty of this dissertation. This proposal marks a new area of study for analyzing more abstract terms, rather than only evaluating the grammatical structure of a sentence to determine if it makes sense or is similar to another one just by how they sound. Understanding and evaluating the abstract meaning of the text would allow us to operate and present information better, eliminating the static presentation of data and making it more accessible and valuable for the general public, contributing to the reduction of the information overload problem.

Appendix A

Individual Results

- **Test dataset for the Generator block**
- **Similarity evaluation questionnaire**
- **USE train set**
- **USE test set**
- **Generator sentences**
- **Test results for the Generator block**
- **General results of the similarity questionnaire**
- **Generated phrases for the finetuned USE test**
- **Comparison on the similarity scores validation**
- **Finetuned USE heatmap of test pairs 1-10**
- **Finetuned USE heatmap of test pairs 11-20**
- **Finetuned USE heatmap of test pairs 21-30**
- **Finetuned USE heatmap of test pairs 31-40**
- **Finetuned USE heatmap of test pairs 41-50**
- **Finetuned USE heatmap of test pairs 51-60**
- **Finetuned USE heatmap of test pairs 61-70**
- **Finetuned USE heatmap of test pairs 71-80**
- **Finetuned USE heatmap of test pairs 81-90**
- **Finetuned USE heatmap of test pairs 91-100**

Table A.1: Test dataset for the Generator block

Pair	Keyword	Descriptive Text	Dataset
1	asylum	(especially formerly) an institution for the maintenance and care of the mentally ill, orphans, or other persons requiring specialized assistance	Colloquial
2	autograph	a person's own signature	Colloquial
3	boy	a male child, from birth to full growth, especially one less than 18 years of age	Colloquial
4	brother	a male offspring having both parents in common with another offspring; a male sibling	Colloquial
5	car	an automobile	Colloquial
6	coast	the land next to the sea; seashore	Colloquial
7	cock	a male chicken; rooster	Colloquial
8	cord	a string or thin rope made of several strands braided, twisted, or woven together	Colloquial
9	crane	any large wading bird of the family Gruidae, characterized by long legs, bill, and neck and an elevated hind toe	Colloquial
10	cushion	a soft bag of cloth, leather, or rubber, filled with feathers, air, foam rubber, etc., on which to sit, kneel, or lie	Colloquial
11	food	any nourishing substance that is eaten, drunk, or otherwise taken into the body to sustain life, provide energy, promote growth, etc	Colloquial
12	furnace	a structure or apparatus in which heat may be generated, as for heating houses, smelting ores, or producing steam	Colloquial
13	gem	a cut and polished precious stone or pearl fine enough for use in jewelry	Colloquial
14	glass	a hard, brittle, noncrystalline, more or less transparent substance produced by fusion, usually consisting of mutually dissolved silica and silicates that also contain soda and lime, as in the ordinary variety used for windows and bottles	Colloquial
15	graveyard	a burial ground, often associated with smaller rural churches, as distinct from a larger urban or public cemetery	Colloquial
16	grin	a broad smile	Colloquial
17	mound	a natural elevation of earth; a hillock or knoll	Colloquial
18	noon	midday	Colloquial

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
19	oracle	(especially in ancient Greece) an utterance, often ambiguous or obscure, given by a priest or priestess at a shrine as the response of a god to an inquiry	Colloquial
20	slave	a person who is the property of and wholly subject to another; a bond servant	Colloquial
21	tool	an implement, especially one held in the hand, as a hammer, saw, or file, for performing or facilitating mechanical operations	Colloquial
22	voyage	a course of travel or passage, especially a long journey by water to a distant place	Colloquial
23	wizard	a person who practices magic; magician or sorcerer	Colloquial
24	woodland	land covered with woods or trees	Colloquial
25	automobile	a passenger vehicle designed for operation on ordinary roads and typically having four wheels and a gasoline or diesel internal-combustion engine	Colloquial
26	bird	any warm-blooded vertebrate of the class Aves, having a body covered with feathers, forelimbs modified into wings, scaly legs, a beak, and no teeth, and bearing young in a hard-shelled egg	Colloquial
27	cemetery	an area set apart for or containing graves, tombs, or funeral urns, especially one that is not a churchyard; burial ground; graveyard	Colloquial
28	forest	a large tract of land covered with trees and underbrush; woodland	Colloquial
29	fruit	any product of plant growth useful to humans or animals	Colloquial
30	hill	a natural elevation of the earth's surface, smaller than a mountain	Colloquial
31	implement	any article used in some activity, especially an instrument, tool, or utensil	Colloquial
32	jewel	a cut and polished precious stone; gem	Colloquial
33	journey	a traveling from one place to another, usually taking a rather long time; trip	Colloquial
34	lad	a boy or youth	Colloquial
35	madhouse	a hospital for the confinement and treatment of mentally disturbed persons	Colloquial

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
36	magician	an entertainer who is skilled in producing illusion by sleight of hand, deceptive devices, etc.; conjurer	Colloquial
37	midday	the middle of the day; noon or the time centering around noon	Colloquial
38	monk	(in Christianity) a man who has withdrawn from the world for religious reasons, especially as a member of an order of cenobites living according to a particular rule and under vows of poverty, chastity, and obedience	Colloquial
39	pillow	a bag or case made of cloth that is filled with feathers, down, or other soft material, and is used to cushion the head during sleep or rest	Colloquial
40	rooster	the male of domestic fowl and certain game birds; cock	Colloquial
41	sage	a profoundly wise person; a person famed for wisdom	Colloquial
42	serf	a person in a condition of servitude, required to render services to a lord, commonly attached to the lord's land and transferred with it from one owner to another	Colloquial
43	shore	the land along the edge of a sea, lake, broad river, etc	Colloquial
44	signature	a person's name, or a mark representing it, as signed personally or by deputy, as in subscribing a letter or other document	Colloquial
45	smile	the act or an instance of smiling; a smiling expression of the face	Colloquial
46	stove	a portable or fixed apparatus that furnishes heat for warmth, cooking, etc., commonly using coal, oil, gas, wood, or electricity as a source of power	Colloquial
47	string	a slender cord or thick thread used for binding or tying;line	Colloquial
48	tumbler	a person who performs leaps, somersaults, and other bodily feats	Colloquial
49	limber_up	make one's body limber or suppler by stretching, as if to prepare for strenuous physical activity	Verb
50	bear	have rightfully; of rights, titles, and offices	Verb

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
51	tickle	touch (a body part) lightly so as to excite the surface nerves and cause uneasiness, laughter, or spasmodic movements	Verb
52	separate	go one's own way; move apart	Verb
53	pout	a disdainful grimace	Verb
54	brush	contact with something dangerous or undesirable	Verb
55	fly	travel in an airplane	Verb
56	squeeze	a twisting squeeze	Verb
57	fall_upon	find unexpectedly	Verb
58	plank	set (something or oneself) down with or as if with a noise	Verb
59	roar	the sound made by a lion	Verb
60	bring_to	return to consciousness	Verb
61	follow	to travel behind, go after, come after	Verb
62	try	make an effort or attempt	Verb
63	plug_in	plug into an outlet	Verb
64	broadcast	a radio or television show	Verb
65	present	hand over formally	Verb
66	slam	throw violently	Verb
67	tissue	part of an organism consisting of an aggregate of cells having a similar structure and function	Verb
68	revolutionize	change radically	Verb
69	learn	gain knowledge or skills	Verb
70	transform	change or alter in form, appearance, or nature	Verb
71	rivet	fasten with a rivet or rivets	Verb
72	hamstring	one of the tendons at the back of the knee	Verb
73	disclose	reveal to view as by removing a cover	Verb
74	incite	provoke or stir up	Verb
75	pipe	a long tube made of metal or plastic that is used to carry water or oil or gas etc.	Verb
76	frazzle	exhaust physically or emotionally	Verb
77	camphorate	treat with camphor	Verb
78	arrange	arrange thoughts, ideas, temporal events	Verb
79	douse	put out, as of a candle or a light	Verb
80	serve	(sports) a stroke that puts the ball in play	Verb
81	jam	crush or bruise	Verb
82	bristle	a stiff hair	Verb
83	toe	drive obliquely	Verb

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
84	web	a fabric (especially a fabric in the process of being woven)	Verb
85	herd	cause to herd, drive, or crowd together	Verb
86	collar	take into custody	Verb
87	interrogate	pose a series of questions to	Verb
88	belt_out	sing loudly and forcefully	Verb
89	character	a written symbol that is used to represent speech	Verb
90	tithe	a levy of one tenth of something	Verb
91	moralize	improve the morals of	Verb
92	restrain	to close within bounds, or otherwise limit or deprive of free movement	Verb
93	begin	begin to speak or say	Verb
94	bode	indicate, as with a sign or an omen	Verb
95	foot	any of various organs of locomotion or attachment in invertebrates	Verb
96	break	come into being	Verb
97	retake	photograph again	Verb
98	canal	provide (a city) with a canal	Verb
99	connect	be or become joined or united or linked	Verb
100	uniformize	make uniform	Verb
101	wind	arrange or or coil around	Verb
102	flux	the rate of flow of energy or particles across a given surface	Verb
103	nest	furniture pieces made to fit close together	Verb
104	geminate	arrange or combine in pairs	Verb
105	mold	sculpture produced by molding	Verb
106	switch	railroad track having two movable rails and necessary connections; used to turn a train from one track to another or to store rolling stock	Verb
107	abstract	consider a concept without thinking of a specific example; consider abstractly or theoretically	Verb
108	transliterate	rewrite in a different script	Verb
109	undergird	lend moral support to	Verb
110	pyramid	(stock market) a series of transactions in which the speculator increases his holdings by using the rising market value of those holdings as margin for further purchases	Verb
111	light	not great in degree or quantity or number	Verb

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
112	specify	decide upon or fix definitely	Verb
113	make	favor the development of	Verb
114	fall	slope downward	Verb
115	wave	set waves in	Verb
116	babble	to talk foolishly	Verb
117	barter_away	trade in in a bartering transaction	Verb
118	cleave	make by cutting into	Verb
119	zap	kill with or as if with a burst of gunfire or electric current or as if by shooting	Verb
120	bunt	(baseball) the act of hitting a baseball lightly without swinging the bat	Verb
121	rough	causing or characterized by jolts and irregular movements	Verb
122	feel	undergo an emotional sensation or be in a particular state of mind	Verb
123	pan_out	be a success	Verb
124	score	make underscoring marks	Verb
125	rigidify	make rigid and set into a conventional pattern	Verb
126	exacerbate	exasperate or irritate	Verb
127	disarray	untidiness (especially of clothing and appearance)	Verb
128	put_on	add to something existing	Verb
129	satisfy	meet the requirements or expectations of	Verb
130	foreshow	foretell by divine inspiration	Verb
131	tape	the finishing line for a foot race	Verb
132	desalinate	remove salt from	Verb
133	strong-arm	be bossy towards	Verb
134	drench	cover with liquid; pour liquid onto	Verb
135	animate	give new life or energy to	Verb
136	sicken	make sick or ill	Verb
137	heave	move or cause to move in a specified way, direction, or position	Verb
138	focus	put (an image) into focus	Verb
139	warn	advise or counsel in terms of someone's behavior	Verb
140	crier	a peddler who shouts to advertise the goods he sells	Noun
141	badminton_court	the court on which badminton is played	Noun
142	american_lobster	flesh of cold-water lobsters having large tender claws; caught from Maine to the Carolinas	Noun

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
143	common_eland	dark fawn-colored eland of southern and eastern Africa	Noun
144	rope	fasten with a rope	Noun
145	picosecond	one trillionth (10≐12) of a second; one thousandth of a nanosecond	Noun
146	vasa_efferentia	the several highly convoluted tubules that lead from the rete testis to the vas deferens and form the head of the epididymis	Noun
147	geomyidae	North American pocket gophers	Noun
148	rhinitis	an inflammation of the mucous membrane lining the nose (usually associated with nasal discharge)	Noun
149	compression	the process or result of becoming smaller or pressed together	Noun
150	descent	the kinship relation between an individual and the individual's progenitors	Noun
151	male_chest	the chest of a man	Noun
152	enceliopsis	small genus of xerophytic herbs of southwestern United States	Noun
153	shoot	produce buds, branches, or germinate	Noun
154	indurated_clay	hardened clay	Noun
155	transmutation	an act that changes the form or character or substance of something	Noun
156	light	of the military or industry; using (or being) relatively small or light arms or equipment	Noun
157	trading_card	a card with a picture on it; collected and traded by children	Noun
158	terminus_ad_quem	final or latest limiting point	Noun
159	fare	the sum charged for riding in a public conveyance	Noun
160	wood_ant	reddish-brown European ant typically living in anthills in woodlands	Noun
161	portuguese	the Romance language spoken in Portugal and Brazil	Noun
162	canebrake_rattlesnake	southern variety	Noun
163	midterm_examination	an examination administered in the middle of an academic term	Noun
164	net	yield as a net profit	Noun
165	openness	characterized by an attitude of ready accessibility (especially about one's actions or purposes); without concealment; not secretive	Noun

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
166	nevoid_elephantiasis	thickening of the skin (usually unilateral on an extremity) caused by congenital enlargement of lymph vessel and lymph vessel obstruction	Noun
167	cyrtomium	small genus of tropical Asiatic greenhouse ferns; in some classifications placed in Polypodiaceae	Noun
168	drive	proceed along in a vehicle	Noun
169	stand	put into an upright position	Noun
170	morphology	the branch of geology that studies the characteristics and configuration and evolution of rocks and land forms	Noun
171	forlorn_hope	a hopeless or desperate enterprise	Noun
172	give	give (as medicine)	Noun
173	propoxyphene	a mildly narcotic analgesic drug (trade name Darvon) related to methadone but less addictive	Noun
174	deceptiveness	the quality of being deceptive	Noun
175	chamois	a soft suede leather formerly from the skin of the chamois antelope but now from sheepskin	Noun
176	shell_bean	a bean plant grown primarily for its edible seed rather than its pod	Noun
177	blue_elder	shrub or small tree of western United States having white flowers and blue berries; fruit used in wines and jellies	Noun
178	journal	a periodical dedicated to a particular subject	Noun
179	coal_black	a very dark black	Noun
180	ravaging	plundering with excessive damage and destruction	Noun
181	pyrochemical_process	processes for chemical reactions at high temperatures	Noun
182	charcoal	of a very dark grey	Noun
183	lactogen	any agent that enhances milk production	Noun
184	papering	the application of wallpaper	Noun
185	lost	not caught with the senses or the mind	Noun
186	thracian	of or relating to Thrace or its people or culture	Noun
187	black_guillemot	northern Atlantic guillemot	Noun
188	stop	cause to stop	Noun
189	dideoxyinosine	an antiviral drug used to combat HIV infection	Noun

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
190	trogoniformes	trogons	Noun
191	delay	time during which some action is awaited	Noun
192	heat	the sensation caused by heat energy	Noun
193	twenty-nine	the cardinal number that is the sum of twenty-eight and one	Noun
194	haggle	an instance of intense argument (as in bargaining)	Noun
195	schoolmarm	a woman schoolteacher (especially one regarded as strict)	Noun
196	systeme_international_d'unites	a complete metric system of units of measurement for scientists; fundamental quantities are length (meter) and mass (kilogram) and time (second) and electric current (ampere) and temperature (kelvin) and amount of matter (mole) and luminous intensity (candela)	Noun
197	child	a human offspring (son or daughter) of any age	Noun
198	lesser_knapweed	a weedy perennial with tough wiry stems and purple flowers; native to Europe but widely naturalized	Noun
199	blastoderm	a layer of cells on the inside of the blastula	Noun
200	hand_glass	light microscope consisting of a single convex lens that is used to produce an enlarged image	Noun
201	estimate	judge tentatively or form an estimate of (quantities or time)	Noun
202	ebenaceae	fruit and timber trees of tropical and warm regions including ebony and persimmon	Noun
203	gusset	a metal plate used to strengthen a joist	Noun
204	physical_anthropology	the branch of anthropology dealing with the genesis and variation of human beings	Noun
205	draw	earn or achieve a base by being walked by the pitcher	Noun
206	western_hemisphere	the hemisphere that includes North America and South America	Noun
207	administration	the persons (or committees or departments etc.) who make up a body for the purpose of administering something	Noun
208	ketonemia	an abnormal increase of ketone bodies in the blood as in diabetes mellitus	Noun

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
209	jack_plane	a carpenter's plane for rough finishing	Noun
210	stain	(microscopy) a dye or other coloring material that is used in microscopy to make structures visible	Noun
211	bowfin	primitive long-bodied carnivorous freshwater fish with a very long dorsal fin; found in sluggish waters of North America	Noun
212	register	a book in which names and transactions are listed	Noun
213	word-painter	a writer of vivid or graphic descriptive power	Noun
214	argonauta	type genus of the family Argonautidae paper nautilus	Noun
215	relational_adjective	an adjective that classifies its noun (e.g., 'a nervous disease' or 'a musical instrument	Noun
216	ocular_muscle	one of the small muscles of the eye that serve to rotate the eyeball	Noun
217	electrostatic_generator	electrical device that produces a high voltage by building up a charge of static electricity	Noun
218	rock_spikemoss	tufted spikemoss forming loose spreading mats; eastern North America	Noun
219	fatalism	a philosophical doctrine holding that all events are predetermined in advance for all time and human beings are powerless to change them	Noun
220	climbing_hydrangea	deciduous climber with aerial roots having white to creamy flowers in fairly flat heads	Noun
221	sight_setting	the adjustment of a gunsight for elevation and windage on a particular range under favorable light conditions	Noun
222	endemic	of or relating to a disease (or anything resembling a disease) constantly present to greater or lesser extent in a particular locality	Noun
223	testudo	a movable protective covering that provided protection from above; used by Roman troops when approaching the walls of a besieged fortification	Noun
224	stub	pull up (weeds) by their roots	Noun
225	charge	financial liabilities (such as a tax)	Noun
226	switch	make a shift in or exchange of	Noun
227	blephilia	small genus of North American herbs wood mints	Noun

Table A.1: (continued)

Pair	Keyword	Descriptive Text	Dataset
228	molarity	concentration measured by the number of moles of solute per liter of solution	Noun
229	bole	a soft oily clay used as a pigment (especially a reddish brown pigment)	Noun
230	procrastination	the act of procrastinating; putting off or delaying or defering an action to a later time	Noun
231	linum	a herbaceous plant genus of the family Linaceae with small sessile leaves	Noun
232	disco	popular dance music (especially in the late 1970s); melodic with a regular bass beat; intended mainly for dancing at discotheques	Noun
233	commune	the smallest administrative district of several European countries	Noun
234	anterior_pituitary	the anterior lobe of the pituitary body; primarily glandular in nature	Noun
235	volume	a relative amount	Noun
236	payment_rate	the amount of money paid out per unit time	Noun
237	stalk	the act of following prey stealthily	Noun
238	bivalvia	oysters; clams; scallops; mussels	Noun
239	chemise	a loose-fitting dress hanging straight from the shoulders without a waist	Noun
240	contentedness	the state of being contented with your situation in life	Noun
241	click	become clear or enter one’s consciousness or emotions	Noun
242	solicitation	request for a sum of money	Noun
243	bowling_pin	a club-shaped wooden object used in bowling; set up in triangular groups of ten as the target	Noun
244	treacle	writing or music that is excessively sweet and sentimental	Noun
245	treaty	a written agreement between two states or sovereigns	Noun
246	romance	relating to languages derived from Latin	Noun

Table A.2: Similarity evaluation questionnaire

Pair	Sentence 1	Sentence 2
1	Two boys run to the entryway of an old building.	A couple of children are eating popsicles at a birthday party.
2	A man and a woman stand in front of a Christmas tree contemplating a single thought.	Two people wonder how a Christmas tree got inside their house in the middle of July.
3	Phone to her ear, a woman bends forward at the side of a busy street.	Phone to her ear, a woman bends forward
4	Two men in hard hats and safety neon vests on scaffolding working outside in the daylight.	Outdoor daytime scene with two workers in safety gear working on scaffolding.
5	Young men drink coffee and read books just outside some buildings on a sunny day; an older bearded man in a dark blue shirt glances at a younger man in a striped shirt, who has a red toy crab on the table next to him.	Gandalf and the red toy crab do battle in the streets.
6	A dark-skinned man in white shirts and a black sleeveless shirt flips his skateboard on a cement surface surrounded by tall buildings and palm trees.	a man flips his skateboard on a cement surface
7	A woman in a red long-sleeved shirt and blue jeans stands at the lectern speaking at or about GLAM U WIKI.	A woman is giving a speech to an interested crowd.
8	A zaftig woman in a tube top and jeans dancing outdoors, with a guitarist behind her.	A woman is reading a book in her house.
9	A young male is running while playing tennis against another person.	An old woman sits on a bench.
10	A woman in a hooded type coat is holding a golden colored axe.	A woman holds a metal axe.
11	Older man wearing dark blue clothing sweeping the ground and a person wearing a red coat and carrying a green bag entering the building.	An old man is sitting at a cafe drinking coffee
12	A female gymnast in black and red being coached on bar skills.	The female gymnast is on her way to the gym.
13	The 3 dogs are cruising down the street.	a family of dogs cruises down the street
14	A man alone crosscountry skis in the wilderness while wearing a huge backpack.	A man skis in the wilderness

Table A.2: (continued)

Pair	Sentence 1	Sentence 2
15	A young boy with close-cropped hair, wearing a red robe, is holding a black kettle as someone is about to pour something in it.	a small boy runs through the corn field
16	A black man in a blue suit is talking on a cellphone while smoking.	The cigarette is lit
17	One many is fishing by a river with some buildings in the background while another man next to him is about to get a drink from the bottle next to him.	A man is in the desert
18	Man sleeping in a couch in a sitting position wearing a blue jeans and a plaided shirt	There is someone sleeping on the couch.
19	A lady and a man in a hat watch baseball from the stands.	Two people, one with a hat, can see baseball from where they are.
20	Biker riding through the forest.	Mountain biker enjoying the local trails.
21	A group of four children poke at two small turtles in the grass.	A group of kids were playing in the grass.
22	A woman working long hours.	A woman is working in a factory.
23	A man pulling items on a cart.	A man is moving items in a cart.
24	Three children are playing on a swing in the garden.	There are children outside.
25	A man talks to two guards as he holds a drink.	The man is holding a drink.
26	A male sitting down enjoy his coffee at a coffee shop.	A woman drinks.
27	A person strolling through an orange tinted woods with a dog.	The person is outside.
28	The team swiftly moves their traditional boat down the river.	The boat has sunk
29	A dog runs.	The dog is running outside
30	A lit girl splashes around in natural water.	The girl is sitting inside.
31	A young boy, wearing a jesters hat is enjoying himself sledding.	A young boy is just sitting down and not sledding.
32	Two guys on a couch, one is looking up the other is looking away with a cup in his hand.	Neither is holding anything.
33	A man in a hard hat, gray t-shirt and holding a cordless drill saluting in front of a large American flag.	A man saluting the American Flag.

Table A.2: (continued)

Pair	Sentence 1	Sentence 2
34	A boy wearing a yellow shirt and blue pants is climbing a tree in a hilly area.	There is one child in this picture, and he is outside.
35	Four young girls playing in the water.	Four girls are swimming.
36	People with bags walk along a dirt road in front of a green building.	Homeless people are walking outside.
37	Family gathered together in a house enjoying each other company.	A family is enjoying a night at home.
38	A person taking pictures of a young brunette girl.	A photographer is working.
39	Two men and a Frisbee	Two men with a toy.
40	A yellow dog is running in a field near a mountain.	A yellow dog is chasing a bunny in the field.
41	Two girls jumping on a trampoline, one upright and the other landing on her back, in a backyard.	The girls are in a rocket ship.
42	A busy street with numerous people interacting and going about their lives.	The morning rush hour fills the streets with busy people.
43	Young person in black next to an older person in black and white polka dot.	A young person in white next to an older person.
44	A woman wearing a red helmet is holding a rope and smiling.	A woman is trying to lasso a cow that escaped.
45	A bearded man, using a wooden pole, stirs bowl-like objects around in a large metal box filled with water.	A man has hair on his face.
46	A woman in an office making a phone call	A woman is making a phone call in an office environment.
47	A woman with a blue jacket around her waist is sitting on the ledge of some stone ruins resting.	A woman sits in front of ancient indian ruins
48	A team of surgeons operate on a female patient.	The surgeons are operating on a male patient.
49	A bunch of people are standing all together in a street with a building in the background.	A group of people are planning something.
50	A group of people are ice skating in a big city.	The people are outside skating.
51	A woman in a bathing suit is sitting on a dock over a body of water.	A woman sits outside on the dock overlooking the ocean.
52	Three sailboats with white sails are seen on a riverbank.	The boats are sailing on the ocean.
53	Men work on a city street.	The men are sleeping

Table A.2: (continued)

Pair	**Sentence 1**	**Sentence 2**
54	A man in a red sweatshirt pushes a giant redwood tree in a snowy forest.	A man pushes a redwood tree in the forest.
55	A dog catching a Frisbee.	An animal is making contact with a toy.
56	The young man in the plaid shirt is sitting on a chair that is out in the yard.	A guy is relaxing outside.
57	The young girl is swinging on a swing, and her hair is flying out behind her.	A girl is sleeping in bed.
58	A man holding a sign of a rainbow elephant.	The person holds a sign.
59	There is one man with light blue jeans on, and one with dark blue jeans on that are walking down the street.	Two men are making pancakes in their swimtrunks.
60	Three people are standing on a lit stage while groups of people sitting at tables face them.	There are only two people.
61	Two young girls hang tinsel on a Christmas tree in a room with blue curtains.	Two girls are decorating their Christmas tree.
62	A group is walking between two giant rock formations.	A group is swimming.
63	A young man in a white hat is playing golf with an older man wearing a blue, long-sleeved shirt and khakis.	There are people of differing ages outside playing golf.
64	A woman doing gymnastics, she is using the balance beam	A woman is doing a handstand on the balance beam.
65	Three soccer players, two in orange one in yellow, running for the ball on a soccer field.	There are some players chasing a ball.
66	Three young men in a kitchen standing around a stove, one of them is stirring something in a pot.	Three people are standing inside a house near each other.
67	A group of people watching a boy getting interviewed by a man.	A group of people are watching a boy and and interviewer.
68	A person in an orange shirt reaching up.	A person is reaching up to grab something.
69	A group of people are camping out on some rocks.	People are outdoors.
70	A boy dressed in an orange shirt and a helmet is riding a dirt bike in the woods.	A boy in orange runs through a local mall.
71	A large audience of people seated at banquet tables is watching the happenings on a stage surrounded by blue curtains.	The large audience is standing and cheering.

Table A.2: (continued)

Pair	Sentence 1	Sentence 2
72	A man in a red shirt is walking towards a blue market stall.	The man wants to purchase some produce.
73	Two people connected to ropes wearing bright orange helmets are sitting on a rocky crag high in the mountains.	Some people go on an adventure.
74	A group of people taking pictures on a walkway in front of a large body of water.	A man cleans up a spill.
75	The firemen and women wearing blue gloves have finished working on a person who is lying on ground bandaged and draped in a blue blanket.	The firefighters have finished a job.
76	Three men are using Washington Mutual ATMs outside near a parking lot.	Three men are playing poker.
77	I am squatting on a dock, looking into a lake.	I am at the lake with my family.
78	A man and a woman embracing on the sidewalk.	Two women embrace on the sidewalk.
79	A woman wearing colonial clothing is sitting on the grass reading to a young girl.	The people are riding in a plane.
80	A woman in a hooded type coat is holding a golden colored axe.	A woman holds an axe.
81	A man standing on a street with a suitcase in front of him while another man bends down to look at what is displayed on top of it.	One man shows the ransom money to the other.
82	Three puppies are in the tub being sprayed with water by a person.	Three puppies are in the tub being sprayed with water by the vet.
83	The man is throwing a toy to a brown and black dog in the park.	A man and a dog are at a park.
84	A woman with a stroller is passing a man walking a dog.	There are people outside.
85	The man in white is playing basketball against the man in blue.	The team wearing white is winning the game.
86	A man in a khaki jacket and ball cap walks down the street.	The man is outdoors walking.
87	A boy is standing in a pool getting splashed with water.	The boy is not in a pool.
88	Two dogs either fighting or playing together.	The dogs are in a meadow.

Table A.2: (continued)

Pair	**Sentence 1**	**Sentence 2**
89	A person on blue pants and a white shirt walks away from a dirt circle on multicolored grass.	a person walks a way from a dirt circle
90	Group of people wearing black shirts riding in an open top vehicle.	The driver was wearing a purple shirt.
91	People walking along the beach on a sunny day.	the beach is empty because of a hurricane.
92	Man straining to climb cliff face.	Man reaching up to secure a handhold on the cliff face.
93	A group of three people are digging a large trench in a yard while another person walks close by.	There is nobody near the trench.
94	A hockey player in blue and red guarding the goal.	a person playing a sport
95	A man in a gray coat is walking on a fallen tree trunk in a forest.	Two girls play tag in a forest.
96	The sun breaks through the trees as a child rides a swing.	A child rides a swing in the daytime.
97	A white-clad soccer player is attempting to intercept a soccer ball that is quickly approaching the goal box.	A soccer player tries to get an interception.
98	A little girl drawing with chalk, on a chalkboard.	A young girl is drawing on a chalkboard.
99	A young woman cooks a meal in a wok while conversing with another woman, as an illuminated shrine to Mr. T looks on.	A young woman stands near a stove.
100	A cowboy riding a horse at a rodeo.	cowboy competing in a rodeo

Table A.3: USE train set

Pair	Left sentence	Right sentence
1	an addiction to nicotine	nicotine_addiction
2	directly and without evasion; not roundabout	squarely
3	a false accusation of an offense or a malicious misrepresentation of someone's words or actions	hatchet_job
4	considered individually	respective
5	crimes created by statutes and not by common law	statutory_offence
6	a portable .30 caliber automatic rifle operated by gas pressure and fed by cartridges from a magazine; used by United States troops in World War I and in World War II and in the Korean War	BAR
7	in a sturdy manner	sturdily
8	preserve (tissue) with plastics, as for teaching and research purposes	plastinate
9	the quality of being extreme	extremeness
10	(chemistry) a reaction in which an elementary substance displaces and sets free a constituent element from a compound	displacement_reaction
11	ectopic pregnancy in the abdominal cavity	abdominal_pregnancy
12	shrubby or herbaceous low-growing evergreen perennials	genus_Armeria
13	a motor (especially an electric motor) that moves or rotates in small discrete steps	stepping_motor
14	a medicine for external application that produces redness of the skin	rubefacient
15	not of natural origin; prepared or made artificially	man-made
16	a soft-finned fish of the genus Alepisaurus	handsaw_fish
17	nonperformance of something distasteful (as by deceit or trickery) that you are supposed to do	evasion
18	a concluding action	closing
19	a genus of tropical American plants have sword-shaped leaves and a fleshy compound fruits composed of the fruits of several flowers (such as pineapples)	Ananas
20	wear away	rub_down
21	understand, usually after some initial difficulty	get_wise
22	New Zealand shrub	New_Zealand_mountain_pine
23	a hang performed on the rings or parallel bars with the body erect and the arms at the sides	straight_hang
24	gleam or glow intermittently	flash
25	the marks used to clarify meaning by indicating separation of words into sentences and clauses and phrases	punctuation
26	an orchid of the genus Liparis having a pair of leaves	twayblade

Table A.3: (continued)

Pair	Left sentence	Right sentence
27	a decorated bier on which a coffin rests in state during a funeral	catafalque
28	a leaf having a notch at the apex	emarginate_leaf
29	the characteristic utterance of an animal	cry
30	one of the two seasons in tropical climates	dry_season
31	English film director noted for his skill in creating suspense (1899-1980)	Hitchcock
32	the omnipotence of a divine being	God's_Will
33	a talker on television who talks directly into the cameras and whose upper body is all that is shown on the screen	talking_head
34	the content of a particular field of knowledge	knowledge_domain
35	plant that is a source of strophanthin	Strophanthus_kombe
36	cut or tear irregularly	lacerate
37	the semantic relation that holds between a whole and its parts	whole_to_part_relation
38	the British royal court	Court_of_Saint_James's
39	(of leaves or petals) having a smooth edge; not broken up into teeth or lobes	entire
40	copiously branched vine of Brazil having deciduous leaves and white flowers tinged with blue	potato_vine
41	require as useful, just, or proper	involve
42	a content word that qualifies the meaning of a noun or verb	qualifier
43	an intellectual who synthesizes or uses synthetic methods	synthesizer
44	(literally an undutiful herb) a variety of cotton rose	Filago_germanica
45	thick cream made from scalded milk	Devonshire_cream
46	a single complete execution of a periodically repeated phenomenon	oscillation
47	the predominant wind direction	prevailing_wind
48	the act of devising something	devisal
49	the quality of being becoming	becomingness
50	pantropical climber having white fragrant nocturnal flowers	moonflower
51	lay eggs	lay
52	a medicine or therapy that cures disease or relieve pain	remedy
53	the blade of a knife	knife_blade
54	the state capital and largest city of Colorado; located in central Colorado on the South Platte river	Denver

Table A.3: (continued)

Pair	Left sentence	Right sentence
55	an article giving opinions or perspectives	editorial
56	capital of the state of Florida; located in northern Florida	Tallahassee
57	a water-base paint that has a latex binder	latex
58	moody and melancholic	glum
59	the ability to identify the pitch of a tone	absolute_pitch
60	type genus of the Crocodylidae	Crocodilus
61	the power or right to give orders or make decisions	dominance
62	a tax that is levied when you are departing a country by land or sea or air	departure_tax
63	a kiln for making bricks	brickkiln
64	supply with railroad lines	railroad
65	become sticky	gum
66	web-footed Australian stilt with reddish-brown pectoral markings	banded_stilt
67	what a communication that is about something is about	substance
68	the entire geographical area drained by a river and its tributaries; an area characterized by all runoff being conveyed to the same outlet	watershed
69	visit informally and spontaneously	drop_in
70	using vehicles	motorized
71	combination of lenses at the viewing end of optical instruments	ocular
72	wild plum of northeastern United States having dark purple fruits with yellow flesh	Allegheny_plum
73	important pest of chrysanthemums	pale_chrysanthemum_aphid
74	accepted or habitual practice	custom
75	capture the attention or imagination of	grab
76	a republic on the southwestern shores of the Arabian Peninsula on the Indian Ocean; formed in 1990	Republic_of_Yemen
77	(mathematics) an attribute of a shape or relation; exact reflection of form on opposite sides of a dividing line or plane	correspondence
78	place where something began and flourished	home
79	a cell of an embryo	formative_cell
80	a mailer for airmail	airmailer
81	pursue someone sexually or romantically	chase_after
82	a hare's-foot fern of the genus Davallia	Australian_hare's_foot
83	run with the ball, in football	rush

Table A.3: (continued)

Pair	Left sentence	Right sentence
84	Israel's elite secret commando unit responsible for counterterrorist and top secret intelligence gathering and hostage rescue missions	Sayeret_Matkal
85	not fulfilling the same grammatical role of any of its constituents	exocentric
86	especially of muscles; bringing together or drawing toward the midline of the body or toward an adjacent part	adducent
87	insuring yourself by setting aside money to cover possible losses rather than by purchasing an insurance policy	self-insurance
88	the position of the head of the Department of Transportation	Secretary_of_Transportation
89	used of the older of two persons of the same name especially used to distinguish a father from his son	elder
90	firm chewy candy made from caramelized sugar and butter and milk	caramel
91	the measurement of viscosity	viscometry
92	expose to a chance of loss or damage	risk
93	cause to be directed or transmitted to another place	mail
94	a genus of evergreen climbers	genus_Ercilla
95	an industrial process for producing sodium carbonate from sodium chloride and ammonia and carbon dioxide	Solvay_process
96	black hornless breed from Scotland	Aberdeen_Angus
97	a building (usually abandoned) where drug addicts buy and use heroin	shooting_gallery
98	aromatic herb of southern Europe; cultivated in Great Britain as a potherb and widely as an ornamental	Salvia_sclarea
99	squeeze like a wedge into a tight space	wedge
100	(Greek mythology) winged goddess of victory; identified with Roman Victoria	Nike
101	put on clothes	get_dressed
102	an endorsement	okay
103	small bushy-tailed South American burrowing rodents	Chinchillidae
104	a syllabic script used in Greece in the 13th century B.C.	Linear_B
105	a specialist assigned to the staff of a diplomatic mission	attache

Table A.3: (continued)

Pair	Left sentence	Right sentence
106	(historical linguistics) an explanation of the historical origins of a word or phrase	etymologizing
107	speak as if delivering a sermon; express moral judgements	moralise
108	a disease blackening the leaves of pear and apple trees	fire_blight
109	a rigid layer of polysaccharides enclosing the membrane of plant and prokaryotic cells; maintains the shape of the cell and serves as a protective barrier	cell_wall
110	a print made using a stencil process in which an image or design is superimposed on a very fine mesh screen and printing ink is squeegeed onto the printing surface through the area of the screen that is not covered by the stencil	silk_screen_print
111	despite anything to the contrary (usually following a concession)	nonetheless
112	indefinite in time or position	one
113	the space in a ship or aircraft for storing cargo	cargo_deck
114	short and thick; as e.g. having short legs and heavy musculature	dumpy
115	one of the system of highways linking major cities in the 48 contiguous states of the United States	interstate
116	perennial erect herb with white flowers; circumboreal	northern_Jacob's_ladder
117	form or join a union	unionise
118	the face veil worn by Muslim women	yashmak
119	a duct formed by the hepatic and cystic ducts; opens into the duodenum	common_bile_duct
120	a white crystalline double sulfate of aluminum	the ammonium double sulfate of aluminum: ammonium_alum
121	a room that is virtually free of dust or bacteria; used in laboratory work and in assembly or repair of precision equipment	clean_room
122	North American songbird whose call resembles a cat's mewing	grey_catbird
123	the quality of conforming to standards of propriety and morality	decency
124	king snakes and milk snakes	genus_Lampropeltis
125	a room in a barracks where soldiers are billeted	squad_room
126	a United Nations agency created to assist developing nations by loans guaranteed by member governments	World_Bank
127	freshly made or left	warm

Table A.3: (continued)

Pair	Left sentence	Right sentence
128	chicken cooked in a sauce made with tomatoes, garlic, and olive oil	chicken_provencale
129	not basic or fundamental	unessential
130	a light shade of blue	sapphire
131	mat-forming lithophytic or terrestrial fern with creeping rootstocks and large pinnatifid fronds found throughout North America and Europe and Africa and east Asia	golden_maidenhair
132	a rich and spectacular ceremony	pageantry
133	large genus of epiphytic or lithophytic unarmed cacti with usually segmented stems and pendulous branches; flowers are small followed by berrylike fruits	Rhipsalis
134	European perennial plant naturalized in United States having triangular ovate leaves and lilac-pink flowers	white_mallow
135	(Greek mythology) a mythical monster with three heads that was slain by Hercules	Geryon
136	of or relating to Nicaragua or is people	Nicaraguan
137	branched candlestick; ornamental; has several lights	candelabra
138	with confidence; in a confident manner	confidently
139	one of two classical Hindu epics; a great collection of poetry worked into and around a central heroic narrative (eight times as large as the Iliad and Odyssey combined)	Hastinapura
140	dried in a kiln	kiln-dried
141	a radiogram made by exposing photographic film to X rays; used in medical diagnosis	X_ray
142	We hope all parties will continue to make efforts and continue the process of dialogue, the Chinese Foreign Ministry said in a statement.	China's foreign ministry said: "We hope all parties will continue to make efforts and continue the process of dialogue."
143	fulmars	genus_Fulmarus
144	wearing footgear	shod
145	belt attaching you to some object as a restraint in order to prevent you from getting hurt	safety_belt
146	an intervening period or episode	interlude
147	a wart occurring on the sole of the foot	plantar_wart
148	United States film maker born in the United States but an Irish citizen after 1964 (1906-1987)	John_Huston
149	convert (an image) into pixels	rasterize

Table A.3: (continued)

Pair	Left sentence	Right sentence
150	most important element	chief
151	a Marxist-Leninist terrorist organization that conducted several attacks in western Europe	PFLP-GC
152	a large cave with basaltic pillars on Staffa island in Scotland	Fingal's_Cave
153	very large carnivorous sea turtle; wide-ranging in warm open seas	loggerhead
154	a unit of magnetic flux equal to 100,000,000 maxwells	Wb
155	the point on the Earth's surface directly above the focus of an earthquake	epicenter
156	having equal dimensions or measurements	isometric
157	slow-growing procumbent evergreen shrublet of northern North America and Japan having white flowers and numerous white fleshy rough and hairy seeds	creeping_snowberry
158	European breed of small dog resembling a terrier with dark wiry hair and a tufted muzzle	monkey_pinscher
159	a space that is contained within another space	subspace
160	evidencing little spirit or courage; overly submissive or compliant; ; - Orville Prescott	spiritless
161	Russian poet who expressed the feelings of the post-Stalinist generation (born in 1933)	Yevgeni_Yevtushenko
162	bottom dwelling marine warm water fishes with two barbels on the chin	mullet
163	the person who holds the secretaryship of the Department of Energy	Secretary_of_Energy
164	specifically design a product, event, or activity for a certain public	calculate
165	a film or TV program presenting the facts about a person or event	docudrama
166	German arms manufacturer and inventor of a repeating rifle and pistol (1838-1914)	Mauser
167	a United States bill worth 5 dollars	five_dollar_bill
168	beat soundly	belabor
169	a French bean variety with light-colored seeds; usually dried	haricot
170	a town on Long Island in New York; site of Belmont Park	Elmont
171	customers collectively	business
172	feelings of excessive pride	self-love
173	a fabric made by knitting	knit

Table A.3: (continued)

Pair	Left sentence	Right sentence
174	exchange people temporarily to fulfill certain jobs and functions	alternate
175	stemming from evil characteristics or forces; wicked or dishonorable; ; ; ; ; ; ; -Thomas Hardy	black
176	Italian pope from 795 to 816 who in 800 crowned Charlemagne emperor of the Romans (750-816)	Leo_III
177	shaped in the form of a spade	spade-like
178	a wandering or disorderly grouping (of things or persons)	straggle
179	perennial gumweed of western and central North America	Grindelia_squarrosa
180	an acute infection of the intestine by shigella bacteria; characterized by diarrhea and fever and abdominal pains	shigellosis
181	a small light typewriter; usually with a case in which it can be carried	portable
182	mottled or spotted bean of southwestern United States; usually dried	pinto_bean
183	by stochastic means	stochastically
184	a person skilled in a particular type of therapy	healer
185	breathe spasmodically, and make a sound	hiccough
186	someone whose career progresses rapidly	ball_of_fire
187	true (leptosporangiate) ferns	order_Polypodiales
188	United States physicist who studied electromagnetic phenomena (1791-1878)	Henry
189	of or relating to the study of the origins and development of human beings	anthropogenetic
190	a heavy silk fabric with a rough surface (or a cotton imitation)	shantung
191	a button that you press to activate the horn of an automobile	horn_button
192	a genus of mammals of the family Bovidae	genus_Anoa
193	someone who plays a musical instrument (as a profession)	instrumentalist
194	European herb somewhat resembling celery widely naturalized in Britain coastal regions and often cultivated as a potherb	Smyrnium_olusatrum
195	the hard durable wood of any oak; used especially for furniture and flooring	oak
196	with hand brought forward and down from above shoulder level	overhand

Table A.3: (continued)

Pair	**Left sentence**	**Right sentence**
197	having achieved a comfortable relation with your environment	adjusted
198	(surgery) the act of puncturing a body cavity or organ with a hollow needle in order to draw out fluid	centesis
199	in a hoarse or husky voice	huskily
200	loud and harsh sounding rock music with a strong beat; lyrics usually involve violent or fantastic imagery	heavy_metal_music
201	small carnivorous freshwater percoid fishes of North America usually having a laterally compressed body and metallic luster	crappies; black bass; bluegills; pumpkinseed: sunfish
202	the angle formed by the inner sides of the legs where they join the human trunk	crotch
203	an act of delaying or interrupting the continuity	interruption
204	swelling from excessive accumulation of watery fluid in cells, tissues, or serous cavities	hydrops
205	the termination of someone's employment (leaving them free to depart)	firing
206	pitched an octave below normal bass instrumental or vocal range	contrabass
207	a contestant who loses the contest	loser
208	an international alliance involving many different countries	world_organisation

Table A.4: USE test set

Pair	Left sentence	Right sentence
1	the cardinal number that is the sum of three and one	quatern
2	the fifth sign of the zodiac; the sun is in this sign from about July 23 to August 22	Leo
3	German composer of many operas; collaborated with librettist Hugo von Hoffmannsthal to produce several operas (1864-1949)	Strauss
4	capable of being diagnosed	diagnosable
5	having or deserving or conferring glory	glorious
6	a critical state; especially the point at which a nuclear reaction is self-sustaining	criticality
7	(botany) a living organism lacking the power of locomotion	plant_life
8	a substance that unites or bonds surfaces together	adhesive
9	in a tantalizing manner	tantalizingly
10	an attractive outfit	bib-and-tucker
11	United States parapsychologist (1895-1980)	J._B._Rhine
12	a legal holiday in the United States	July_4
13	an inclination to support or be loyal to or to agree with an opinion	sympathy
14	Canadian hockey player who holds the record for playing the most games (born 1928)	Gordon_Howe
15	European perennial	wild_liquorice
16	highly attractive and able to arouse hope or desire	tempting
17	a print made by an impression of the ridges in the skin of a finger; often used for biometric identification in criminal investigations	fingerprint
18	an impairment in understanding spoken language that is not attributable to hearing loss	word_deafness
19	a stock made with chicken	chicken_broth
20	deciduous shrubby tree of Europe and western Asia having grey leaves and small yellow fruits covered in silvery scales; sometimes spiny	silver_berry
21	New Zealand with pinnate fronds and a densely woolly stalks; sometimes included in genus Todea	Leptopteris_superba
22	a city in southwestern Tanzania	Mbeya
23	the state of being heterozygous; having two different alleles of the same gene	heterozygosity
24	a U-shaped curve in a stream	oxbow
25	large genus of chiefly tropical herbs with showy flowers and mostly globose fruits: globe mallows	genus_Sphaeralcea
26	a hand-operated lever that controls the throttle valve	hand_throttle

Table A.4: (continued)

Pair	Left sentence	Right sentence
27	free-swimming tadpole-shaped pelagic tunicate resembling larvae of other tunicates	appendicularia
28	an imaginary place for lost or neglected things	limbo
29	a region in southeastern Italy on the Adriatic	Puglia
30	act in opposition to	counteract
31	large genus of epiphytic and lithophytic orchids of tropical and subtropical Americas and West Indies; formerly included in genus Epidendrum	Encyclia
32	a mass of undifferentiated cells from which an organ or body part develops	blastema
33	make imperfect	spoil
34	when a wrestler's shoulders are forced to the mat	pin
35	excessive thirst (as in cases of diabetes or kidney dysfunction)	polydipsia
36	express willingness to have in one's home or environs	receive
37	in a wry manner	wryly
38	a region of northeastern France famous for its wines	Alsatia
39	a specialist in oncology	oncologist
40	plant with mostly basal leaves and slender open racemes of white or pale pink flowers; prairies and open forest of northwestern United States to British Columbia and Alberta	prairie_star
41	important food fish on both sides of the Atlantic; related to cod but usually smaller	haddock
42	a bit that is used for beveling	chamfer_bit
43	not bent	unbent
44	the mystical theological doctrine of Jakob Boehme that influenced the Quakers	Boehmenism
45	a voter who votes illegally at different polling places in the same election	floater
46	a tray (or large plate) for serving food or drinks; usually made of silver	salver
47	a thief who steals from the pockets or purses of others in public places	pickpocket
48	a witness whose testimony is both relevant to the matter at issue and required in order to resolve the matter	material_witness
49	a mischievous elf in Irish folklore	leprechaun
50	the first coherent school of American art; active from 1825 to 1870; painted wilderness landscapes of the Hudson River valley and surrounding New England	Hudson_River_school

Table A.4: (continued)

Pair	Left sentence	Right sentence
51	small Eurasian shrub having clusters of yellow flowers that yield a dye; common as a weed in Britain and the United States; sometimes grown as an ornamental	dyer's-broom
52	the mechanical advantage gained by being in a position to use a lever	purchase
53	stick to firmly	bond
54	fungi having a zygote or a single cell developing directly into an ascus	order_Endomycetales
55	bottled or freshly squeezed juice of oranges	orange_juice
56	(computer science) a reference or value that is passed to a function, procedure, subroutine, command, or program	argument
57	a day set aside for doing household laundry	washing_day
58	any of various trees having yellowish wood or yielding a yellow extract	yellowwood
59	United States electrical engineer noted for his work on the theory of alternating currents; independently of Oliver Heaviside he discovered the existence of an atmospheric layer that reflects radio waves back to earth (1861-1939)	Arthur_Edwin_Kennelly
60	characterized by assertion of unproved or unprovable principles	dogmatical
61	freeing from error or corruption	antiseptic
62	a temporary stay (e.g., as a guest)	sojourn
63	a fungus family of loose smuts	Ustilaginaceae
64	of or relating to or belonging to mammals of the order Artiodactyla	artiodactyl
65	a pump that draws air or another gas through a liquid	aspirator
66	in another and different manner	other_than
67	an inhabitant of an eastern area; especially of the U.S.	easterner
68	bees	superfamily_Apoidea
69	an Apocryphal book that was a popular novel for several centuries	Tobit
70	(nautical) a line that holds an object (especially a boat) in place	mooring
71	the edible white meat of a coconut; often shredded for use in e.g. cakes and curries	coconut
72	the scads (particularly horse mackerels)	Trachurus
73	the side upon which the use of a thing depends (usually the most prominent surface of an object)	face

Table A.4: (continued)

Pair	Left sentence	Right sentence
74	a public instance of reciting or repeating (from memory) something prepared in advance	reading
75	any of the animal viruses that cause painful blisters on the skin	herpes_virus
76	a typeface with thick heavy lines	boldface
77	widespread European weed with spiny tongue-shaped leaves and yellow flowers; naturalized in United States	oxtongue
78	arrange (open windows) on a computer desktop so that they overlap each other, with the title bars visible	cascade
79	keen and shared excitement	electricity
80	drawing of fluid or inflammation away from a diseased part of the body	derivation
81	the energy that an atomic system must acquire before a process (such as an emission or reaction) can occur	energy_of_activation
82	Australian mound bird; incubates eggs naturally in sandy mounds	mallee_fowl
83	shrub of eastern North America closely resembling silky cornel	Cornus_obliqua
84	remorse caused by feeling responsible for some offense	guilt_trip
85	a hemoprotein composed of globin and heme that gives red blood cells their characteristic color; function primarily to transport oxygen from the lungs to the body tissues	hemoglobin
86	soft creamy candy	fudge
87	headdress that protects the head from bad weather; has shaped crown and usually a brim	lid
88	the phenomenon of floating (remaining on the surface of a liquid without sinking)	flotation
89	tropical American trees	genus_Cedrela
90	making or willing to make concessions	compromising
91	any of several herbs of the genus Browallia cultivated for their blue or violet or white flowers	browallia
92	a legislator who gives long speeches in an effort to delay or obstruct legislation that he (or she) opposes	filibuster
93	any of several erect biennial herbs of temperate Eurasia having stout taproots and producing burs	clotbur
94	tropical African climbing plant having yellow flowers with a dark purple center	black-eyed_Susan_vine

Table A.4: (continued)

Pair	**Left sentence**	**Right sentence**
95	plant hoppers: lantern flies	Fulgoridae
96	annual of tropical South America having edible purple fruits	cape_gooseberry
97	type genus of the family Rhodymeniaceae	Rhodymenia
98	throw or toss with a light motion	toss
99	in a clumsy manner	clumsily
100	not lessened or diminished	unrelieved

Table A.5: Generator sentences

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
1	instrumentality created	0.2242	0.2168	0.1972	0.1406	0.1620	0.6505	-0.3914	-0.4435	-0.1992	-0.0714
2	persons signature	0.2855	0.2807	-0.2993	-0.6392	0.3018	-0.3865	-0.1098	-0.1506	-0.2459	0.0769
3	male child	0.6041	0.4301	-0.0806	-0.0838	0.1777	-0.2443	0.4508	-0.3845	0.2986	0.2568
4	male offspring	-0.0501	0.4261	0.9568	-0.1363	0.3927	0.2659	0.8169	-0.2178	0.6020	-0.4026
5	car powered	-0.3587	0.0835	0.7150	-0.3555	0.1489	0.1148	0.2128	0.1335	0.3210	-0.1590
6	small piece land	-0.2999	0.2689	0.5196	0.1173	0.4909	1.5794	1.1814	0.0066	0.1758	0.6039
7	male chicken ro	-0.0034	0.1514	0.4490	-0.0007	0.4467	0.8135	1.0072	-0.1214	0.2908	0.1870
8	$<$ |None| $>$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	small wading bird	-0.3151	0.2719	0.6063	0.0356	0.6042	1.5036	1.2520	0.0854	0.2611	0.7313
10	soft bag cloth	0.6811	0.4018	0.3937	0.3573	0.2714	0.3982	1.2458	-0.6624	0.5078	-0.7750
11	foodstuff obtained plant	0.6793	0.4044	0.3831	0.3556	0.2810	0.4008	1.2497	-0.6622	0.5093	-0.7755
12	iron gate	0.0566	0.2028	0.3565	0.0964	0.3979	0.8924	0.9531	-0.1149	0.2471	0.0550
13	ornamental carving	0.6237	0.2997	0.1786	0.3037	0.3530	0.3597	1.2076	-0.7556	0.5627	-0.7885
14	hard brittle	-0.4210	0.1504	0.1462	0.1614	0.1369	0.7450	0.2935	0.1098	-0.0510	0.2795
15	burial ground often	-0.6194	0.0571	0.2949	-0.0452	0.3885	1.3486	0.5786	0.3529	-0.0249	1.0134
16	sheep	0.2719	0.1594	0.2906	0.0780	0.1873	0.1180	0.7055	-0.3373	0.2780	-0.2690
17	natural elevation earth	-0.2506	0.1647	-0.1161	-0.0246	0.4024	1.5549	0.0886	0.1037	0.3298	0.9110
18	day	-0.3013	0.0032	0.1205	-0.0301	0.2357	0.6949	0.3017	0.2017	0.0095	0.4724

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
19	often used combination	0.2244	0.2169	0.1973	0.1412	0.1620	0.6502	-0.3913	-0.4436	-0.1989	-0.0716
20	person	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
21	device consisting	0.3426	0.2008	0.1951	0.1786	0.1359	0.1979	0.6248	-0.3316	0.2559	-0.3844
22	course travel	-0.5863	0.0716	0.2459	-0.0399	0.4553	1.3607	0.5825	0.4171	0.0153	1.0032
23	person practices magic	0.0054	0.2031	-0.3311	-0.0223	0.2448	0.7014	-0.2879	0.4762	-0.4064	0.2211
24	place refuge	-0.0135	0.2477	0.3447	0.1585	0.2851	0.8705	0.8902	-0.1675	0.2286	0.1496
25	passenger vehicle designed	0.4785	0.4111	0.3853	0.6622	-0.1266	0.2020	0.5993	-0.3770	-0.2797	-0.4434
26	large family birds	0.1956	0.3793	0.6326	0.1994	0.5619	1.4811	0.6074	-0.5564	0.1137	0.1307
27	place bones	-0.0071	0.2479	0.3469	0.1592	0.2775	0.8646	0.8913	-0.1688	0.2273	0.1525
28	large tract land	-0.3078	0.2516	0.4661	0.1290	0.5154	1.5600	1.1916	0.0318	0.2357	0.6221
29	plant yields	0.0379	0.2053	0.3146	0.1480	0.3841	0.8966	0.9295	-0.1279	0.2643	0.0853
30	natural elevation	0.0692	0.1109	-0.2887	-0.0092	0.2516	0.9010	-0.1862	-0.0480	0.3638	0.3692
31	piece brass	-0.6283	0.0106	0.3164	-0.1036	0.3848	1.3676	0.6165	0.3646	-0.0347	0.9571
32	unbroken chain	0.2242	0.2168	0.1971	0.1402	0.1619	0.6506	-0.3915	-0.4436	-0.1990	-0.0716
33	traveling place	-0.6369	0.1149	0.2749	-0.0298	0.3588	1.3324	0.5449	0.3785	-0.0224	1.0682
34	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
35	hospital confinement	0.0370	0.2056	0.3158	0.1483	0.3844	0.8973	0.9281	-0.1284	0.2661	0.0851
36	person	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
37	time day	-0.6009	0.0065	0.2429	-0.0599	0.4735	1.3895	0.6055	0.4030	0.0187	0.9440

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
38	healer skilled	0.2634	0.2281	0.0241	0.1240	0.2640	0.0020	0.1081	-0.5649	0.3793	0.4201
39	bag case made	0.0436	0.2042	0.3174	0.1488	0.3731	0.8931	0.9277	-0.1307	0.2656	0.0862
40	male domestic f	0.1670	0.3522	-0.2206	0.2852	0.0310	1.0969	1.1903	-0.4539	0.2394	0.0054
41	profoundly wise person	-0.3014	0.0039	0.1233	-0.0298	0.2370	0.6939	0.3037	0.2017	0.0078	0.4716
42	person condition	-0.3012	0.0042	0.1233	-0.0295	0.2377	0.6941	0.3034	0.2013	0.0073	0.4711
43	area lying directly	-0.6473	0.0484	0.2739	-0.0499	0.3826	1.3605	0.5699	0.3663	-0.0167	1.0095
44	persons	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
45	facial expression resembling	-0.6009	0.0072	0.2459	-0.0594	0.4759	1.3885	0.6079	0.4034	0.0168	0.9424
46	portable fixed apparatus	0.6863	0.4021	0.3896	0.3565	0.2726	0.3964	1.2486	-0.6643	0.5134	-0.7698
47	thread cord	0.6839	0.4006	0.3902	0.3570	0.2716	0.4009	1.2523	-0.6688	0.5013	-0.7662
48	jumper	0.3402	-0.3145	0.1342	0.1208	0.0346	0.3046	0.4506	-0.2923	0.2753	-0.4795
49	provide firm	0.1605	-0.0060	-0.2803	0.0300	-0.1394	0.0561	-0.4532	-0.3791	-0.6987	-0.1296
50	undeserving	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
51	tend pull	-0.2996	0.0033	0.1223	-0.0299	0.2377	0.6946	0.3039	0.2013	0.0091	0.4716
52	disent	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
53	disdainful grimace	-0.3005	-0.0004	0.1228	-0.0300	0.2371	0.6934	0.3038	0.2026	0.0114	0.4725
54	message sent	-0.6013	0.0020	0.2424	-0.0604	0.4741	1.3890	0.6052	0.4043	0.0201	0.9438
55	intercept	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
56	tw	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
57	discover	0.2908	-0.4037	0.0939	0.0950	0.0828	-0.5514	-0.2114	0.3696	0.2363	0.4275
58	flat	-0.3267	0.0559	0.1754	-0.0142	0.1535	0.6602	0.2767	0.1567	-0.0403	0.5440
59	sound made	-0.3014	0.0039	0.1232	-0.0296	0.2370	0.6939	0.3037	0.2019	0.0080	0.4716
60	wake_up	-0.2338	0.1165	0.4601	0.1812	0.1159	0.0857	0.0681	0.0566	0.2400	-0.1812
61	move place	-0.6498	0.0498	0.2737	-0.0496	0.3790	1.3610	0.5702	0.3640	-0.0195	1.0083
62	make_tries	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
63	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
64	short	-0.3480	0.0453	0.1518	-0.0191	0.1440	0.6668	0.2642	0.1629	-0.0273	0.5393
65	person holds	-0.2996	0.0033	0.1224	-0.0299	0.2377	0.6946	0.3039	0.2013	0.0091	0.4716
66	throw_	-0.2996	0.0033	0.1223	-0.0299	0.2377	0.6946	0.3039	0.2013	0.0091	0.4716
67	part organism	-0.0780	0.1599	0.2956	0.0340	0.3536	0.7820	0.8421	-0.0577	0.2241	0.1341
68	start move	-0.5993	0.0065	0.2447	-0.0597	0.4755	1.3892	0.6077	0.4026	0.0182	0.9433
69	know	-0.2994	0.0033	0.1221	-0.0299	0.2378	0.6944	0.3040	0.2015	0.0089	0.4719
70	form sclerot	-0.2923	-0.0078	0.1573	-0.0828	0.2528	0.6941	0.3318	0.2049	0.0013	0.4302
71	secure rivet	0.6175	-0.1860	0.2868	0.2654	0.1825	-0.3424	0.3978	0.0545	0.4776	0.0712
72	part leg	-0.5929	0.0362	0.3277	-0.1254	0.4755	1.3921	0.6098	0.3890	-0.0499	0.8999
73	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
74	cause happen happen	-0.2997	0.0033	0.1224	-0.0299	0.2377	0.6946	0.3038	0.2013	0.0091	0.4716
75	long tube made	0.3413	0.2035	0.1957	0.1784	0.1369	0.1972	0.6243	-0.3302	0.2582	-0.3857
76	exhaust physically emotionally	-0.3146	0.2047	0.0860	0.1370	0.1236	0.8175	0.3233	0.0220	0.0285	0.2038
77	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
78	put put_in	-0.3491	0.4366	0.3866	0.1158	0.3828	0.8174	0.4240	0.2813	-0.5816	1.0109
79	fire	-0.2997	0.0033	0.1224	-0.0298	0.2377	0.6946	0.3038	0.2013	0.0091	0.4716

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
80	stroke puts	-0.5876	0.0684	0.2457	-0.0401	0.4547	1.3594	0.5824	0.4182	0.0177	1.0041
81	fragment	-0.2991	0.0433	0.1702	-0.0443	0.2236	0.7008	0.2762	0.1831	-0.0461	0.4721
82	stiffening hair	0.0684	0.2732	0.3114	0.1646	0.3697	0.8588	0.8964	-0.1074	0.2613	0.1459
83	make wry	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
84	fabric especially	0.3379	0.2009	0.1987	0.1787	0.1361	0.1997	0.6219	-0.3304	0.2513	-0.3895
85	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
86	custody	-0.3012	0.0040	0.1233	-0.0295	0.2376	0.6942	0.3032	0.2015	0.0073	0.4714
87	prepare prepare	-0.4675	0.2329	0.9202	-0.4425	0.1587	0.1281	0.1075	0.0927	-0.5691	0.1613
88	move clap	-0.5993	0.0065	0.2447	-0.0597	0.4755	1.3892	0.6077	0.4026	0.0183	0.9432
89	written letter	0.3388	0.2014	0.1945	0.1781	0.1303	0.1983	0.6258	-0.3291	0.2556	-0.3917
90	tax head	-0.6016	0.0466	0.2926	-0.0724	0.4595	1.3961	0.5831	0.3862	-0.0434	0.9403
91	cause morally sound	-0.6011	0.0072	0.2455	-0.0595	0.4747	1.3885	0.6075	0.4032	0.0171	0.9432
92	restrain body	-0.0611	0.4221	0.1685	0.4693	-0.4056	0.1091	0.5199	-0.4450	0.5092	0.1520
93	give voice	-0.5469	0.2140	0.3555	0.1047	0.3863	1.2700	0.5530	0.5433	-0.0132	1.0102
94	indicate	0.2754	-0.3888	0.0915	0.0903	0.0471	-0.5422	-0.2266	0.3848	0.2237	0.4592
95	various organs	-0.2714	-0.0035	0.1645	-0.0868	0.2505	0.6990	0.3147	0.2025	0.0134	0.4471
96	material	-0.2848	-0.0069	0.1583	-0.0829	0.2521	0.6960	0.3298	0.2005	-0.0053	0.4246
97	field soft sil	-0.3481	0.0453	0.1517	-0.0193	0.1439	0.6666	0.2645	0.1628	-0.0271	0.5393
98	provide	-0.0637	-0.2229	-0.4774	-0.1107	-0.3014	-0.5944	-0.0618	0.0644	-0.4995	-0.0582
99	become part	-0.2036	-0.3756	-0.2479	-0.2084	0.0743	0.2805	0.2584	0.3766	-0.5220	0.6200
100	unconfined state	-0.3474	0.0437	0.1500	-0.0207	0.1427	0.6676	0.2665	0.1633	-0.0274	0.5384
101	form web web	-0.9586	0.1005	0.4980	-0.1105	0.4955	2.0144	0.8416	0.5471	-0.0251	1.5180
102	rate flow	-0.5856	0.0644	0.2429	-0.0648	0.4627	1.3589	0.5817	0.4275	0.0076	0.9988

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
103	several ornate	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
104	pair	0.2245	0.2169	0.1972	0.1410	0.1619	0.6502	-0.3913	-0.4438	-0.1991	-0.0716
105	piece wood	-0.6248	0.0160	0.3185	-0.1011	0.3833	1.3765	0.6120	0.3574	-0.0364	0.9525
106	conic section	-0.6023	0.0040	0.2461	-0.0607	0.4744	1.3872	0.6067	0.4038	0.0186	0.9444
107	consider abstractly theoretically	0.2754	-0.3888	0.0915	0.0903	0.0471	-0.5422	-0.2266	0.3848	0.2237	0.4592
108	novel Edgar Rice	-0.3009	0.0000	0.1222	-0.0301	0.2371	0.6933	0.3039	0.2023	0.0115	0.4726
109	$< \|None\| >$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
110	exchange buying	-0.5839	0.0677	0.2196	-0.0605	0.4630	1.3684	0.5655	0.4159	0.0299	1.0052
111	much common	-0.6492	0.0479	0.2722	-0.0491	0.3818	1.3618	0.5661	0.3650	-0.0184	1.0106
112	make	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
113	treat respect	-0.7199	0.1536	0.2683	0.1322	0.3747	1.4383	0.5969	0.3119	-0.0420	0.7517
114	descend	0.2753	-0.3888	0.0915	0.0903	0.0470	-0.5422	-0.2266	0.3848	0.2237	0.4592
115	make whizzing	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
116	talk_out_of	0.2754	-0.3888	0.0915	0.0903	0.0471	-0.5422	-0.2266	0.3848	0.2237	0.4592
117	price	-0.3012	0.0040	0.1232	-0.0296	0.2381	0.6938	0.3040	0.2023	0.0076	0.4709
118	fastener	0.3427	0.2006	0.1953	0.1784	0.1347	0.2011	0.6249	-0.3353	0.2515	-0.3842
119	throw stream	-0.5936	0.0508	0.2951	-0.0585	0.4732	1.3907	0.5773	0.4000	-0.0385	0.9410
120	swat	-0.2996	0.0033	0.1223	-0.0298	0.2377	0.6946	0.3039	0.2013	0.0091	0.4716
121	marked resulting	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
122	patient suffer	-0.2998	-0.0596	-0.4506	-0.7303	0.1891	-0.2375	0.0716	0.0639	-0.1654	-0.4257
123	make success	-0.5991	0.0066	0.2446	-0.0597	0.4756	1.3890	0.6078	0.4028	0.0179	0.9435

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
124	emphasize	0.2754	-0.3888	0.0915	0.0903	0.0471	-0.5422	-0.2267	0.3848	0.2237	0.4592
125	make rigid set	-0.0750	0.2202	0.3194	0.1111	0.3997	1.3446	-0.0873	-0.2423	-0.1902	0.4003
126	beg help	-0.3334	0.4650	0.3887	0.1254	0.3990	0.8282	0.4049	0.2765	-0.5964	0.9856
127	quality unatt	-0.3012	0.0039	0.1232	-0.0296	0.2381	0.6937	0.3039	0.2022	0.0077	0.4708
128	become fixed	0.0879	-0.3681	-0.4057	-0.1252	-0.1773	-0.4111	-0.0724	0.1720	-0.5226	0.1907
129	within certain requirements	-0.2997	0.0032	0.1224	-0.0299	0.2377	0.6946	0.3039	0.2012	0.0091	0.4716
130	tell future	-0.0261	-0.3848	0.2144	0.0607	0.2846	0.1519	0.0771	0.5868	0.2309	0.9300
131	finishing line	-0.0769	0.2206	0.3206	0.1106	0.4003	1.3444	-0.0874	-0.2416	-0.1913	0.3992
132	dec	-0.3488	0.0473	0.1488	-0.0190	0.1443	0.6676	0.2632	0.1684	-0.0225	0.5375
133	lose temper become	-0.2398	0.0774	-0.0184	0.0088	0.2118	0.4127	0.3448	0.4387	-1.1156	1.1805
134	drain head	-0.6012	0.0469	0.2918	-0.0718	0.4614	1.3950	0.5826	0.3858	-0.0413	0.9421
135	give new life energy	-0.8311	0.2788	0.4535	0.0728	0.6113	1.9444	0.8145	0.7566	0.0067	1.5433
136	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
137	move cause move	-0.8989	0.0098	0.3670	-0.0896	0.7132	2.0838	0.9116	0.6039	0.0273	1.4149
138	put image	-0.6003	0.0034	0.2443	-0.0600	0.4750	1.3877	0.6078	0.4039	0.0203	0.9445
139	express opinion hope	-0.2966	0.2033	-0.2079	-0.0518	0.4822	1.3946	0.0152	0.6782	-0.3967	0.6933
140	peddler shouts	-0.0342	0.2656	0.2277	0.1189	0.3651	0.7055	-0.2641	0.6213	-0.3297	0.9126
141	court bad	-0.0769	0.2208	0.3205	0.1113	0.4001	1.3442	-0.0873	-0.2412	-0.1912	0.3994
142	warm-water lobsters	-0.5349	-0.2165	-0.1836	-0.3243	0.2793	0.9418	0.5287	0.5590	-0.4854	0.9004
143	genus family	-0.0753	0.2201	0.3194	0.1107	0.3999	1.3449	-0.0874	-0.2420	-0.1904	0.4005
144	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table A.5: (continued)

Pair	**Sentence**	**d1**	**d2**	**d3**	**d4**	**d5**	**d6**	**d7**	**d8**	**d9**	**d10**
145	trillionth10	-0.6016	0.0049	0.2404	-0.0607	0.4738	1.3914	0.6031	0.4045	0.0167	0.9426
146	segment penis	-0.6422	0.0710	0.3325	-0.0576	0.3585	1.3770	0.5569	0.3391	-0.0819	0.9944
147	small common gophers	-0.4274	0.2362	0.5828	0.0059	0.5143	1.5084	1.2705	-0.0184	0.2243	0.7624
148	inflammation	-0.2874	0.0679	0.1232	-0.0123	0.2164	0.6660	0.2778	0.2159	0.0039	0.5322
149	process result	-0.5705	0.1357	0.2259	-0.0429	0.4339	1.3403	0.5413	0.4312	0.0187	1.0618
150	kinship relation	-0.5997	0.0055	0.2465	-0.0596	0.4718	1.3901	0.6086	0.4047	0.0149	0.9398
151	mans chest	-0.0773	0.2481	0.3442	0.1069	-0.1515	1.1384	0.0016	0.2398	0.3637	0.0954
152	genus plants	0.0381	0.2053	0.3144	0.1480	0.3843	0.8964	0.9296	-0.1277	0.2640	0.0856
153	grow	-0.0373	-0.1956	-0.4955	-0.1303	-0.2498	-0.5353	-0.0826	0.0746	-0.4330	0.0917
154	solid substance resembling clay	-0.8463	-0.0206	0.4856	-0.2549	0.7508	2.0751	0.9725	0.6094	0.0070	1.3020
155	act changes	-0.6005	0.0033	0.2446	-0.0600	0.4748	1.3879	0.6076	0.4038	0.0206	0.9442
156	weapon mass destruction	-0.8674	0.4141	0.4446	0.2428	0.5114	2.0878	0.8738	0.5624	0.0239	1.2086
157	card picture	0.0536	0.1961	0.3524	0.0950	0.3830	0.8974	0.9573	-0.1258	0.2494	0.0336
158	highest important	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
159	sum charged riding	-0.5868	0.0713	0.2468	-0.0403	0.4531	1.3616	0.5830	0.4175	0.0133	1.0018
160	large genus shrubs	-0.3130	0.0622	0.2644	0.0509	0.6547	1.5441	1.2444	-0.0073	0.3107	0.5181
161	dialect indigenous	-0.3009	0.0000	0.1228	-0.0307	0.2371	0.6934	0.3029	0.2018	0.0111	0.4730

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
162	Old World herb	0.2786	0.0552	0.0269	0.1170	0.1993	0.1627	0.6352	-0.4072	0.2813	-0.4180
163	study nature	-0.6008	0.0072	0.2455	-0.0594	0.4758	1.3885	0.6078	0.4034	0.0168	0.9424
164	business profitable	0.2242	0.2168	0.1971	0.1407	0.1620	0.6505	-0.3914	-0.4435	-0.1992	-0.0714
165	unmasked	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
166	disease skin	0.0701	0.2755	0.3051	0.1352	0.3797	0.8490	0.8886	-0.1192	0.2665	0.1643
167	family_Polypodi	0.2242	0.2168	0.1972	0.1406	0.1620	0.6505	-0.3914	-0.4435	-0.1992	-0.0713
168	$< \vert None \vert >$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
169	$< \vert None \vert >$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
170	branch geology	-0.0753	0.2201	0.3193	0.1108	0.3998	1.3449	-0.0874	-0.2420	-0.1904	0.4006
171	hopeless desperate enterprise	-0.1474	-0.2975	0.2072	-0.1718	0.2490	1.3600	0.5853	0.1742	0.0449	0.8516
172	provide	-0.0637	-0.2229	-0.4774	-0.1107	-0.3014	-0.5944	-0.0618	0.0644	-0.4995	-0.0582
173	mildly narcotic analges	-0.3892	0.1653	0.1579	0.2176	0.0665	0.7258	0.2955	-0.0453	-0.0614	0.2619
174	quality deceives	-0.0258	-0.3849	0.2147	0.0607	0.2851	0.1515	0.0773	0.5870	0.2314	0.9300
175	soft suede leather formerly	-0.0034	0.2234	0.3570	0.1618	0.2690	0.8784	0.9040	-0.1761	0.2211	0.1341
176	pea plant grown primarily	0.0666	0.1984	0.3570	0.0908	0.3969	0.9010	0.9413	-0.1272	0.2691	0.0608

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
177	small genus tropical American	-0.2737	-0.2734	0.4159	-0.3420	0.3473	1.7774	0.2572	0.3674	0.2371	0.5151
178	periodical devoted	0.3399	0.2022	0.1935	0.1782	0.1315	0.1995	0.6258	-0.3304	0.2543	-0.3897
179	$< \vert\text{None}\vert >$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
180	woody	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
181	thermostat	0.3429	0.1999	0.1994	0.1746	0.1363	0.2013	0.6258	-0.3329	0.2551	-0.3816
182	$< \vert\text{None}\vert >$	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
183	agent enhances milk	-0.6237	0.1870	-0.1735	0.0847	0.1038	0.7656	0.5427	0.5569	-0.6251	0.6660
184	method application	-0.5991	0.0065	0.2445	-0.0597	0.4754	1.3890	0.6078	0.4028	0.0180	0.9435
185	caught senses	-0.2994	0.0033	0.1221	-0.0299	0.2378	0.6944	0.3040	0.2015	0.0089	0.4719
186	relating	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
187	genus Olig	-0.2994	0.0033	0.1222	-0.0299	0.2379	0.6944	0.3040	0.2015	0.0088	0.4719
188	stop	-0.2996	0.0033	0.1223	-0.0298	0.2377	0.6946	0.3039	0.2013	0.0091	0.4716
189	small hole	-0.6495	0.0923	0.3193	-0.0618	0.3681	1.3685	0.5406	0.3530	-0.0717	1.0074
190	genus small tree	-0.3111	0.1038	0.3122	0.0382	0.6382	1.5546	1.2159	-0.0270	0.2557	0.5167
191	period	-0.2408	0.2113	0.2345	0.1343	0.1446	0.5827	0.2435	0.3397	-0.0175	0.5370
192	sensation caused heat	-0.5835	0.0681	0.2202	-0.0617	0.4622	1.3688	0.5659	0.4151	0.0290	1.0054
193	cardinal number	-0.1316	0.2391	0.3085	0.1257	0.3902	0.9076	0.4475	0.3476	0.2665	-0.3525
194	dispute point	-0.5991	0.0066	0.2445	-0.0597	0.4756	1.3890	0.6078	0.4028	0.0180	0.9435

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
195	woman schoolteacher	0.4310	0.4564	0.5188	-0.2042	-0.3152	-0.2694	0.3699	0.0497	0.6792	-0.0573
196	complete metric system	0.3068	0.2235	-0.1981	0.0791	0.7254	1.1782	1.1077	-0.1735	-0.0171	-0.1691
197	human offspring	-0.0932	0.4881	0.8699	-0.0822	0.3583	0.8022	-0.2669	-0.3378	0.1261	-0.2141
198	weedy perennial tough	1.0033	0.3878	0.2787	0.2810	0.3021	0.2719	1.0194	-0.6653	0.1519	-0.4595
199	layer cells	-0.0057	0.2461	0.3388	0.1614	0.2821	0.8683	0.8867	-0.1635	0.2295	0.1536
200	low-grade heroin	-0.4550	0.4239	0.4532	0.3381	0.4557	2.0736	0.1763	-0.2646	-0.2067	0.7462
201	act determine advance	-0.2915	-0.3321	0.3368	0.0219	0.5442	0.8365	0.3374	0.8179	0.2482	1.4306
202	shrubby tree	0.2897	0.0556	0.0197	0.1110	0.1767	0.1594	0.6335	-0.4129	0.2968	-0.4244
203	metal plate used	0.0630	0.1989	0.3585	0.0947	0.3884	0.8891	0.9507	-0.1159	0.2466	0.0492
204	branch anthropology	-0.0753	0.2201	0.3193	0.1108	0.3998	1.3449	-0.0874	-0.2420	-0.1904	0.4006
205	< |None| >	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
206	southern hemisphere contains	-0.760	0.267	0.140	0.274	-0.413	0.582	0.164	0.043	0.220	1.077
207	persons committees	0.224	0.217	0.197	0.141	0.162	0.651	-0.391	-0.443	-0.199	-0.071
208	state ketone	-0.589	0.280	0.390	0.124	0.284	1.230	0.488	0.509	-0.041	1.087
209	carpenters plane	0.646	0.299	-0.202	0.202	-0.554	0.392	0.709	-0.665	0.480	-0.625

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
210	medicine prescribed	-0.299	0.003	0.122	-0.030	0.238	0.694	0.304	0.201	0.009	0.472
211	type genus	-0.599	0.007	0.244	-0.060	0.476	1.389	0.608	0.403	0.018	0.944
212	list items	-0.594	-0.007	0.281	-0.115	0.491	1.385	0.635	0.405	0.011	0.904
213	writer vivid	0.214	0.212	0.182	-0.040	0.269	-0.028	-0.409	0.393	-0.451	0.327
214	freshwater tortoise	0.017	0.140	0.475	-0.014	0.431	0.841	1.009	-0.104	0.319	0.166
215	class nouns	-0.068	0.209	0.355	0.058	0.415	1.344	-0.059	-0.239	-0.198	0.359
216	triangular muscle front	-0.646	0.095	0.324	-0.058	0.372	1.367	0.541	0.352	-0.074	1.009
217	car drives	-0.658	0.087	0.837	-0.385	0.387	0.809	0.517	0.335	0.330	0.313
218	pantropical tree	0.290	0.056	0.020	0.111	0.177	0.159	0.633	-0.413	0.297	-0.424
219	psychology	-0.299	0.003	0.122	-0.030	0.238	0.694	0.304	0.201	0.009	0.472
220	giant sequoia	-0.027	0.140	0.484	-0.028	0.402	0.831	1.004	-0.154	0.275	0.150
221	gun	-0.324	0.201	0.090	0.138	0.124	0.817	0.321	0.020	0.034	0.199
222	relating	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
223	movable protective covering	0.682	0.397	0.384	0.362	0.275	0.401	1.247	-0.663	0.510	-0.771
224	remove remove	-0.599	0.006	0.247	-0.060	0.472	1.391	0.609	0.404	0.015	0.941
225	business fails	0.214	-0.001	-0.332	-0.098	-0.035	0.182	-0.476	-0.279	-0.615	0.055
226	shift	-0.300	0.003	0.122	-0.030	0.238	0.695	0.304	0.201	0.009	0.472
227	various plants	0.338	0.202	0.192	0.178	0.146	0.202	0.626	-0.329	0.255	-0.386
228	amount solution	-0.665	0.198	0.280	0.163	0.336	1.468	0.612	0.166	0.020	0.720

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
229	soft oily clay used	-0.292	-0.007	0.167	-0.083	0.253	0.696	0.332	0.203	-0.006	0.425
230	act procrast	-0.301	0.000	0.122	-0.030	0.237	0.693	0.304	0.202	0.011	0.473
231	small genus eastern North	-0.601	0.048	0.293	-0.073	0.462	1.395	0.582	0.386	-0.041	0.941
232	percussion instrument consisting	0.043	0.204	0.317	0.148	0.374	0.893	0.929	-0.130	0.265	0.087
233	smallest administrative district	-0.347	0.044	0.150	-0.021	0.143	0.667	0.266	0.164	-0.027	0.539
234	anterior lobe	0.057	0.252	0.359	0.132	0.379	0.891	0.899	-0.138	0.207	0.085
235	amount	-0.300	0.003	0.123	-0.030	0.236	0.696	0.304	0.202	0.007	0.470
236	amount paid	-0.300	0.003	0.123	-0.030	0.236	0.696	0.304	0.202	0.007	0.470
237	act following prey	0.252	-0.108	0.272	0.187	0.430	1.948	0.163	0.085	-0.270	0.851
238	small bright-colored amph	0.000	-0.275	0.279	-0.109	0.118	0.251	0.096	-0.196	-0.438	0.799
239	loose-fitting dress	0.225	0.342	0.434	0.255	0.466	0.974	0.597	-0.221	0.553	0.455
240	state content	-0.123	0.261	0.347	0.120	0.305	1.318	-0.125	-0.280	-0.226	0.467
241	mind	-0.299	0.003	0.122	-0.030	0.238	0.694	0.304	0.201	0.009	0.472
242	demand sum	-0.601	0.004	0.246	-0.061	0.474	1.388	0.608	0.404	0.017	0.944
243	club-shaped wooden	0.224	0.217	0.197	0.141	0.162	0.651	-0.391	-0.443	-0.199	-0.071
244	work popular culture	-0.075	0.220	0.320	0.111	0.400	1.345	-0.087	-0.242	-0.190	0.400

Table A.5: (continued)

Pair	Sentence	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10
245	written agreement two	-0.602	0.003	0.244	-0.061	0.476	1.388	0.605	0.404	0.019	0.944
246	relating languages derived	-0.300	-0.001	0.123	-0.030	0.237	0.693	0.305	0.202	0.012	0.472

Table A.6: Test results for the Generator block

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
1	0.01784235	0.20378414	0.88653655	0.57786458	0.41639773	0.96731852	0.21702238	0.51472025	0.91098314
2	-0.32999102	0.58903068	0.88822698	1	0.99999997	1.00000004	-0.32999102	0.58903068	0.88822698
3	-0.13697018	0.30317879	0.42427279	0.07602285	0.67128062	0.97885666	-0.04145296	0.59116827	0.455703
4	0.82289723	0.28748361	0.90344475	0.94651225	0.84727857	0.98436822	0.74789079	0.48636921	0.91397373
5	1.00000000	0.79062698	0.96596821	1	0.7898493	0.96275937	1	0.65744371	0.96091961
6	0.69966143	0.43967881	0.96771143	0.95001764	0.67872464	0.98118284	0.88217349	0.50113931	0.97146951
7	0.90502079	0.50192406	0.90150983	0.95865984	0.89068482	0.98546543	0.74656688	0.55318633	0.91522575
8	0.95055611	0.19775944	0.92263266	Nan	0.28749751	0.94062883	Nan	0.6362091	0.94934611
9	0.55483403	0.21887584	0.95085453	0.97809328	0.65724249	0.98227871	0.42869701	0.3711278	0.96171511
10	0.60317897	0.26563946	0.94855338	0.60306828	0.66813919	0.97605592	0.99998626	0.43508198	0.95806901
11	0.63494509	0.25841903	0.93247496	-0.07511655	0.57771514	0.96699848	0.14643418	0.52161563	0.94460697
12	0.72303716	0.17752212	0.93200065	0.99838883	0.3689675	0.96951051	0.7297803	0.56360315	0.94934277
13	0.61192381	0.25428497	0.87471897	0.58901848	0.58164423	0.98338189	0.98922264	0.55779025	0.89811149
14	0.77478103	0.19923535	0.86602825	0.86140397	0.53579433	0.95970177	0.91854079	0.56225714	0.85699854
15	0.95434504	0.284158	0.91297712	0.9559294	0.76431594	0.97262187	0.99677428	0.51256771	0.92170688
16	0.73031016	0.69621616	0.954514	0.1478876	0.52931945	0.92740598	0.09987688	0.7512513	0.94138674
17	-0.00117735	0.31266703	0.93937914	0.96829471	0.74016011	0.98209999	0.0192582	0.46691362	0.94671279
18	1.00000000	0.91431114	0.98306021	0.99999213	0.59037135	0.9196193	0.99999213	0.59406522	0.93721397
19	-0.31784646	0.2681729	0.70353201	0.38033568	0.476658	0.9714725	-0.07914718	0.49421776	0.74010735
20	-0.53285263	0.39406285	0.8286779	0	0.67293015	0.97389954	0	0.55240981	0.84928554
21	0.16796896	0.18086067	0.59529432	0.16857465	0.55442248	0.94612498	0.99999823	0.5412881	0.65142993
22	0.99788675	0.32444417	0.83719564	0.99972141	0.83821676	0.9838032	0.99868952	0.46998363	0.85441947
23	-0.08109437	0.32152503	0.72270616	0.68699065	0.87115205	0.97918483	-0.20702332	0.49164907	0.72941129
24	0.04461160	0.5154797	0.95403214	0.68022978	0.45059491	0.97866763	0.70988243	0.50437259	0.9514153
25	0.39882894	0.37392552	0.93894868	0.56411721	0.72503358	0.9746287	-0.02595569	0.58880757	0.94383078
26	0.74504164	0.28311028	0.91819736	0.95510877	0.56941652	0.9588872	0.6230509	0.59352155	0.94687184

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
27	0.97946124	0.26871269	0.90879557	0.79892871	0.53438127	0.94977821	0.69379038	0.45914503	0.9238497
28	0.50265919	0.40896544	0.92958076	0.96183439	0.7171548	0.9800811	0.27925033	0.52112534	0.94828481
29	0.89229986	0.40815806	0.89785264	0.89628855	0.6550847	0.97330762	0.7095446	0.64377325	0.89426776
30	0.88497717	0.23184454	0.94034073	0.79841572	0.60744684	0.97811436	0.61138802	0.48222072	0.94644291
31	0.97060371	0.21061242	0.90459539	0.25934477	0.40169903	0.97503044	0.01988734	0.49819989	0.91427639
32	0.89808636	0.44299656	0.92485477	0.27455015	0.58891928	0.97712496	0.21468547	0.62877247	0.92854924
33	0.99878060	0.34106304	0.92954968	0.99857244	0.7219073	0.98333495	0.99807251	0.59900169	0.9448642
34	0.52135431	0.57018656	0.96051951	Nan	0.53987934	0.93593587	Nan	0.81494577	0.95157653
35	0.46186723	0.38243522	0.92221694	0.94726588	0.71541783	0.98001412	0.72012992	0.55585797	0.92736141
36	0.66648576	0.40694158	0.90418955	0	0.52807164	0.97436624	0	0.53217212	0.9087539
37	0.99967578	0.44030425	0.91572268	0.99969873	0.79973794	0.98294865	0.99999792	0.54146683	0.91697613
38	0.41468499	0.25844189	0.91254551	0.09782577	0.5841441	0.96047558	-0.06676789	0.48508114	0.93790067
39	0.64414161	0.28113852	0.93272522	0.99449309	0.69575359	0.96489298	0.71745087	0.54227631	0.93945783
40	0.95650852	0.28556587	0.9353204	0.8761487	0.62874966	0.98514576	0.69903374	0.59563384	0.94989677
41	-0.32886751	0.38794948	0.97350255	0.99999906	0.87904925	0.98237295	-0.32913884	0.48608385	0.96500145
42	-0.23157166	0.44357647	0.96509396	0.69066903	0.61360704	0.97416375	-0.19583038	0.48234305	0.98098262
43	0.91203528	0.3054548	0.89320493	0.90652478	0.65799209	0.97201837	0.99562474	0.55683279	0.9075102
44	-0.22651355	0.25165195	0.92208517	0	0.71920969	0.98535543	0	0.42575092	0.93060397
45	0.99999754	0.3744616	0.95908961	0.99999492	0.7059064	0.98059628	0.99998547	0.62563139	0.96014314
46	0.20390600	0.18986569	0.91758167	0.20628841	0.7466759	0.98345514	0.99997586	0.37800302	0.92860789
47	0.70824949	0.36377445	0.93281778	0.70823121	0.73431628	0.98232598	0.99999934	0.53182479	0.9347064
48	-0.10532391	0.20786707	0.92924148	-0.1208723	0.16154539	0.92910725	0.99137705	0.70673207	0.9572938
49	-0.36897096	0.46172608	0.98311702	0.78241914	0.4426146	0.96404521	-0.45698114	0.54454625	0.96594597
50	-0.37217326	0.29146115	0.9022497	0.49097791	0.47474774	0.97610846	-0.19682362	0.62729806	0.92061403
51	-0.18292783	0.42995767	0.97065404	0.89386374	0.55866421	0.95623298	-0.44495998	0.57144314	0.96276309
52	0.50027373	0.47269406	0.94119756	0.99641933	0.50181399	0.96448927	0.49228849	0.73654705	0.94692989

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
53	-0.2102223	0.60598068	0.97143221	0	0.99999995	1.00000015	0	0.60598068	0.97143221
54	0.4945837	0.35032227	0.94870592	0.67256006	0.58767826	0.95055946	0.49548013	0.47411572	0.92791461
55	0.28835926	0.53268064	0.94873758	0.99739928	0.45712914	0.87012514	0.24633677	0.81793195	0.89528495
56	0.45068376	0.7531455	0.91573763	0.99870431	0.52041891	0.97998334	0.47936435	0.65862344	0.90889353
57	-0.21349144	0.6939215	0.88067536	0.02855025	0.81641376	0.90918287	-0.44407623	0.57174388	0.82895728
58	0.46866013	0.29879419	0.90863005	0.78022676	0.30756391	0.80643874	0.51187917	0.85270494	0.8578176
59	-0.38071525	0.55538427	0.93090051	0.73943089	0.87179281	0.99500106	-0.20279548	0.56700035	0.92877459
60	0.13481648	0.59342133	0.96086387	0.63862326	0.55617488	0.95211396	0.0319843	0.73532697	0.96799477
61	-0.43752188	0.50336794	0.93374339	0.98888894	0.67611984	0.96790513	-0.39068688	0.529068	0.93180178
62	-0.21345696	0.54402634	0.92164884	0.55217472	0.53888942	0.96421095	-0.50235019	0.72154216	0.91831121
63	-0.42705052	0.57491663	0.95160821	0.97299875	0.44749215	0.96089338	-0.33091596	0.46097282	0.95544114
64	-0.48273027	0.53781026	0.91957908	0.44860549	0.56264222	0.96769708	-0.21444946	0.70717155	0.93788917
65	-0.25160516	0.59666131	0.91142018	0.9007147	0.32113733	0.93864073	-0.39446723	0.35546076	0.89725359
66	0.4571205	0.64158759	0.9674498	0.4745562	0.68058465	0.92022639	0.13434373	0.71792257	0.90576534
67	-0.44819358	0.30593202	0.91649909	0.65763441	0.77323205	0.97845356	-0.23823761	0.53456613	0.93276657
68	-0.01546545	0.85292095	0.89541919	0.99499226	0.60815967	0.91409491	-0.02400499	0.56773338	0.86283609
69	-0.21341082	0.53282434	0.8835476	0.71663125	0.48486128	0.91386185	-0.45333345	0.88141402	0.94593965
70	0.02236581	0.21976061	0.93706375	0.99177194	0.43141922	0.98333856	0.01820929	0.55062635	0.93554564
71	0.08236054	0.50027542	0.98662252	0.07820276	0.80297637	0.98718707	-0.08836946	0.63736031	0.98574672
72	0.09217773	0.59346276	0.88329918	0.99689944	0.58840499	0.98738055	0.10563764	0.52181231	0.89200628
73	0.26223037	0.53298907	0.8944372	0.29138274	0.39836678	0.95755917	0.99761257	0.57325564	0.89949391
74	0.27363744	0.70808233	0.92601495	0.64266222	0.69041546	0.98282217	0.47430296	0.55949489	0.92845544
75	0.42530688	0.25459607	0.92229252	0.19040014	0.72925421	0.96414782	0.38930673	0.53470376	0.9412161
76	-0.5105276	0.46524892	0.93945533	0.64304976	1.00000002	0.99999991	-0.61498844	0.46524892	0.93945533
77	0.22793314	0.54165303	0.96444709	0.56859302	0.41810895	0.95407851	0.28590221	0.45258219	0.95565294
78	0.66833229	0.44269626	0.91465823	0.93251427	0.52878018	0.95950286	0.46636557	0.41832993	0.89381418

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
79	-0.04821552	0.40231841	0.94981762	0.90358112	0.39963754	0.94059888	-0.19807331	0.72498946	0.96216715
80	0.17038383	0.31522606	0.85182546	0.87025975	0.75104709	0.97063419	0.21573327	0.50328521	0.85281764
81	-0.23068908	0.60466881	0.94337641	0	0.61892886	0.9719645	0	0.73695904	0.93415909
82	-0.42751988	0.71387382	0.78146814	0	0.78950682	0.98634658	0	0.57174993	0.74996756
83	0.45170319	0.29226545	0.92840702	0.03010398	0.36164693	0.94355788	-0.11862892	0.48849225	0.965632
84	-0.45812563	0.31351427	0.92317198	0.99910842	0.71659744	0.9768747	-0.46482956	0.4799145	0.92247269
85	0.48083535	0.32555668	0.92700276	0.11552772	0.34945085	0.94746032	-0.4691315	0.6146701	0.94506763
86	-0.41874434	0.45858914	0.93125868	0.57112918	0.50496958	0.95814134	0.07927709	0.90680312	0.95107594
87	-0.38362198	0.64841718	0.94863758	-0.3032786	0.56012824	0.93347816	0.05539968	0.43807685	0.94423533
88	1	0.49589971	0.94583059	0.05534611	0.65766567	0.94741011	0.05534283	0.56937314	0.95601013
89	0.3856791	0.3518901	0.88284658	0.26015732	0.63291765	0.97679463	-0.03103431	0.54773128	0.89089222
90	-0.55812308	0.46515016	0.97372262	0.20005902	0.78171103	0.98169294	-0.50869436	0.57187376	0.97100638
91	-0.31749207	0.79111376	0.90247633	0.75821674	0.72259299	0.96661786	-0.47327006	0.76803248	0.92053062
92	0.14243693	0.29524267	0.88796727	0.66555368	0.53417349	0.9716873	-0.0975623	0.64472238	0.90852173
93	0.87733249	0.5493516	0.89331911	-0.3903628	0.67478661	0.97141324	-0.30538383	0.70573011	0.90349734
94	0.28242476	0.34980183	0.96896237	0.54415226	0.64274331	0.96902979	-0.07339332	0.50659844	0.96931721
95	-0.67582099	0.22879989	0.71119363	0.98507361	0.64323178	0.98170351	-0.70954481	0.51936982	0.73682104
96	Nan	0.59590481	0.95919121	Nan	0.59102058	0.91563309	0.03034303	0.78555282	0.94586523
97	-0.44994776	0.73274934	0.95225761	0.99989987	0.56776673	0.93084452	-0.45294149	0.58571901	0.95396046
98	-0.04934165	0.5552937	0.92486458	0.84924403	0.63140482	0.97874851	-0.55855669	0.57174556	0.93260799
99	0.4144099	0.37122461	0.89433508	-0.3368428	0.76114827	0.96932714	-0.21792996	0.5641034	0.9020382
100	0.11943162	0.81845797	0.90984229	0.70185398	0.48063728	0.92908767	-0.50691318	0.46916042	0.86209982
101	0.67650015	0.50072819	0.91220767	Nan	0.6754038	0.9815394	Nan	0.50802949	0.91589339
102	-0.52589	0.27042849	0.94689609	0.98208831	0.78574173	0.98570827	-0.45688437	0.41606162	0.95106836
103	-0.32761292	0.34962303	0.9353934	0.95280558	0.52012193	0.96041839	-0.20145448	0.58813803	0.94653356
104	0.56234838	0.52425405	0.90325296	0.3551246	0.55192116	0.90069541	0.47347233	0.55873621	0.89177677

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
105	-0.42608808	0.44019765	0.91793921	-0.36881449	0.59512749	0.97844616	0.40866527	0.51600356	0.92692027
106	-0.6137718	0.26105235	0.88890024	0.85837109	0.49891868	0.96031034	-0.43815801	0.50712362	0.90526494
107	0.28159096	0.33669039	0.82944694	0.88394407	0.7950055	0.978134	-0.1973573	0.54129715	0.86150708
108	-0.21062102	0.58807537	0.93291266	0.55215843	0.50488545	0.97935757	-0.19814804	0.45350277	0.93766398
109	0.11260922	0.56650986	0.95454115	0.74987177	0.50773474	0.95465352	0.35759256	0.55036719	0.95812046
110	-0.2856756	0.22432178	0.96195254	-0.00058058	0.64675194	0.96157965	-0.44168607	0.52111178	0.95508003
111	-0.51070933	0.46697136	0.93774248	0.84487214	0.62235004	0.97838436	-0.66287874	0.50800178	0.94109874
112	-0.44797437	0.61745641	0.86468628	0.72645249	0.61682743	0.91573893	-0.21040575	0.91866458	0.90874732
113	0.33844186	0.55454786	0.92153417	0.90585871	0.66365522	0.95571726	0.33952229	0.52994261	0.93443381
114	-0.27964331	0.72774271	0.96877683	0.4261466	0.70068369	0.97748708	-0.57108081	0.83347751	0.97123239
115	-0.43652929	0.62918709	0.94744073	0.64900795	0.49016128	0.96267344	-0.44927972	0.44917183	0.93648277
116	-0.21052346	0.61856773	0.97233695	0.99729649	0.52044495	0.96955732	-0.23047649	0.52576622	0.97860495
117	-0.55930694	0.62144451	0.97437193	0.99999289	0.58132778	0.96880379	-0.55834433	0.59963945	0.96338712
118	0.44778085	0.51460204	0.95944961	0.99328944	0.65353025	0.96977573	0.50929553	0.48526227	0.95955157
119	-0.51027941	0.25787763	0.98024717	Nan	0.53052089	0.96839845	Nan	0.5590359	0.96808714
120	0.4332564	0.45703667	0.97992719	0.83111622	0.27131385	0.9795295	0.24713146	0.61720584	0.98635527
121	-0.14139402	0.38260679	0.87107555	0.85579709	0.61873179	0.98445563	-0.08477299	0.42387216	0.8692911
122	-0.27589965	0.37734789	0.85506346	Nan	0.65873218	0.97316895	Nan	0.60379987	0.88121529
123	-0.37312553	0.54100404	0.96684825	-0.21348973	0.92529392	0.97904314	0.20947432	0.52052583	0.96324945
124	0.45182581	0.45874651	0.95376354	0.71748425	0.46718104	0.9152111	0.2464552	0.7881671	0.93951515
125	-0.5135873	0.61354248	0.94176957	0.9067644	0.87039918	0.9856259	-0.38934854	0.74719886	0.94455364
126	0.69753446	0.6093322	0.98167148	-0.42051795	0.55840857	0.96822766	-0.5316037	0.56746476	0.97323444
127	0.21675541	0.57219552	0.92621374	0.89768721	0.64045224	0.97839181	0.09576186	0.53408726	0.92845869
128	0.98422603	0.63317444	0.93366447	-0.45751652	0.70882493	0.96439392	-0.50483884	0.64550517	0.94474376
129	-0.37476198	0.41611108	0.94090742	0.78275858	0.86197937	0.98037461	-0.68996173	0.46886543	0.93810531
130	0.31290456	0.59250626	0.97551014	0.52562379	0.67686889	0.98608401	-0.2102454	0.48779962	0.96834138

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
131	-0.34149186	0.46939269	0.94883139	0.08737655	0.90582777	0.98853946	-0.44959938	0.55785024	0.94372332
132	-0.5688264	0.62020386	0.97256839	0.82009893	0.57124184	0.97028235	-0.65157337	0.56663491	0.97940079
133	0	0.59354169	0.94302833	0	0.68773651	0.93855957	-0.4362779	0.62744049	0.91699003
134	0.01537241	0.54144045	0.96838257	0.34808534	0.65122473	0.97623093	-0.25825648	0.60657593	0.97447679
135	-0.15375303	0.57449233	0.97124164	0.76363013	0.96464559	0.98837878	-0.4236879	0.55255864	0.97486489
136	-0.54575621	0.7133874	0.96602622	0.56992181	0.49392537	0.94481147	0.05648846	0.61100307	0.96675728
137	0.47830541	0.30049772	0.96100961	0.77498178	0.77091166	0.98530394	0.47293608	0.46138111	0.96132148
138	0.52918609	0.50785618	0.94969897	Nan	0.86255563	0.9908267	Nan	0.55781956	0.9465597
139	0.45836511	0.26965054	0.802359	0.41782334	0.72654818	0.9849446	-0.2138329	0.50563241	0.81281518
140	0.61456779	0.28922193	0.97534385	0.62780519	0.82737398	0.99084168	0.56259106	0.46742419	0.97677559
141	0.77160235	0.55917381	0.9346181	-0.25944688	0.58881012	0.96934591	-0.33924911	0.52607042	0.92660362
142	0.9840052	0.31071763	0.976605	0.65358966	0.74035667	0.97949546	0.64156622	0.48529786	0.97912042
143	0.47815613	0.37431821	0.97413296	0.35514612	0.50093198	0.97890402	0.65894627	0.43775972	0.96669103
144	0.53633653	0.65840388	0.95987274	0.54019893	0.44791552	0.96237166	0.99996279	0.57053286	0.96409848
145	0.96844601	0.56945794	0.98084244	0.80842108	0.63956883	0.96996018	0.72481307	0.54680807	0.97187981
146	0.99960541	0.47057961	0.9772527	Nan	0.70487519	0.97845485	Nan	0.47312796	0.97886518
147	0.91184	0.58525849	0.96160542	0.76421054	0.85361559	0.98480935	0.89642445	0.60130035	0.96046409
148	0.88002205	0.59973189	0.98219196	0.88327939	0.51811971	0.97860599	0.92259116	0.58352184	0.97916963
149	0.99480151	0.3101241	0.93875853	0.43886633	0.77842414	0.98734973	0.47317888	0.5238066	0.94325263
150	0.90469874	0.29303608	0.94221027	0.91904611	0.71103409	0.98838342	0.99534392	0.57584137	0.94893065
151	0.9999993	0.55713641	0.97545588	0.46778137	0.87735474	0.98304742	0.46815456	0.57882092	0.98355222
152	0.39400313	0.71852672	0.98432136	0.89247649	0.7577516	0.98612295	0.65537985	0.6826649	0.9797132
153	0.53631247	0.287656	0.94467359	0.45556024	0.27777643	0.88957092	0.12497933	0.80994283	0.95464188
154	0.99995055	0.6003853	0.92706323	0.99053705	0.71172248	0.93314699	0.99046211	0.55317827	0.96238666
155	0.99945005	0.54204963	0.97746567	0.99940543	0.77249119	0.98725144	0.99999195	0.51270774	0.97476374
156	0.72574944	0.28298721	0.92533559	0.77907781	0.58203172	0.97711445	0.91619024	0.4307258	0.93470547

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
157	0.72856927	0.4827999	0.96197308	0.70010148	0.81008333	0.98693629	0.99588782	0.58631751	0.9628061
158	0.99999122	0.55165815	0.93466738	Nan	0.47422872	0.95727274	Nan	0.27461904	0.95864706
159	0.82339637	0.41640681	0.87546973	0.4603823	0.85887205	0.96880759	0.02701533	0.57672488	0.86929227
160	0.3105544	0.32863348	0.95395303	0.82258274	0.61993796	0.98079163	0.74571264	0.36796565	0.96329045
161	0.99999995	0.58405631	0.9565896	0.99933703	0.72074097	0.98372296	0.99934364	0.54590332	0.96699476
162	0.10869917	0.36247237	0.79609546	0.51155196	0.58033874	0.82652211	-0.06124559	0.48455455	0.97244891
163	0.94820159	0.62496581	0.96297035	0.95516024	0.61358726	0.98189566	0.99663155	0.48846543	0.96290286
164	0.46865056	0.39968282	0.86418429	0.78730482	0.66167473	0.97438412	0.67639249	0.48863229	0.84464006
165	0.99999496	0.52623034	0.91064225	0.9143606	0.47446029	0.95838627	0.91475058	0.57994301	0.93494846
166	0.75008738	0.51843957	0.96997215	Nan	0.64049268	0.97263428	Nan	0.57034721	0.98293153
167	0.6292554	0.6796234	0.97949039	0.93191502	0.5915766	0.97172539	0.38079339	0.60827613	0.97758302
168	-0.17337083	0.56266556	0.91572795	-0.27959672	0.48052835	0.94397765	0.87030754	0.65414544	0.93404296
169	-0.00813595	0.56317234	0.91252086	0.78099956	0.41361783	0.95559354	0.47785327	0.62739616	0.92458614
170	0.91003691	0.43988239	0.94208196	Nan	0.69082957	0.97858934	Nan	0.61705553	0.95933743
171	0.94694957	0.57330477	0.97742218	0.73456239	1.00000001	1.00000002	0.71724147	0.57330477	0.97742218
172	0.98303907	0.57233727	0.92709426	0.9902119	0.54093916	0.91187572	0.9535375	0.89248961	0.94805551
173	0.98709486	0.65935094	0.96854904	-0.01871307	0.83976842	0.96614586	0.04719809	0.65463936	0.97799925
174	0.99999998	0.74831188	0.96059665	0.90221615	0.78131503	0.97804149	0.90222532	0.72437395	0.94910515
175	0.71103909	0.51953477	0.98118705	0	0.66825107	0.97808885	0	0.51797874	0.97322625
176	0.68567981	0.47689011	0.92943271	0.66166393	0.71594244	0.9743194	-0.0400871	0.53595332	0.94429634
177	0.61926741	0.29967229	0.96844668	-0.62014383	0.5869478	0.98253827	-0.32494505	0.37291346	0.97283865
178	0.89666087	0.49331257	0.81048851	0.81057797	0.76999742	0.97819476	0.46755555	0.65766195	0.82990941
179	0.9999997	0.55955963	0.9030774	0.02916492	0.50322159	0.96420056	0.02908854	0.49416095	0.92225307
180	1	0.70055552	0.97162027	0.99999632	0.43772938	0.98437218	0.99999623	0.35746215	0.96975543
181	0.98580166	0.61118426	0.97166639	0.98335057	0.60481086	0.97520494	0.99502983	0.60510798	0.97130387
182	0.93576053	0.54241793	0.93448589	0.99571484	0.44650329	0.95571432	0.94427269	0.62130431	0.95491385

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
183	0.8176506	0.58750613	0.97824269	0	0.90493967	0.99001065	0	0.65647021	0.97961794
184	0.99902494	0.6414954	0.8774571	0.71291099	0.70448464	0.98204061	0.70709737	0.64798309	0.87081317
185	0.20010521	0.31360972	0.89449262	0.02986638	0.9185191	0.9834177	0.21509649	0.38065974	0.90697471
186	0.51271096	0.56533868	0.97862108	0.29080315	0.75734712	0.96742261	-0.08959474	0.39168207	0.95904398
187	1	0.61903451	0.98014218	0.78796884	0.50184693	0.98228364	0.78796892	0.50502314	0.96714531
188	1	0.69615867	0.9286806	0.20018804	0.69615867	0.9286806	0.20017655	0.99999997	0.99999999
189	0.92433269	0.53946522	0.97250132	-0.38966307	0.46529093	0.96281835	-0.58736011	0.31605407	0.95203541
190	0.92982723	0.76539608	0.98884284	0.6060278	0.60674092	0.97448862	0.56428078	0.56373756	0.97454171
191	1	0.46276298	0.86675239	0.27100414	0.7761269	0.97774487	0.27103104	0.57144039	0.88138653
192	0.99938859	0.49628743	0.96074958	0.31746058	0.96509035	0.98873234	0.30550886	0.53713854	0.9511354
193	0.9443607	0.50246092	0.95893341	0.93823054	0.85615405	0.9759535	0.90069558	0.40796817	0.96619638
194	0.99999845	0.42103778	0.96406807	0.7353198	0.75041473	0.97769076	0.73489477	0.49581256	0.96545091
195	0.69239401	0.56006705	0.95362238	Nan	0.75813422	0.96659513	Nan	0.62612679	0.95350905
196	0.98823388	0.49933242	0.95601401	0.95596852	0.66234733	0.96271401	0.90770331	0.515366	0.96843961
197	-0.43361226	0.5089612	0.91051647	0.30117113	0.74884512	0.97977535	-0.36890798	0.66251726	0.92460679
198	0.75939414	0.51594009	0.94038248	0.82062847	0.78754449	0.97991959	0.96404665	0.57510252	0.94259795
199	0.73885284	0.51210725	0.95582106	0.98687374	0.80098213	0.98303856	0.70208711	0.55829816	0.95971273
200	0.46657778	0.30973049	0.92964992	0.43771392	0.44945444	0.97167532	0.98969568	0.43942667	0.93844697
201	0.94863906	0.48813451	0.8703498	0.88095537	0.66352389	0.96613681	0.93226069	0.51541467	0.8992072
202	0.89993617	0.56877304	0.98081407	0.38734449	0.72329154	0.98152957	0.06556597	0.65978679	0.98396502
203	0.51617689	0.3512212	0.97170174	0.52959028	0.7537453	0.98115744	0.97202467	0.57949255	0.96929935
204	0.93997187	0.51518847	0.94964828	0.93301298	0.76333248	0.9897352	0.90044238	0.63136328	0.95212191
205	0.09336281	0.21034456	0.93332783	0.07406864	0.32814461	0.95046698	0.98689711	0.5347101	0.94712844
206	0.52609377	0.51491334	0.91759529	-0.25762803	0.71423107	0.97267148	0.20675808	0.6793027	0.91265931
207	0.71445408	0.42705787	0.88556589	0	0.77458371	0.98028894	0	0.52754189	0.89181232
208	0.99843546	0.4957013	0.98407578	0.01174834	0.41998271	0.98124747	-0.00340014	0.69740073	0.98600274

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
209	0.71049662	0.52134881	0.9628523	0.72796456	0.82277013	0.99179262	0.99319507	0.52303527	0.96530455
210	0.97709196	0.25083069	0.91610373	0.4755452	0.62155083	0.95608362	0.55393976	0.47483246	0.90741599
211	0.3365037	0.34979344	0.91337613	0.86523911	0.48592832	0.98084848	0.73081031	0.45220165	0.92163659
212	0.89679278	0.43155036	0.92858882	0.27444738	0.73177016	0.97206805	-0.08432095	0.59759048	0.93801667
213	0.35049736	0.64054307	0.96465129	0.82081119	0.88262259	0.9879036	-0.09525854	0.6954589	0.9674749
214	0.41931413	0.71541389	0.98664469	0.90233503	0.66449281	0.97791929	0.62320125	0.64802419	0.98104044
215	0.84969222	0.47988976	0.97788512	0.98593803	0.72141981	0.98229794	0.9136051	0.64086614	0.98288546
216	0.98576144	0.55632432	0.98348835	0.68749179	0.67358105	0.98231009	0.70953372	0.5586525	0.97928617
217	0.33345477	0.56737514	0.96724192	0.63894643	0.60929692	0.97148885	0.04560815	0.51500808	0.96317927
218	0.78654396	0.37801471	0.95226586	0.47716553	0.60564706	0.96839484	-0.02943676	0.45991969	0.96333761
219	0.965072	0.4337583	0.90003752	0.40363874	0.41401683	0.97601086	0.1998709	0.60984362	0.91701164
220	0.336694	0.47288009	0.95337302	0.90697186	0.45187724	0.97714041	-0.08647588	0.55574199	0.96914014
221	0.14703126	0.45632781	0.95081074	Nan	0.52603318	0.97760326	Nan	0.53863447	0.94638621
222	0.99781238	0.35158662	0.92862097	0.70036258	0.71418261	0.97201253	0.68304772	0.50277445	0.93993953
223	0.83548508	0.45148407	0.96914926	0	0.70207814	0.97323539	0	0.60040666	0.97043439
224	0.46164525	0.43199435	0.90775175	0.89714026	0.55756216	0.97793843	0.02252436	0.58492605	0.9019414
225	0.99999311	0.38351586	0.92197911	0.11173779	0.66108764	0.97548636	0.11429646	0.56309642	0.91767203
226	0.01557775	0.50419282	0.90737617	0.70667363	0.52327999	0.92391465	0.71810824	0.88032529	0.97044082
227	0.45960213	0.52506601	0.97967649	0.75032646	0.62678344	0.97619778	0.71818032	0.53392934	0.97948615
228	0.96926917	0.55379153	0.97454595	0.9692624	0.73490195	0.98198291	0.99999989	0.64979258	0.97344753
229	0.98155658	0.34245182	0.96774569	0.98232085	0.8694	0.97612854	0.92944504	0.4590182	0.9836936
230	0.99967915	0.58455756	0.97107878	0.9980066	0.64885035	0.98428691	0.99714355	0.67056525	0.9761972
231	0.46467048	0.55405063	0.97527372	0	0.57305606	0.97856457	0	0.62013946	0.98008072
232	0.77921972	0.3705408	0.92783639	0.78069935	0.56395203	0.96528615	0.99998565	0.48422805	0.93832666
233	0.71883287	0.44887921	0.92420536	-0.07157476	0.77189005	0.98195158	-0.0351994	0.52744768	0.93306308
234	0.83243886	0.72943451	0.97372237	0.57131525	0.81991863	0.98572417	0.02216792	0.67518129	0.98002017

Table A.6: (continued)

Pair	Group 1			Group 2			Group 3		
	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net	GPT-2	BERT	XL-Net
235	0.84243609	0.65518813	0.18493978	0.07146501	0.73412289	0.97434358	-0.0176817	0.45648729	0.18074844
236	0.99999599	0.55020891	0.91832003	0.03590529	0.81132866	0.98386315	0.03726692	0.63935835	0.92856709
237	0.11614717	0.35029289	0.91248332	Nan	0.86345299	0.98937405	Nan	0.49657663	0.90688208
238	0.6199518	0.38695886	0.96337331	0.77550856	0.42635699	0.96457519	0.42789586	0.52110554	0.97359577
239	0.53214675	0.60778313	0.97192763	0.53809828	0.86339586	0.97592251	0.57346866	0.7055667	0.97353384
240	0.99873446	0.6835272	0.96303616	0.13356143	0.77900999	0.98441352	0.14581532	0.66628477	0.96982031
241	0.91523891	0.27513289	0.91043311	-0.2368302	0.5156043	0.96829831	0.02851678	0.41637301	0.94087645
242	0.99999803	0.65223393	0.95802402	Nan	0.75766292	0.98298156	Nan	0.63417163	0.95368172
243	0.12899196	0.46712779	0.9400365	-0.15028882	0.69808997	0.973262	0.87929021	0.47945346	0.94568247
244	0.96801296	0.31587734	0.97789385	0.69885213	0.58290276	0.98267129	0.73203447	0.61533421	0.97441801
245	0.95047044	0.47162816	0.87553751	0.46179143	0.84412382	0.98874505	0.47218261	0.56367245	0.8897655
246	0.99999992	0.51691669	0.85208705	0.99914703	0.73596539	0.98025417	0.99914767	0.67089206	0.85748308
Mean	0.41738092	0.47036051	0.92556685	0.5729076	0.64573869	0.96816335	0.26965478	0.56846558	0.93061628
Std. Dev.	0.5141558	0.1507291	0.0760888	0.40125779	0.15343756	0.02433545	0.51311105	0.10957588	0.07339418

Table A.7: General results of the similarity questionnaire

Pair	USE	BERT	XL-Net	Hyperbolic	Survey
1	0.20507446	0.76554413	0.98189549	0.78741279	0.06173913
2	0.5315156	0.77560055	0.98234811	0.94393238	0.32434783
3	0.8228531	0.87391693	0.97660076	0.99520004	0.60434783
4	0.6453475	0.80804356	0.98543142	0.98887259	0.70608696
5	0.3005308	0.65888754	0.97592409	0.97936867	0.08521739
6	0.4545987	0.84995316	0.96651561	0.9299972	0.64434783
7	0.26251554	0.77587195	0.98311639	0.64476768	0.50434783
8	0.33757654	0.76771621	0.98076357	0.92574109	0.05043478
9	0.23559931	0.7476288	0.98919629	0.68960966	0.02173913
10	0.47510535	0.85678646	0.99360321	0.9348734	0.61043478
11	0.33316642	0.69953999	0.96438015	0.98748969	0.04782609
12	0.527152	0.72585064	0.98559536	0.8986168	0.1626087
13	0.63287306	0.82051377	0.97599764	0.93027885	0.63565217
14	0.68878347	0.82948216	0.98178679	0.99885157	0.60347826
15	0.1767056	0.68206765	0.96741453	0.93509589	0.09304348
16	0.21797678	0.50390303	0.91714856	0.91782074	0.21826087
17	0.14432636	0.65793445	0.97431734	0.90770017	0.02782609
18	0.4600697	0.65356742	0.9865122	0.96783913	0.62086957
19	0.55096316	0.8289564	0.99173366	0.47019851	0.66434783
20	0.60081244	0.75564151	0.98330886	0.75963936	0.54608696
21	0.5437585	0.85719678	0.98163596	0.77867734	0.48173913
22	0.53497845	0.72925698	0.98438245	0.76180527	0.4026087
23	0.84277743	0.91269652	0.99696492	0.95008657	0.59043478
24	0.40674514	0.72329364	0.95537683	0.93070998	0.4573913
25	0.5935366	0.81090791	0.98502465	0.96949356	0.44869565
26	0.31347254	0.67132856	0.96303137	0.48556016	0.06173913
27	0.21214408	0.63043208	0.98714079	0.76462022	0.26608696
28	0.4434235	0.58615393	0.95062965	0.46396705	0.06086957
29	0.7172539	0.82209759	0.97047654	0.99712947	0.60695652
30	0.13884786	0.66417728	0.98701348	0.91170628	0.07391304
31	0.5920586	0.87660464	0.98366615	0.96207688	0.06782609
32	0.18668789	0.58772375	0.97513121	0.84862461	0.06347826
33	0.6385894	0.82274927	0.98085411	0.52499448	0.53130435
34	0.15580799	0.66573033	0.98825792	0.77063448	0.25217391
35	0.7156794	0.72904997	0.98013408	0.95142463	0.53913043
36	0.42080283	0.83098976	0.98677145	0.83767462	0.25913043
37	0.5597972	0.83716404	0.98695576	0.51291525	0.58521739
38	0.2837236	0.71375839	0.96434075	0.61734006	0.46434783
39	0.5117048	0.731655	0.98233063	0.71721107	0.56
40	0.6921946	0.95749762	0.99271859	0.86937794	0.49043478
41	0.24754027	0.68807106	0.97105789	0.43589485	0.07652174

Table A.7: (continued)

Pair	USE	BERT	XL-Net	Hyperbolic	Survey
42	0.4077363	0.74351953	0.98363284	0.70396215	0.51130435
43	0.67545885	0.90304867	0.98961625	0.86063349	0.40434783
44	0.35875997	0.76250101	0.98430869	0.5286675	0.26173913
45	0.1738242	0.69815288	0.95432615	0.41417352	0.12695652
46	0.8736828	0.88892288	0.99076963	0.9680284	0.71652174
47	0.44114164	0.81040076	0.97977488	0.57023176	0.54956522
48	0.72833323	0.89765857	0.98885432	0.84173476	0.36
49	0.35308102	0.80255965	0.98655367	0.72915446	0.22086957
50	0.5412669	0.80143991	0.97988111	0.9938891	0.53652174
51	0.70135975	0.88339037	0.99427245	0.78676131	0.54608696
52	0.5220903	0.77703068	0.98569323	0.99266122	0.31826087
53	0.29149276	0.66316097	0.89277733	0.90568043	0.05043478
54	0.772733	0.91017059	0.99355803	0.94817193	0.63130435
55	0.3921284	0.70826174	0.98738605	0.75078724	0.59391304
56	0.31193784	0.7407005	0.97878645	0.46067163	0.54695652
57	0.26591083	0.76554785	0.98310548	0.95192613	0.0826087
58	0.28608897	0.76742381	0.99210256	0.95335995	0.52956522
59	0.2706507	0.70539507	0.98455171	0.86347532	0.07130435
60	0.21402316	0.62571769	0.97062895	0.98488749	0.02347826
61	0.711674	0.86703893	0.98809983	0.90418441	0.56956522
62	0.29292613	0.76361773	0.98357856	0.87115028	0.04956522
63	0.33511588	0.57628326	0.96074195	0.85884163	0.60869565
64	0.82606673	0.87501674	0.98819967	0.99545871	0.62
65	0.4234625	0.69763287	0.97421053	0.10929467	0.59391304
66	0.39387867	0.8426355	0.98810697	0.87476937	0.40869565
67	0.7268636	0.85358935	0.99078203	0.91276942	0.7
68	0.5584279	0.80083717	0.99142693	0.89204857	0.64
69	0.37381035	0.58325256	0.91974673	0.87557437	0.43391304
70	0.39456356	0.81276032	0.99341074	0.9828704	0.14608696
71	0.28812262	0.74657023	0.98132381	0.94923249	0.27217391
72	0.2669924	0.7642128	0.97419503	0.90363457	0.32086957
73	0.08884905	0.61651452	0.97500622	0.93159781	0.39304348
74	0.25579572	0.67583205	0.98766463	0.74449311	0.06
75	0.3261479	0.71397764	0.96467539	0.31286835	0.48173913
76	0.37574124	0.79976089	0.95825413	0.99967624	0.03913043
77	0.39632735	0.80988336	0.98499202	0.77267913	0.24608696
78	0.73684096	0.8825914	0.98170465	0.8141251	0.34608696
79	0.11623742	0.72762875	0.98534557	0.79571148	0.06608696
80	0.41878688	0.83327837	0.99055522	0.93742463	0.66956522
81	0.38313597	0.74411244	0.98395661	0.83226529	0.27304348
82	0.9042159	0.92265345	0.99426041	0.92979861	0.6773913

Table A.7: (continued)

Pair	USE	BERT	XL-Net	Hyperbolic	Survey
83	0.6489723	0.84431475	0.98919688	0.93025122	0.56434783
84	0.12230691	0.61910127	0.97182425	0.68168189	0.26173913
85	0.4759945	0.77460287	0.98768248	0.97398519	0.20347826
86	0.37982762	0.72868093	0.97534291	0.1805727	0.48086957
87	0.6232217	0.81542367	0.98204947	0.6829504	0.07565217
88	0.3128978	0.72764807	0.97047628	0.91513539	0.33565217
89	0.4405168	0.75580777	0.96919552	0.8617979	0.58173913
90	0.4070181	0.77577745	0.97976106	0.94345542	0.06608696
91	0.35985032	0.65355165	0.98235857	0.88935539	0.03217391
92	0.6571269	0.83381826	0.99060732	0.94646619	0.46956522
93	0.44263938	0.64164946	0.97626481	-0.05439499	0.05304348
94	0.21657833	0.64732385	0.96284876	0.97259526	0.47217391
95	0.3243384	0.81012653	0.9910684	0.63226314	0.04869565
96	0.708892	0.87640657	0.98800734	0.94031094	0.5426087
97	0.7367654	0.78521997	0.9823588	0.97935933	0.65130435
98	0.8229939	0.84502306	0.9899968	0.96057154	0.68695652
99	0.46358114	0.79450877	0.97934185	0.88930981	0.37913043
100	0.7523612	0.81903834	0.98113427	0.85694062	0.59217391
Mean	0.45426637	0.76398648	0.97887818	0.81247165	0.36767826
Std. Dev.	0.20294657	0.09059829	0.01570697	0.20624668	0.22828823

Table A.8: Generated phrases for the finetuned USE test

Pair	Generated sentence
1	three_quarters
2	the fifth sign of the zodiac; the sun
3	composer_of_songs
4	describable
5	praiseworthy
6	immediacy
7	vegetative_state
8	adhesive_material
9	tantalizingly
10	look_alike
11	Rhinestia
12	Fourth_of_July
13	agree
14	Howe
15	genus_Lampridae
16	attractive
17	fingerprint
18	someme_deafness
19	chicken_stock
20	Solanum_coccinea
21	genus_Leptopteris
22	Dikangale_Tanzania
23	heterozygosity
24	oxbow_river
25	Sphaeral
26	throttle
27	genus_Plectrophenax
28	nowhere
29	Pisa
30	counter
31	genus_Epidendrum
32	blastoma
33	blight
34	pin
35	polydipsia
36	adopt
37	wryly
38	Alsatia
39	oncologist
40	genus_Rhubarb
41	Sarcophaga

Table A.8: (continued)

Pair	**Generated sentence**
42	chamfer_bit
43	unbend
44	Bloemism
45	nonc
46	saltwater_trough
47	pocket_monkey
48	probate
49	leprechaun
50	Old_Fashion
51	Eriophorum
52	lever
53	seal
54	fungi having a single cytoplasm and a
55	orange_juice
56	value
57	washday
58	yellowwood
59	Kennelly
60	unproved
61	antiseptic
62	visit
63	Ustilaginaceae
64	artiodactyl
65	aspirator
66	other_than
67	easterner
68	genus_Protea
69	Tobit
70	mooring
71	coconut
72	genus_Trachurus
73	side
74	recitation
75	herpes_simplex
76	Basterill
77	genus_Buckleya
78	cross-correspond
79	elation
80	effacement
81	alpha_and_beta_energy
82	Calliphora

Table A.8: (continued)

Pair	Generated sentence
83	genus_Cornus
84	arousal
85	hematocoele
86	fudge_lily
87	lancet
88	floatation
89	genus_Cedrela
90	soft
91	genus_Browallia
92	filibusterer
93	genus_Caulobacterium
94	genus_Echinocerca
95	genus_Fulgorina
96	Carduus
97	genus_Rhodymenia
98	toss
99	clumsily
100	unimpeded

Table A.9: Comparison on the similarity scores validation

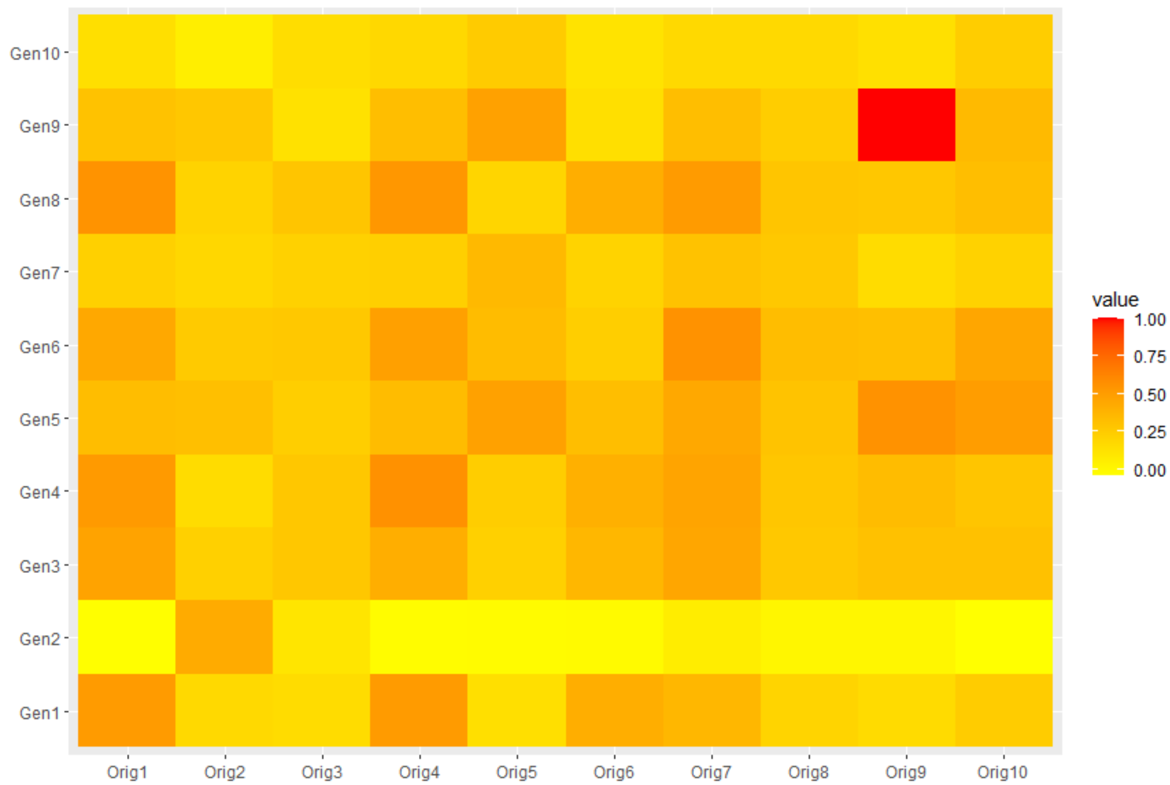
Pair	Before finetuning	After finentuning	Score difference
1	52.6213%	96.2029%	43.5816%
2	44.2574%	79.8925%	35.6350%
3	29.3864%	98.5118%	69.1254%
4	57.4997%	98.3089%	40.8092%
5	49.3037%	64.4720%	15.1683%
6	25.6244%	97.1760%	71.5516%
7	31.9883%	4.0767%	-27.9116%
8	30.7653%	6.4215%	-24.3438%
9	100.0000%	100.0000%	0.0000%
10	26.1455%	3.1999%	-22.9456%
11	35.6721%	87.8573%	52.1853%
12	51.1945%	95.1120%	43.9175%
13	22.3702%	98.2007%	75.8305%
14	32.5116%	96.4660%	63.9544%
15	37.8987%	98.3684%	60.4697%
16	48.0306%	12.3078%	-35.7228%
17	100.0000%	100.0000%	0.0000%
18	52.4613%	18.0238%	-34.4375%
19	51.9828%	98.7807%	46.7979%
20	46.2968%	99.5915%	53.2948%
21	28.4994%	12.2105%	-16.2889%
22	54.8069%	96.1462%	41.3393%
23	100.0000%	100.0000%	0.0000%
24	9.1592%	6.6018%	-2.5574%
25	38.0960%	96.3487%	58.2527%
26	24.3187%	78.8076%	54.4889%
27	34.7631%	98.0085%	63.2454%
28	25.6224%	92.9493%	67.3269%
29	41.1683%	96.2135%	55.0452%
30	54.5918%	34.1714%	-20.4203%
31	44.1268%	99.1796%	55.0528%
32	53.1106%	99.6379%	46.5273%
33	42.3409%	53.8239%	11.4831%
34	100.0000%	100.0000%	0.0000%
35	100.0000%	100.0000%	0.0000%
36	34.2002%	77.4743%	43.2741%
37	100.0000%	100.0000%	0.0000%
38	100.0000%	100.0000%	0.0000%
39	100.0000%	100.0000%	0.0000%
40	60.0604%	99.4705%	39.4100%
41	31.4635%	34.2881%	2.8246%

Table A.9: (continued)

Pair	Before finetuning	After finentuning	Score difference
42	100.0000%	100.0000%	0.0000%
43	56.8690%	96.7408%	39.8718%
44	50.3049%	98.0826%	47.7777%
45	26.3063%	30.6042%	4.2979%
46	38.5336%	99.5263%	60.9927%
47	25.7627%	1.4646%	-24.2981%
48	17.0602%	0.6153%	-16.4449%
49	100.0000%	100.0000%	0.0000%
50	31.6399%	97.7015%	66.0616%
51	50.2886%	99.7344%	49.4459%
52	28.5974%	20.2665%	-8.3309%
53	36.3230%	94.7790%	58.4560%
54	0.0000%	0.0000%	0.0000%
55	100.0000%	100.0000%	0.0000%
56	34.5117%	80.3177%	45.8061%
57	53.4790%	99.4903%	46.0113%
58	100.0000%	100.0000%	0.0000%
59	31.1989%	59.2519%	28.0530%
60	39.6260%	98.0712%	58.4452%
61	100.0000%	100.0000%	0.0000%
62	43.4312%	97.0394%	53.6082%
63	100.0000%	100.0000%	0.0000%
64	100.0000%	100.0000%	0.0000%
65	100.0000%	100.0000%	0.0000%
66	100.0000%	100.0000%	0.0000%
67	100.0000%	100.0000%	0.0000%
68	52.1839%	96.9923%	44.8085%
69	100.0000%	100.0000%	0.0000%
70	100.0000%	100.0000%	0.0000%
71	100.0000%	100.0000%	0.0000%
72	52.8472%	93.2128%	40.3656%
73	57.1018%	77.2139%	20.1121%
74	51.9206%	79.2852%	27.3646%
75	44.3350%	95.3965%	51.0614%
76	40.7907%	3.5534%	-37.2373%
77	40.3949%	5.6103%	-34.7847%
78	27.4970%	17.9474%	-9.5496%
79	25.7258%	5.5681%	-20.1577%
80	25.9188%	14.8284%	-11.0904%
81	57.3523%	94.7328%	37.3805%
82	43.6939%	75.2778%	31.5840%

Table A.9: (continued)

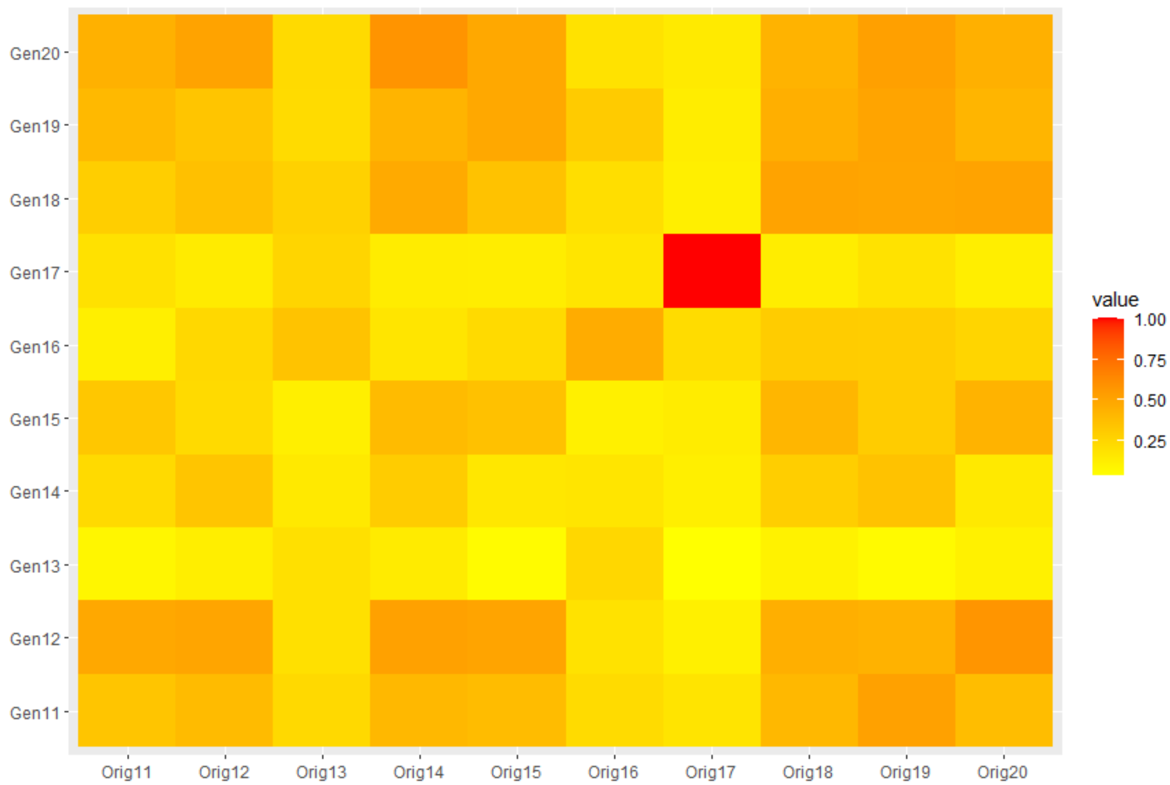
Pair	**Before finetuning**	**After finentuning**	**Score difference**
83	22.3751%	96.4351%	74.0599%
84	20.1291%	9.8386%	-10.2904%
85	15.1694%	0.9242%	-14.2453%
86	26.2015%	39.2001%	12.9986%
87	27.3895%	55.7233%	28.3338%
88	78.6052%	93.6777%	15.0725%
89	100.0000%	100.0000%	0.0000%
90	21.8961%	87.0487%	65.1526%
91	53.3363%	99.6424%	46.3061%
92	18.2224%	10.2204%	-8.0020%
93	48.1800%	98.0156%	49.8356%
94	45.6813%	71.0850%	25.4037%
95	57.5373%	99.2191%	41.6819%
96	29.8778%	94.4161%	64.5383%
97	57.9640%	95.1584%	37.1944%
98	100.0000%	100.0000%	0.0000%
99	100.0000%	100.0000%	0.0000%
100	45.8019%	93.8586%	48.0567%
Mean score improvement		21.5167%	



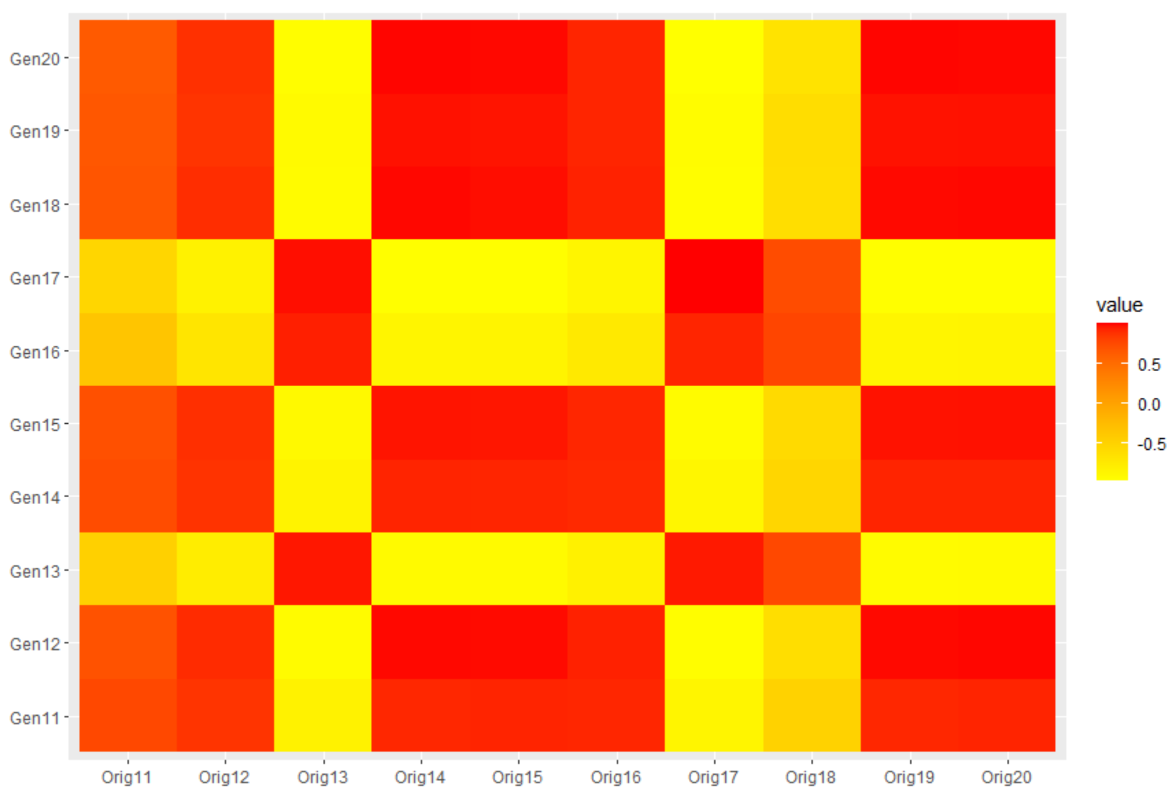
(a) Before finetuning

(b) After finetuning

Figure A.1: Finetuned USE heatmap of test pairs 1-10

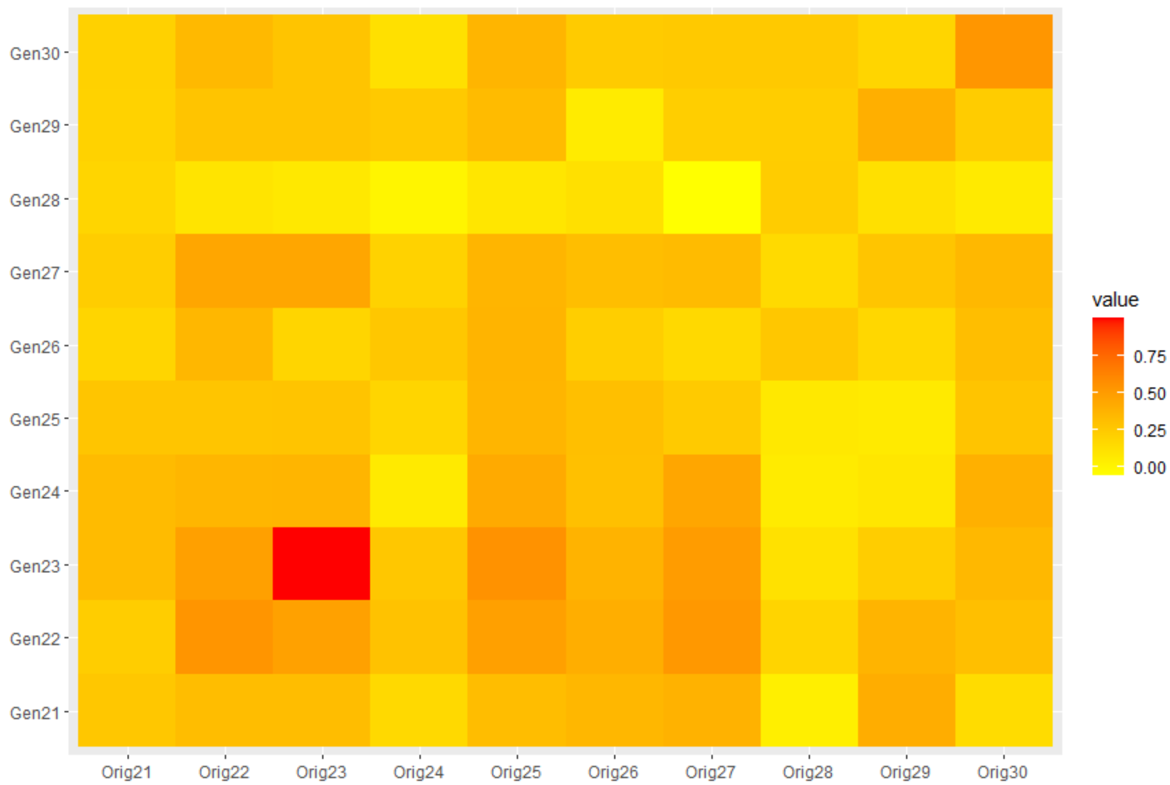


(a) Before finetuning

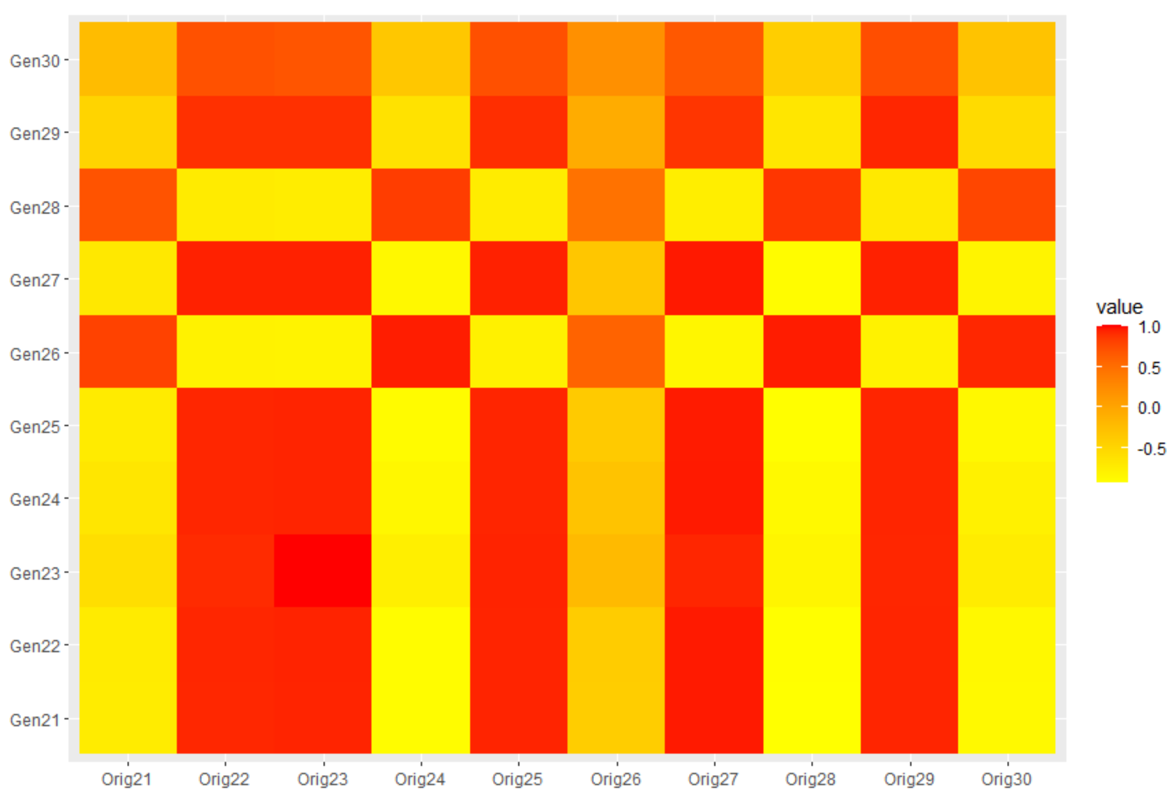


(b) After finetuning

Figure A.2: Finetuned USE heatmap of test pairs 11-20

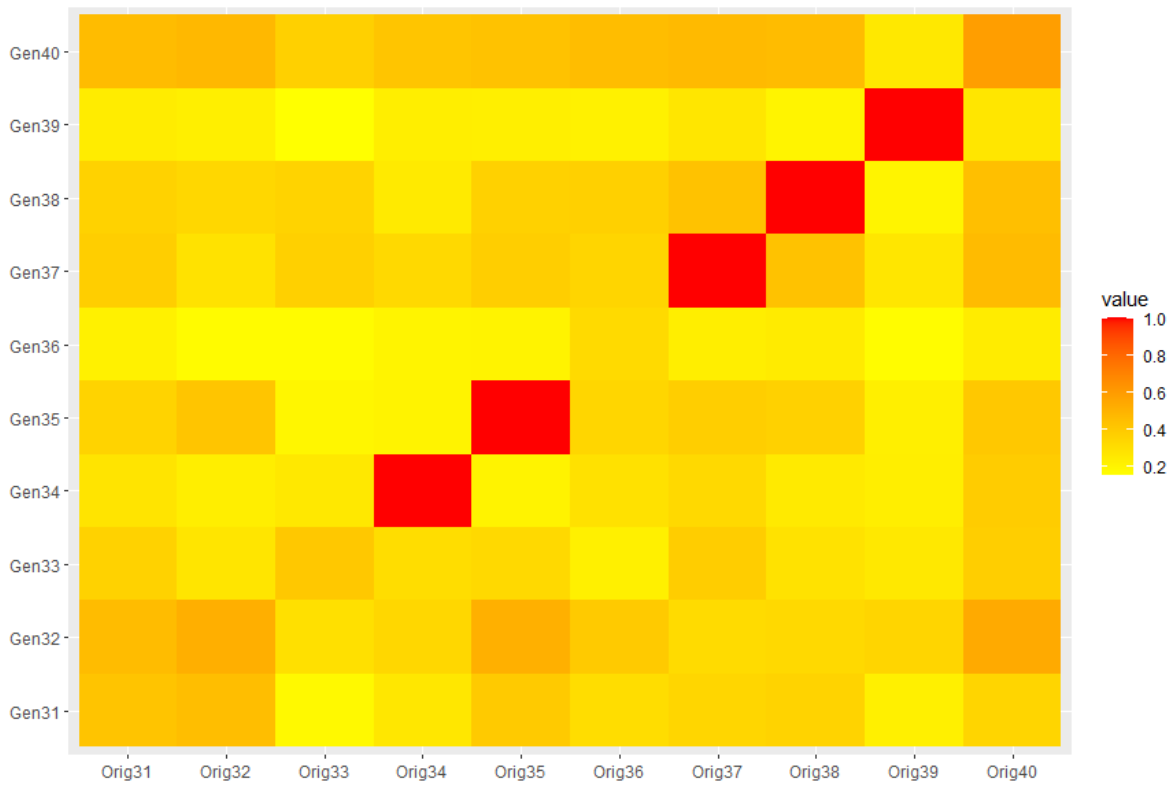


(a) Before finetuning

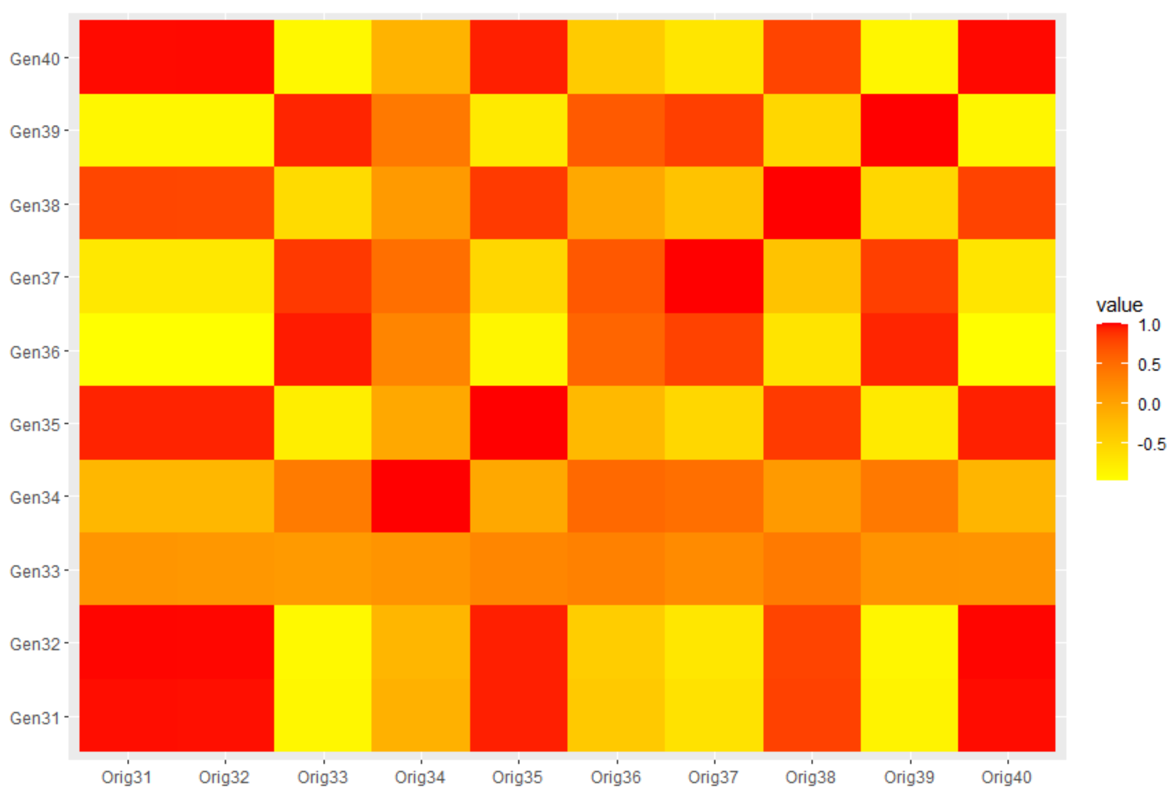


(b) After finetuning

Figure A.3: Finetuned USE heatmap of test pairs 21-30

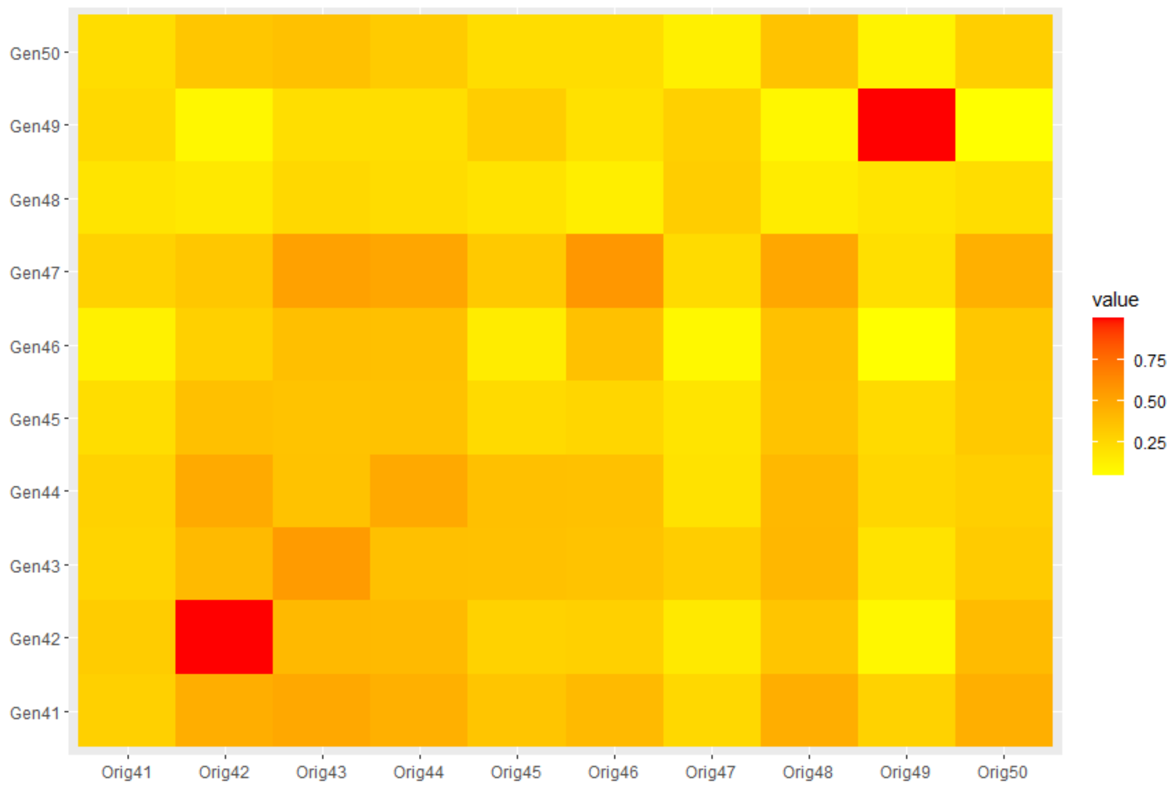


(a) Before finetuning

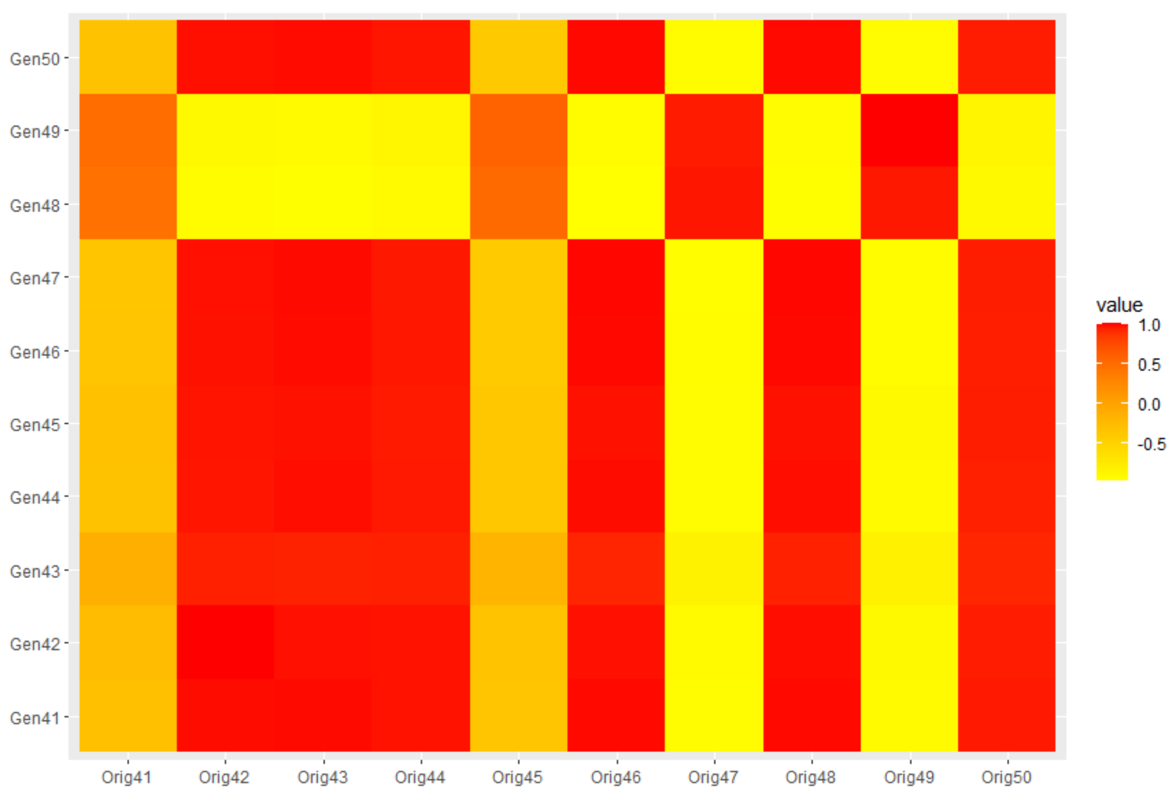


(b) After finetuning

Figure A.4: Finetuned USE heatmap of test pairs 31-40

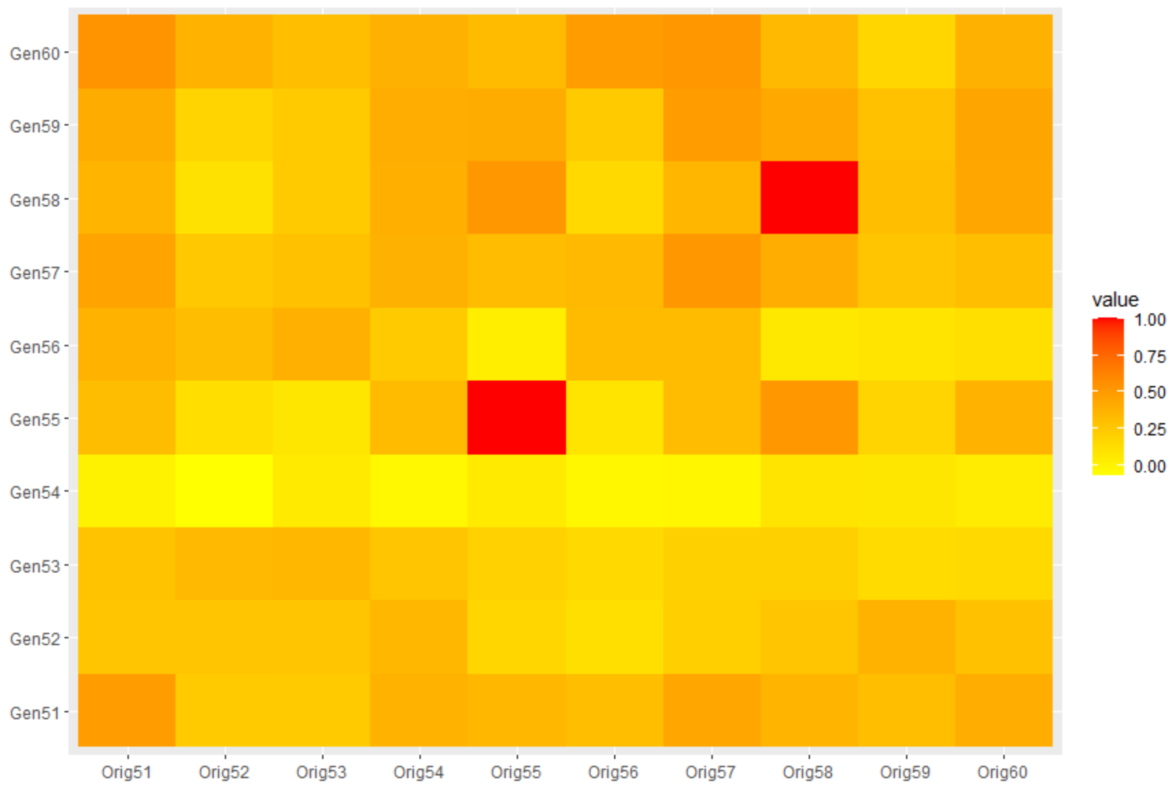


(a) Before finetuning

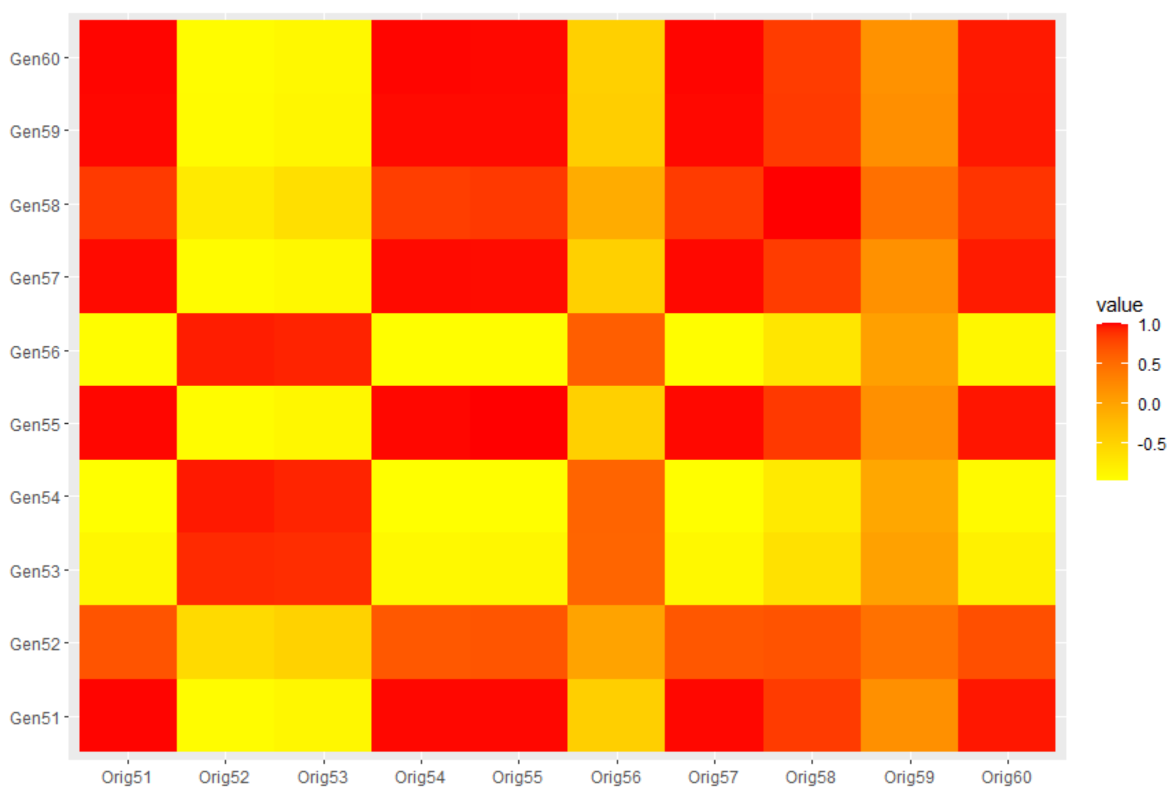


(b) After finetuning

Figure A.5: Finetuned USE heatmap of test pairs 41-50

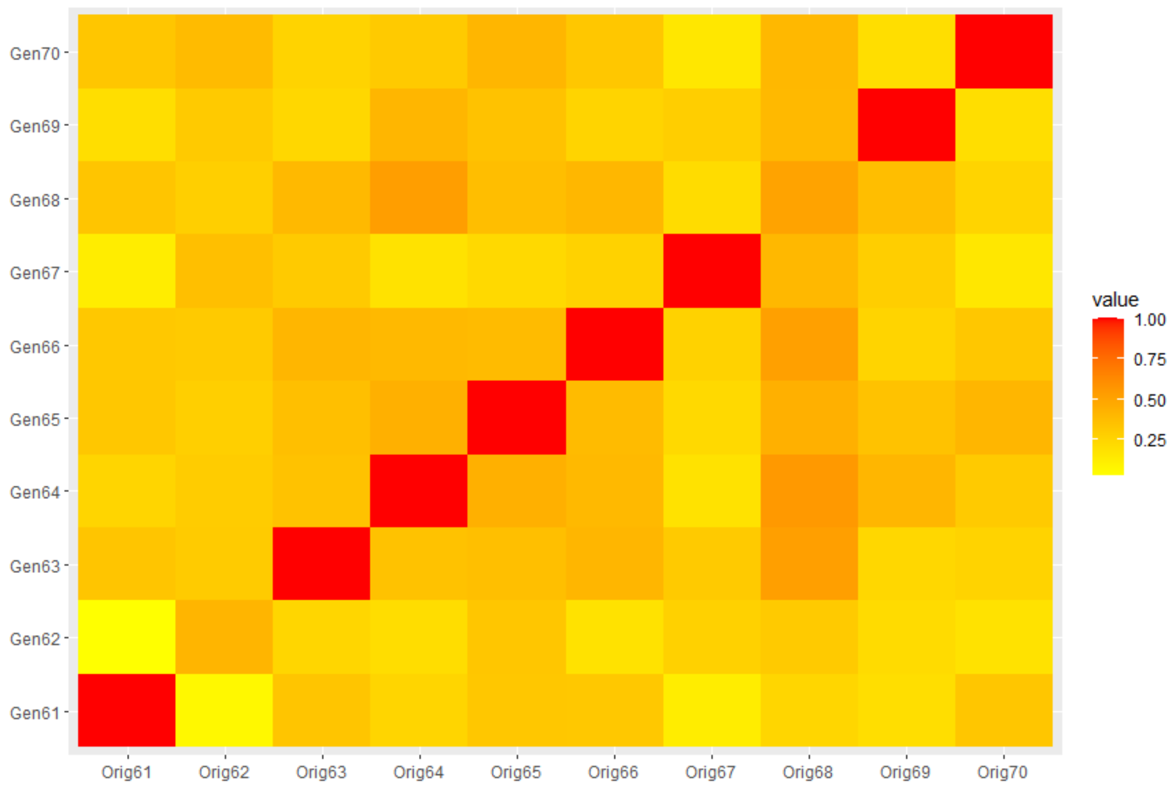


(a) Before finetuning

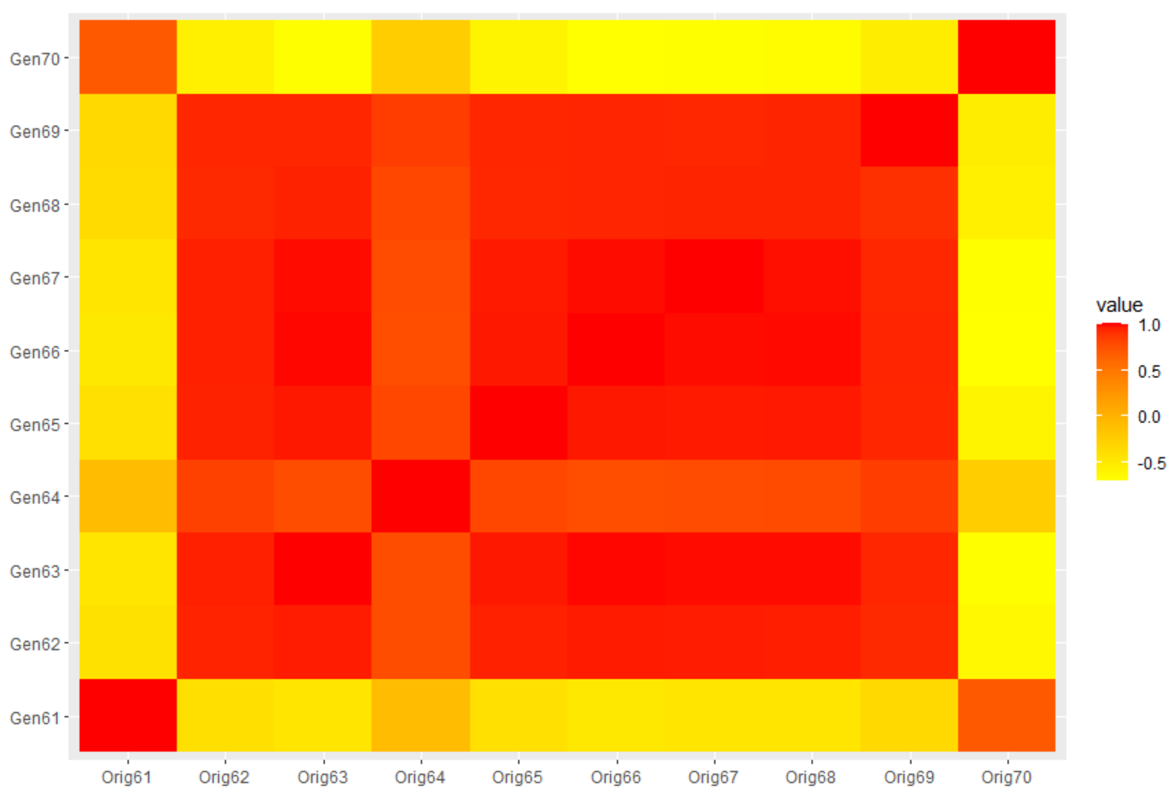


(b) After finetuning

Figure A.6: Finetuned USE heatmap of test pairs 51-60

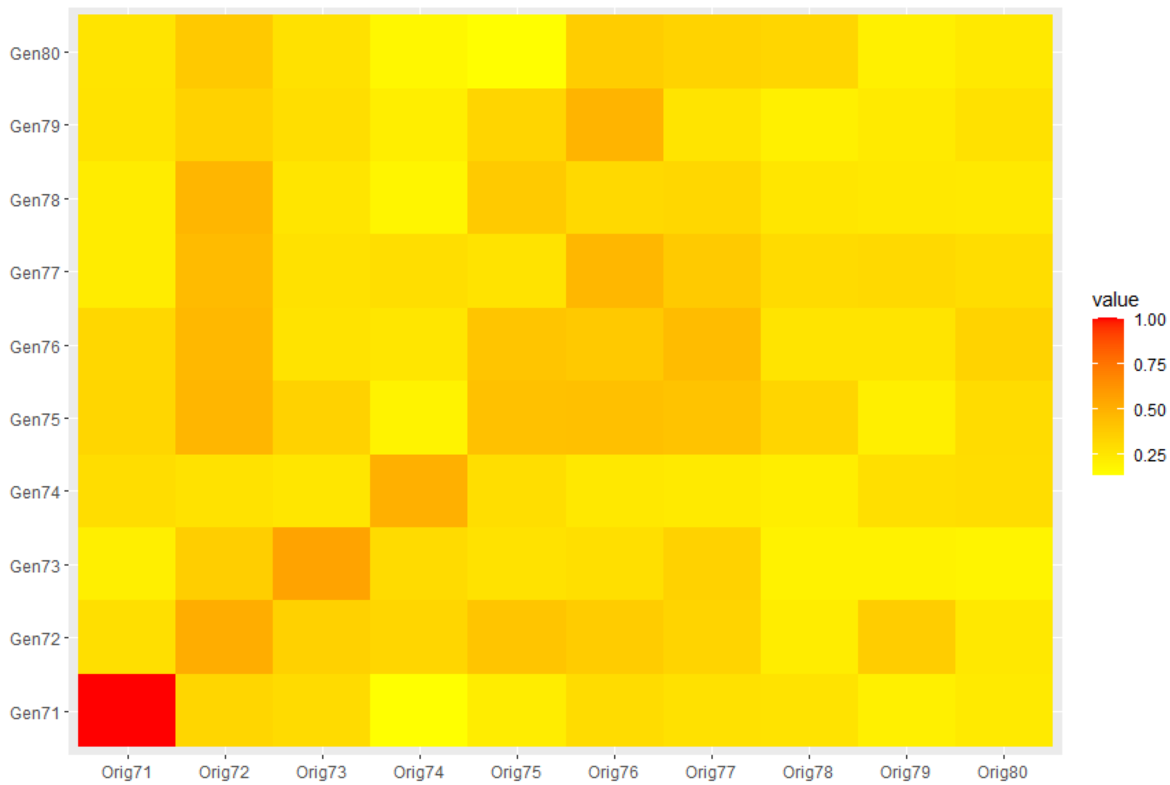


(a) Before finetuning

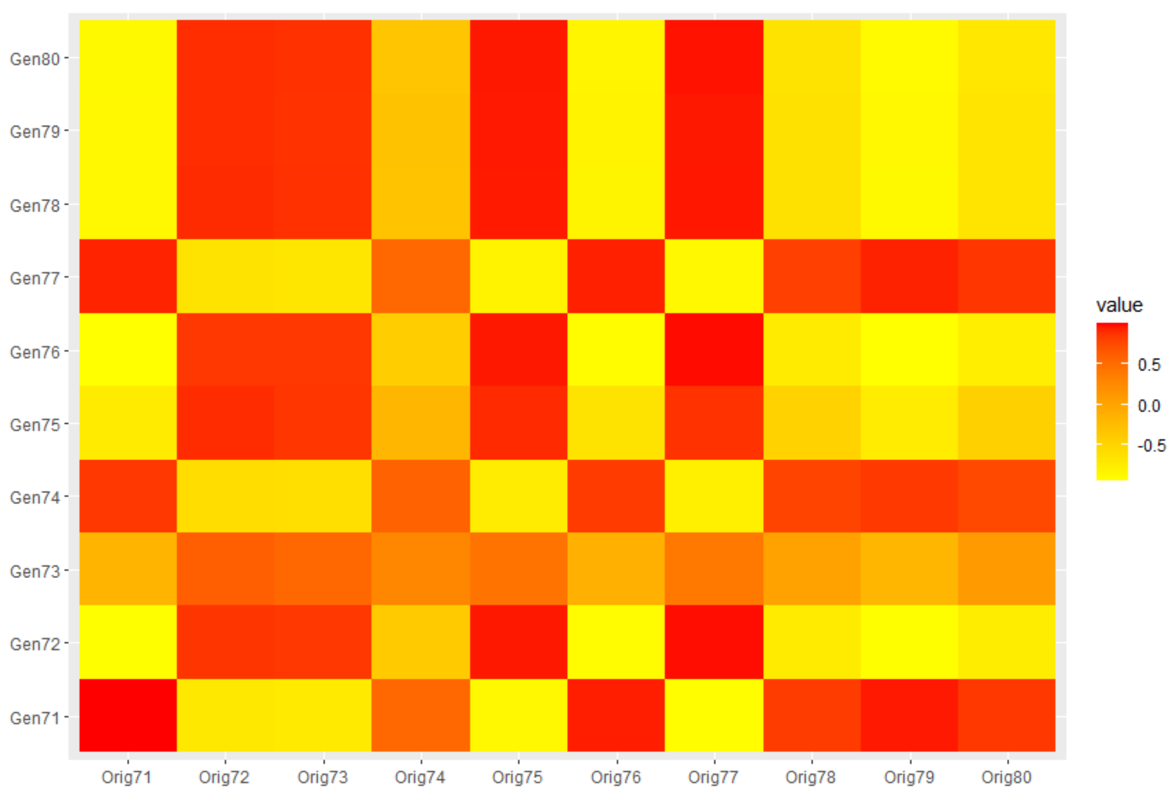


(b) After finetuning

Figure A.7: Finetuned USE heatmap of test pairs 61-70

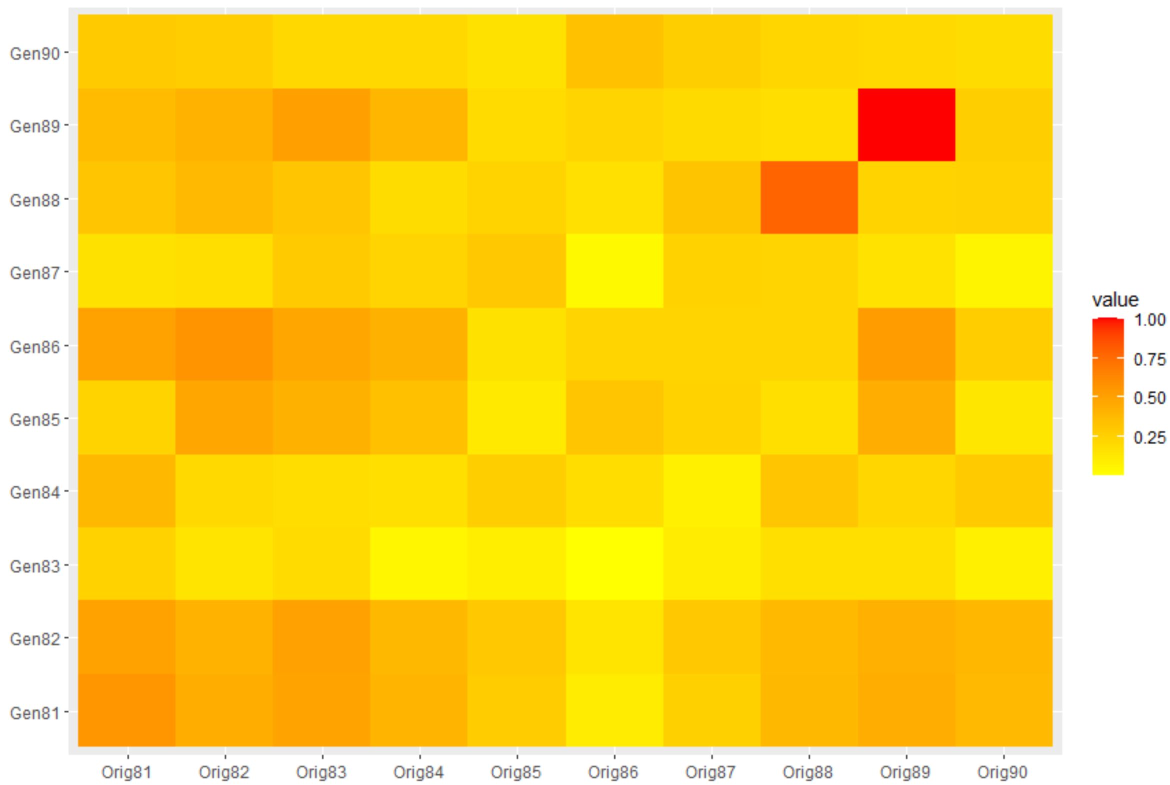


(a) Before finetuning

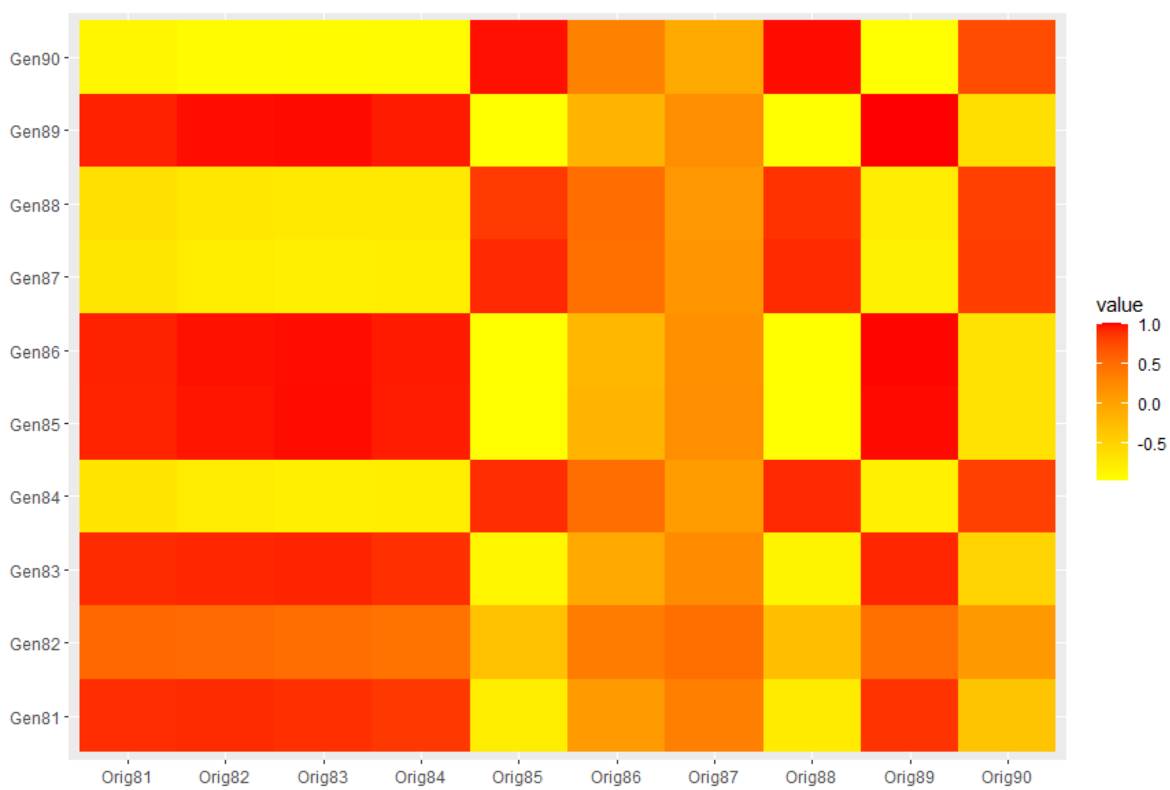


(b) After finetuning

Figure A.8: Finetuned USE heatmap of test pairs 71-80

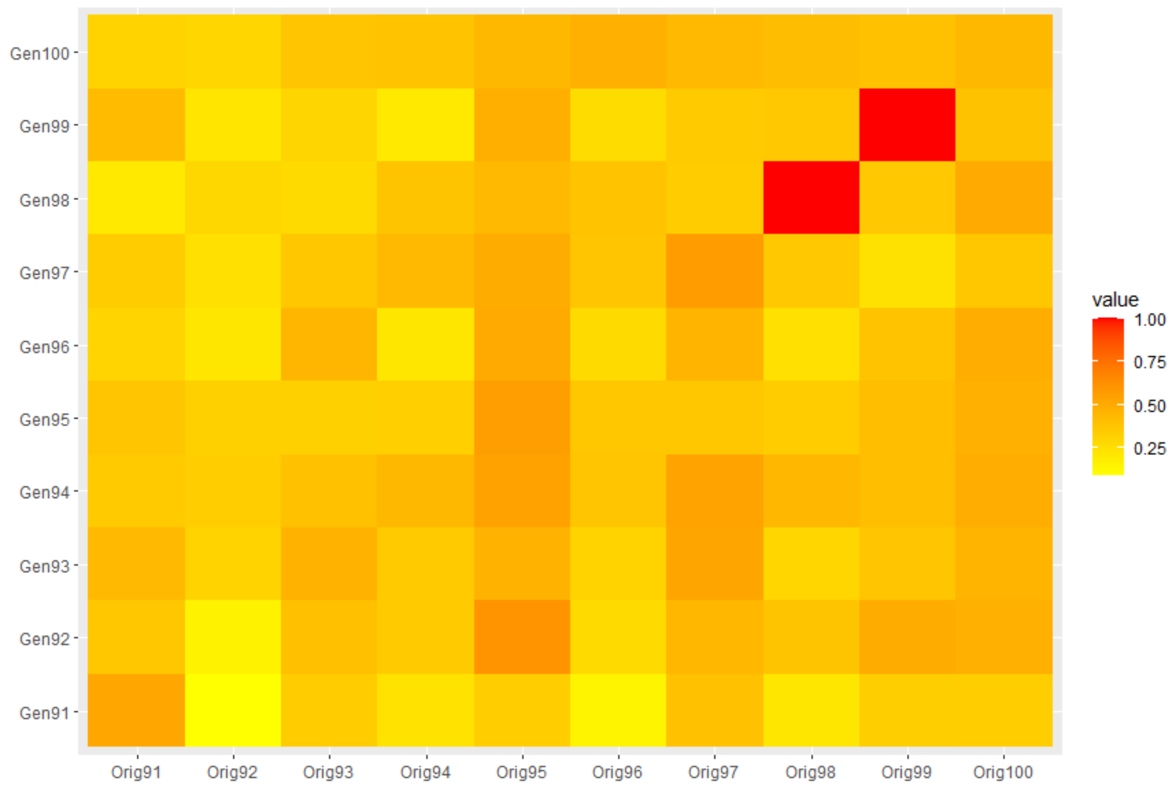


(a) Before finetuning

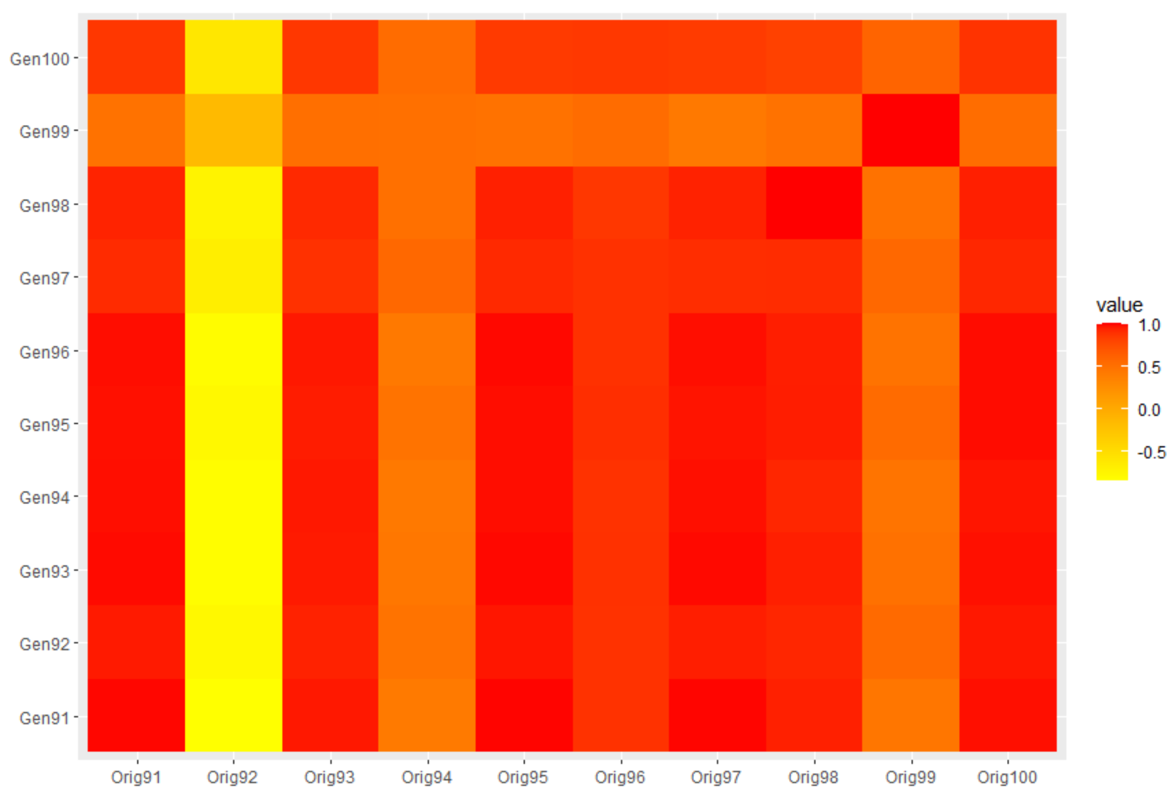


(b) After finetuning

Figure A.9: Finetuned USE heatmap of test pairs 81-90



(a) Before finetuning



(b) After finetuning

Figure A.10: Finetuned USE heatmap of test pairs 91-100

Appendix B

Publications

Conference Papers

Accepted Papers

- Baltazar, G., Ponce, P., López, E., & Molina, A. (2021). **Analysis of Text Definitions based on a Semantic Word Knowledge-Based Representation and a Transformer Neural Network Model**. In 20th Mexican International Conference on Artificial Intelligence.
- Ponce, P., Molina, A., Morales, B., & Baltazar, G. (2021). **Social Robotics for Children with Autism Based on Fuzzy Logic**. In International Engineering Congress (CONIIN 2021).

Published Papers

- Lopez-Caudana, E., Ponce, P., Mazon, N., Marquez, L., Mejia, I., & Baltazar, G. (2019, December). **Improving the Attention Span of Elementary School Children in Mexico Through a S4 Technology Platform**. In International Conference on Smart Multimedia (pp. 525-532). Springer, Cham.
- Reyes, G. E. B., Ponce, P., & Ayyanar, R. (2019, December). **Power Electronics in the Engineering Field: A Perception Comparison between Undergraduate and Graduate Students Using Fuzzy Logic Type 2 Signal Detection Theory**. In 2019 International Conference on Mechatronics, Robotics and Systems Engineering (MoRSE) (pp. 128-133). IEEE.
- Pedro, P., German, B. R., Barbara, C. A., Jhonattan, M., Arturo, M., & Maria, P. R. (2019). **Sensing, smart and sustainable product analysis methodology through EEG evaluation**. IFAC-PapersOnLine, 52(13), 2378-2383.
- Caudana, E. L., Reyes, G. B., Acevedo, R. G., Ponce, P., Mazon, N., & Hernandez, J. M. (2019, June). **RoboTICs: Implementation of a robotic assistive platform in a mathematics high school class**. In 2019 IEEE 28th International Symposium on Industrial Electronics (ISIE) (pp. 1589-1594). IEEE.

- Ponce, P., Molina, A., Mata, O., & Baltazar, G. (2019, March). **LEGO® EV3 Platform for STEM Education in Elementary School**. In Proceedings of the 2019 8th International Conference on Educational and Information Technology (pp. 177-184).

Journal Papers

Papers Under Review

- Baltazar Reyes, G. E., Ponce, P., López Caudana, E. O., & Molina, A. (2021). **Analysis and Use of Textual Definitions through a Transformer Neural Network Model and Natural Language Processing**. Applied Sciences.

- Ponce, P., López, C., Baltazar, G., López, E., Mazon, N., & Molina, A. (2021). **Use of Robotic Platforms as a Tool to Support STEM and Physical Education in Developed Countries**. Sensors.

Accepted Papers

- López, E., & Baltazar, G. (2021). **Improving the Attention Span of Elementary School Children for Physical Education through an NAO Robotics Platform in Developed Countries**. International Journal on Interactive Design and Manufacturing (IJIDEM).

- Baltazar, G., López, E., Tlalpan, P., Jiménez, B., Mazon, N., & Ponce, P. (2021). **Design of a Novel High School Mathematics Class through the Usability Analysis of a Robot Implementation**. International Journal on Interactive Design and Manufacturing (IJIDeM).

- Baltazar, G., Ponce, P., Molina, A., & Avendaño, L. A. (2021). **Design of an Open Innovation Laboratory Based on the Student's Class Perception and the Knowledge Economy**. Transactions on Computer Systems and Networks.

- Mendoza, C., López, C., Baltazar, G., López, E., & Chong, E. (2021). **Use of robotics to teach on real case scenarios. A project-based learning case of study: physical therapy**. Australasian Journal of Educational Technology.

Published Papers

- Reyes, G.E.B.; Ponce, P.; Castellanos, S.; Hernández, J.A.G.; Cruz, U.S.; MacDaniel, T.; Molina, A. Driver's Personality and Behavior for Boosting Automobile Security and Sensing Health Problems Through Fuzzy Signal Detection Case Study: Mexico City. Sensors 2021, 21, 7350. https://doi.org/10.3390/s21217350

- Reyes, G. E. B., López, E., Ponce, P., & Mazón, N. (2021). **Role Assignment Analysis of an Assistive Robotic Platform in a High School Mathematics Class, Through a Gamification and Usability Evaluation**. International Journal of Social Robotics, 13(5), 1063-1078.

- Mendez, E., Baltazar-Reyes, G., Macias, I., Vargas-Martinez, A., de Jesus Lozoya-Santos, J., Ramirez-Mendoza, R., ... & Molina, A. (2020). **ANN Based MRAC-PID Controller Implementation for a Furuta Pendu-lum System Stabilization**. Advances in Science, Technology and Engineering Systems Journal, 5(3), 324-333.

- Ponce, P., Molina, A., Caudana, E.O.L. et al. **Improving education in developing countries using robotic platforms**. Int J Interact Des Manuf 13, 1401–1422 (2019). https://doi.org/10.1007/s12008-019-00576-5.

- Molina, A., Ponce, P., Reyes, G. E. B., & Soriano, L. A. (2019). **Learning perceptions of Smart Grid class with laboratory for undergraduate students**. International Journal on Interactive Design and Manufacturing (IJIDeM), 13(4), 1423-1439.

Book Chapters

Published Chapters

- Lopez-Caudana, E., Reyes, G. E. B., & Cruz, P. P. (2020). **Socially Assistive Robotics: State-of-the-Art Scenarios in Mexico**. In Industrial Robotics-New Paradigms. IntechOpen.

Analysis of Text Definitions based on a Semantic Word Representation and a GPT-2 Model

Germán Eduardo Baltazar Reyes[1][0000−0002−2358−2002], Pedro Ponce[1][0000−0001−7035−5286], Edgar Lopez-Caudana[1][0000−0002−1216−4219], and Arturo Molina[1][0000−0001−5461−2879]

[1] Tecnologico de Monterrey, Puente St. 14380, CDMX, Mexico a01331329@itesm.com
[2] Tecnologico de Monterrey, Puente St. 14380, CDMX, Mexico pedro.ponce@tec.mx
[3] Tecnologico de Monterrey, Puente St. 14380, CDMX, Mexico edlopez@tec.mx
[4] Tecnologico de Monterrey, Puente St. 14380, CDMX, Mexico armolina@tec.mx

Abstract. Nowadays, information access has generated a global problem of information overload that mainly results from information presented statically. On the other hand, the recent use of Natural Language Processing (NLP) has been trying to solve this problem using different approaches to resume, paraphrase or even change the way textual information is presented. However, those algorithms are being used merely to copy the writing structure of a human being without analyzing if the resulting text follows a semantically correct structure. This paper uses a GPT-2 neural network to generate textual definitions of different words based on their semantical similarity with the ground-truth definition. The results obtained show that a neural network model can generate text while analyzing the semantical structure of the chosen words, but further validation analysis is needed to determine if the generated text is truly comprehensible or similar to the understanding of the user.

Keywords: Natural Language Processing · Semantic Analysis · Transformer · Definition Analysis.

1 Introduction

The world we live in has moved into a state that needs to generate and obtain information faster. Thanks to the constant use of the Internet and the Web 2.0 model, the amount of information that people have access to has considerably increased. Google's index increased from 45 million web pages to more than 60 million [1] from 2017 to 2019. Regarding the publication of scientific content, just in 2018, Web of Science included 70 million articles in one month, while 3 million papers were published by 33,100 English-language journals [2]. The tremendous amount of information has produced a problem related to the management of data, information, and knowledge [1], known as information overload.

Three main components cause information overload [3]: the existence of multiple data sources, which generates an excessive amount of information; due to a large amount of information the user needs to evaluate, it is complicated to

Social Robotics for Autistic Children Based on Fuzzy Logic

Pedro Ponce
School of Engineering and Sciences
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

Arturo Molina
School of Engineering and Sciences
Tecnologico de Monterrey
Mexico City, Mexico
armolina@tec.mx

Berenice Morales González
Benemérita Escuela Normal Veracruzana "Enrique C. Rébsamen"
berenice.morales.g@gmail.com

Germán E. Baltazar Reyes
School of Engineering and Sciences
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

***Abstract*— Technology is a valuable tool for the treatment of autism in children. It can enhance the therapy, providing a new social interaction channel in a friendly and controlled environment. This paper proposes an integral technological platform based on two main elements. The first one is a robot, specifically designed to collaborate in the intervention process eliciting pro-social behavior, and a tablet application program explicitly intended to complement the intervention activities with the robot, giving at the same time administration support to the therapist offering information about the statistical performance of the child during the intervention. On the other hand, fuzzy logic systems are implemented in the technological platform to increase the performance of the robot and the tablet. Fuzzy logic includes an inference system based on linguistic rules that help the therapist interact with the children when they are taking the integral therapy because the inference system can cover different social situations automatically. This paper shows the complete design of the robot and tablet application as the main supported tools. Social skill therapy is designed and implemented for different children with autism. Moreover, fuzzy logic demonstrated that it is a robust system for dealing with uncertainties so that the robot can generate a correct response according to the input signal from the child. Finally, the preliminary results give a good insight into the potential of this proposed integral therapy. Also, fuzzy logic is a principal component for designing the treatment.**

***Keywords*— robot; tablet; autism; therapy; fuzzy logic; educational innovation; higher education**

I. INTRODUCTION

It is considered that the most appropriate robots to work with humans are humanoid robots [1]. There is an increasing interest in combining external phenomena with internal activities (seeing with "thinking") to create humanoid robots with facial expressions that make them appealing to humans and helpful in various environments. In this paper, we present a technological platform, part of which is a humanoid robot, TEC-O, designed to appear "sociable" and helpful in the therapy of children with Autism Spectrum Disorder (ASD).

When designing a humanoid robot, a complex one, that is required to interact with humans, particularly with children with ASD, several ethical issues need to be taken into consideration, such as morality and trust as far as the degree of autonomy is concerned. Vincent Wiegel [2] argued that robots need to feature constraining behavior to be most helpful and acceptable in their interaction with humans. Coeckelbergh [3], having similar concerns, proposed that robo-ethics should be built around one central question: if the humans will benefit from their interaction with the robots or not. Considering these ethical concerns in the design, TEC-O is a robot that assists the therapist, featuring the degree of constraining behavior that will allow him to feel comfortable and help him with certain aspects of his work with children with ASD. At the same time, TEC-O has specific characteristics (facial expressions, voice, arm movement) that appeal to children with ASD, attract their attention, and will help them increase their social interaction.

Several advances have been made in the use of robots in the therapy of children with autism [4, 5], and the development of detailed requirements has the potential to help improve upon the effectiveness of clinical robots for using in the treatment of children with autism [6, 7]. As mentioned above, many robots have been created with significant variations in shape, size, and style. The evaluation of their effectiveness is primarily based on the judgment and experience of expert clinicians and engineers [8]. Furthermore, it has been suggested that a robot must be robust, easily reprogrammable, affordable [9, 10], and appealing to children with autism to be helpful in therapy. Other requirements that have been proposed for a robot include having aspects familiar to the child, providing choices, having a modular design that can easily be customized, being simple.

Hence, this paper proposes using a robot as a collaborative alternative for the therapist during intervention activities in the child's session. Also, a tablet application is proposed to offer real-time information to the therapist regarding the child's performance during the intervention. Although there are several papers, which present robots and tables for autism therapy [11], there are very few that design a technological platform according to the specific needs of children with autism because they use a generic robot or a developed program for a tablet, but the tablet and the robot might not be linked as a single platform [12].

Improving the Attention Span of Elementary School Children in Mexico Through a S4 Technology Platform

Edgar Lopez-Caudana[1](✉), Pedro Ponce[2], Nancy Mazon[3], Luis Marquez[1], Ivan Mejia[1], and Germán Baltazar[1]

[1] School of Engineering and Sciences, Tecnologico de Monterrey, Mexico City, Mexico
edlopez@tec.mx
[2] Writing Lab, TecLabs, Tecnologico de Monterrey, Monterrey, Mexico
[3] Psychology Faculty, UNAM, Mexico City, Mexico

Abstract. Today's education faces a powerful enemy: lack of interest from students, who, even when attending class, find themselves distracted. This enemy has been in our schools for a long time, where technology has made it worse. This investigation intends to turn technology back to our side by proposing the use of an assistive robot, proving that it is capable of attracting student's attention and increasing their motivation for a Physical Education (PE) class. This paper demonstrates that for the use of this robot, the characteristics of a sensitive, sustainable, intelligent, and social service/product (S^4 products) need to be covered. The obtained data was analyzed from both an engineering and a psychological background. This study concludes that the attention span of children improves while their motivation increases as a result of participating in a robot-assisted PE class.

Keywords: Social robotics · Assistive education · S^4 products · Educational innovation · Higher education

1 Introduction

Nowadays, education professionals are constantly confronted with the fact that their students have a lack of attention and behavioral problems in the classroom. From [1], it was observed that usual outdoor physical education activities are inefficient, and children are quickly discouraged. According to this study, Mexico is the most inactive country in levels of physical-sport activities with a higher dropout rate [2]. Because of this, there is a need to develop open science strategies that guarantee the improvement of educational achievements among the school-age population, without forgetting to emphasize groups at risk of being excluded in opportunities to participate and learn [3, 4]. To improve the learning and cognitive processes (like memory and executive functions), it is critical to stimulate the attentional resources of children during their first years of school [5]. Technology is defined as "the set of scientifically ordered technical knowledge that allows for the design and creation of goods and services" [6]. In this sense, the use of

T. McDaniel et al. (Eds.): ICSM 2019, LNCS 12015, pp. 525–532, 2020.
https://doi.org/10.1007/978-3-030-54407-2_44

Power Electronics in the Engineering Field: A Perception Comparison Between Undergraduate and Graduate Students Using Fuzzy Logic Type 2 Signal Detection Theory

Germán E. Baltazar Reyes
School of Engineering and Sciences
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

Pedro Ponce
School of Engineering and Sciences
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

Raja Ayyanar
Arizona State University
Tempe Arizona, United States
rayyanar@asu.edu

***Abstract*—This paper proposes the implementation and analysis of a student survey to evaluate the perception of undergraduate and graduate engineering students have regarding the use of power electronics in the new electrical field. Also, a comparison between the undergraduate power electronics program of Tecnologico de Monterrey and the graduate program of Arizona State University is presented. The objective of this work is to use Fuzzy Logic Type 2 Signal Detection Theory to evaluate the results of the surveys and find the qualities of both educational models and propose a general educational methodology to standardize the general concepts required for future generations of engineers that intend to work in the electrical field. The results obtained demonstrate that both graduate and undergraduate students are sure that their power electronics programs are adequate to their specialization necessities, considering even the uncertainty levels when answering the perception survey.**

***Index Terms*—Educational Innovation, Fuzzy Logic, Perception, Power Electronics, Signal Detection Theory**

I. Introduction

During this last decade, the global demand for electrical energy has increased considerably. Also, many countries and international associations have noticed the importance of implementing renewable and clean energy sources to reduce the greenhouse effect on the planet and reduce the consumption of fossil fuels. Both circumstances require the use of new technologies in the electrical field, as well as qualified human personnel that operates this new, changing electrical model and equipment [1]. However, the number of engineers associated with the electrical field has decreased considerably [2], making it crucial to train a new generation of electrical engineers [3]. Many universities have created partnership relationships [4] with electrical firms to study, evaluate, and improve the educational models given to engineering students to improve their understanding of the new electrical grid [5].

The Smart Grid concept englobes the new technological, financial, environmental, and social aspects of the new electrical grid model. It integrates communication and informatics technologies to monitor the generation, transmission, distribution, and consumption of electrical energy. The consideration of these new concepts demonstrates the importance of specializing their engineering students into different multidisciplinary skills in order for them to operate effectively the new technologies used in the electrical grid [6].

Even though universities are aware of the necessity of educating their students into the Smart Grid implementation, there are no specific norms nor well-defined study programs and design methodologies that universities could use for educating their students [7]. This problematic causes that every institution teaches and sees the Smart Grid contents from different perspectives, making it more difficult for future engineers to understand what the future of electrical engineering is [8].

In this work, a comparative study between an undergraduate class of electrical engineering from Tecnologico de Monterrey (ITESM) and a graduate class of power electronics from Arizona State (ASU) University is made. The objective of this work is to analyze the perspective of the students from both universities regarding the contents seen in their classes and the relationship they see between the Smart Grid and the use of power electronics. The results obtained from this comparison will help to determine if the contents seen in both classes are well-fitted to educate the new generation of electrical engineers into the implementation of the Smart Grid. Section II will present the generalities of both ITESM and ASU classes, followed by the explanation of how the perception was measured in both classes using Fuzzy type 2 Signal Detection (F2SDT) in Section III. The case of study is described in Section IV, while the results and its discussion are discussed in Sections V and VI, respectively. Finally, the conclusions of this work are presented in Section VII.

II. Current Class Methodologies

There has been difficulties when presenting novel and useful contents to the students that are involved with electrical engineering studies, specially with undergraduate students. This problem is perceived by the low retention rates and the lack of knowledge of the students regarding the usefulness

Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 52-13 (2019) 2378–2383

Sensing, smart and sustainable product analysis methodology through EEG evaluation.

Ponce Pedro * **Baltazar R. German** ** **Castillo A. Barbara** ***
Miranda Jhonattan * **Molina Arturo** * **Perez R. Maria** ****

* *Tecnologico de Monterrey, School of Engineering and Science, Puente St., CDMX 14380 Mexico ({pedro.ponce, jhonattan.miranda, armolina}@tec.mx).*
** *Tecnologico de Monterrey, School of Engineering and Science, Puente St., CDMX 14380 Mexico (a01331329@itesm.mx)*
*** *Universidad de Huelva, Dr. Cantero Cuadrado 6, Huelva 21004 Spain (barbaracastilloabdul@gmail.com)*
**** *Universidad de Huelva, Dr. Cantero Cuadrado 6, Huelva 21004 Spain (amor@perez-rodriguez.es)*

Abstract: Thanks to the technological proliferation of digital platforms, the amount of people that shares content regarding any topic has increased considerably. This phenomenon has allowed fashion and product designers to promote their merchandise more easily. However, these approaches require new methodologies for developing their products, as well as consider the sustainable aspects in their campaigns. In the era of Information 4.0, it is possible to create a product that is customized to the user's needs by following a sensing, smart and sustainable S^3 paradigm. This paper proposes a product analysis methodology for the fashion industry, based on the S^3 model, for evaluating how useful internet videos influence the customer's emotional response and the tendency for using certain products. All of this while analyzing their responses through monitoring their brain activity during the view of the videos using electroencephalography (EEG) readings, and the emotional response they generate in them by implementing the "Emocard" model, all of this to analyze the media influence for using or not the promoted product.

Keywords: Product strategy, design systems, sensing, smart and sustainable systems, user response.

1. INTRODUCTION

In the globalized world we live in, it is necessary for the industry to find ways of connecting with their final users to ensure their loyalty to the brand (Cho and Fiore (2015)). In the fashion industry, where product design and usage tendencies change considerably fast, is no exception. One of the main objectives of this industry is to develop, produce and deliver products or services with multiple customization options, enabling their clients to find what they want at an acceptable price (Pine et al. (1993)). However, achieving this goal is quite complex due to the difficulty of getting the grip of the client's emotional fit (Desmet et al. (2001)).

For this reason, multiple researchers have investigated ways of evaluating the best methodologies for fashion product design and marketing. Lavidge and Steiner (1961) developed the Hierarchy of Effects (HOE) model for analyzing the cognition-affect-conation sequence of consumers regarding mass customized products. Park and Yoo (2018) continued using this methodology for determining if the emotional response of the client is related to such sequence. Townsend et al. (2016) developed a product design methodology for creating fashionable clothing for mature women based on their emotional fit, while Desmet et al. (2001) implemented the use of "Emocards" to evaluate the emotional fit of their clients for the first stages of designing new products. Weihua (2009a) promotes the benefits and necessity for the fashion industry to consider and evaluate the final client's emotional response when thinking about new products. As observed, the importance of assessing the emotional response of users has become an incredibly important task in product design for the fashion industry.

With the arrival of web 2.0, people have been able to generate and share content without requiring a deep understanding of the topic they talk about (Jones et al. (2009)), but only expressing their opinion. From all the people that share their minds and create original content with it, the most recognized globally are called "influencers." The reason why these people are so famous among the rest of the population is their proximity to the final user, which gives a more familiar perspective about the use of certain products that cannot be seen in the apparels or being used by models.

Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2019.11.562

RoboTICs: Implementation of a Robotic Assistive Platform in a Mathematics High School Class

Edgar Lopez Caudana
School of Engineering and Science
Tecnologico de Monterrey
Mexico City, Mexico
edlopez@tec.mx

German Baltazar Reyes
School of Engineering and Science
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

Rodrigo Garcia Acevedo
School of Engineering and Science
Tecnologico de Monterrey
Mexico City, Mexico
a01333171@itesm.mx

Pedro Ponce
School of Engineering and Science
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

Nancy Mazon
Faculty of Psychology
National Autonomous University of Mexico
Mexico City, Mexico
nomy1329@gmail.com

Jorge Membrillo Hernandez
School of Engineering and Science
Writing Lab, TecLabs
Tecnologico de Monterrey
Mexico City, Mexico
jmembrillo@tec.mx

***Abstract*—The Mathematics program created by the Mexican Ministry of Public Education (SEP) aims to have high school students with active development of a critical, logical thinking, creativity, and broad general knowledge to be able to argue and structure their ideas practically and logically. Although there have been constant changes in educational programs to improve, Mexico has consistently placed itself in the last places of The Program for International Student Assessment (PISA), a triennial international survey which aims to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. One of the reasons given is that Mexican students are not sufficiently motivated during classes, resulting in a decline in academic qualifications. In the present report, we developed and implemented an experimental protocol that allows us to improve the levels of focusing on high school students during a math class using an NAO robot. Three different scenarios were studied: one with a mathematics teacher traditionally giving a complete class, another where the robot teaches the course and finally one where both the teacher and the robot cooperate to provide explanations and exercises. Electroencephalogram (EEG) readings were used in the students to observe levels of attention, while a group of Psychology students monitored their physical behavior through a body language analysis protocol.**

***Index Terms*—Behavior, EEG, Human-Robot Interface, Humanoid Robots, Interdisciplinary Research, Mathematics, NAO Robot, Teaching**

I. Introduction

In 2012, Mexico ranked 53th among the 65 countries that con-ducted the PISA assessment [1]–[4]. This review was developed by the Organization for Economic Co-operation and Development (OECD) to analyze the level of mathematical, linguistic and scientific understanding that high school students have around the world. In this examination, students must be able to identify and understand the use of mathematics in real-life scenarios [3]. This test is performed every three years, evaluating only one of the three main themes mentioned above.

According to the study Scholar Failure and Transition to the Working World, the leading causes that generate school failure in high school students are boredom and lack of interest. The reason for this is that their current classes lack innovative activities that invite them to participate. During these last years, the Ministry of Education (SEP) of Mexico encouraged the use of Information and Communication Technologies (ICTs), believing that the implementation of new technologies generates a better inclusion of students [5], [6].

There have been different approaches around the world that use a robotic platform for educational purposes. Most of them rely on STEM (Science, Technology, Engineering, and Mathematics) education using Arduino and other open source modules [7]. In Chile, a BEAM and LEGO Mindstorms kit were used with the same purpose [8]. Around rural areas of Costa Rica, a group of teachers created and implemented a series of robotics workshops for low-income secondary and high school students [9]. However, none of these projects sought to improve a regular class.

The Georgia Institute of Technology analyzed a student's behavior when a Darwin robot interacted with him using verbal, non-verbal, both or any of these languages while responding to a math test. The results showed that the use of a robotic platform improves the time taken to complete the test, at the same time that the student felt more comfortable responding [10]. A study by the University of Plymouth showed that a student achieves better grades when a teacher teaches them, rather than a robotic platform. In this study, 11 children between eight and nine years of age were taught with only one NAO robot, while 11 other children had the same class with a human math teacher. In both cases, the children responded to a previous and final test to assess the development of children's knowledge [11].

This project seeks to improve levels of attention for Mexican high school students during a math class. The team worked with a group of high school teachers to develop and apply an experimental protocol that included the use of an NAO robot as an auxiliary tool for the teacher during the class.

LEGO® EV3 Platform for STEM Education in Elementary School

Pedro Ponce[1], Arturo Molina[1], Omar Mata[1], Germán Baltazar[1]
Tecnologico de Monterrey Campus Ciudad de Mexico
ITESM
Mexico City, Mexico
pedro.ponce@tec.mx, armolina@tec.mx, omar.mata@tec.mx, a01331329@itesm.mx

ABSTRACT
In this paper, describes the implementation of a LEGO® robotic kit inside an elementary school. This platform was programmed using NI LabVIEW to create 26 different diligences used for STEM teaching with students from fourth to sixth grade. Such diligences covered mathematics and renewable energies topics that the students used to play and learn about the respective topics remarked by the school educational program. After using the robotic kit, every student answered a test that included questions regarding the topic treated. The evaluation of the student's results was made using fuzzy logic through a NI LabVIEW interface. This program was evaluated in an elementary school in Mexico City, showing that the use of this platform helped secure their knowledge of the topics seen.

CCS Concepts
• Applied computing →Education →Interactive learning environments • Social and professional topics →Professional topics →Computing education →Student assessment • Human-centered computing →Ubiquitous and mobile computing

Keywords
Education; LEGO®; NI LabVIEW; STEM; artificial intelligence, elementary school.

1. INTRODUCTION
In Mexico, the methodology used for educating in elementary schools has followed a very traditional approach: the professor gives all the theoretical explanations to the students, while they listen and memorize the content given. However, this approach makes students lose interest in their classes, affecting their academic performance [1]. The lack of more dynamic approaches influences negatively the children's ability to comprehend the theoretical concepts seen in class. This also affects their design, reasoning and measurement competencies, making it more challenging to understand future concepts that are seen in further years.

ICEIT 2019, March 2–4, 2019, Cambridge, United Kingdom

ACM ISBN 978-1-4503-6267-2/19/03…$15.00
http://doi.org/10.1145/3318396.3318426

With the purpose of generating greater inclusion with the students during their classes, the Educational Secretary of Mexico (SEP) started implementing in elementary and high schools the use of Information and Communications Technologies (ICTs). This approach looked for creating more compelling explanations and exercises that could help students retain the information.

The first ICT used with educational purposes was implemented by Seymour Papert with the LOGO programming language [2] to teach mathematics. Nonetheless, in the technologically advanced era we live in, the use of computers became less efficient with younger generations due to their familiarity with such technologies.

For this reason, the use of robotic kits started to be implemented during these last years to create more enthusiasm in children when dealing with Science, Technology, Engineering and Mathematics (STEM) topics [3-7]. One of the most common platforms used is the LEGO® robotic kit, being the EV3 the latest version. This kit provides engaging, hands-on experiences required to comprehend STEM concepts and link them to real-life applications [8].

Thanks to the collaboration between LEGO® and NI LabVIEW, it is possible to create programs and diligences that generate integral educational content according to the educational programs used in Mexico.

2. EDUCATION SYSTEM IN MEXICO
The Mexican government divides educational levels into four levels created and supervised by the SEP [9] as follows:

- Starter Education. Conformed by students younger than 6 years.
- Basic Education. An educational period that teaches basic concepts to children between 6 and 15 years old.
- Highschool. Preparatory education taught before entering the professional studies.
- Superior Education. Education focused on a specific area for the student's professional life and possible postgraduate studies.

Thanks to the Integral Reform of Basic Education, SEP started to invest in the distribution and use of ICTs with the objective of promoting "learning by reception" methodologies. These methods guide to active and participative construction of knowledge from the students [10].

This paper focuses on the implementation of the LEGO EV3 platform for STEM-oriented diligences in the basic education level, with students from 4th, 5th, and 6th grade.

Article

Analysis and Use of Textual Definitions through a Transformer Neural Network Model and Natural Language Processing

Germán Eduardo Baltazar Reyes [1,*], Pedro Ponce [1], Edgar López Caudana [1], and Arturo Molina [1]

[1] School of Engineering and Sciences, Tecnologico de Monterrey National Department of Research, Mexico City 14380, Mexico; A01331329@tec.mx (G.E.B.R.); pedro.ponce@tec.mx (P.P.); edlopez@tec.mx (E.O.L.C.); armolina@tec.mx (A.M.)

* Correspondence: a01331329@1tec.mx

Abstract: There is currently an information overload problem, where data is excessive, disorganized, and presented statically. These three problems are deeply related to the vocabulary used in each document since the usefulness of a document is directly related to the number of understood vocabulary. This model uses a GPT-2 Transformer neural network to interpret a descriptive input phrase and generate a new phrase that intends to speak about the same abstract concept, similar to a particular keyword. The validation of the generated text is in charge of a Universal Sentence Encoder network, which was finetuned for properly relating the semantical similitude between the total sum of words of a sentence and its corresponding keyword. The results demonstrated that the proposal could generate new phrases that resemble the general context of the descriptive input sentence and the ground truth keyword. At the same time, the validation of the generated text was able to assign a higher similarity score between these phrase-word pairs. Nevertheless, this process also showed that it is still needed deeper analysis to ponderate and separate the context of different pairs of textual inputs.

Keywords: Natural Language Processing; Transformer; Semantic Similarity; Text Analysis; Phrase-Word Analysis

Citation: Baltazar Reyes, G.E.; Ponce, P.; López Caudana, E.; Molina, A. Analysis of Textual Definitions through Natural Language Processing. *Journal Not Specified* **2021**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

We currently live in an environment where it is crucial to find, analyze, and use reliable information daily [1]. At the same time, there have been multiple technical advances that allow us to have access to more affluent, complex information sources in different formats and types [2]. One of the most noticeable sources of information has been the Internet, which after the Web 2.0 model, allowed any person to publish, share, and comment on new content. However, the amount of information available to everyone has surpassed the human capacity to analyze and use it completely.

This phenomenon is called information overload, referring to the difficulty a person faces when deciding in the presence of excessive information [3], the overabundance of relevant information that exceeds the human processing capability [1,4], or the burden of equally excessive unsolicited information [5]. Thus, it can be said that information overload is a problem related to the management of data, information, and knowledge [1].

Information overload is mainly caused by the existence of multiple sources of information, the overabundance of information, difficulty in managing it, the vast amount of irrelevant data, and the scarcity of time on the part of information users to analyze and understand all the received data [6]. All these problems can be reduced to three particular issues:

- Information is **excessive**. Nobody can analyze every existent document.

Design of a Novel High School Mathematics Class through the Usability Analysis of a Robot Implementation

Germán E. Baltazar Reyes
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

Edgar Omar Lopez
Tecnologico de Monterrey
Mexico City, Mexico
edlopez@tec.mx

Patricia Tlalpan
Tecnologico de Monterrey
Mexico City, Mexico
prt@tec.mx

Brenda Jiménez
Tecnologico de Monterrey
Mexico City, Mexico

Nancy Mazon
Universidad Autónoma de México
Mexico City, Mexico
nancymaz01@yahoo.com.mx

Pedro Ponce
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

***Abstract*—This paper proposes the implementation of a robotic platform that is capable of improving the attention levels of students in high school math class. This proposal found that the topics of the math classes taught with the support of a robotic platform were more attractive and dynamic compared with a traditional class model with only the professor's guidance. The primary goal was to transform a conventional class into an experimental one to improve the students' grades. To assess the robotic platform, an interdisciplinary research work was conducted that integrated psychology metrics, as well as an usability and gamification evaluation of the platform. The proposed platform was an NAO humanoid robot that followed the approved teaching routines of the group's teachers. The students' attention levels were measured, scores were recorded, and groups (classes) were compared using observation scales. The evaluation results showed that the humanoid robot empowered the proposed platform's effectiveness when implemented as a complementary educational tool. Teaching classes with this platform increased student motivation, provided an interactive space, and developed new skills in both students and teachers, such as teamwork.**

***Keywords*—Social Robotics, Assistive Education, Robot NAO, Educational Innovation, Higher Education, Math Education, Usability, Gamification**

I.INTRODUCTION

In Mexico, teaching at the high school level must have an immediate application to everyday life for the students to significantly impact society. Therefore, the use of Information and Communication Technologies (ICTs) within the classroom can improve teaching-learning processes in engineering education [1]. New technologies generate greater inclusion as everyone interactively participates in classes without being propped up by the teacher or their companions [2,3].

In most classes at the national level, teachers do not use teaching materials due to lack of creativity, time, proper training, or planning [4].

It is essential to know how mathematics classes are developed in Mexico within the social and environmental context. The results in [5] showed that students should have better learning activities and assessment experiences in Mexico. The study found that the index of socio-economic and cultural status significantly and positively predicted achievement in mathematics. However, it did not change the relationship that learning opportunities have with learning mathematics. For example, the anxiety of Mexican students towards mathematics can be reduced with adequate motivation, so the tool developed to improve learning outcomes also aimed to address the issue of motivation [6-8].

It is interesting to consider that the use of technological tools help students achieve more efficient learning skills from different perspectives [1]. Competencies such as acquiring meaningful learning, collaborative learning, and empathy within the social environment give students desirable characteristics in their long-term training process. All could be of great help, but not all are effective if not used timely and moderately. Some educational scenarios that include ICT have better results than others. However, it is not always possible to observe the effects of the most disruptive technologies (robotics or virtual laboratories, for example), as seen in this work.

Educational robotics is a new teaching system that seeks to develop creativity, organization and collaborative work. This system is based on pedagogical constructivism that promotes creation, innovation, and self-design [9]. Besides strengthening knowledge, this system enables students to adapt to current production processes. Incorporating robotics into the class design helps the

Improving the Attention Span of Elementary School Children in Developed Countries for Physical Education through an NAO Robotics Platform

Abstract

Today's education faces a powerful enemy, lack of interest from students, who even if attending class, find themselves distracted, this enemy has been in our schools for a long time, but it has never been as strong as it is now, technology has made it strong, phone bearing children has become its ally. This investigation intends to turn technology back in our side, by proposing the use of an assistive robot, proving that it attracts student's attention and increases motivation for *Physical Education (PE)* class, so children learn how to carry a healthy life, avoiding diseases and conditions such as obesity, diabetes and heart problems. To prove this the levels of attention and motivation from the students were measured on two primary school groups, once with a traditional class and once with a robot-assisted class. The data was analyzed from both an engineering and a psychological background. This study concludes that the attention span of children improves while their motivation increases, a result of using a *NAO* robot, so a robot-assisted *PE* class can if applied decrease diabetes, obesity, and heart condition levels, by having children learning effectively how to live a healthy life.

Keywords
Social Robotics, Asistive Education, Robot NAO, Educational Innovation, Higher Education.

1. Introduction

Nowadays, education professionals are constantly confronted with the fact that their students have a lack of attention and behavior problems in the classroom. From [1], it was observed that usual outdoor physical education activities are inefficient, and children are quickly discouraged. According to this study, Mexico is the most inactive country in levels of physical-sport activities with a higher dropout rate [2]. Because of this, there is a need to develop open science strategies to guarantee the improvement of educational achievements among the school-age population, without forgetting to emphasize groups at risk of being excluded in opportunities to participate and learn. To improve the learning and cognitive

Design of a Novel High School Mathematics Class through the Usability Analysis of a Robot Implementation

Germán E. Baltazar Reyes
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

Edgar Omar Lopez
Tecnologico de Monterrey
Mexico City, Mexico
edlopez@tec.mx

Patricia Tlalpan
Tecnologico de Monterrey
Mexico City, Mexico
prt@tec.mx

Brenda Jiménez
Tecnologico de Monterrey
Mexico City, Mexico

Nancy Mazon
Universidad Autónoma de México
Mexico City, Mexico
nancymaz01@yahoo.com.mx

Pedro Ponce
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

***Abstract*—This paper proposes implementing a robotic platform to improve students' attention levels in high school math classes. This proposal found that the topics of the math classes taught with the support of a robotic platform were more attractive and dynamic compared with a traditional class model with only the professor's guidance. The primary goal was to transform a conventional class into an experimental one to improve the students' grades. An interdisciplinary research work integrated psychology metrics and a usability and gamification evaluation to assess the robotic platform. The proposed platform was an NAO humanoid robot that followed the approved teaching routines of the group's teachers. The students' attention levels were measured, scores were recorded, and groups (classes) were compared using observation scales. The evaluation results showed that the humanoid robot empowered the proposed platform's effectiveness when implemented as a complementary educational tool. Teaching classes with this platform provided an interactive space and developed a novel way of collaboration between students and teachers.**

***Keywords*—Social Robotics, Assistive Education, Robot NAO, Educational Innovation, Higher Education, Math Education, Usability, Gamification**

I.INTRODUCTION

In Mexico, teaching at the high school level must immediately apply to everyday life to significantly impact society. Therefore, using Information and Communication Technologies (ICTs) within the classroom can improve teaching-learning processes in engineering education [1]. New technologies generate greater inclusion as everyone interactively participates in classes without being propped up by the teacher or their companions [2,3].

In most classes at the national level, teachers do not use teaching materials due to a lack of creativity, time, proper training, or planning [4].

It is essential to know how mathematics classes are developed in Mexico within the social and environmental context. Bazán et al. [5] showed that students should have better learning activities and assessment experiences in Mexico. The study found that the index of socio-economic and cultural status significantly and positively predicted achievement in mathematics. However, it did not change the negative relationships that learning opportunities have with learning mathematics. For example, the anxiety of Mexican students towards mathematics can be reduced with adequate motivation, so the tool developed to improve learning outcomes also aimed to address the issue of motivation [6-8].

It is interesting to consider that technological tools help students achieve more efficient learning skills from different perspectives [1]. Competencies such as acquiring meaningful learning, collaborative learning, and empathy within the social environment give students desirable characteristics in their long-term training process. All could be of great help, but not all are effective if not used timely and moderately. Some educational scenarios that include ICT have better results than others. However, as seen in this work, it is not always possible to observe the effects of the most disruptive technologies (robotics or virtual laboratories, for example).

Educational robotics is a new teaching system that seeks to develop creativity, organization, and collaborative work. This system is based on pedagogical constructivism that promotes creating, innovation, and self-design [9]. Besides strengthening knowledge, this system enables students to adapt to current production processes. Incorporating robotics into the class design helps the teacher solve problems where it costs him to concentrate. The use of robotics in education involves putting new activities in place for students to develop disciplinary and transversal competencies.

However, it is not enough to implement the ICTs as it is. It is required to evaluate and question which aspects of the class model are better implemented by a professor and the platform [10]. According to Goodrich et al., it is essential to measure the robot's level of autonomy, its ability to exchange information, the structure of the model

Design of an Open Innovation Laboratory Based on the Student's Class Perception and the Knowledge Economy

Germán E. Baltazar Reyes
Tecnologico de Monterrey
Mexico City, Mexico
a01331329@itesm.mx

Pedro Ponce
Tecnologico de Monterrey
Mexico City, Mexico
pedro.ponce@tec.mx

Arturo Molina
Tecnologico de Monterrey
Mexico City, Mexico
armolina@tec.mx

Luis Arturo Avendaño
Tecnologico de Monterrey
Mexico City, Mexico

***Abstract*—Due to the modernization of the electrical grid and the commitments to it made by several governments and industries worldwide, the work of engineers specialized in the electrical field is necessary more than ever. However, in recent years, the number of engineers working in this area has decreased, while almost half of their current population is prone to retirement. To solve this problem, universities began to modify their electrical engineering programs and courses, focusing on Smart Grid technology implementation. Simultaneously, it is not enough to only evaluate if the given classes are considered appropriate by the students, but to the industry, since one of the main objectives of new class implementations is to generate marketable products in collaboration with different firms. This paper proposes a new syllabus and new Smart Grid class based on hands-on experiments in a Smart Grid Laboratory and a novel concept of innovation laboratory that focuses on particular requirements that the industry has nowadays. In this paper, the student's perception about the mentioned class is evaluated through Signal Detection Theory and Fuzzy Logic Types 1 and 2, while the structure of the open innovation laboratory is described based on four innovation facilities. The results showed that the students acquired a synthesis of learning and analytical thinking to equip them with the competencies to solve electrical grid modernization's various challenges. Additionally, the proposed new class methodology utilized innovative hands-on activities in Laboratory practices that reinforced learning the most relevant theoretical concepts of the Smart Grid technology while evaluating current electrical industry problems.**

Keywords—Educational Innovation, Fuzzy Logic, Perception, Signal Detection Theory, Smart Grid, Open Innovation Laboratory, Knowledge Economy

I. INTRODUCTION

The Smart Grid implementation requires effort from several sectors, including governments, industry, education, and end-users. The traditional electrical grid modernization involves the implementation and development of advanced technology. Because of this modernization, the electrical power stages of generation, transmission, distribution, and consumption have increased their functionality and participation. Integrating communication and information technologies in Smart Grids enable bidirectional communication between stakeholders and the main grid. Similarly, the Smart Grid promotes the high penetration of renewable resources in the main grid through distributed generation installation. Thus, according to Tuballa & Abundo [1], the grid's modernization's main functional benefits are reliability, security, and efficiency.

In [2], the National Institute of Standards and Technology (NIST) formulate the Smart Grid in seven domains: Customers, Markets, Service Providers, Operations, Bulk Generation, Transmission, and Distribution. These domains communicate with each other to achieve the same objectives. Hence, Smart Grid's implementation requires integrating several technologies, new standards, and a new engineering workforce with multidisciplinary skills that render it capable of solving problems in all the fields related to Smart Grid implementation.

However, developing and implementing these new technologies is difficult when the engineers hired have a traditional formation and are not aware of the specific problems brought about by technology integration. Meanwhile, established firms dominate their sectors with well-known technological models. This fact demonstrates that employees' multidisciplinary skills in engineering firms have not been adequately taught in the universities [3], resulting in a lack of specialization in engineering fields [4].

Another urgency that needs attention is that we currently live in a world where products and production are constantly changing. Demand forces shorter and more rapid innovation cycles, pushing engineers to develop and comprehend different disciplines, apply essential information and communications technology skills, and interact with people from other careers and cultures [5] to deal with the world's complex challenges.

Recognizing these problems, some universities started developing educational models that connect the engineering classes and laboratories with real-world firms that give students true-life challenges to solve, helping them develop needed competencies. The strengthening of relationships among universities and industries [6] allows collaborative efforts to define specific innovation areas in educational methodologies. In this way, engineering colleges can implement industry inputs to the design and teach entrepreneurship programs [3] for the new generations of engineers, benefitting from the industry's partnerships.

Even though it is known that countries' economic growth and development depend mainly on the knowledge economy [7], there is still little information and studies regarding the different ways of possible collaboration between universities and industry to achieve this [8,9]. For this reason, it is precise to understand how to implement the four pillars of the knowledge economy: economic and institutional regime, education and skills, information and communication

Use of robotics to teach on real case scenarios. A project-based learning case of study in physical therapy.

The dynamic society we live in requires constant adaptation and innovation on all the techniques known, leaving room to improve and assist the necessities of future generations. For this particular study, a new approach regarding project-based learning in higher education goes beyond the typical school environment and brings solutions to real-life scenarios. This project was developed with the engineering undergraduate students collaborating with a rehabilitation institute in Mexico City to design a physical therapy routine using the NAO robot. This proposal allows interaction with infant patients in real-time and fosters empathy during the process of developing a final usable product. This proposal measured the usability of the robotic platform during the sessions and the reproductivity of the proposal through Cronbach's alpha evaluation. The usability results show a higher interest in the project while constructing the material needed to develop a product matching the standards given by the rehabilitation institute.

Implications for practice or policy

- *Professors could change the traditional approaches of teaching while including new methodologies with real case scenarios.*
- *Higher education students could learn about their school topics and create empathy while helping people with physical disabilities.*
- *Schools might need to consider creating alliances within a wide range of institutions to create common goals and provide technological solutions.*

Keywords
Higher education, Social robotics, Project-based learning, Physical therapy, Educational innovation

Introduction

Nowadays, the labor market requires students to possess a wide range of skills, often referred to as 21st-century skills, beyond mastering the basic skills of their specific disciplines (Zarouk et al., 2020). Therefore, universities should focus on the technical tools they provide to their students and create a wide range of soft skills to allow their future alumni success in society. Furthermore, creating real-life approaches during their professional preparation should be a must in planning their curricula, allowing students to create projects to get a passing grade inside the classroom, interact with society, and give a higher meaning to their work. In order to match higher education with the current needs required from professionals, teachers need to develop an instructional model that focuses on enabling students to develop practical skills which they can apply to learn real-life situations (Sakulvirikitkul et al., 2020). This model would allow students to apply their knowledge on the specific project, interact with people from different professional areas, and work together to achieve a mutual goal.

This study demonstrates the importance and benefits of project-based learning to solve real-life problems by designing, programming, and applying a physical therapy routine using the NAO robot. This platform is a humanoid robot capable of interacting with the user through different human-like movements and programmed conversation lines. At the same time, the robot allows the user to interact with it through a variety of touch sensors, microphones, voice recognition, or even image detection through a camera.

Undergraduate students were asked to collaborate with the Foundation to Help the Mentally Weak (FADEM) in Mexico City to generate physical therapies using robotics. This institute is a non-profit organization committed to improving the quality of life of children, young people, and adults with intellectual disabilities.

sensors

MDPI

Article

Driver's Personality and Behavior for Boosting Automobile Security and Sensing Health Problems Through Fuzzy Signal Detection Case Study: Mexico City

Germán E. Baltazar Reyes [1], Pedro Ponce [1,*], Sergio Castellanos [2], José Alberto Galván Hernández [1], Uriel Sierra Cruz [1], Troy MacDaniel [3] and Arturo Molina [1]

[1] School of Engineering and Sciences, Tecnologico de Monterrey National Department of Research, Mexico City 14380, Mexico; A01331329@itesm.mx (G.E.B.R.); josegalvan_94@hotmail.com (J.A.G.H.); A01171504@itesm.mx (U.S.C.); armolina@tec.mx (A.M.)

[2] Civil, Architectural and Environmental Engineering, Cockrell School of Engineering, University of Texas at Austin, Austin, TX 78712, USA; sergioc@utexas.edu

[3] Center for Cognitive Ubiquitous Computing (CUbiC), Arizona State University, Tempe, AZ 85281, USA; troy.mcdaniel@asu.edu

* Correspondence: pedro.ponce@tec.mx

Abstract: Automobile security became an essential theme over the last years, and some automakers invested much money for collision avoidance systems, but personalization of their driving systems based on the user's behavior was not explored in detail. Furthermore, efficiency gains could be had with tailored systems. In Mexico, 80% of automobile accidents are caused by human beings; the remaining 20% are related to other issues such as mechanical problems. Thus, 80% represents a significant opportunity to improve safety and explore driving efficiency gains. Moreover, when driving aggressively, it could be connected with mental health as a post-traumatic stress disorder. This paper proposes a Tailored Collision Mitigation Braking System, which evaluates the driver's personality driving treats through signal detection theory to create a cognitive map that understands the driving personality of the driver. In this way, aggressive driving can be detected; the system is then trained to recognize the personality trait of the driver and select the appropriate stimuli to achieve the optimal driving output. As a result, when aggressive driving is detected continuously, an automatic alert could be sent to the health specialists regarding particular risky behavior linked with mental problems or drug consumption. Thus, the driving profile test could also be used as a detector for health problems.

Keywords: signal detection; fuzzy logic; personality; health problems

check for updates

Citation: Reyes, G.E.B.; Ponce, P.; Castellanos, S.; Galván, J. A.; Sierra, U.; McDaniel, T.; Molina, A. Driver's Personality and Behavior for Boosting Automobile Security and Sensing Health Problems Through Fuzzy Signal Detection Case Study: Mexico City. *Sensors* **2021**, *21*, 7350. https://doi.org/10.3390/s21217350

Academic Editor: Giovanni Pau

Received: 1 September 2021
Accepted: 12 October 2021
Published: 5 November 2021

1. Introduction

Early automotive warning systems were implemented since the late 1950s to increase vehicle safety [1]. One of the first prototypes created was the Cadillac Cyclone [1], which used radar technology to detect objects in front of the car. In 1995, a research team of Hughes Research Laboratories (HRL) developed the Forewarn, a radar-based forward collision avoidance system [2]. Unfortunately, these models' expensive manufacturing costs prevented their mass production and marketing.

At the beginning of the 2000s, several investigations were published regarding the viability and usability of frontal collision warning systems. For example, the Insurance Institute for Highway Safety [3] discovered that the use of autonomous avoidance and adaptive headlights regarding the driver's steers considerably reduced insurance coverage in car accidents. However, the U.S. National Highway Traffic Safety Administration (NHTSA) considered that it is still not mandatory the usage of these systems in commercial vehicles [4].

International Journal of Social Robotics
https://doi.org/10.1007/s12369-020-00698-x

Role Assignment Analysis of an Assistive Robotic Platform in a High School Mathematics Class, Through a Gamification and Usability Evaluation

Germán E. Baltazar Reyes[1] · Edgar López[1] · Pedro Ponce[1] · Nancy Mazón[2]

Accepted: 5 September 2020

Abstract
This project implements an assistive robotic platform in a mathematics high school class to support the professor's teaching process while analyzing its performance by using a gamification approach and the Octalysis framework. The results obtained from this study evaluated both the platform usability and the general class perception from the students and the professors involved in the experimentation. This paper demonstrates that the implementation of a robotic platform as a supportive tool for the professor improves the class' dynamism and the cooperative behavior of the students, by following the Octalysis approach. Finally, the results helped to determine a better role assignment for future cases where the professor implements a robotic platform during class, leaving the theoretical explanations and class control to the professor, and giving the robot the role of enhancing the activities and review exercises.

Keywords Social robotics · Educational innovation · Higher education · Gamification · Usability

1 Introduction

The use of robotic platforms in different human-labor tasks has increased during the last years. Most of their applications are focused on the industrial field, or the implementation of mechatronics, artificial intelligence, among others. However, their use in more social scenarios has also increased considerably, like entertainment, comfort, assistance for children and elderly, autistic, and disabled persons, and education [40].

Regarding the latter one, the use of robotic platforms has been mainly focused on teaching programming techniques and artificial intelligence for undergraduate students [2,12, 30,33,42,43], being the purpose of the platform to reinforce the topics seen in class while adequating the implementation of the platform to the social and cultural queues of the region [28]. Another aspect that began to take a more significant role in robotics research, and that still needs further investigation [38,40], is the use of the platform as a supportive tool for the professor [2,39] in topics that are not directly implied to robotics per se.

There have been some approaches regarding this last area. Chang et al. proved that the implementation of a humanoid robotic platform generates an engaging and interactive environment for children to learn a second language [3]. Other researchers demonstrated that early interactions with robotic platforms improve students' STEM knowledge and thinking processes, like Sullivan et al. [44]. Schweikardt et al. observed that the use of robotic kits provides a novel idea for teaching complex STEM topics to students [35], while fomenting problem-solving and teamwork activities between them [23]. It has also been noticed that the robot could not have the teaching role entirely. It is also necessary that the professor performs as an authority figure for the students, acting as the mediator between them and the platform, and helping them understand how to interact with the robot appropriately [26].

On the other hand, these previous studies demonstrate that the use of a robotic platform improves the students learning process and acquisition of new information, especially in STEM-related classes. Nevertheless, it is still necessary to evaluate which the robotic platform improved class aspects and which ones by the professor's intervention. Evaluating

Germán E. Baltazar Reyes
a01331329@itesm.mx

1 School of Engineering and Sciences, Tecnologico de Monterrey, Calle del Puente 222, Mexico City, Mexico

2 Faculty of Psychology, UNAM, Avenida Universidad 3004, Mexico City, Mexico

Advances in Science, Technology and Engineering Systems Journal
Vol. 5, No. 2, 324-333 (2020)
www.astesj.com

ASTES Journal
ISSN: 2415-6698

ANN Based MRAC-PID Controller Implementation for a Furuta Pendulum System Stabilization

Efrain Mendez*, German Baltazar-Reyes, Israel Macias, Adriana Vargas-Martinez, Jorge de Jesus Lozoya-Santos, Ricardo Ramirez-Mendoza, Ruben Morales-Menendez and Arturo Molina

Tecnologico de Monterrey, School of Engineering and Sciences, Mexico

A R T I C L E I N F O

Article history:
Received: 15 January, 2020
Accepted: 09 May, 2020
Online: 30 May, 2020

Keywords:
Furuta Pendulum
Robotics
Robotic systems
UMS
Artificial Intelligence
ANN
MRAC-PID
Non-linear Systems

A B S T R A C T

Nowadays, process automation and smart systems have gained increasing importance in a wide variety of sectors, and robotics have a fundamental role in it. Therefore, it has attracted greater research interests; among them, Underactuated Mechanical Systems (UMS) have been the subject of many studies, due to their application capabilities in different disciplines. Nevertheless, control of UMS is remarkably more difficult compared to other mechanical systems, owing to their non-linearities caused by the presence of fewer independent control actuators with respect to the degrees of freedom of the mechanism (which characterizes the UMS). Among them, the Furuta Pendulum has been frequently listed as an ideal showcase for different controller models, controlled often through non-lineal controllers like Sliding-Mode and Model Reference Adaptive controllers (SMC and MRAC respectively). In the case of SMC the chattering is the price to be paid, meanwhile issues regarding the coupling between control and the adaptation loops are the main drawbacks for MRAC approaches; coupled with the obvious complexity of implementation of both controllers. Hence, recovering the best features of the MRAC, an Artificial Neural Network (ANN) is implemented in this work, in order to take advantage of their classification capabilities for non-linear systems, their low computational cost and therefore, their suitability for simple implementations. The proposal in this work, shows an improved behavior for the stabilization of the system in the upright position, compared to a typical MRAC-PID structure, managing to keep the pendulum in the desired position with reduced oscillations. This work, is oriented to the real implementation of the embedded controller system for the Furuta pendulum, through a Microcontroller Unit (MCU). Results in this work, shows an average 58.39% *improvement regarding the error through time and the effort from the controller.*

1 Introduction

In recent decades, robotics have gained increasing importance in endless applications for different disciplines (as explained in [1]); such as, robotic manipulators for industrial automation, precision robots to perform surgeries, automation of assembly lines, among others. Thus, robotics have attracted greater interest multidisciplinary researchers.

Moreover, according to [2], it is defined in robotics that an Underactuated Mechanical System (UMS), is a scheme with more Degrees of Freedom (DOF) compared to its control actuators. Additionally, [3] explains that the control of UMS's is a very active research topic, due to their broad applications in Robotics, Aerospace, and Marine Vehicles.

Therefore, in spite of the complexity caused by the lack of actuators to control the movement of the system, it is precisely the low number of required actuators that makes UMS's ideal for applications where energy efficiency is sought. Additionally, UMS's are systems that allow to decrease the size of the manipulators, and even simplify the amount of elements of a more complex system. The above, resulting also in cost reduction with an increased process efficiency, which is the key element that attracts the most interest for its development in the industry.

The non-linearities caused by the relation between the actuators and the DOF to be controlled, makes the complexity of these systems attractive as testbeds, for the research of different control structures. Among them, this work is focused in the rotatory pendulum, also known as the *Furuta Pendulum* (in honor of its inventor *K.*

*Corresponding Author: Efrain Mendez Flores; Tecnologico de Monterrey, School of Engineering and Sciences, Mexico City 14380, Mexico; Correspondence: efrain@tec.mx; Tel.: +52-5521-333-361.

International Journal on Interactive Design and Manufacturing (IJIDeM)
https://doi.org/10.1007/s12008-019-00576-5

ORIGINAL PAPER

Improving education in developing countries using robotic platforms

Pedro Ponce[1] · Arturo Molina[1] · Edgar Omar Lopez Caudana[1] · German Baltazar Reyes[1] · Nancy Mazon Parra[2]

Received: 23 January 2019 / Accepted: 3 May 2019

Abstract
In developing countries, educational programs have been using an out of date teaching methodology, where students' only work is to listen to the professors' lessons without having practical applications or hands-on experience, thereby provoking a decrease in student attention span, motivation, and boredom with learning processes. This work proposes the use of robotic platforms inside elementary schools and universities to improve and evaluate the usability and effectiveness of robots on students' attention spans, motivation, and knowledge acquirement during their classes. This work mentions a case of study in Mexico where the use of the robotic platforms was evaluated. To evaluate the learning process of undergraduate students, a hexapod robot was used throughout a basic programming course to help the students learn about robotics programming and path planning algorithms. At the end of the course, several surveys were given to the students to evaluate their perceptions about the course and the use of the robotic platform. In the elementary school, a NAO Robot was used to give students four different diligences in physics, mathematics and physical education classes. During these experimentations, the attention span of the students and their ability to use the robotic platforms were observed, using a behavior observation protocol; also, their knowledge acquisition before and after class was evaluated. The results showed that the use of a robotic platform in class helps the students improve their knowledge acquisition and increases their motivation and attention span. Also, the surveys and usability analysis demonstrate that the design of the diligences and course projects were sufficient to generate greater interest among the students in the topics taught in school.

Keywords Assistive robotics · Educational innovation · Educational platform · Learning robotics · NAO robot · Path planning

1 Introduction

To prepare students for the challenging future, it is necessary to question what we need to teach and what kinds of abilities they need to develop through their academic life, from their first years of elementary school through their undergraduate years.

To achieve this, it is necessary to determine which skills the students need to develop in proper education. According to many sources dedicated to the improvement of education around the world, the three most essential groupings of skills a student needs to develop in school are:

- Communication and multidisciplinary teamwork skills, life-long learning and awareness of new economic, social, and environmental concerns [1].
- The ability to think both critically and creatively and to have a systems perspective [2].
- Ingenuity, creativity, business leadership and flexibility [3].

With this in mind, the present challenge is to know how we will teach and how students will learn. The model we propose features a continuous flow between practice and

✉ German Baltazar Reyes
a01331329@itesm.mx

Pedro Ponce
pedro.ponce@tec.mx

Arturo Molina
armolina@tec.mx

Edgar Omar Lopez Caudana
edlopez@tec.mx

Nancy Mazon Parra
nancymaz01@yahoo.com.mx

1 School of Engineering and Sciences, Tecnologico de Monterrey, Mexico City, Mexico

2 Universidad Autonoma de Mexico, Mexico City, Mexico

International Journal on Interactive Design and Manufacturing (IJIDeM)
https://doi.org/10.1007/s12008-019-00603-5

ORIGINAL PAPER

Learning perceptions of Smart Grid class with laboratory for undergraduate students

Arturo Molina[1] · Pedro Ponce[1] · Germán Eduardo Baltazar Reyes[1] · Luis Arturo Soriano[1]

Received: 23 January 2019 / Accepted: 7 June 2019

Abstract
Due to the modernization of the electrical grid and the commitments to it made by several governments and industries around the world, the work of engineers specialized in the electrical field is necessary more than ever. However, in recent years, the number of engineers working in this area has been decreasing, while almost half of their current population is prone to retirement. To solve this problem, universities began to modify their electrical engineering programs and courses, giving more focus to the implementation of Smart Grid technology. Although various approaches have been used in teaching methodologies to educate new engineers, it is also necessary to evaluate if the contents given in such classes are being properly taught. This paper proposes a new syllabus and new Smart Grid class, which is based on hand on experiments in a Smart Grid laboratory. This proposal promotes and trains undergraduate students in the use of the new technologies that are being deployed in the electrical industry nowadays, and it includes a discussion of the social, economic and environmental implications of the new ways to generate and distribute electrical power. To evaluate if the class methodology in our project was successfully implemented, a student perception survey was applied to analyze the way the undergraduate students perceived the Smart Grid class given to them. Additionally, signal detection theory and fuzzy logic type 1 and type 2 were used to compare their answers with the ones given by the professor as part of assessing the efficiency of the class syllabus and the teaching methodology for the purpose of improving their quality in future courses. The results obtained showed that the students acquired a synthesis of learning and analytical thinking to equip them with the competencies to solve the various challenges of electrical grid modernization. Additionally, the proposed new class methodology utilized innovative hands-on activities in laboratory practices that reinforced the learning of the most relevant theoretical concepts of the Smart Grid technology.

Keywords Educational Innovation · Fuzzy logic type 1 · Fuzzy logic type 2 · Perception · Signal detection theory · Smart Grid

1 Introduction

The Smart Grid implementation requires effort from several sectors, including governments, industry, education, and end users, among others. The traditional electrical grid modernization involves the implementation and development of advanced technology. Because of this modernization, the electrical power stages of generation, transmission, distribution, and consumption have increased their functionality and participation. The integration of communication and information technologies in Smart Grids enables bidirectional communication between stakeholders and the main grid. Similarly, the Smart Grid promotes the high penetration of renewable resources in the main grid through distributed generation installation. Thus, according to [1], the main functional benefits acquired by the modernization of the grid are reliability, security, and efficiency.

In [2], the National Institute of Standards and Technology (NIST) formulate the Smart Grid in seven domains; namely, Customers, Markets, Service Providers, Operations, Bulk Generation, Transmission and Distribution. These domains

Germán Eduardo Baltazar Reyes
a01331329@itesm.mx

Arturo Molina
armolina@tec.mx

Pedro Ponce
pedro.ponce@tec.mx

Luis Arturo Soriano
larturo.soriano@tec.mx

[1] School of Engineering and Sciences, Tecnologico de Monterrey, Mexico City, Mexico

Chapter

Socially Assistive Robotics: State-of-the-Art Scenarios in Mexico

Edgar Lopez-Caudana, Germán Eduardo Baltazar Reyes and Pedro Ponce Cruz

Abstract

In this chapter, we describe the experience about the use of a humanoid robotic platform, in scenarios such as education and health in Mexico. The results obtained are commented on through the perspective of cultural, technological, and social aspects in the frameworks of education (from elementary to high school) and training of health professionals. The opening towards humanoid robotic systems in elementary school children, as well as health professionals, is not far from the acceptance due not only for the technological advancement but also for different social aspects. These two considerations influenced the results obtained and experiences achieved. At the same time, this chapter shows how humanoid robotics has functioned as a tool for final projects of undergraduate students.

Keywords: social robotics, assistive education, higher education, educational innovation

1. Introduction

The implementation of biped and humanoid robots has increased during the last decade around the world. Depending on its characteristics, these robots can be implemented in different areas and domains, such as manufacture, agriculture, health, and education, among others. For the rest of this chapter, every reference will treat the last two applications mentioned, particularly rehabilitation, patient assistance, and elementary children's education.

Depending on the functionality of the robot, it can be classified into three different types [1], as seen in **Figure 1**. The first type of robot is called an assistance robot (AR). Its functionality is based only on giving physiological assistance to a patient that presents a physical disability or is recovering from a surgical operation. This type of robots regularly presents a basic structure, since they have one single task and the work environment does not vary too often.

On the other hand, as the name suggests, socially interactive robots (SIR) have the task of interacting with the end user with the sole purpose of generating a certain level of communication and entertainment to him. Generally, these platforms have a more complex design, since they need to detect and use human behavior patterns to achieve an efficient interaction.

Bibliography

[1] Thesaurus dictionary. `https://www.thesaurus.com/`. Accessed: 2021-11-02.

[2] ACHINSTEIN, P. What is an explanation? *American Philosophical Quarterly 14*, 1 (1977), 1–15.

[3] ACHINSTEIN, P. *The nature of explanation*. Oxford University Press on Demand, 1983.

[4] AGIRRE, E., CER, D., DIAB, M., GONZALEZ-AGIRRE, A., AND GUO, W. * sem 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity* (2013), pp. 32–43.

[5] ASHER, N., ASHER, N. M., AND LASCARIDES, A. *Logics of conversation*. Cambridge University Press, 2003.

[6] AUSTIN, J. L. *How to do things with words*. Oxford university press, 1975.

[7] BAHDANAU, D., CHO, K., AND BENGIO, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[8] BARSALOU, L. W. Ad hoc categories. *Memory & cognition 11*, 3 (1983), 211–227.

[9] BAWDEN, D., AND ROBINSON, L. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science 35*, 2 (2009), 180–191.

[10] BERLEANT, D., AND BERGHEL, H. Customizing information: Part 1, getting what we need, when we. *Computer 27*, 9 (1994), 96–98.

[11] BERLEANT, D., AND BERGHEL, H. Customizing information: Part 2, how successful are we so far? *Computer 27*, 10 (1994), 76–78.

[12] BEX, F. J., PRAKKEN, H., AND VERHEIJ, B. Formalising argumentative story-based analysis of evidence. In *Proceedings of the 11th international conference on artificial intelligence and law* (2007), pp. 1–10.

[13] BIERWISCH, M., AND SCHREUDER, R. From concepts to lexical items. *Cognition 42*, 1-3 (1992), 23–60.

[14] BIRD, S., KLEIN, E., AND LOPER, E. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[15] BOWMAN, S., POTTS, C., AND MANNING, C. D. Recursive neural networks can learn logical semantics. In *Proceedings of the 3rd workshop on continuous vector space models and their compositionality* (2015), pp. 12–21.

[16] BOWMAN, S. R., ANGELI, G., POTTS, C., AND MANNING, C. D. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326* (2015).

[17] BROWN, T. B., MANN, B., RYDER, N., SUBBIAH, M., KAPLAN, J., DHARIWAL, P., NEELAKANTAN, A., SHYAM, P., SASTRY, G., ASKELL, A., ET AL. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[18] BURKOV, A. *The Hundred-Page Machine Learning Book.* Andriy Burkov, 2019.

[19] CAMBRIA, E., AND WHITE, B. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine 9*, 2 (2014), 48–57.

[20] CER, D., DIAB, M., AGIRRE, E., LOPEZ-GAZPIO, I., AND SPECIA, L. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055* (2017).

[21] CER, D., YANG, Y., KONG, S.-Y., HUA, N., LIMTIACO, N., JOHN, R. S., CONSTANT, N., GUAJARDO-CÉSPEDES, M., YUAN, S., TAR, C., ET AL. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[22] CHEN, C.-M., LEE, H.-M., AND CHEN, Y.-H. Personalized e-learning system using item response theory. *Computers & Education 44*, 3 (2005), 237–255.

[23] CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H., AND BENGIO, Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[24] CHURCH, K., AND HANKS, P. Word association norms, mutual information, and lexicography. *Computational linguistics 16*, 1 (1990), 22–29.

[25] CONNEAU, A., KIELA, D., SCHWENK, H., BARRAULT, L., AND BORDES, A. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).

[26] COVINGTON, M. A. A fundamental algorithm for dependency parsing. In *Proceedings of the 39th annual ACM southeast conference* (2001), vol. 1, Citeseer.

[27] DAI, Z., YANG, Z., YANG, Y., CARBONELL, J., LE, Q. V., AND SALAKHUTDINOV, R. Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860* (2019).

[28] DANGETI, P. *Statistics for machine learning*. Packt Publishing Ltd, 2017.

[29] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[30] DEVORE, J. L. *Probability and Statistics for Engineering and the Sciences*, 8 ed. Cengage learning, 2011.

[31] DICTIONARY, O. E. Oxford english dictionary. *Simpson, JA & Weiner, ESC* (2021).

[32] DOWTY, D. R., WALL, R., AND PETERS, S. *Introduction to Montague semantics*, vol. 11. Springer Science & Business Media, 2012.

[33] EDMUNDS, A., AND MORRIS, A. The problem of information overload in business organisations: a review of the literature. *International journal of information management 20*, 1 (2000), 17–28.

[34] FODOR, J. A. *LOT 2: The language of thought revisited*. Oxford University Press on Demand, 2008.

[35] FRANCOIS, C. *Deep learning with Python*. Manning Publications Company, 2017.

[36] GAZDAR, G., AND MELLISH, C. Natural language processing in prolog: an introduction to computational linguistics, 1990.

[37] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A., AND BENGIO, Y. Generative adversarial nets. *Advances in neural information processing systems 27* (2014).

[38] HAMDI, M. S. Information overload and customization. *IEEE Potentials 25*, 5 (2006), 9–12.

[39] HOCHREITER, S., AND SCHMIDHUBER, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[40] HODOSH, M., YOUNG, P., AND HOCKENMAIER, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research 47* (2013), 853–899.

[41] HOFMANN, T. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), pp. 50–57.

[42] HOFMANN, T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning 42*, 1 (2001), 177–196.

[43] HON, G. The why and how of explanation: An analytical exposition. In *Explanation*. Springer, 2001, pp. 1–39.

[44] HOQ, K. M. G. Information overload: causes, consequences and remedies-a study. *Philosophy and Progress* (2014), 49–68.

[45] HSUEH-CHAO, M. H., AND NATION, P. Unknown vocabulary density and reading comprehension. *Reading in a foreign language 13*, 1 (2000), 403–30.

[46] HUANG, Z., XU, W., AND YU, K. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991* (2015).

[47] IYYER, M., MANJUNATHA, V., BOYD-GRABER, J., AND DAUMÉ III, H. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)* (2015), pp. 1681–1691.

[48] JOHNSON, R., WATKINSON, A., AND MABE, M. The stm report: An overview of scientific and scholarly publishing. *International Association of Scientific, Technical and Medical Publishers* (2018).

[49] JOHNSON-LAIRD, P. N. Mental models.

[50] JOOS, M. Description of language design. *The Journal of the Acoustical Society of America 22*, 6 (1950), 701–707.

[51] KAMATH, U., LIU, J., AND WHITAKER, J. *Deep learning for NLP and speech recognition*, vol. 84. Springer, 2019.

[52] KHAN, A., SALIM, N., FARMAN, H., KHAN, M., JAN, B., AHMAD, A., AHMED, I., AND PAUL, A. Abstractive text summarization based on improved semantic graph approach. *International Journal of Parallel Programming 46*, 5 (2018), 992–1016.

[53] KHAN, A., SALIM, N., REAFEE, W., SUKPRASERT, A., AND KUMAR, Y. J. A clustered semantic graph approach for multi-document abstractive summarization. *Jurnal Teknologi 77*, 18 (2015).

[54] KINTSCH, W. *The representation of meaning in memory (PLE: Memory)*. Psychology Press, 2014.

[55] LAN, Z., CHEN, M., GOODMAN, S., GIMPEL, K., SHARMA, P., AND SORICUT, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).

[56] LANDAUER, T. K. Lsa as a theory of meaning. In *Handbook of latent semantic analysis*. Psychology Press, 2007, pp. 15–46.

[57] LANDAUER, T. K., MCNAMARA, D. S., DENNIS, S., AND KINTSCH, W. Mathematical foundations behind latent semantic analysis. In *Handbook of latent semantic analysis*. Psychology Press, 2007, pp. 47–68.

[58] LE, H., VIAL, L., FREJ, J., SEGONNE, V., COAVOUX, M., LECOUTEUX, B., ALLAUZEN, A., CRABBÉ, B., BESACIER, L., AND SCHWAB, D. Flaubert: Unsupervised language model pre-training for french. *arXiv preprint arXiv:1912.05372* (2019).

[59] LIPTON, P. What good is an explanation? In *Explanation*. Springer, 2001, pp. 43–59.

[60] LIU, P. J., SALEH, M., POT, E., GOODRICH, B., SEPASSI, R., KAISER, L., AND SHAZEER, N. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198* (2018).

[61] LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[62] LOMBROZO, T. The structure and function of explanations. *Trends in cognitive sciences 10*, 10 (2006), 464–470.

[63] LUONG, M.-T., PHAM, H., AND MANNING, C. D. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[64] MACCARTNEY, B., AND MANNING, C. D. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* (2008), pp. 521–528.

[65] MANNING, C., AND SCHUTZE, H. *Foundations of statistical natural language processing*. MIT press, 1999.

[66] MARCUS, M., SANTORINI, B., AND MARCINKIEWICZ, M. A. Building a large annotated corpus of english: The penn treebank.

[67] MARTIN, L., MULLER, B., SUÁREZ, P. J. O., DUPONT, Y., ROMARY, L., DE LA CLERGERIE, É. V., SEDDAH, D., AND SAGOT, B. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894* (2019).

[68] MEYER, J.-A. Information overload in marketing management. *Marketing Intelligence & Planning 16*, 3 (1998), 200–209.

[69] MIKOLOV, T., CHEN, K., CORRADO, G., AND DEAN, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[70] MILLER, G. A. Wordnet: a lexical database for english. *Communications of the ACM 38*, 11 (1995), 39–41.

[71] MOHLER, M., AND MIHALCEA, R. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)* (2009), pp. 567–575.

[72] MONTAGUE, R., AND THOMASON, R. H. Formal philosophy. selected papers of richard montague. *Erkenntnis 9*, 2 (1975).

[73] MURPHY, G. L., AND ALLOPENNA, P. D. The locus of knowledge effects in concept learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition 20*, 4 (1994), 904.

[74] MURPHY, G. L., AND MEDIN, D. L. The role of theories in conceptual coherence. *Psychological review 92*, 3 (1985), 289.

[75] MURPHY, M. L. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press, 2003.

[76] NICKEL, M., AND KIELA, D. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems 30* (2017), 6338–6347.

[77] PAGE, L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab, 1999.

[78] PATALANO, A. L., CHIN-PARKER, S., AND ROSS, B. H. The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of memory and language 54*, 3 (2006), 407–424.

[79] PAWAR, A., AND MAGO, V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv preprint arXiv:1802.05667* (2018).

[80] PENNINGTON, J., SOCHER, R., AND MANNING, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014), pp. 1532–1543.

[81] PRESS, C. U. Cambridge online dictionary. https://dictionary.cambridge.org, 2008.

[82] RADFORD, A., NARASIMHAN, K., SALIMANS, T., AND SUTSKEVER, I. Improving language understanding by generative pre-training.

[83] RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., SUTSKEVER, I., ET AL. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8 (2019), 9.

[84] RICH, E., KNIGHT, K., GONZÁLEZ CALERO, P. A., BODEGA, T., ET AL. *Inteligencia artificial*, 2 ed. 1994.

[85] RUBENSTEIN, H., AND GOODENOUGH, J. B. Contextual correlates of synonymy. *Communications of the ACM 8*, 10 (1965), 627–633.

[86] RUMELHART, D. E., LINDSAY, P. H., AND NORMAN, D. A. A process model for long-term memory.

[87] SALA, F., DE SA, C., GU, A., AND RÉ, C. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning* (2018), PMLR, pp. 4460–4469.

[88] SANH, V., DEBUT, L., CHAUMOND, J., AND WOLF, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[89] SIZE, W. W. W. The size of the world wide web (the internet). `https://www.worldwidewebsize.com`.

[90] SONG, S., HUANG, H., AND RUAN, T. Abstractive text summarization using lstm-cnn based deep learning. *Multimedia Tools and Applications 78*, 1 (2019), 857–875.

[91] STROTHER, J. B., ULIJN, J. M., AND FAZAL, Z. *Information overload: an international challenge for professional engineers and technical communicators*, vol. 2. John Wiley & Sons, 2012.

[92] SULEMAN, R. M., AND KORKONTZELOS, I. Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications 165* (2021), 114130.

[93] SUTSKEVER, I., MARTENS, J., AND HINTON, G. E. Generating text with recurrent neural networks. In *ICML* (2011).

[94] TAI, K. S., SOCHER, R., AND MANNING, C. D. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075* (2015).

[95] TIFREA, A., BÉCIGNEUL, G., AND GANEA, O.-E. Poincar\'e glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546* (2018).

[96] TOFFLER, A. *Future shock*, vol. 553. Bantam, 1984.

[97] TURC, I., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2* (2019).

[98] TURING, A. M. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, 2009, pp. 23–65.

[99] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, Ł., AND POLOSUKHIN, I. Attention is all you need. In *Advances in neural information processing systems* (2017), pp. 5998–6008.

[100] WANG, R., AND NEUMANN, G. Recognizing textual entailment using sentence similarity based on dependency tree skeletons. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* (2007), pp. 36–41.

[101] WEST, M. P. *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans, Green, 1953.

[102] WIEMER-HASTINGS, P., AND ZIPITRIA, I. Rules for syntax, vectors for semantics. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (2001), vol. 23.

[103] WINSTON, M. E., CHAFFIN, R., AND HERRMANN, D. A taxonomy of part-whole relations. *Cognitive science 11*, 4 (1987), 417–444.

[104] WU, H., HUANG, H.-Y., JIAN, P., GUO, Y., AND SU, C. Bit at semeval-2017 task 1: Using semantic information space to evaluate semantic textual similarity. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (2017), pp. 77–84.

[105] WU, Y., SCHUSTER, M., CHEN, Z., LE, Q. V., NOROUZI, M., MACHEREY, W., KRIKUN, M., CAO, Y., GAO, Q., MACHEREY, K., ET AL. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[106] XIE, S., RASTOGI, R., AND CHANG, M. Deep poetry: Word-level and character-level language models for shakespearean sonnet generation. In *Natural Lang. Process. Deep Learn.* 2017.

[107] YANG, Y., YUAN, S., CER, D., KONG, S.-Y., CONSTANT, N., PILAR, P., GE, H., SUNG, Y.-H., STROPE, B., AND KURZWEIL, R. Learning semantic textual similarity from conversations. *arXiv preprint arXiv:1804.07754* (2018).

[108] YANG, Z., DAI, Z., YANG, Y., CARBONELL, J., SALAKHUTDINOV, R., AND LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237* (2019).

[109] YOUNG, P., LAI, A., HODOSH, M., AND HOCKENMAIER, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics 2* (2014), 67–78.

[110] ZHAI, C., AND MASSUNG, S. *Text data management and analysis: a practical introduction to information retrieval and text mining*. Morgan & Claypool, 2016.

[111] ZHU, Y., KIROS, R., ZEMEL, R., SALAKHUTDINOV, R., URTASUN, R., TORRALBA, A., AND FIDLER, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision* (2015), pp. 19–27.

Curriculum Vitae

Germán Eduardo Baltazar Reyes was born in Oaxaca, México, on January 29th, 1994. He earned the Engineering on Telecommunications and Electronic Systems degree from the Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico City Campus in June 2017. During his undergraduate studies, he collaborated with the campus robotics team Nao Team CCM in multiple national contests and as the team's captain in 2017. The team achieved first place in the contest's medal table for two consecutive years during his participation.

He was accepted into the Ph.D. in Engineering Science program of the same institution in August 2017. He collaborated in multiple international congresses and events where he showed the results of his research and use of technology for educational purposes.

His research interests are focused on the use of Natural Language Processing and Social Robotics for educational purposes. While doing his graduate studies, he served as a teacher on Digital Systems in the Instituto Tecnológico y de Estudios Superiores de Monterrey, Mexico City Campus since 2020.