

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS
SUPERIORES DE MONTERREY
CAMPUS MONTERREY
SCHOOL OF ENGINEERING AND INFORMATION
TECHNOLOGIES GRADUATE PROGRAMS



DOCTOR OF PHILOSOPHY
in
INFORMATION TECHNOLOGIES AND
COMMUNICATIONS MAJOR IN INTELLIGENT SYSTEMS

*The Impact of Statistical Word Alignment Quality and
Structure in Phrase Based Statistical Machine Translation*

by

Francisco Javier Guzmán Herrera

DECEMBER 2011

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY
CAMPUS MONTERREY

SCHOOL OF ENGINEERING AND INFORMATION TECHNOLOGIES
GRADUATE PROGRAMS



DOCTOR OF PHILOSOPHY

in

INFORMATION TECHNOLOGIES AND COMMUNICATIONS
MAJOR IN INTELLIGENT SYSTEMS

**The Impact of Statistical Word Alignment Quality and Structure in
Phrase Based Statistical Machine Translation**

By

Francisco Javier Guzmán Herrera

DECEMBER 2011

INSTITUTO TECNOLÓGICO Y DE ESTUDIOS SUPERIORES DE MONTERREY
CAMPUS MONTERREY

SCHOOL OF ENGINEERING AND INFORMATION TECHNOLOGIES
GRADUATE PROGRAMS



DOCTOR OF PHILOSOPHY

in

INFORMATION TECHNOLOGIES AND COMMUNICATIONS
MAJOR IN INTELLIGENT SYSTEMS

**The Impact of Statistical Word Alignment Quality and Structure in
Phrase Based Statistical Machine Translation**

By

Francisco Javier Guzmán Herrera

DECEMBER 2011

The Impact of Statistical Word Alignment Quality and Structure in Phrase Based Statistical Machine Translation

A dissertation presented by

Francisco Javier Guzmán Herrera

Submitted to the
Graduate Programs in Information Technologies
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Information Technologies and Communications
Major in Intelligent Systems



Thesis Committee:

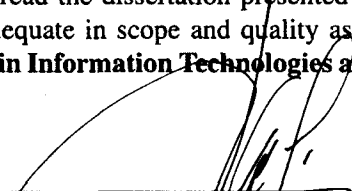
Dr. Leonardo Garrido Luna	-	Tecnológico de Monterrey
Dr. Stephan Vogel	-	Carnegie Mellon University - Qatar Computing Research Institute
Dr. Juan Arturo Nolasco Flores	-	Tecnológico de Monterrey
Dr. Ramón Brena Pinero	-	Tecnológico de Monterrey
Dr. Eduardo Uresti Charre	-	Tecnológico de Monterrey

Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey
December 2011

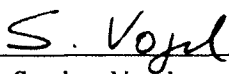
Instituto Tecnológico y de Estudios Superiores de Monterrey
Campus Monterrey

School of Engineering and Information Technologies
Graduate Program

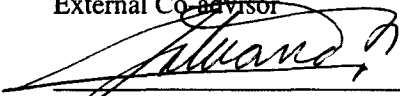
The committee members, hereby, certify that have read the dissertation presented by Francisco Javier Guzmán Herrera and that it is fully adequate in scope and quality as a partial requirement for the degree of **Doctor of Philosophy in Information Technologies and Communications**, with a major in **Intelligent Systems**.



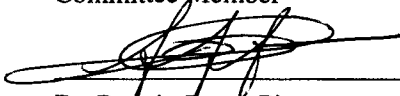
Dr. Leonardo Garrido Luna
Tecnológico de Monterrey
Institutional Co-advisor



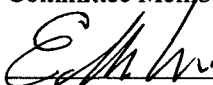
Dr. Stephan Vogel
Carnegie Mellon University
Qatar Computing Research Institute
External Co-advisor




Dr. Juan Arturo Nolasco Flores
Tecnológico de Monterrey
Committee Member



Dr. Ramón Brena Pinero
Tecnológico de Monterrey
Committee Member



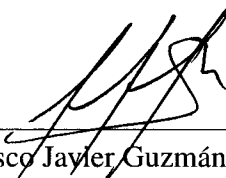
Dr. Eduardo Uresti Charre
Tecnológico de Monterrey
Committee Member



Dr. José Luis Gordillo Moscoso
Director of Doctoral Program
Information Technologies and Communications
School of Engineering and Information Technologies

Copyright Declaration

I, hereby, declare that I wrote this dissertation entirely by myself and, that, it exclusively describes my own research.



Francisco Javier Guzmán Herrera
Monterrey, N.L., México
December 2011

©2011 by Francisco Javier Guzmán Herrera
All Rights Reserved

Dedication

A mi esposa Gloria.
Porque nunca dejaste de creer en mí.
Por apoyarme y acompañarme en este viaje desde el inicio.

A mis padres.
Por impulsarme a trazar metas más altas.
Porque siempre estuvieron ahí cuando los necesité.

A mis hermanos.
Por su cariño incondicional.

Acknowledgements

I would like to express my deepest gratitude to all those who have been part of this long journey.

- To Dr. Leonardo Garrido,
who invited me to join the program, and always looked after me.
- To Dr. Stephan Vogel,
who accepted me into his team, shared his knowledge and guidance.
- To Dr. Juan Nolzco,
who welcomed me into his team and granted me the opportunity to finish this work.
- To my Thesis Committee,
for their time spent reading my work and their insightful reviews.
- To Dr. Hugo Terashima,
who always looked after my well-being as a PhD student.
- To Dr. Alon Lavie,
for inviting me to a research stay at CMU and coordinating the whole process.
- To Qin Gao,
who was always eager to help when I most needed it.

I would also like to acknowledge all the people who made this journey more enjoyable. To the CMU-ISL translation team, specially to Silja Hildebrand, Sanjika Hewavitharana, Alok Parlikar, Thuylinh Nguyen and Narges Razavian for sharing ideas, knowledge and friendship. To the CMU-LTI translation team, specially to Jonathan Clark whose acquaintance in Aguascalientes changed direction my research path. To the CMU-crew, for making my stay in Pittsburgh as smooth as possible. Specially, thanks to Linda Hager, Lisa Mauti, Isaac Harris and Kelley Widmaier.

To the ITESM-TEChila speech team, for their support and patience to teach me how speaker verification works. To my ITESM-Robocup labmates, for the time spent together and the friendship. To the CS-ITESM staff, for always helping me with all the never-ending paperwork. Special thanks to Dora Elia Juárez, Isabel Cerda and Sandra Domínguez.

The Impact of Statistical Word Alignment Quality and Structure in Phrase Based Statistical Machine Translation

by

Francisco Javier Guzmán Herrera

Abstract

Statistical Word Alignments represent lexical word-to-word translations between source and target language sentences. They are considered the starting point for many state of the art Statistical Machine Translation (SMT) systems. In phrase-based systems, word alignments are loosely linked to the translation model. Despite the improvements reached in word alignment quality, there has been a modest improvement in the end-to-end translation. Until recently, little or no attention was paid to the structural characteristics of word-alignments (e.g. unaligned words) and their impact in further stages of the phrase-based SMT pipeline. A better understanding of the relationship between word alignment and the entailing processes will help to identify the variables across the pipeline that most influence translation performance and can be controlled by modifying word alignment's characteristics.

In this dissertation, we perform an in-depth study of the impact of word alignments at different stages of the phrase-based statistical machine translation pipeline, namely word alignment, phrase extraction, phrase scoring and decoding. Moreover, we establish a multivariate prediction model for different variables of word alignments, phrase tables and translation hypotheses. Based on those models, we identify the most important alignment variables and propose two alternatives to provide more control over alignment structure and thus improve SMT. Our results show that using alignment structure into decoding, via alignment gap features yields significant improvements, specially in situations where translation data is limited.

During the development of this dissertation we discovered how different characteristics of the alignment impact Machine Translation. We observed that while good quality alignments yield good phrase-pairs, the consolidation of a translation model is dependent on the alignment structure, not quality. Human-alignments are more dense than the computer generated counterparts, which trend to be more sparse and precision-oriented. Trying to emulate human-like alignment structure resulted in poorer systems, because the resulting translation models trend to be more compact and lack translation options. On the other hand, more translation options, even if they are noisier, help to improve the quality of the translation. This is due to the fact that translation does not rely only on the translation model, but also other factors that help to discriminate the noise from bad translations (e.g. the language model). Lastly, when we provide the decoder with features that help it to make "more informed decisions" we observe a clear improvement in translation quality. This was specially true for the discriminative alignments which inherently leave more unaligned words. The result is more evident in low-resource settings where having larger translation lexicons represent more translation options. Using simple features to help the decoder discriminate translation hypotheses, clearly showed consistent improvements.

List of Abbreviations

AC	Acquis Communautaire set
AER	Alignment Error Rate
BA	Balanced Accuracy
BLEU	Bi Lingual Evaluation Understudy
BP	BLEU points
CPER	Consistent Phrase Error Rate Metric
CRF	Conditional Random Fields
DEV	Development (cross-validation) document set
DWA	Discriminative Word Alignment
EM	Expectation Maximization algorithm
EMNLP	Conference on Empirical Methods in Natural Language Processing
EPPS	European Parliament Plenary Sessions
GALE	Global Autonomous Language Exploitation
HMM	Hidden Markov Models
IBM	International Business Machines
LDC	Linguistic Data Consortium
LM	Language Model
MERT	Minimum Error Rate Training
METEOR	Automatic Machine Translation Evaluation System
MIRA	Margin Infused Relaxed Algorithm
MLE	Maximum Likelihood Estimation
MT	Machine Translation
NC	News Commentary document set
NIST	National Institute of Standards and Technology
NW	News Wire document set
PCA	Principal Components Analysis
PT	Phrase table
SEM	Structural Equation Modeling
SMT	Statistical Machine Translation
TER	Translation Error Rate
TM	Translation Model
TST	Test document set
UN	United Nations
WA	Word Alignment
WMT	Workshop on Statistical Machine Translation

List of Figures

2.1	Components of SMT	13
2.2	Phrase-based Statistical Machine Translation	15
2.3	Example of a phrase-based translation	16
2.4	Training pipeline for phrase-based SMT	17
2.5	A preprocessing example	18
2.6	A simple example of a word alignment	19
2.7	Heuristic symmetrization	21
2.8	Phrase consistency	24
2.9	Step-by-step phrase-extraction	25
2.10	Decoding overview	26
2.11	Translation search space	28
2.12	Hypothesis recombination	29
3.1	Alignment quality metrics for different alignments	37
3.2	Statistics of unaligned words	40
3.3	Distribution of length of the source phrases extracted from different alignments.	42
3.4	Number of extracted phrase pairs vs. unique phrase pairs per system	43
3.5	Phrase pair quality evaluation results	46
4.1	Snippet from a phrase table	50
4.2	Alignment to phrase-tables experimental outline	54
4.3	Correlation map for different variables	55
4.4	Number of phrases and monotonicity	59
4.5	Phrase-length and monotonicity	60
5.1	Example of a first-best hypothesis	66
5.2	General model for BLEU	71
5.3	General model for METEOR	73
5.4	General model for TER	74
5.5	Variables densities	76
5.6	Model of BLEU for easy set	77
5.7	Model of METEOR for easy set	78
5.8	Model of TER for easy set	79
5.9	Model of BLEU for medium-hard set	80
5.10	Model of METEOR for medium-hard set	82
5.11	Model of TER for medium-hard set	83

6.1	Different representations of alignment gaps	87
6.2	Bleu gains by test-set	94
6.3	BLEU gains by system	95
6.4	Average BLEU gains by system vs. the normalized weight change for the $p(f e)$ feature $\Delta\hat{w}_{p(f e)}$	96
A.1	Example of preprocessing	107

List of Tables

2.1	Summary of generative alignment models	20
2.2	An example of translation model estimation	24
2.3	Most frequent decoding features in modern systems	27
2.4	Example of a n-best list	30
3.1	Data Statistics of the alignment training set	35
3.2	Alignment quality metrics for different alignments	36
3.3	Link density statistics	38
3.4	Alignment gaps statistics	39
3.5	Phrase length statistics	41
3.6	Phrase gaps statistics	44
3.7	ANCOVA table for phrase-quality experiment	45
4.1	Translation probabilities in translation model	50
4.2	Alignment and phrase-table variables considered in regression analysis	52
4.3	Hand alignment datasets	53
4.4	Regression results for phrase-table entries	58
4.5	Regression coefficients for phrase-table phrase length	59
4.6	Regression coefficients for translation model entropies	61
5.1	Variables used in the translation quality prediction study	67
5.2	Effects of several variables in translation quality	75
5.3	Classification of datasets	76
5.4	Summary of effects for the easy document set	79
5.5	Summary of effects for the hard document set	84
5.6	General summary of the regression model with the translation model and hypotheses effects.	84
6.1	Alignment quality metrics for each of the different alignments	88
6.2	Alignment structure metrics for each alignment	89
6.3	Translation results using four different models	90
6.4	MERT-tuned weights for the FTG feature by system	92
6.5	Best scores per document set obtained by system/feature combination	93
6.6	Translation results for the baseline systems (base) and the gap-feature enhanced systems (gaps) built upon different alignments.	97
A.1	Statistics for Raw and preprocessed data	108

A.2	Datasets used for translation	109
A.3	Data Statistics for Spanish-English hand alignments	110

Contents

Abstract	ix
List of Abbreviations	xi
List of Figures	xiv
List of Tables	xvi
1 Introduction	1
1.1 General Problem Statement	2
1.1.1 Alignment Quality vs. Translation Quality	3
1.1.2 Alignment Quality and the SMT Pipeline	4
1.1.3 Impact of Alignment Quality and Alignment Structure on Translation Quality	5
1.2 Hypothesis and Research Questions	5
1.3 Contributions	6
1.4 Methodology	7
1.4.1 Exploratory Analysis	7
1.4.2 Predicting the Translation Model	7
1.4.3 Predicting Translation Quality	7
1.4.4 Improving Translation Using Alignment Structure	8
1.5 Limitations	8
1.6 Thesis Organization	9
2 Background	11
2.1 Introduction to Statistical Machine Translation	11
2.2 Principles of Statistical Machine Translation	12
2.2.1 Noisy Channel Model	12
2.3 Language Model Estimation	13
2.3.1 N-grams	13
2.3.2 Smoothing Techniques	14
2.3.3 Estimation	14
2.4 Phrase Based Translation Model	14
2.4.1 Phrase Based Motivation	15
2.4.2 PBSMT Principles	15
2.4.3 Estimating the Model	16

2.5	PBSMT Training Pipeline	17
2.5.1	Preprocessing	18
2.5.2	Word Alignment	18
2.5.3	Phrase Extraction	23
2.5.4	Phrase Scoring	23
2.6	Decoding	24
2.6.1	Log-Linear Models	25
2.6.2	The Process of Decoding	27
2.6.3	Translation Evaluation	29
3	The Effects of Word Alignments on Phrase-Extraction	33
3.1	Alignments and Phrase-Pairs	34
3.2	Experimental Setup	34
3.3	Characterization of Word Alignments	35
3.3.1	Analysis of Alignment Quality	36
3.3.2	Analysis of Alignment Structure	36
3.3.3	Unaligned Words	38
3.3.4	Summary	38
3.4	Analysis of Phrase-Pairs	39
3.4.1	Analysis of Number and Length of Phrase-Pairs	40
3.4.2	Analysis of Phrase Alignment Gaps	42
3.4.3	Summary	43
3.5	The effect of Alignment Gaps in Phrase-Pair Quality	44
3.6	Conclusions	46
4	The Effects of Alignments on the Phrase-Based Translation Model	49
4.1	The Phrase Translation Model	49
4.1.1	Translation and Lexical Probabilities	50
4.2	Experimental Setup	51
4.2.1	Data and Sampling	51
4.3	Correlation Analysis	53
4.4	Regression Models	57
4.4.1	Entries in the Phrase-Table	57
4.4.2	Length of Phrases	59
4.4.3	Translation Entropies	60
4.5	Conclusions	62
5	Predicting Translation Quality	65
5.1	Methodology	65
5.1.1	Key Terms	66
5.1.2	Measuring Variables	66
5.1.3	Finding a Regression Model	67
5.2	Experimental Setup	68
5.2.1	Data and Sampling	68
5.2.2	Data Measurement and Consolidation	70

5.3	Experimental Results	70
5.3.1	General Model	70
5.3.2	Refined Models	75
5.3.3	Refined Models: Easy Set	75
5.3.4	Refined Models: Hard	80
5.3.5	Summary	81
5.4	Conclusions	82
6	Improving Machine Translation with Alignment Structure	85
6.1	Improving Word Alignment for Phrase-Extraction	85
6.1.1	Alignment Results	88
6.1.2	Translation Experiments	90
6.1.3	Conclusions	90
6.2	Improving Translation Using Alignment Gap Features	91
6.2.1	Tuning Weights for Quality	91
6.2.2	Translation Results	92
6.3	Translation Gaps for Limited Scenarios	94
6.3.1	Setup	95
6.4	Conclusions	96
7	Conclusions	99
7.1	Summary of our Findings	99
7.1.1	The Effects of Alignments on the Phrase-Based Translation Model	99
7.1.2	The Characteristics of the Translation Model	100
7.1.3	Predicting Translation Quality	100
7.1.4	Improving Translation Using Alignment Structure	101
7.2	Hypothesis and Research Questions Revisited	102
7.3	Final Remarks	104
7.4	Future Work	104
A	Baseline System	107
A.1	Translation Training Data	107
A.1.1	Data Sources	107
A.1.2	Preprocessing	107
A.2	Translation Test Data	108
A.3	Hand Aligned Data	109
B	Variables	111
B.1	Alignment Variables	111
B.1.1	Alignment Dimension	111
B.1.2	Alignment Density	111
B.1.3	Alignment Distortion	112
B.1.4	Alignment Quality	113
B.2	Phrase-table Variables	113
B.2.1	Alignment Variables	113

B.2.2	Phrase-table Dimension	114
B.2.3	Phrase-table Entries	114
B.2.4	Coverage Variables	114
B.2.5	Translation Entropies	115
B.3	Translation (First-best) Variables	115
B.3.1	Alignment Variables	115
B.3.2	Cost Variables	115
B.3.3	Translation Quality	116
	Bibliography	121

Chapter 1

Introduction

In the second half of the twentieth century, as people looked for computer assistance in several intellectual tasks, human attention turned to the automation of natural language processing. Translating documents between any two languages by a computer was one of the first and most relevant goals of this field. The aim was to reduce the expensive and time consuming human translations. Moreover, the dramatic increase in the quantity of produced content, which has been witnessed in the last couple of decades, has set the role of machine translation (MT) as a crucial tool for equal information access.

While no current machine translation system is capable of successfully imitating the behavior of a human translator, various types of existing systems do help to reduce language barriers. The recent availability of vast amount of human-generated translations and powerful computers, has enabled the development of new research directions in Machine Translation. This is specially the case for statistical methods (Brown et al., 1990), which allow the analysis of parallel text corpora and the automatic construction of machine translation systems. Statistical Machine Translation (SMT) has proven to be very effective, this is a reason why it has become the predominant paradigm in the research community.

Although the earlier word-based generative Statistical Machine Translation systems (Brown et al., 1993) are no longer used, their by-products (word alignments, lexica, fertility tables) constitute the base for many of the newer models. This is specially the case for word alignments, which represent the correspondences between the source and target words of a sentence pair and are used indirectly for estimating the translation model in phrase-based (Och, 2003; Koehn et al., 2003; Marcu and Wong, 2002) and hierarchical (Chiang, 2005) systems.

Statistical Word Alignment is the task of finding the correspondences between words in a source language sentence and the words in a target language sentence. The alignment A of this pair is simply a set of these correspondences.

In recent years, the increasing availability of human generated word alignments, has made possible the assessment of alignment quality, using metrics such as Alignment Error Rate (AER), F-score (Och and Ney, 2003), etc. As a result, there have been several efforts to improve the performance of word alignments. Moreover, those annotated alignments have also enabled the development of discriminative word alignment models (Ittycheriah and Roukos, 2005; Blunsom and Cohn, 2006; Niehues and Vogel, 2008; Lambert et al., 2009) which are tuned towards these quality metrics.

Despite the improvements reached in alignment quality, there has been a modest improvement in the performance of end-to-end translation. Some authors (Fraser and Marcu, 2007) have found that AER and translation quality (as measured by the BLEU metric (Papineni et al., 2002)) do not correlate well. Moreover, it has been observed that under certain situations, lower alignment quality can lead to improvements in translation quality (Vilar and Ney, 2006). However, little attention has been given to the fact that in phrase-based systems, word alignments undergo a series of steps (e.g. heuristic-based phrase extraction) which result in a loose link between the word-alignment and the phrase-based translation models.

For phrase-based SMT, there have been few efforts to understand how alignment quality affects the translation models (Ayan and Dorr, 2006; Lopez and Resnik, 2006). Until recently (Guzman et al., 2009; Lambert et al., 2009), little or no attention was paid to the structural characteristics of word alignments (e.g. unaligned words, number of links, etc.) and their impact in further stages of the phrase-based SMT pipeline. As a result, the role of word alignment on phrase-based systems has been only partially understood.

A better understanding of the relationship between word alignment and the entailing processes will help to identify the variables across the pipeline that most influence translation performance and can be controlled by modifying word alignment's characteristics. Additionally, it will allow to develop better alignment assessment metrics, enabling for word alignment developments to carry through the pipeline. By doing so, it will help to close the gap between word alignment and translation performance. In summary, the better the understanding of the processes, the better the resulting translation quality.

In this document, we advocate for the analysis of alignment structure as one important contributor to Machine Translation performance. To that end, we perform an in-depth deep study of the impact of the characteristics of statistical word alignments at different stages of the phrase-based Statistical Machine Translation pipeline, namely word alignment, phrase extraction, phrase scoring and decoding. Moreover, we estimate multivariate prediction models for different variables phrase tables and translation hypotheses using alignment structure and other alignment variables as predictors. These models describe how variations in characteristics of word alignments will affect a phrase-based translation model and how those variations in the model will influence translation. Finally, we use the knowledge obtained from those models to propose alternatives for better alignment training, decoding and ultimately translation quality.

The remainder of this chapter is structured as follows: In Section 1.1, we provide a closer look to the problem of discrepancies between word alignment quality and translation quality that have been observed in previous studies. In Section 1.2, we present our initial hypothesis and a set of research questions. Next, in Section 1.2 we present a summary of the contributions of this dissertation, followed by the proposed methodology in 1.4. Later in Section 1.5 we state the limitations of this work. Finally in Section 1.6 we provide the organization of the body of this dissertation.

1.1 General Problem Statement

Statistical word alignments serve as a starting point for the Statistical Machine Translation (SMT) pipeline. Improving their quality has been a major focus of research in the SMT

community. One of the first attempts to bring attention to this matter was (Och and Ney, 2003) where they evaluated several word alignments produced by different aligners. To that end, they proposed metric AER which remained as the standard for reporting improvements in alignments. What followed then, was a series of developments from several teams aiming for better alignment quality. In this section, we introduce different works that have analyzed statistical word alignment quality and its impact in Machine Translation. First, we present the work that focused in describing Machine Translation quality in terms of alignment quality. Then, we present studies aiming to explain the effects of alignment quality in the phrase-based MT pipeline. Finally, we present some recent studies which aimed to describe the effects of alignment structure in Machine Translation.

1.1.1 Alignment Quality vs. Translation Quality

The increasing availability of human annotated data has made possible to develop supervised or discriminative algorithms that maximize the alignment quality as prescribed by AER. Many of such methods emerged and continue to emerge (Ittycheriah and Roukos, 2005; Blunsom and Cohn, 2006; Niehues and Vogel, 2008; Lambert et al., 2009) achieving good results in alignment quality. Despite the ever increasing alignment quality achieved by several systems, end-to-end translation quality improvements derived from such increasing quality remained small in the best of cases. This was first pointed out by Fraser and Marcu (2007), who observed that SMT translation quality as measured by BLEU metric (Papineni et al., 2002), did not correlate well with AER. In their study, they performed several regression tests and concluded that there were several flaws with the AER (e.g. the inclusion of possible links), that prevented achieving a good correlation between AER and BLEU. As a consequence, they encouraged the use of variations of the F-measure (instead of AER) that could be tuned to favor precision or recall to achieve a better correlation. Since their results were sensitive to the different language pairs they studied (i.e. French-English, Arabic-English and Romanian-English), they proposed to tune the modified F metric accordingly to obtain better correlation with BLEU. Although this study was indeed revealing, the proposed solution to the underlying problem is too cumbersome to implement in day-to-day developments. It also falls short to take into account other structural characteristics of the alignment (which could explain the variation across the language pairs). One of the recommendations they made, which was quickly adopted across the community, was to report a translation quality assessment in addition to word alignment quality.

Similarly, other studies revealed mismatches between BLEU and AER. Vilar and Ney (2006) had observed that improvements in translation quality could be achieved by degrading alignment quality. For instance, alignments that were modified by hand to take into account specific linguistic phenomena of German, scored lower AER but helped to improve translation quality in the long run. Although some of the conclusions from this study might seem flawed given its premise (i.e. lower AER gives better BLEU) there are some interesting remarks that are worth mentioning. First, they back up the use of AER as a alignment quality metric. Second, they point out that the mismatch between alignment quality and translation quality is due to the mismatch between alignment and translation models. This is one of the first acknowledgements that there might be something in between alignment and translation that could be accountable for the disparity.

The main criticism that can be raised to these previous studies is that they regard the SMT pipeline as a "black-box" where there is one input (i.e. alignment quality) and one output (i.e. translation quality). That assumption ignores the series of processes that occur after a word alignment is obtained, which oversimplifies the current situation. For example, in phrase-based translation, those entailing processes are model estimation and decoding.

1.1.2 Alignment Quality and the SMT Pipeline

The first studies to analyze in detail the impact of alignment quality in the model estimation and decoding phases are (Ayan and Dorr, 2006; Lopez and Resnik, 2006).

Ayan and Dorr (2006) presented an in-depth analysis of the quality of the alignments as well its effect in resulting phrase tables. Their analysis compared several types of alignments, their quality, and the translations from the resulting phrase tables in different scenarios. They also took into account the indirect effects of alignment structure through lexical weightings. They also performed an extensive analysis on the length of phrases used by the decoder and the phrase-table coverage. In their study, they shed light on the behavior of different types of alignments. For example, they realized that recall oriented alignments (which tend to have more links) yield smaller phrase tables. Similarly, precision oriented alignments, which are sparser, yield larger phrase tables. Furthermore, they analyzed different phrase extraction configurations based on the structure of phrase pairs (i.e. tight vs. loose phrases). In their view, loose phrases refer to phrase-pairs whose either source or target boundary words (either beginning or end of phrase) are unaligned. Tight phrases are the opposite. In addition, they proposed the Consistent Phrase Error Rate Metric (CPER) which is similar to AER but operates at the phrase level. CPER compares the phrase table extracted from an alignment to the one generated by a hand alignment. However, the underlying assumption, that the extracted phrases from the hand aligned data using the current phrase extraction algorithms is perfect, could be challenged.

While being one of the most complete studies done to analyze the SMT pipeline, they do not directly analyze structural characteristics of the alignments that impact the configuration of a phrase-table such as the number of unaligned words or number of links. Instead, these alignment variables get obfuscated inside of quality measurements such as precision or recall. Therefore many of their conclusions could be challenged (e.g. more links in the alignment equals fewer extracted phrase pairs in the phrase table). However, they are the first to suggest that the characteristics in the alignment play a crucial role in the upcoming stages of the SMT pipeline.

In a different study, Lopez and Resnik (2006) analyze variations in the translation search space of the decoder by having alignments of gradually degraded quality. Their alignments are all obtained using the same aligner (GIZA++) and training data. However, they achieve variations in quality (precision, recall, F-score) by segmenting the data into chunks of varying sizes to obtain noisy alignments. While their study proposes an interesting methodology for evaluating the impact of alignments in the configuration of a decoder's search space, features and weightings; their conclusions are limited by the initial setup of the experiment. In that sense, what they end up measuring is not the effect of quality (F-score, precision or recall) into the decoder variables but rather the effect of training sizes in the latter (alignment quality being a result of that).

1.1.3 Impact of Alignment Quality and Alignment Structure on Translation Quality

Later studies (Guzman et al., 2009; Lambert et al., 2009) have brought closer attention to the structure of word alignment and their implications in following stages of the SMT pipeline. For instance Lambert et al. (2009), analyze the effect of the number of links of different types of alignments including its repercussions on the size of phrase tables and the ambiguity of the translation model. They also propose new structural metrics for alignments such as link length, distortion and crossings.

Likewise, the work we presented in (Guzman et al., 2009) analyzes several alignment characteristics and their impact on the extracted phrase pairs. In this study we observed that alignment structure has a large impact on the phrase-translation model. For instance, we discovered that the number of unaligned words in an alignment has an important effect in the size and configuration of the phrase-table. Furthermore, we showed that the distribution of unaligned words inside a phrase pair (also known as gaps) and the distribution of unaligned words in the alignment are highly correlated and is not affected by the heuristic extraction. Additionally, by performing a manual evaluation of Chinese-English phrase pairs, we made two interesting observations: First, better quality alignments (i.e. human annotated data) yield better human perceived quality phrase pairs. Secondly, phrase pairs with a higher proportion of unaligned words show lower human perceived quality. Also, by including a set of features to the decoder that take into account the number of gaps inside a phrase-pair, we were able to obtain significant improvement in translation quality.

In this dissertation we extend those studies and focus on describing how alignment structure impacts Machine Translation. For instance, we use a myriad of measures for alignment structure that represent alignment sparsity (Guzman et al., 2009; Lambert et al., 2009) and distortion (Lambert et al., 2010). We also analyze variables that comprehend the coverage and model ambiguity of a translation model (Ayan and Dorr, 2006) and other characteristics. Finally, based on our observations, we propose new alignment metrics and design new decoding features as suggested by Lopez and Resnik (2006) based on the alignment structure.

1.2 Hypothesis and Research Questions

The main hypothesis of this study can be summarized in the following way:

Alignment structure has a large impact on the characteristics of the resulting translation model. Hence, it should also have a large impact on Machine Translation performance. Thus, by controlling the impact of alignment structure we will be able to improve Machine Translation performance.

By dissecting this hypothesis, we can identify three parts:

1. The impact of alignment structure on the translation model

The first part of our hypothesis states that alignment structure determines greatly the characteristics of the resulting translation model. In other words, we hold that translation models are dependent upon the alignment structure.

2. The impact of alignment structure of the translation model on MT performance

The second part of our hypothesis states that alignment structure differences yield different translation models that result in differences in translation quality.

3. Providing means to control alignment structure will result in improvements in MT performance

The last part states that by controlling alignment structure, we will be able to improve machine translation quality.

Following our hypothesis, there are several questions that arise and need to be answered as we progress in our study.

- Which variables describing word alignments, translation models and translation hypotheses that we are going to include in this study?
- Which are the structural variables that are most important to our translation model? Which ones are more important to translation quality? Do the importance of the variables describing translation quality vary depending on the translation task?
- Which model or multivariate technique should be used to build our model? Should it be linear only or should we use higher order modeling? Should we allow for latent variables?
- How do we control alignment structure for Machine Translation?
- Which type of multivariate analysis better suits our scenario?
- How are we going to deal with undesired effects in multivariate analysis such as collinearity?
- How do we compare predictive models? How do we evaluate their robustness?

1.3 Contributions

In this dissertation, we perform an in-depth study of the impact the structure of word alignments at different stages of the phrase-based Statistical Machine Translation pipeline, i.e. phrase extraction, phrase scoring and in decoding. We present different multivariate models that highlight the **impact of alignment structure on phrase-based translation model estimation**. Furthermore we test their robustness against unseen data from different language pairs. By using a multivariate approach (as opposed to simple correlation analysis) we take into account the effects of several variables simultaneously.

Moreover, we also establish a multivariate model including different phrase tables and first-best translation hypotheses, that predict how **variations in the translation model predict translation quality**. Highlighting the importance of the translation model and the alignment structure associated with it.

Finally, we identify the most important structural alignment features that influence translation quality and use the information to **provide better alignment training and translation modeling** which ultimately results in better translation quality.

1.4 Methodology

In this section, we briefly summarize the methodology used in this dissertation. It is divided into four parts, each reflecting the contents of a chapter of this dissertation.

1.4.1 Exploratory Analysis

In Chapter 3, we present the results of a study we presented in (Guzman et al., 2009) and some of the intuitions developed in (Guzman et al., 2011). In this chapter we perform an exploratory analysis of different alignment variables, including quality and structure, and develop an intuition of how they impact the phrase-based translation model.

- Analyze the different variables in the alignment (quality and structure) taking into account alignment density characteristics.
- Analyze the phrase-tables (translation-model) and their characteristics.
- Observe the relationship between alignment structure and the phrase-translation model.
- Perform a user-based study to determine the impact of alignment structure on phrase-quality.

1.4.2 Predicting the Translation Model

In Chapter 4, we present a study built upon linear regression models. This study, allowed us to identify the most important structural variables and their effects in the consolidation of a translation model. In this chapter we perform the following:

- Analyze different alignment quality and structure variables, including alignment sparsity.
- Perform a multivariate correlation analysis using clustering.
- Build linear regression models to predict the most important characteristics of a translation model.
- Test our models against unseen data and evaluate the results.

1.4.3 Predicting Translation Quality

In Chapter 5, we present a study where we identify the most important variables of the translation models and the translations, and how are they related with translation performance.

- Perform multiple translation experiments and measure translation quality and structure characteristics of the phrase-translation models.
- Build predictive models using feature selection to discriminate the most important features.
- Test the models against unseen data.
- Build alternate models for specific translation task domains.
- Identify the variables that most impact translation performance.

1.4.4 Improving Translation Using Alignment Structure

In Chapter 6, we propose new training alternatives to incorporate alignment structure in alignment training and decoding.

- Propose several alignment tuning metrics that take into account more of the alignment structure.
- Benchmark the translation improvements resulting from those metrics.
- Propose new decoding features that incorporate alignment structure.
- Benchmark the translation improvement resulting from those decoding features.
- Use the features in low-resource situations to improve translation performance.

1.5 Limitations

In order to keep the development of this study in a manageable time frame, there are several restrictions that we applied.

- The set of language pairs used for experiments will be fixed.

Given the difficulty and time consuming process for SMT training, the only language pairs that will be addressed will be Spanish-English, for estimating the multivariate predictive models and Chinese-English, Arabic-English, to evaluate the generality of such model (i.e. how robust it is to different language pairs).
- We will use a set of generative and discriminative alignments for analysis.

In this analysis we will only employ standard symmetrized alignments, provided by GIZA++ (Och, 2000) plus heuristics, and discriminative alignments with varying settings provided by the DWA (Niehues and Vogel, 2008).
- Phrase-extraction heuristic parameter will be fixed (to default max-length 7).

While the max-length parameter has been observed to have an effect on the size of the phrase table, allowing for longer phrases is too expensive memory wise.
- For the decoding experiments, we will be using the Moses decoder (Koehn et al., 2007).
- We will use linear models because we want to privilege interpretation rather than higher accuracy. Furthermore, by using feature selection (via stepwise regression) instead of other mechanisms that cope with higher dimensionality, we will also privilege the model interpretation.
- Other particular settings of the experiments will be specified in the Experimental Setup in Appendix A.

1.6 Thesis Organization

The remainder of this dissertation is distributed as follows: in Chapter 2, we introduce some of the concepts of Statistical Machine Translation and Phrase Based Machine Translation that we will be using in this document. We pay special attention to statistical word alignment and to the training pipeline of phrase-based Statistical Machine Translation.

Next, in Chapter 3, we present an exploratory study where we observe the effects of statistical word alignments in the estimation of the phrase-translation model. There, we present experiments where we perform alignment experiments with different aligners and compare the output. Then using the same aligners, we obtain translation models and compare them. Additionally, we perform a hand-assessment of the generated phrase-pairs to measure their quality.

In Chapter 4 we present a predictive model for the phrase-table characteristics based on the alignment variables. We highlight the importance of alignment structure in determining the characteristics of the translation model, from structure (size, coverage, etc.) to translation entropies.

In Chapter 5 we use the characteristics of a phrase-table and the characteristics of the first-best translations to predict translation quality. We perform translation experiments for different translation models, and evaluate the translation quality using different metrics. Furthermore, we build models that predict translation quality depending on the translation task.

Next, in Chapter 6 we use the information from alignment structure to improve Statistical Machine Translation. For instance, we take into account alignment structure to propose two new alignment metrics. Then we train aligners to maximize those metrics and compare the generated translation models. Furthermore, we compare the translation outputs of such aligners. Additionally, we propose new decoding features based on the alignment gaps. We use them in different settings and compare the translation results.

In Chapter 7, we summarize the findings of this dissertation. We revisit the hypothesis and the research questions and provide answers. Finally, we discuss the future work, where we discuss enhancements to the current study.

Chapter 2

Background

In this section, we introduce some concepts that are essential to the understanding and development of this proposal. First, we will start by presenting the basics of Machine Translation (MT). Then, we will introduce the main components of a Machine Translation system: language model, translation model and decoder. Next, we will discuss the main steps involved in model the phrase-based translation model estimation. Finally we will discuss the topic of performance in Machine Translation.

2.1 Introduction to Statistical Machine Translation

Machine Translation (MT) or computer assisted translation is the automatic translation of text or speech from a source natural language like Spanish to a target language like English, using a computer system. Despite being one of the most important applications of Natural Language Processing, it has long been considered a hard problem (Manning and Schütze, 1999). Although the ideal goal of MT systems is to produce high-quality translation without human intervention at any stage, in practice this is not possible except in constrained situations (Hutchins, 2003) where translation tasks are limited to small sub-language domains (Jurafsky and Martin, 2007).

Historically, MT systems have been classified in accordance to the level of analysis of the source language. Roughly, three categories have been long defined: Direct Systems which include little or no syntactic analysis; Transfer Systems, based on syntactic parses; and Interlingua Systems based on a deep semantic analysis. For a more detailed description of each one of these architectures, please refer to: (Jurafsky and Martin, 2007; Hutchins, 2003; Somers, 2003).

Along with the different architectures, there are different paradigms that have been adopted by researchers in order to build their systems. Most of the systems using the transfer architecture are known as Rule Based systems, because of the linguistic rules used to translate one source language syntactic tree into a target language syntactic tree. This was the dominant paradigm until the 1990s, when the availability of large amounts of human translated texts gave rise to the corpora-based or empirical paradigms. Examples of these paradigms are the Example Based Machine Translation (EBMT) and the Statistical Machine Translation (SMT), which will be the focus of this thesis. More specifically, we will address the

translation model estimation in the phrase-based Statistical Machine Translation (PBSMT).

2.2 Principles of Statistical Machine Translation

The core idea of Statistical Machine Translation (SMT) is that we can estimate the probability of a source language sentence of being translated into another a target language sentence by analyzing parallel data (i.e. collections of human generated translations) . It was first proposed by Brown et al. (1990) at IBM and became widely popular because it outperformed other MT paradigms.

SMT has proven to be very robust and flexible, because it is not bound to any specific source-target language pair. However, it relies heavily on the training procedures used to estimate its models as well as the nature and availability of the training data.

In its pure form, SMT makes no use of linguistic data. Instead, it models the probability of a source language sentence f of being translated into a target language sentence e ¹ and looks for the translation that maximizes such probability. This can be depicted in the following equation:

$$\hat{e} = \arg \max_e p(e|f) \quad (2.1)$$

Using Bayes rule, this formula can be transformed into:

$$\hat{e} = \arg \max_e \frac{p(f|e)p(e)}{p(f)} \quad (2.2)$$

Note that $p(f)$ is a normalization factor, and because of the $\arg \max$ this equation can be rewritten as:

$$\hat{e} = \arg \max_e p(f|e)p(e) \quad (2.3)$$

This equation is known as the noisy channel model for translation.

2.2.1 Noisy Channel Model

The noisy channel model has been long used in areas such as speech recognition. In SMT, the channel metaphor is applied as follows: we pretend that we originally had a target language sentence e which then was corrupted by noise and transformed to a source sentence f . Therefore, we need to find the target sentence e that f is more likely to have arisen from.

Although modeling the translation probability $p(f|e)$ is not easier than modeling $p(e|f)$, this rewriting ensures that our target translation e has both *fluency* (i.e. that it is *good* English, ensured by the language model $p(e)$) and has *fidelity* (i.e. that it is true to the original meaning in the source language sentence f , as modeled by the translation model $p(f|e)$). Meeting both criteria is difficult. In some cases, to achieve more fluency, the fidelity is sacrificed. Therefore we have to focus in maximizing both criteria simultaneously. This search procedure is widely known as decoding and is taken into account by the $\arg \max$ of the equation. Figure 2.1

¹for historical reasons, f was used to denote French and e to denote English

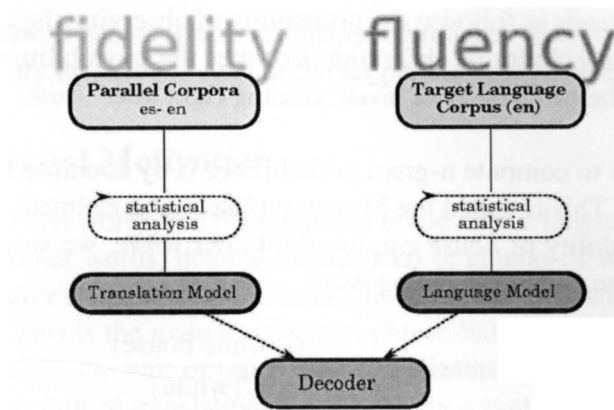


Figure 2.1: Three components in Classical SMT: Language Model, Translation Model and Decoder

depicts the interaction between these three components for a scenario where we want to translate from Spanish to English. The Language Model is estimated by doing statistical analysis on some monolingual corpus (e.g. large collections of English documents). The Translation Model is estimated based in bi-texts or collections of translations (e.g. Spanish/English). The Decoder is a piece of software that uses both models to provide English translations to unseen Spanish translations.

In the following sections, we will describe the main characteristics of the Language Model and the Translation Model (more specifically, the phrase-based). Later, in Section 2.6, we will discuss in depth the decoding procedure for phrase-based models.

2.3 Language Model Estimation

As we mentioned before, the language model (lm), gives us an idea of how fluent a target language sentence e is. It can be regarded as the probability that the words are correctly combined in the target language. They are usually computed via the probabilistic grammars called n-grams.

2.3.1 N-grams

N-gram models are closely related to the problem of word prediction where given a sequence of words, we need to predict the most likely word to complete the succession. Making use of the Markov assumption, which states that only the local context affects the next word, n-gram models use the previous $N - 1$ words to predict the next one (Jurafsky and Martin, 2007). Therefore a bi-gram (or 2-gram) would use the previous word to predict the next word. A trigram, would use the previous two words, and so on. For instance, to calculate the language model probability for the English phrase *the white house* with a bi-gram model would be:

$$p(\text{the white house}) = p(\text{the}) \times p(\text{white}|\text{the}) \times p(\text{house}|\text{white}) \quad (2.4)$$

The previous example reads as follows: the probability of observing the construction *the white house* is equal to the probability of observing *the* times the probability of observing *the* followed by *white* times the probability of *house* coming right after *white*.

The simplest way to compute n-gram probabilities is by counting the frequencies of the sequences in a corpus. This is called the Maximum likelihood estimate (MLE). For instance, to determine the probability of *house* coming right after *white*, we would have to count the occurrences in the corpus in the following way:

$$p(\text{house}|\text{white}) = \frac{C(\text{white house})}{C(\text{white})} \quad (2.5)$$

where $C(x)$ stands for counts of word x .

2.3.2 Smoothing Techniques

Unfortunately MLE has the problem of not being able to deal with sparse data. Therefore, if during training we did not observe the occurrence of a given n-gram, it would be given a zero probability. To overcome this, some methods of “discounting” or “smoothing” are applied. These methods arrange for a certain amount of probability to be distributed among the unseen events, so the distribution is smoother. Some examples are the Good-Turing estimate (Chen and Goodman, 1998) and the Kneser-Ney discounting (Ney et al., 1994).

Higher n-gram models will be a better predictors of the following words because they have more context information. Thus, in MT it is desirable to use higher n-gram models. The major drawback is that the higher the n-gram size, the more sparse the distribution is. To overcome this problem, we can make use of back-off models. Simplifying, back-off models allow us to turn to a lower n-gram model whenever a higher n-gram model fails (i.e. we had no count for such n-gram). Another strategy that helps us to combine several n-gram models is called the deleted interpolation, which linearly combines the probability of several n-grams models under the restriction that the linear coefficients add up to 1.

2.3.3 Estimation

While Language Models are crucial to Machine Translation, estimating a language model is not a frequent practice due to the amount of time and computational resources involved in the process. Instead, we often compute a single target language model (i.e. for English) and use the same for different MT Systems only updating the model when strictly necessary (large amounts of new monolingual data available). For instance, in several of the experiments in this thesis are carried using the same language model and only changing the translation model.

2.4 Phrase Based Translation Model

Under the noisy channel framework, the translation model $p(f|e)$ is the probability that a target language e sentence has arisen from the source language sentence f . Translation models give us an idea of the faithfulness of a translation, and it constitutes the core component in

a Machine Translation System. There are many approaches to calculate this model. In this section, we will briefly discuss the state of the art phrase-based models.

2.4.1 Phrase Based Motivation

It has been observed that words that form phrases in the source language tend to cohere in the target language. In other words, there are collections or chunks of words that appear next to each other in the source language, whose translations appear as chunks of words in the target language (fig. 2.2). This is the main idea behind phrase-based statistical machine translation. In phrase-based models, the unit of translation is a contiguous sequence of words known as phrase. However, this unit of translation does not imply a syntactic structure.

During decoding, phrases are translated as a whole, and then moved to their final positions in the target language by a reordering process. By using phrases instead of words, we capture some local context (e.g. literal translations, idioms), and reduce the number of permutations needed in the final reordering.

The principles of phrase-based SMT (PBSMT) models can be traced back to the template alignments proposed by Och et al. (1999). However the term was coined later by the method proposed by Koehn et al. (2003), which was popularized by the Pharaoh decoder (Koehn, 2004a) and its later successor Moses (Koehn et al., 2007).

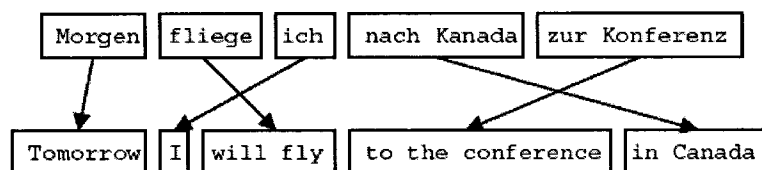


Figure 2.2: Phrase-based models take advantage of the fact that some words often move as units during translation. This is the principle that motivated PBSMT. From (Koehn, 2004a)

2.4.2 PBSMT Principles

The generative story of phrase-based models can be summarized in the following way: The sentence f is segmented into a sequence of I phrases \bar{f}_1^I . Every possible segmentation is assumed to be equally likely. Then each foreign phrase \bar{f}_i in \bar{f}_1^I is translated to an English phrase \bar{e}_i with the phrase translation probability $\phi(\bar{f}_i|\bar{e}_i)$. Also, the English phrase may be reordered with a distortion probability d .

A distortion model d is used to penalize large reorderings by giving them lower probabilities. For example, in a relative distortion model (there are many others), distortion probability refers to the probability of the translations of two consecutive source language phrases being separated in the target language by a span of a particular length. Thus, relative distortion is parameterized by $d(a_i - b_{i-1})$ where a_i is the start position of the target phrase generated by \bar{f}_i and b_{i-1} is the end position of the phrase generated by \bar{f}_{i-1} .

Summarizing, the original phrase-based translation model using this settings can be decomposed into:

$$p(\bar{f}_1^I | \bar{e}_1^I) = \prod_{i=1}^I \phi(\bar{f}_i | \bar{e}_i) d(a_i - b_{i-1}) \quad (2.6)$$

To better illustrate this, consider the translation example in fig. 2.3.

Position	1	2	3	4
Spanish	la casa	en donde vivo	es	blanca
English	the house	where I live	is	white

Figure 2.3: An example of a translation generated from the Spanish sentence *la casa en donde vivo es blanca*

In this example, the segmentation proposed divides the source sentence into four phrases. Each phrase is then translated to English and reordered. For this particular example, every English phrase falls into the same position as its Spanish counterpart, leaving all the distortions equal to 1. Therefore the translation probability for this segmentation can be computed as follows:

$$p(\bar{f}_1^I | \bar{e}_1^I) = \phi(\text{la casa} | \text{the house}) \times d(1) \times \phi(\text{en donde vivo} | \text{where I live}) \times d(1) \\ \times \phi(\text{es} | \text{is}) \times d(1) \times \phi(\text{blanca} | \text{white}) \times d(1)$$

Whether this translation or this segmentation are chosen, depends on how good is its probability in regard to other translation hypotheses.

There are many other ways to calculate a translation model. Variations in how we model distortion or ϕ can make a difference. In fact, the use of a log-linear framework, can allow for a translation model to use many other arbitrary phrase-features. This shall be discussed later, in the decoding section (Section 2.6). At this moment, we will discuss how the phrasal translation probability ϕ is estimated.

2.4.3 Estimating the Model

One issue that comes with the formulation of the phrase-based approach is how to calculate the phrase probabilities like $\phi(\bar{f} | \bar{e})$. If we had a phrase-level aligned bi-text (where chunks of source words were explicitly associated to chunks of target words), it would suffice to count frequencies and establish a maximum likelihood estimate. However this is not the case. We need to find the phrases which are translation of each other directly from the corpus.

There are different types of phrase-based models (e.g. (Marcu and Wong, 2002; Koehn et al., 2003)) and each has its own way of estimating phrasal probability ϕ . On one hand Marcu and Wong (2002) estimates phrasal probability directly from the corpus using unsupervised methods. On the other hand Koehn et al. (2003) uses heuristics to extract phrases from word alignments. In this work, we will be using the latter approach, which has been well spread through

the community thanks to the open source phrase-based Pharaoh system (Koehn, 2004a) and its later successor Moses decoder (Koehn et al., 2007). In the following section we will introduce the main steps involved in the estimation of this type phrase based translation model.

2.5 PBSMT Training Pipeline

From raw training data to translation model, the entire pipeline can be pictured in Figure 2.4. There are at least four important steps: Preprocessing, Word Alignment, Phrase Extraction and Phrase Scoring. Below, we provide a description of each of these stages.

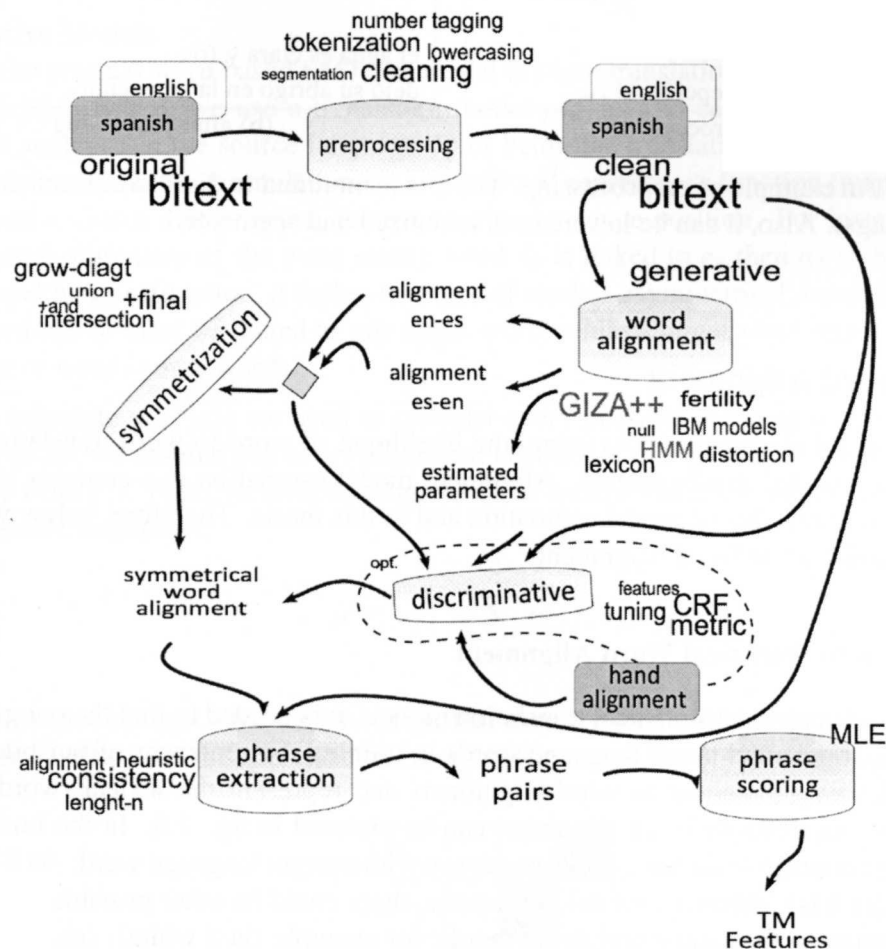


Figure 2.4: Pipeline of Phrase-Based Translation Model Estimation. First we clean the raw training data by preprocessing. Then we estimate statistical word alignments (in a generative or discriminative way) to generate symmetrized alignments. Then we use phrase-extraction to obtain phrase-pairs that are consistent with the word alignments. Finally we estimate the probabilities of phrase-pairs using maximum likelihood.

2.5.1 Preprocessing

The preprocessing step focuses on cleaning the bilingual corpus and preparing it for the model estimation. This is an important step, as it allows to filter for undesired characters, html tags, etc. Also some engineering choices can be carried: tokenization (i.e. how to split words), segmentation (i.e. which morphological criteria to use), case (i.e. use truecase or lowercase), number tagging, etc.; to help models to carry a better estimation. In Figure 2.5 we observe an example of preprocessing.

<pre>< td > La casa es blanca. < /td > El teatro queda a dos calles hacia el norte. Ella tiene un < em > libro de política < /em > internacional. El agua es clara y fría. Dejó su abrigo en la recepción.</pre>	<pre>la casa es blanca . el teatro queda a dos calles hacia el norte . ella tiene un libro de política internacional . el agua es clara y fría . dejó su abrigo en la recepción .</pre>
(a) before processing	(b) after processing

Figure 2.5: An example of preprocessing. The text is modified to discard unwanted characters (e.g. html tags). Also, it can be lowercased, tokenized and segmented.

2.5.2 Word Alignment

During the word alignment (WA) stage, the likelihood of word-to-word translations are estimated and encoded as alignments. Alignment model estimation is a complex process but crucial to the phrase-based model estimation and to this thesis. Therefore, below we provide a brief introduction to Word Alignment.

Introduction to Statistical Word Alignment

In order to estimate most statistical translation models, it is needed to find the correspondence between the source and target language words in training the sentences of our bitext. These correspondences are known as word alignments and represent the lexical (word-by-word) translations. An example of an alignment can be pictured in fig. 2.6. In the image shown, each source language word has a correspondence with a target language word: $\langle \text{la} \parallel \text{he} \rangle$, $\langle \text{casa} \parallel \text{house} \rangle$, $\langle \text{es} \parallel \text{is} \rangle$, $\langle \text{blanca} \parallel \text{white} \rangle$. Of course, there could be other possible alignments or mappings between the source and target words, for example $\langle \text{la} \parallel \text{white} \rangle$, $\langle \text{casa} \parallel \text{the} \rangle$, $\langle \text{es} \parallel \text{house} \rangle$, $\langle \text{blanca} \parallel \text{is} \rangle$.

More formally, Statistical Word Alignment is the task of finding the correspondences between words in a source language sentence ($f_1^I = f_1 \dots f_I$) and the words in a target language sentence ($e_1^J = e_1 \dots e_J$) in which the source and target sentences contain I and J words, respectively. The alignment A of this pair is simply a set of these correspondences. We say that $A \subset \{1, 2, \dots, I\} \times \{1, 2, \dots, J\}$. If $(i, j) \in A$, then the i th source word is aligned to the j th target word. Also, it is said that there exists a *link* between words (i, j) .

Alignment Models

There have been several approaches to find alignments in a corpus. First, the unsupervised or generative models (Brown et al., 1993; Vogel et al., 1996), which model the alignment as a hidden variable, and try to discover their likelihoods using machine learning algorithms such as the Expectation Maximization or the forward backward algorithm. Second, the heuristic models (Melamed, 2000; Kobdani et al., 2009), which estimate the likelihood of an alignment according to word-distance and other metrics. Lastly, supervised or discriminative models (Ittycheriah and Roukos, 2005; Blunsom and Cohn, 2006; Niehues and Vogel, 2008) which make use of human annotated data (word alignments done by humans also known as hand alignments) to learn a model according to specific features.

Generative Models

In the original word based SMT, we model the best translation using the noisy channel approach, where we use a translation model $p(f_1^I | e_1^J)$ to calculate the probability of one sentence in the source language f_1^I of being the translation of the target language sentence e_1^J . In this formulation, an alignment a_1^I represents a function (mapping) which given a source side position, returns a target language position. For instance if in the current alignment a_1^I the word source word f_3 is linked to e_8 then $a_3 = 8$. While this model helps estimation, it makes this type of models asymmetrical, because one source word can at most be linked to one target word, while a target word might be linked to two or more source words.

To calculate $p(f_1^I | e_1^J)$ we need to consider every possible mapping in our calculation. To do so, we assume that every alignment is possible (i.e. any word f_i can be linked to any word e_j). Then, we treat the alignments as hidden variables and we sum over every possible alignment.

$$p(f_1^I | e_1^J) = \sum_{a_1^I} p(f_1^I, a_1^I | e_1^J) \quad (2.7)$$

Generative models have the advantage that they are well suited for a noisy-channel approach. They use unsupervised training to estimate $p(f_1^I, a_1^I | e_1^J)$ using large amount of

blanca				■
es			■	■
casa		■		■
la	■			
	the house	is	white	

Figure 2.6: A simple example of word-alignment. Here, each source language word has a correspondence with a target language word: $\langle la \parallel he \rangle$, $\langle casa \parallel house \rangle$, $\langle es \parallel is \rangle$, $\langle blanca \parallel white \rangle$

unlabeled parallel data. There have been several generative models that were proposed by Brown and his colleagues at IBM (Brown et al., 1993). Therefore they are widely known as the IBM models. Another widespread model is the HMM alignment model proposed by (Vogel et al., 1996). Each model has an increasing level of complexity. In the Table 2.1, we show the most important features of each model.

Model	Main features
IBM1	lexical probabilities only
IBM2	lexicon plus absolute position distortion
HMM	lexicon plus relative position distortion
IBM3	plus fertilities
IBM4	inverted relative position alignment
IBM5	non-deficient version of model 4

Table 2.1: Summary of main generative alignment models and their main characteristics

Lexical probabilities are the likelihood that one word in the source language be translated to another word in the target language. Distortion is a penalization for less likely alignments which try to align words that are too far apart. Fertility helps us to provide a mechanism with which a single target language word can be regarded as generating several source language words.

Since higher order models are more complex, a common procedure is to use the estimations from lower order models as priors. E.g. to train IBM1 to determine lexical probabilities and then use them as initialization for HMM, and so on. This is known as bootstrapping.

GIZA++

All standard alignment models (IBM1...IBM5, HMM) have been implemented in the GIZA++ toolkit. This toolkit was started at John Hopkins University workshop 1998 and later extended by Och (2000). It is widely used by many groups and included in many translation toolkits such as Moses Decoder. Newer multi-threaded adaptations (Gao and Vogel, 2008) are also now in use.

Viterbi Alignments

In order to obtain a word alignment for the I words in f_1^I to the J words in e_1^J using the estimated models, one needs to perform a search among all possible mappings a_1^I for the that maximizes the likelihood of the model \mathcal{M} in question.

$$\hat{a}_1^I = \arg \max_{a_1^I \in A} p_{\mathcal{M}}(a_1^I, f_1^I | e_1^J) \quad (2.8)$$

In other words, we choose the alignment between that maximizes the probability of the sentence e_1^J of being the translation of the f_1^I , given the probability model $p_{\mathcal{M}}$ we estimated during training. Such alignment \hat{a}_1^I is known as the Viterbi alignment for e_1^J and f_1^I .

Symmetrization

Because of how generative approaches model probabilities, certain restrictions apply. For example, many source to one target word alignments are allowed (modeled by the fertility), but only one target per source word alignments are permitted. This results in an inherent asymmetry depending on the direction of the alignment we are performing. For instance, the alignment obtained from English to Spanish is substantially different from the alignment obtained from Spanish to English.

Fortunately, there are some heuristic processes that can be performed using the information in the source-to-target (s2t) and target-to-source (t2s) alignments to boost the accuracy of an alignment. These heuristics help to overcome the original weaknesses of generative models and are very helpful for phrase extraction. Some of these heuristics include intersection, union, and some other refined heuristics such as grow-diag, grow-diag-final, grow-diag-final-and (Och et al., 1999; Koehn et al., 2003). Symmetrization is a challenging procedure. Because of this, different research efforts have focused on it (Matusov et al., 2004; Liang et al., 2006).

An example of a symmetrization using a union heuristic is depicted in fig. 2.7.

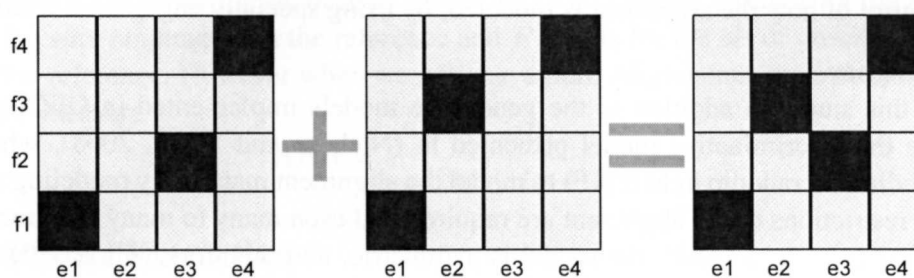


Figure 2.7: Example of a union type symmetrization: Alignments of different directions (left, middle) are combined to include the links that are present in both (right).

The most used symmetrization heuristics are the grow-diag family of heuristics. We describe them below.

grow-diag. This heuristic starts by using the intersection of the s2t and t2s alignments. Then visits each of the links $(i, j)_I$ in the intersection and adds any link $(i \pm 1, j \pm 1)_U$ in the union of s2t and t2s that lies in the neighborhood of the first link provided that either source i or target j words were left unaligned.

grow-diag-final. This heuristic is based on the *grow-diag* but uses the *final* procedure which consists on visiting each of the unaligned words in e and f . If there is a link in either s2t or t2s that contains i or j , add it to the final alignment.

grow-diag-final-and This heuristic is based on the previous one, but requires that both words in i and j are left unaligned before adding a link from s2t or t2s.

We expect to have a natural order in terms of number of links between these heuristics. In other words: $A_I \subset A_{GD} \subset A_{GDFA} \subset A_{GDF} \subseteq A_U$.

Discriminative Models

Generative models are very robust, however they have a major disadvantage: they can hardly make use of the increasingly available manual alignments. Also, given their complexity, to incorporate other sources of information such as POS tags, word frequencies etc., is a non-trivial task.

In recent years several authors (Moore, 2005; Taskar and Lacoste, 2005; Blunsom and Cohn, 2006) have proposed discriminative word alignment frameworks and showed that this leads to improved alignment quality. In the discriminative alignment approach, we model $p(a|f, e)$ directly. Therefore, the decision rule for this approach becomes:

$$\hat{a} = \arg \max_{a_k \in A} p_{\mathcal{M}}(a_k | f_1^I, e_1^J) \quad (2.9)$$

where a_k represents an alignment among the set of all possible alignments A between the source f_1^I and target e_1^J sentences.

One main advantage of the discriminative framework is the ability to use all available knowledge sources by introducing additional features. Also, these models tend to be symmetric in nature, making them suitable for phrase-based machine translation without the need of symmetrization. Finally, discriminative approaches allow to have more control of how the alignment is modeled, by using specially engineered features.

DWA Aligner

In this study, in addition to the generative models implemented in GIZA++ we will use the discriminative model presented in (Niehues and Vogel, 2008), which uses a conditional random field (CRF) to model the alignment matrix. By modeling the matrix, no restrictions to the alignment are required and even many to many alignments can be generated. As a result, the model is symmetric, and therefore will produce the same alignment regardless of the direction in which it is used.

Word Alignment Quality

When analyzing the errors made by the automatically generated word alignments we compare the aligner output (e.g. Viterbi alignments generated by GIZA++) against a gold standard reference of consisting of human hand aligned data. In some gold standards, there is a distinction between Sure and Possible links. Sure links represent the hand alignments made by the annotator for which he is sure of the alignment. Possible links are those which represent a degree of uncertainty, e.g. were different annotators differ in the manual alignment.

Based on the differences between output alignments and hand alignments, there are three basic quantities that we can measure: the number of links in which these two alignments agree, i.e. true positives (tp); the number of links that are present in the output of the aligner but not in the gold standard, i.e. false positives (fp); and the number of links that are present in the gold standard, but not in the output of the aligner, i.e. false negatives (fn). There are several metrics that are used for measuring the quality of a word alignment, but most of them are based on these three quantities.

Precision

This measure gives us a notion of how accurate is the output of the aligner. It is the ratio

of the correct to all generated links.

$$Precision = \frac{tp}{tp + fp} \quad (2.10)$$

Recall

This measures how well we cover the desired links, i.e. those in the hand alignments, with the automatically generated ones. That is, of all the links in the gold standard, what is the amount of links that are also present in the aligner output.

$$Recall = \frac{tp}{tp + fn} \quad (2.11)$$

Alignment Error Rate

AER, as defined in (Och and Ney, 2003), takes Sure and Possible links into account.

$$AER = 1 - \frac{|A \cup S| + |A \cup P|}{A + S} \quad (2.12)$$

Where A stands for the set of links in the output alignment, S stands for the set of links in the sure alignments in the reference and P stands for the set of possible alignments in the reference. However when we only have Sure Alignments, the metric is related to the F measure :

$$AER = 1 - \frac{2tp}{2tp + fp + fn} = 1 - F \quad (2.13)$$

2.5.3 Phrase Extraction

In order to estimate phrasal probabilities, we need to find out which phrases are translation of each other. This is done using word alignments as guidelines to obtain all phrase-pairs $\langle \bar{f}_i \parallel \bar{e}_i \rangle$ that are consistent. Consistency is defined as follows: if the source word f_i is aligned with the target word e_j , then a phrase-pair containing f_i must also contain e_j ; likewise, a phrase-pair containing e_j must also contain f_i . Phrase pairs containing neither f_i nor e_j are not constrained in any way by the alignment point (i, j) . In Figure 2.8, we observe an example of this criterion.

The process of phrase extraction is done incrementally, starting from phrase-pairs of length 1 up to a predefined maximal length (usually 7). In Figure 2.9 we observe the process of phrase extraction incrementally up to length 4.

2.5.4 Phrase Scoring

After all phrase-pairs have been extracted, the phrase-based translation model is computed using a maximum likelihood estimate with no smoothing, as shown in the equation below.

$$p(\bar{e}|\bar{f}) = \frac{\text{count}(\bar{e}, \bar{f})}{\text{count}(\bar{f})} \quad (2.14)$$

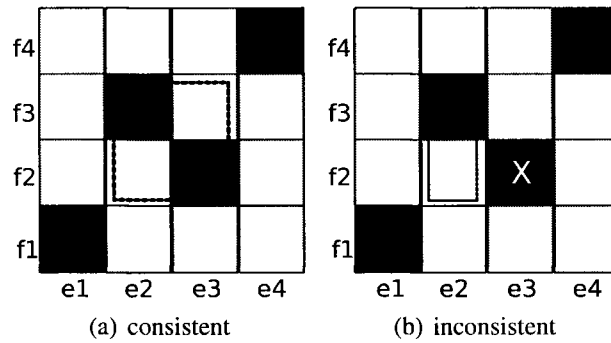


Figure 2.8: Examples of a consistent (a) and an inconsistent (b) phrase-pair. In (a) the phrase-pair (a square in black dotted lines) $\langle f2\ f3 \parallel e2\ e3 \rangle$ has all its words in the source language aligned to target language words which are part of the same phrase-pair. In the second picture (b) the phrase-pair $\langle f2\ f3 \parallel e2 \rangle$ is inconsistent given that the word $f2$ is aligned to $e3$ which lies outside the phrase-pair.

In this equation, the direct phrasal translation probability feature $p(\bar{e}|\bar{f})$ is calculated by counting the number of times that the source phrase \bar{f} and the target phrase \bar{e} belong to the same phrase-pair, and normalizing by the number of times the source phrase \bar{f} is extracted paired to any other target phrase. The learned phrases are then stored in a data-structure widely known as phrase-table. Phrase-tables usually include the source and target language phrases, as well as their phrase-translation probabilities, lexical weighting, etc. This works as a database for the decoder.

source phrase \bar{f}	target phrase \bar{e}	count (\bar{f}, \bar{e})
la casa	the house	17
la casa	the household	5
casa es	house is	8
casa blanca	white house	30
casa blanca	house white	11
la cámara	the house	120
cámara de diputados	house of representatives	125

Table 2.2: A toy example of conditional phrasal translation estimation. In order to estimate the likelihood of the phrase *la casa* to translate into *the house* or $p(\hat{e}|\hat{f})$ we need to count all the times we observed such pairing (17) and divide it by the times that we observed *la casa* (17 + 5), which results in $p(\text{the house} | \text{la casa}) = 17/(17 + 5) = 0.77$. Similarly, $p(\text{la casa} | \text{the house}) = 17/(17 + 120) = 0.12$

2.6 Decoding

Decoding is the process through which we can use the translation and language models to translate unseen source language sentences into target language. The process consists in a

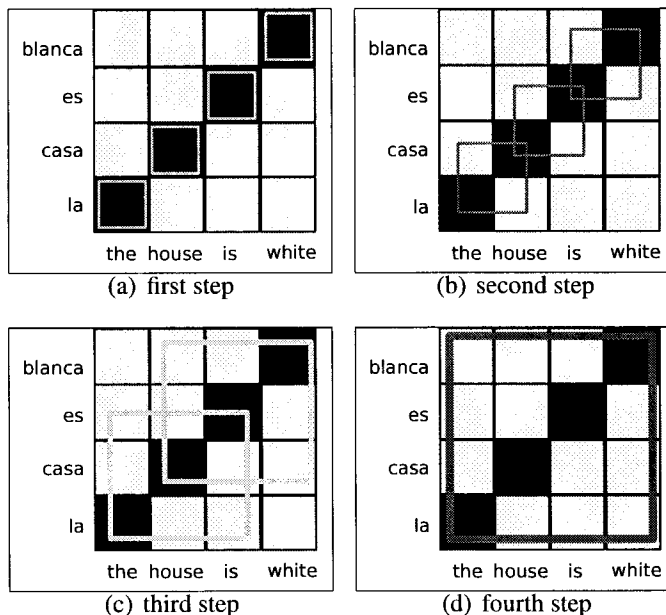


Figure 2.9: Phrase extraction example. At each step, the heuristic extracts the phrase-pairs that are coherent with the alignment. First step (a): $\langle la \parallel the \rangle$ $\langle casa \parallel house \rangle$ $\langle es \parallel is \rangle$ $\langle white \parallel blanca \rangle$. Second step (b): $\langle la \ casa \parallel the \ house \rangle$ $\langle casa \ es \parallel house \ is \rangle$ $\langle es \ blanca \parallel is \ white \rangle$. Third step (c): $\langle la \ casa \ es \parallel the \ house \ is \rangle$ $\langle casa \ es \ blanca \parallel house \ is \ white \rangle$. Fourth step (d): $\langle la \ casa \ es \ blanca \parallel the \ house \ is \ white \rangle$.

search for the best candidate translation. Modern decoders use the log-linear framework (Och and Ney, 2002) to score different hypotheses. Under this framework, translation models and language models are viewed as features. Other features are also available for use. The decoder, uses all the information available to translate input sentences. The output of the decoder can be (and usually is) evaluated to keep track of performance. A summarized picture of decoding is presented in Figure 2.10.

In the next section, we will explain the log-linear framework, with its features and weights. Later, we will introduce some concepts of the search procedure used by the decoder. Finally, we will discuss the different evaluation metrics used to measure the performance of a decoder.

2.6.1 Log-Linear Models

While early statistical MT was based on the noisy channel model, most recent systems make use of a discriminative log linear model (Och and Ney, 2002). In this model we directly model the posterior probability $p(e|f)$ and we search for the sentence with the highest posterior probability:

$$\hat{e} = \arg \max_e p(e|f) \quad (2.15)$$

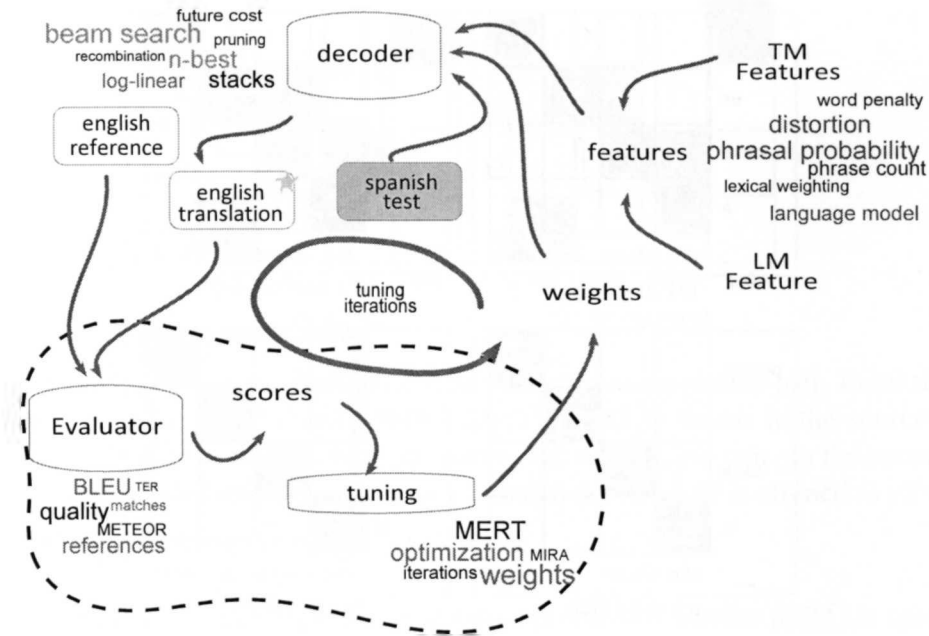


Figure 2.10: Decoding overview. The decoder uses features (e.g. translation model, language model) and weights for those features, to translate a source sentence (Spanish) into a target language (English). Translations from the decoder can be compared to references to evaluate a decoder’s performance. During parameter tuning, the feature weights are optimized towards the evaluation metric.

A well suited framework for doing this modeling is maximum entropy (Berger et al., 1996). In this framework we model $p(e|f)$ through a set of M feature functions $h_m(e, f)$ each of which has a parameter λ_m which acts as a weighting. The translation probability is then:

$$p(e|f) = \frac{\exp \left[\sum_{m=1}^M \lambda_m h_m(e, f) \right]}{\sum_{e'} \exp \left[\sum_{m=1}^M \lambda_m h_m(e', f) \right]} \quad (2.16)$$

Where $h_m(e, f)$ is the m th local feature (depends only on e and f) and λ_m is its scaling factor or feature weight. On the bottom part of the equation, we have a normalization factor which assures that the exponential term is a true probability. Fortunately, when doing the search, we can omit the normalization denominator. Thus, we obtain the following decision rule:

$$\hat{e} = \arg \max_e \sum_{m=1}^M \lambda_m h_m(e, f) \quad (2.17)$$

In fact, the noisy channel approach is a special case of this kind of optimization, with $h_1(e, f) = \log p_{LM}(e)$ and $h_2(e, f) = \log p_{TM}(f|e)$ with $\lambda_1 = \lambda_2 = 1$. In practice, the language model and translation model are still the most important feature functions in the log-linear model, but the architecture has the advantage of allowing the inclusion of other arbitrary features as well.

Since the optimal weights for the model are unknown, log linear models are trained to directly optimize evaluation metrics using a procedure such as minimum error rate training (MERT) (Och, 2003). This is known as parameter tuning.

Some of the most used features in state-of-the-art PBSMT systems are summarized in Table 2.3.

Feature	Description
$\phi(\bar{f} \bar{e})$	An inverse conditional phrase-to-phrase probability model.
$\phi(\bar{e} \bar{f})$	A direct conditional phrase-to-phrase probability model.
$p_{lex}(\bar{f} \bar{e})$	An inverse lexical weighting feature (a word by word version of ϕ).
$p_{lex}(\bar{e} \bar{f})$	A direct lexical weighting feature (a word by word version of ϕ).
p_d	A distortion model.
$p_{lm}(e)$	A language model .
w	A word penalty, which provides means to ensure that the translations do not get too long or too short.

Table 2.3: Most frequent decoding features in modern systems

Most of these features are estimated using phrase scoring and then stored in a dictionary-like database called phrase-table. Then, phrase-tables are loaded by the decoder at runtime.

2.6.2 The Process of Decoding

Given an input source language sentence, there are several possible translations that can be achieved. Each of those translations is called a translation option or translation hypothesis and has a cost (the log-probability). We want to find the translation that has a lower cost. During decoding, the input sentence is segmented and all translation options of the possible segmentations are collected. This allows a quicker lookup, for faster decoding.

Search

The search for the best translation is done using a beam search algorithm. Each state is defined by the 'covered' words, (i.e. the number of words that have already been translated) and the target language string that has been generated. In the initial state, no source language words have been translated and no target language words have been generated. The goal state is where all the source language words are covered. From each state, the transitions are defined by the translation options. In Figure 2.11

For example, if our source sentence is “la casa blanca”, the initial state has three uncovered words, no English translation and no cost associated $\{la\ casa\ blanca\ ||\ -\ ||\ 1\}$. When a translation option that covers the first two words, using $\langle la\ casa\ ||\ the\ house \rangle$ would allow us to transition from the initial state to an state where there are two words covered , generating the partial hypothesis “the house” with a cost (hypothetical) of $-\log p(0.75)$ (i.e. the state $\{*\ bla\ nca\ ||\ the\ house\ ||\ 0.75\}$).

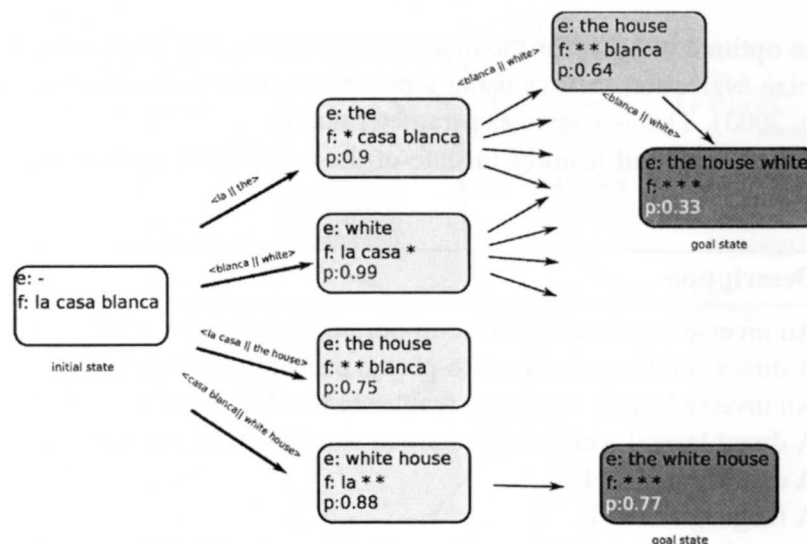


Figure 2.11: Example of different states and transitions during decoding

The current cost of the new state is the cost of the original state multiplied with the translation, distortion and language model costs of the added phrasal translation. Among these the hypothesis with the lowest cost (highest probability) is selected as best translation.

Recombination

When there are two paths that lead to two or more hypotheses that are equivalent to the decoder, we keep only the hypothesis the least cost so far. The other hypothesis will always have higher cost regardless of the future paths, so we can safely discard (prune) it. This is known as hypothesis recombination. For example the following paths, lead to hypotheses that can be recombined:

1. $\{ \text{la casa blanca} \parallel - \parallel 1 \} \rightarrow \{ * \text{ casa blanca} \parallel \text{the} \parallel 0.9 \} \rightarrow \{ * * \text{ blanca} \parallel \text{the house} \parallel 0.64 \}$
2. $\{ \text{la casa blanca} \parallel - \parallel 1 \} \rightarrow \{ * * \text{ blanca} \parallel \text{the house} \parallel 0.75 \}$

In this example, we would choose the hypothesis b) that has a lower cost (higher probability) and discard hypothesis a).

Pruning

While the recombination of hypotheses as described above reduces the size of the search space, this is not sufficient. Therefore, several types of pruning can be implemented (e.g. threshold or histogram). This is done taking into account the cost of the hypothesis so far and a future cost estimation. Future cost estimation is similar to the heuristic distance estimate in a A^* search. Since during decoding we are dealing with partial hypotheses (i.e. hypotheses that cover/translate different parts of the source sentence), future cost estimation helps to prevent

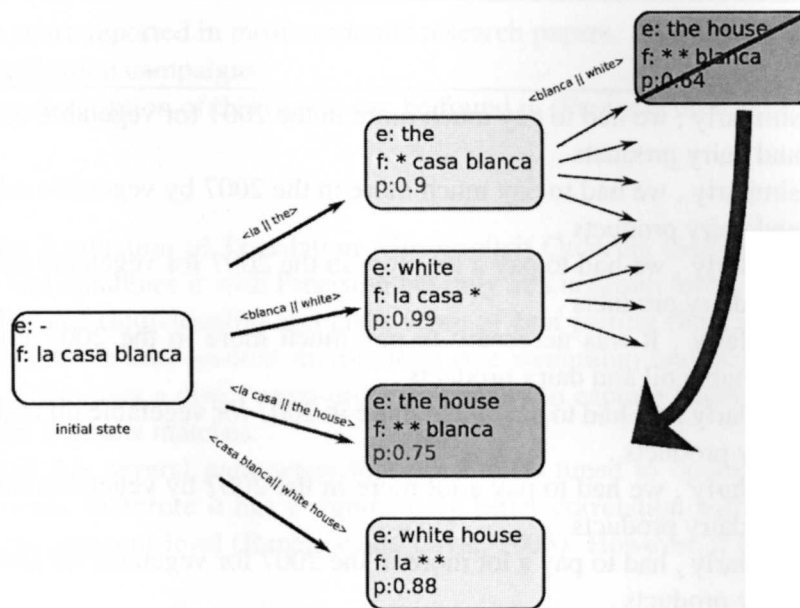


Figure 2.12: Example of hypothesis recombination

the early pruning of partial hypotheses with “difficult” segments (i.e. that cover parts with inherent high translation cost) in favor of partial hypotheses with lower cost that are yet to cover those high-cost segments.

Note that this type of pruning is not risk-free. If the future cost estimates are inadequate, we may prune out hypotheses on the path to the best scoring translation.

N-best lists

Usually, we expect the decoder to give us the best translation for a given input according to the model. However, search and model errors can occur, and the best-cost or first-best translation hypothesis could not be the best option. Therefore sometimes is useful to have more options to choose from. An n-best list is one way to represent multiple candidate translations. N-best lists are formed by a set of N translation hypotheses sorted by their model score. Table 2.4 depicts an example of an n-best list.

2.6.3 Translation Evaluation

The widespread development and usage of MT systems created a need for cheap and fast evaluation tools that could help to differentiate their quality and allow for development testing on an ongoing basis. Therefore several automatic quality metrics have been proposed.

While current metrics are still very crude, translation evaluation is a very active research field. One of the most widespread metrics is BLEU (Papineni et al., 2002), but there are others that are starting to gain adoption (Meteor, TER). In this section we will briefly introduce them.

Model rank	Hypothesis	Model Score
1	similarly , we had to pay much more in the 2007 for vegetable oil and dairy products .	-7.397
2	similarly , we had to pay much more in the 2007 by vegetable oil and dairy products .	-7.405
3	similarly , we had to pay a lot more in the 2007 for vegetable oil and dairy products .	-7.413
4	similarly , it was necessary to pay much more in the 2007 for vegetable oil and dairy products .	-7.419
5	similarly , we had to pay much more in 2007 for vegetable oil and dairy products .	-7.420
6	similarly , we had to pay a lot more in the 2007 by vegetable oil and dairy products .	-7.421
7	similarly , had to pay a lot more in the 2007 for vegetable oil and dairy products .	-7.422
8	similarly , we had to pay much more in 2007 by vegetable oil and dairy products .	-7.425
9	similarly , it was necessary to pay much more in the 2007 by vegetable oil and dairy products .	-7.427
10	similarly , had to pay a lot more in the 2007 by vegetable oil and dairy products .	-7.431
Source:	de igual modo, hubo que pagar mucho más en el 2007 por el aceite vegetal y por los productos lácteos .	
Ref:	consumers also have had to pay significantly more for vegetable oils and dairy products in 2007 .	

Table 2.4: Example of a n-best list

BLEU

The BiLingual Evaluation Understudy (BLEU) is a widely used automatic evaluation tool for Machine Translation. It ranks from 0 to 1, 1 being the best translation. Its main goal is to compare n-grams of the candidate with the n-grams of the reference translations and to count the number of matches. These matches are position-independent. The more the matches, the better the candidate translation.

Since BLEU is precision oriented, shorter translations with more n-gram matches are privileged. To avoid this, a brevity penalty is introduced. This brevity penalty deters from having translations that are shorter than the reference. BLEU has two nice properties. It accounts for adequacy by looking at word precision and it accounts for fluency by calculating n-gram precisions. However, since the final score is the weighted geometric average, it is deficient at the sentence level. Instead, it is an aggregate score over a large test set. Although recent studies suggest that BLEU's correlation with human judgments is not as strong as previously thought (Callison-Burch et al., 2006), and other metrics are available, BLEU is

still the main score reported in most academic research papers. Also, is the main indicator in many SMT evaluation campaigns.

A clearer description of this metric can be found in (Papineni et al., 2002).

METEOR

The Metric for Evaluation of Translation with Explicit Ordering (METEOR) takes into account Recall and combines it with Precision but only at a unigram level. It aligns MT output with each reference (individually) and takes score of best pairing (best alignment). The unigram matching takes into account morphology (via stemming) and also includes semantic equivalents. It also uses a direct word-ordering penalty to capture fluency instead of relying on higher order n-grams matches.

METEOR has several parameters that need to be tuned to optimize correlation with human judgments, therefore it has a significantly better correlation with human judgments, especially at the segment-level (Banerjee and Lavie, 2005). However, it is more expensive to compute.

TER

Translation Error Rate is an error metric for machine translation that measures the number of edits required to change a system output into one of the references (Snover et al., 2006). This metric is widely used in evaluation campaigns such as GALE².

²<http://projects.ldc.upenn.edu/gale/>

Chapter 3

The Effects of Word Alignments on Phrase-Extraction

Statistical word alignments serve as a starting point for the Statistical Machine Translation (SMT) training pipeline. Improving their quality has been a major focus of research in the SMT community. However, due to the amount of processing that a word alignment undergoes before being used in translation (e.g. phrase extraction), the quality of word alignments is not necessarily related to the quality of translation.

The goal of better understanding the relationship between the alignment quality metrics (AER, precision, recall) and translation quality, is to make improvements in word alignment carry over to improvements in the end-to-end system performance. This is especially important in the case of discriminative word alignment (Niehues and Vogel, 2008), where alignments are designed to maximize a desired quality metric based on hand labeled data.

In this chapter we present the results of our study published in (Guzman et al., 2009). Here, we explore in detail the dependencies between the word alignment and the phrase extraction, as an effort to better understand the role of word alignments in phrase extraction. By studying the relationship of these two processes, we aim to shed light into the role that statistical word alignments play into the SMT training procedure.

Furthermore, we explore not only alignment quality, but also structural characteristics of the alignment such as link density and number of unaligned words, and their impact in the phrase-pairs extracted from them. Our findings suggest that these structural characteristics have a strong influence on how our translation models are estimated. We found that along with alignment quality, alignment structure affects the quality of our translation model.

This chapter is organized as follows: First, we start with a brief discussion of the importance of alignments in SMT. Next, we cover the details of the experimental setup of our study. We present the different types of alignments used and their characteristics. Later we analyze the phrase-pairs extracted from these alignments and discuss the results. Finally, we present the findings of a study we performed to assess the quality of phrase-pairs and observe how the structural characteristics of our alignments impact phrase-pair quality.

3.1 Alignments and Phrase-Pairs

Statistical word alignments are the basis for many SMT paradigms. They can be pictured as 'links' between the source and target language words of a sentence pair. They represent the lexical (word-to-word) translation of our sentences. Depending on the approach used to build our translation model, their application in SMT may vary. For phrase-based SMT, we use those alignments as input to build a phrase-lexicon (phrase-level dictionary) which is latter used to translate unseen documents. In consequence, word alignments determine which phrase-pairs are extracted and which are ignored. Furthermore, phrase-pair translation probabilities are determined based on frequencies. If a defective phrase-pair is extracted in several occasions during training, it will be prone to be wrongly used as a reliable translation. Hence the importance of having good, accurate word alignments to start with.

Traditionally, high quality alignments are those that the most alike to a human generated alignment. While this notion is itself debatable (two annotators can differ in their lexical translations of a sentence pair), much research has been done in augmenting the quality of automatic alignments, having as an objective to improve machine translation. Conversely, there have been studies that have found that alignment quality is not correlated with translation quality (Fraser and Marcu, 2007), or even prejudicial (Ayan and Dorr, 2006). However, little has been done to measure the impact of alignments in a step-by-step procedure. If high quality alignments are the goal, how does their quality impact our translation models?

In this study, we focus in describing how different alignments, with varying alignment qualities, produce phrase-pairs. We pay attention to the quantity and quality of these generated phrase-pairs. In the following sections, we describe the details of this study.

3.2 Experimental Setup

There are several types of alignments. On one hand, we have the generative alignments which are discovered automatically given a set of sentence-aligned bilingual corpus. These models are asymmetrical, which means that the alignment discovered in one direction (e.g. source-to-target (S2T)) are different than the alignments discovered in their reverse counterpart. To alleviate this problem, there are several heuristics that symmetrize the two directions by adding the common links first, and then the remaining links that do not cause conflict. One example of such heuristic is the commonly used grow-diag-final heuristic (Och and Ney, 2004). On the other hand, we have the discriminative alignments which are trained to maximize a certain alignment metric. The results of these alignments are symmetrical and sometimes use the generative model alignments as input features.

For the following analysis, we used a small set of different automatically generated alignments along with human annotated data. We used both types of alignments: generative and discriminative to align a Chinese (S) English (T) corpus. For the generative alignment, we used the Viterbi alignments resulting from performing training through the standard sequence of word alignment models IBM1, HMM, IBM3 and finally IBM4, in both directions, i.e. source to target (S2T) and target to source (T2S). We used implementation provided by the modified GIZA toolkit (Gao and Vogel, 2008). In addition, we generated the symmetrized alignment (SYM), using the grow-diag-final heuristic implemented and used in the MOSES

package (Koehn et al., 2007).

For the discriminative alignments, we used the approach described in (Niehues and Vogel, 2008), because the output alignment matrix generated by such a system is composed of continuous values representing the alignment strength between source and target word. This allows to easily control the density of the alignment matrix, by using different intensity thresholds, without having to recalculate the alignment. The different probability thresholds used throughout this study are $p = \{0.1, 0.2, \dots, 0.9\}$.

In the following experiments, the discriminative word aligner (DWA) uses the models from the GIZA training (lexicon, fertility) as well as the GIZA S2T and T2S Viterbi alignments as features. It is tuned to minimize the alignment error rate (AER) on the hand-aligned data using the alignment with threshold $p = 0.5$ as output. In Table 3.1, we show the sizes of the training sets for each of the aligners. We also show the size of our testing set. Note that for the tuning and the evaluation test sets the number of English words is about 20% higher than the number of Chinese words. For the training data the ratio is closer to 1 : 1.13.

	Corpus Statistics	
	#Sentences	#Words
GIZA Training		
Chinese	11.0 M	273M
English	11.0 M	309M
DWA Tuning		
Chinese	500	10,285
English	500	12,632
Alignment Test		
Chinese	2,000	39,052
English	2,000	48,655

Table 3.1: Data Statistics of the alignment training set

3.3 Characterization of Word Alignments

When describing an alignment, there are two types of measurements we can use. On one hand, there are the alignment quality measures like AER, precision and recall, which describe how close our output is to a gold standard in terms of the number of common links in the alignment. On the other hand, we have different structural measurements that can be computed over alignments, i.e. number of unaligned words, number of links, etc., which allow us to better understand the inner structure of an aligner’s output. In the following sections, we explore the differences in the quality and structure of the different alignments. Later we discuss how these differences impact phrase-extraction.

3.3.1 Analysis of Alignment Quality

In this part, we compare each of the automatically generated alignments against the hand-aligned data. We used precision, recall and alignment error rate (AER) to evaluate the quality of the alignments.

In Table 3.2 we display the quality measurements for the different word alignment approaches. First, the one-sided GIZA alignments and the symmetrized (grow-diag-final) alignments are listed. For the discriminative word alignments the results for different thresholds are shown. Notice that the lowest error (AER) is achieved using the DWA-5. Changing the threshold allows us to cover a wide variety of alignments, from high precision (DWA-9) to high recall (DWA-1). Observe that the best discriminative alignments give a lower AER than the symmetrized alignment from the generative models.

Aligner	Quality Measurements		
	Precision	Recall	AER
GIZA S2T	51.67	34.91	58.33
GIZA T2S	66.48	56.92	38.67
SYM	67.98	56.29	38.41
DWA-1	45.16	71.96	44.51
DWA-2	57.35	65.26	38.95
DWA-3	64.59	61.64	36.92
DWA-4	69.47	59.22	36.06
DWA-5	72.99	57.38	35.75
DWA-6	76.04	55.22	36.02
DWA-7	79.26	52.33	36.96
DWA-8	83.26	47.91	39.18
DWA-9	89.19	38.58	46.13

Table 3.2: Precision, Recall and AER for the different alignments. We show in bold the best results for each column.

The same results are summarized in the Figure 3.1. Here we get a better sense of the behavior of the alignments. First, for the DWA alignments, we get a balance between precision and recall with the DWA-3 alignment. However, the best AER is obtained by a slightly more precise alignment (DWA-5). For the GIZA alignments and the heuristically symmetrized alignment, we observe more precision than recall balance, the symmetrization certainly reduces the overall error rate.

3.3.2 Analysis of Alignment Structure

In addition to quality, other statistics related to the structure of the alignment were also computed. In this part of the study we also include the hand aligned data to have a better sense of which alignments are closer to the human generated data in terms of structure. The characteristics measured are:

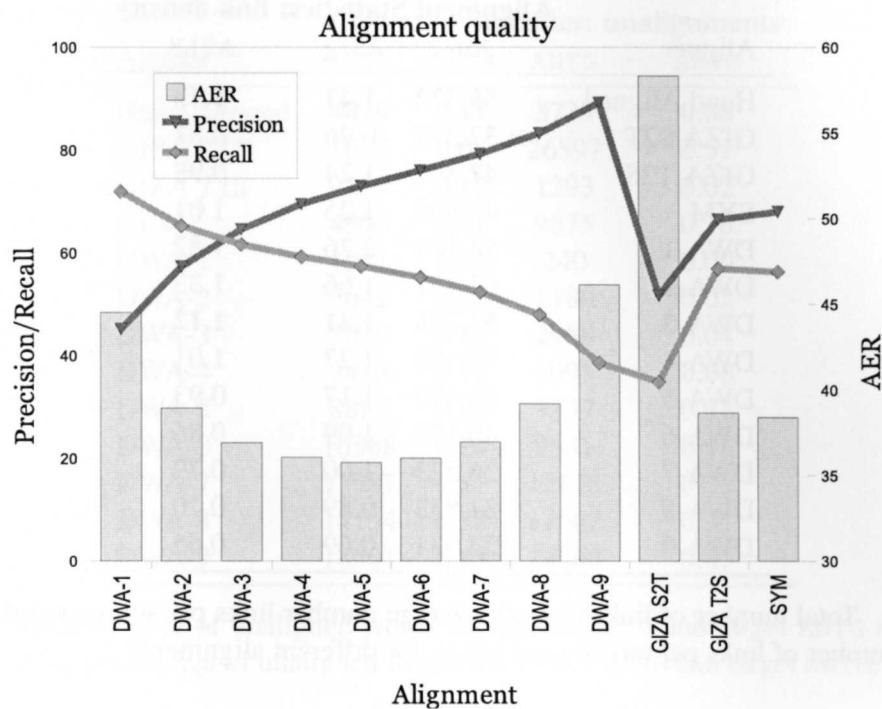


Figure 3.1: Precision, Recall and AER for the different alignments

- the total number of links ($ANLK$) in an alignment;
- the average density of an alignment, i.e. the average number of links per word in a sentence (source $ASLK$, or target $ATLK$).
- the number of unaligned words (gaps) in the source ($ANSG$), and target ($ANTG$), and their corresponding per word averages ($AWSG$, $AWTG$).

Table 3.3 displays the source and target densities of the alignments resulting from human and different automatic alignments. For the hand aligned data we see that on average, one English word is aligned to 1.13 Chinese words, while the reverse case is almost one and a half. This discrepancy can be explained due to the difference in sentence lengths of Chinese and English test data.

The GIZA alignments have the characteristic that in one direction each word is aligned exactly to one word in the other language (source and target change their role in different directions). Since some words, in our case 2-4%, are explicitly aligned to the NULL word, the density of links per proper word is slightly below 1 (0.96 and 0.98 for S2T and T2S, respectively). For the discriminative aligner the number of links decreases when the threshold is increased. The threshold DWA-3 gives a density closer to the hand-aligned data. This alignment has the best balance between precision and recall. Nevertheless as shown in Table 3.2 its quality is not the best. A higher threshold gives a sparse alignment which results in a higher precision. This makes evident that human alignments are denser than our best quality alignments. In other words, there are many “good links” that we are missing.

Aligner	Alignment Statistics: link density		
	ANLK	ASLK	ATLK
Hand Aligned	55,322	1.41	1.13
GIZA S2T	37,377	0.96	0.81
GIZA T2S	47,362	1.24	0.98
SYM	45,805	1.25	1.01
DWA-1	88,164	2.26	1.82
DWA-2	62,951	1.66	1.33
DWA-3	52,796	1.41	1.12
DWA-4	47,160	1.27	1.01
DWA-5	43,493	1.17	0.93
DWA-6	40,176	1.09	0.86
DWA-7	36,523	1.00	0.79
DWA-8	31,833	0.89	0.70
DWA-9	23,931	0.69	0.55

Table 3.3: Total number of links (ANLK), average number links per source word (ASLK) and average number of links per target word ATLK, for different alignments

3.3.3 Unaligned Words

More interesting yet, is to look at the summary of unaligned words in Table 3.4.

In general, there is a tendency to leave more Chinese words than English words unaligned. Note that the hand alignment has the most symmetric distribution (about 10% in both sides).

The GIZA alignments have the most striking disparity. While S2T leaves more than half of the English words unaligned, T2S leaves many Chinese words unaligned.

For the DWA case, we observe more Chinese words unaligned at first. But as we increase the threshold, the situation is reversed (remark DWA-9). The situation is better appreciated in Figure 3.2 where we can observe the increase in the number of unaligned words as we move from lower to higher thresholds.

Comparing the different alignments to the human generated alignment, we find that for the source side the symmetrized alignment is very similar to the gold standard on the number of words left unaligned. On the target side, the closest match is given by DWA-3 (totals) and DWA-4 (percentage). This shows that even our best quality alignments (AER-wise) are leaving too many words unaligned.

3.3.4 Summary

So far we have discussed several different statistics that can be used to describe alignments. They give us different perspectives on the nature of an alignment. And how does it compare to a human reference.

Aligner	Alignment Statistics: unalignments			
	ANSG	AWSG	ANTG	AWTG
Hand Aligned	4629	0.11	3739	0.08
GIZA S2T	1675	0.04	26597	0.51
GIZA T2S	9309	0.22	1293	0.02
SYM	4905	0.11	9675	0.16
DWA-1	1241	0.03	240	0.00
DWA-2	3642	0.07	1180	0.02
DWA-3	5676	0.12	2988	0.04
DWA-4	7418	0.16	5095	0.08
DWA-5	8882	0.20	7137	0.12
DWA-6	10368	0.24	9531	0.16
DWA-7	11987	0.27	12623	0.22
DWA-8	14154	0.32	17002	0.31
DWA-9	18591	0.43	24768	0.45

Table 3.4: Total number of unaligned words for source ANSG and target ANTG sides of the alignment. Also percentage of unaligned words for source AWSG and target AWTG.

- We presented the different measures of alignment quality for the alignments. We observed that we have a myriad of alignments in term of balance of precision and recall. However, we discovered that precision-oriented alignments prevail over their recall counterparts.
- We also presented two alternatives to measure the structure of the alignments in terms of the density of an alignment. First, we presented the link density for the different alignments. Then we measured the number of unaligned words. We discovered that the most-human like alignments in terms of density are not the better quality ones. Our better quality alignments trend to be less-dense than the human counterpart. This outlines that there are many links in the hand alignment that we are leaving out.

In the following section, we will analyze the output of the phrase-extraction algorithm using the analyzed alignments as input. Our objective is to determine which of the characteristics in the alignment might have more impact on the generated phrase-pairs.

3.4 Analysis of Phrase-Pairs

After generating symmetrized word alignments, the usual step in the pipeline is to extract phrase pairs. In the experiments described in this section, we used phrase-extract heuristic (Och and Ney, 2004) as implemented in the Moses package (Koehn et al., 2007), with a maximum phrase length of 7.

As opposed to word alignments, there is no gold standard human generated phrase table. While some metrics as CPER (Ayan and Dorr, 2006) have been proposed, they rely heavily in

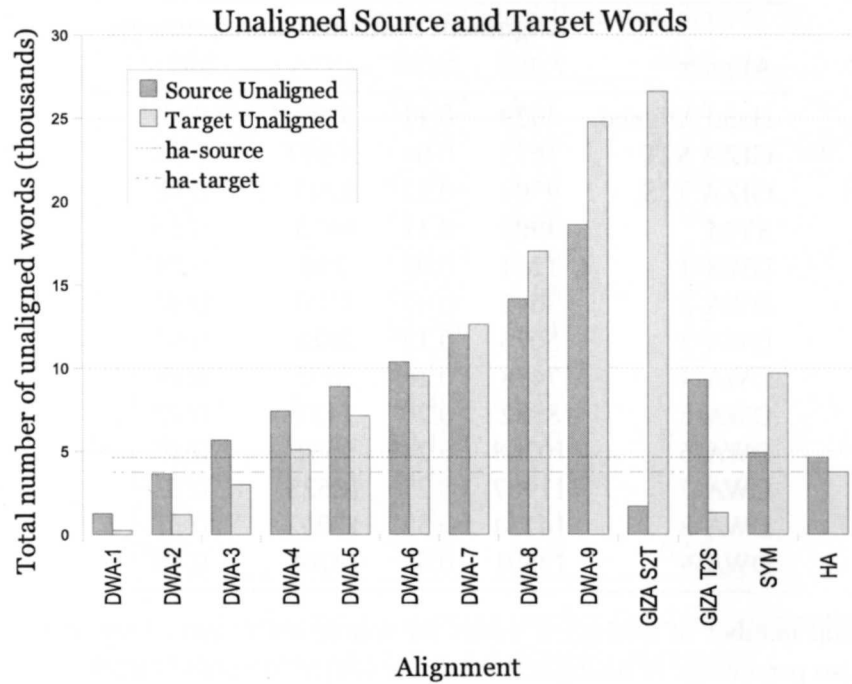


Figure 3.2: Total number of unaligned words by system.

the phrase-extraction algorithm to generate gold-standard phrase-pairs from a hand alignment. By doing so, the metric obscures the effect that the phrase-extraction heuristic may have on the quality of the phrase-table. In practice, measurements such as coverage, number of generated phrase pairs, size of the phrase table (i.e. unique phrase pairs), are used. In this study, we analyzed the following characteristics of the output of the extraction heuristic:

- The number of phrase-pair instances PNI generated by an alignment a .
- The percentage of singletons PNU , i.e. unique phrase-pairs.
- The average source PSL and target PTL phrase lengths per phrase pair.
- The average number of source PSG and target PTG “gaps” or unaligned words inside a phrase-pair.
- The per-word average number of unaligned words in the source ($PWSG$) and target ($PWTG$) sides of the phrases extracted from an alignment.

In the following subsections, we analyze the impact of the alignments in each of these characteristics of the phrase table.

3.4.1 Analysis of Number and Length of Phrase-Pairs

The number of generated phrases is an important characteristic of an alignment. The more unique phrase-pairs represent more translation options during decoding. However, the more

repeated phrases represent that our models will have smooth probabilities and be better estimated. Additionally, the longer our phrases are, usually means that we will be able to translate our input sentences using fewer phrases (which reduces the need for reordering). In this part, we start off by measuring these characteristics in the phrase-pairs generated by our alignment.

Aligner	Phrase Statistics: number & length			
	PNI	PNU (%)	PSL	PTL
Hand Aligned	111K	78.5	2.90	3.29
GIZA S2T	112K	87.6	2.13	3.38
GIZA T2S	90K	77.3	2.74	2.60
Symmetrized	136K	84.8	2.80	3.35
DWA-0.1	19K	65.4	2.28	2.38
DWA-0.2	51K	74.4	2.63	2.70
DWA-0.3	86K	79.0	2.81	2.92
DWA-0.4	126K	82.7	2.96	3.13
DWA-0.5	171K	85.8	3.13	3.33
DWA-0.6	231K	88.4	3.28	3.56
DWA-0.7	321K	90.8	3.45	3.79
DWA-0.8	467K	93.3	3.64	4.01
DWA-0.9	829K	95.7	3.91	4.31

Table 3.5: Different statistics for the phrase-pairs according to their length and number. We have the total number of phrase-pair instances (PNI), the percentage of singletons (PNU) and the average source (PSL) and target (PTL) phrase lengths

In Table 3.5, we summarize the statistics of the phrases according to their length, and number. The first piece of information that we observe is that as the DWA alignments gets sparser, the number of phrase-pairs increases steadily. Furthermore, the percentage of singletons also increases.

This is a result of the behavior of the phrase extraction algorithm. As the DWA alignments become less dense, the number of phrase-pairs that are consistent with the alignment increases. This is similar with the results reported by (Ayan and Dorr, 2006), where they found that the size of the phrase table increases dramatically as the number of links in the initial alignment gets smaller. However not all the alignments exhibit the same behavior. For example, take DWA-7 and GIZA-S2T alignments. They have about the same number of links. Nonetheless, the number of generated phrase-pairs is almost three times larger for DWA-7 than for the S2T. Instead, the number of phrases generated seems to be influenced by an interaction between the number of links and the number of unaligned words.

This result is more evident when we look at Figure 3.3. In this graph, we overlay the total number of unaligned words (source and target) on the bar graph for the total number of extracted phrases for each alignment. Observe that there is an interaction between the number of unaligned words on the source and target side of the alignment and the number of extracted phrases. While this rule is not perfect, it describes much better the quantity of

extracted phrases than the number of links.

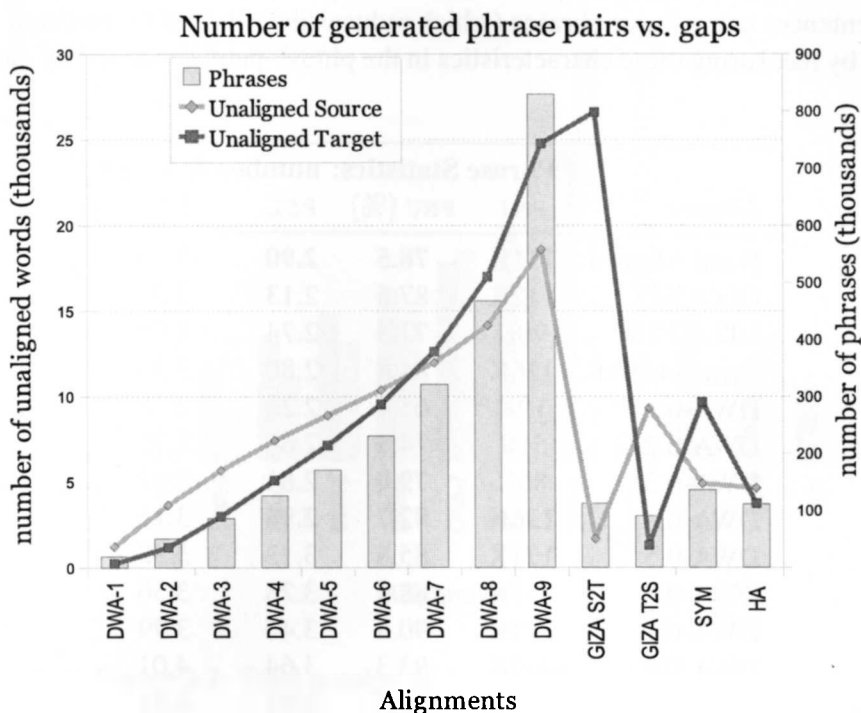


Figure 3.3: Distribution of length of the source phrases extracted from different alignments.

Also remark that as we increase the sparsity of our alignments, the number of unique phrase-pairs also increases, but not nearly at the same rate. In Figure 3.4 we overlay the graphs of the two statistics. Observe how the number of extracted phrases increases quasi-exponentially, while the growth in the unique phrases is more damped, almost logarithmic.

This showcases the impact of sparsity (more specifically, unaligned words) in the composition of our translation model. The more sparse is our alignment, the more phrase pairs we will be able to obtain. Unfortunately, many of the entries in our phrase-table will be seen only once (unique), which will result in a un-smoothed estimation.

Another interesting piece of information is the distribution of lengths of the extracted phrase-pairs. We observe that as an alignment gets sparser the number of short source phrases remains about the same. However, the phrase extraction algorithm is able of finding longer phrases; many of which include a larger number of unaligned words, as we will show in the following subsection.

3.4.2 Analysis of Phrase Alignment Gaps

In Table 3.6, we summarize the gap statistics for the phrase-pairs extracted from the different alignments. Observe that the average number of gaps in both source and target sides of a phrase-pair increases as the sparsity of an alignment increases. For instance, most of the phrase-pairs of most-dense alignment (DWA-1) are gap-less (90% for source and 98.9% for

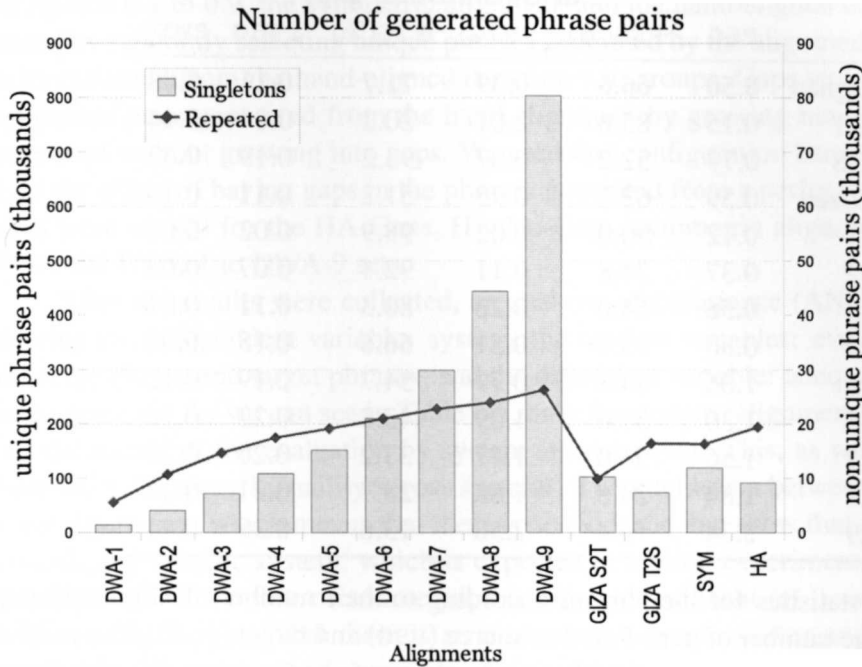


Figure 3.4: Number of extracted phrase pairs vs. unique phrase pairs per system

target). In contrast for the DWA-9, the gap-less phrase pairs for source and target side account for 16.4% and 13.6% respectively.

Notice also that the number of gap-less phrase-pairs tends to be higher in the English side, than in the Chinese side (not taking into account GIZA alignments). In fact, when we look at the average per-word number of phrase alignment gaps (PWSG, PWTG), we observe that is extremely close to the per-word average of unaligned words of the alignments (AWSG, AWTG). In fact, the correlation between these two statistics is very high (0.93 for source and 0.94 for target) suggesting that the distribution of unaligned words in our alignment carries into the phrase-pairs even after phrase-extraction.

3.4.3 Summary

In this section, we have observed how the structure of the different alignments has an impact on the phrase-pairs that are extracted from them. First, we observed that the number of extracted phrase-pairs is related to the number of unaligned words in the generating alignment. Furthermore, we observed that many of the generated phrase-pairs are unique, and tend to be longer.

Furthermore, we discovered that the average per-word number of alignment gaps are preserved after phrase extraction. The correlation between the gap-measuring variables is very high. Therefore, we wonder if the gaps in an alignment also affect the quality of a phrase-pair. In the next section, we present a study in which we evaluated the human-assessed quality of phrase-pairs.

Aligner	Phrase Statistics: gaps					
	PSG	\neg PSG(%)	PNTG	\neg PTG(%)	PWSG	PWTG
Hand Aligned	0.50	66.6	0.39	71.7	0.11	0.08
GIZA S2T	0.15	85.6	2.01	20.2	0.04	0.44
GIZA T2S	0.79	52.2	0.07	93.2	0.19	0.02
Symmetrized	0.59	62.3	0.95	51.2	0.11	0.16
DWA-0.1	0.12	90.0	0.02	98.9	0.02	0.00
DWA-0.2	0.37	74.8	0.11	92.4	0.07	0.02
DWA-0.3	0.56	64.0	0.26	80.3	0.11	0.05
DWA-0.4	0.80	53.7	0.51	66.6	0.15	0.08
DWA-0.5	1.06	44.8	0.79	54.7	0.19	0.12
DWA-0.6	1.30	37.6	1.09	43.8	0.22	0.16
DWA-0.7	1.56	31.0	1.47	33.2	0.26	0.22
DWA-0.8	1.84	24.8	1.88	23.7	0.31	0.29
DWA-0.9	2.34	16.4	2.56	13.6	0.39	0.41

Table 3.6: Different statistics for the phrases according to their number of alignment gaps. We display the average number of gaps found in source (PSG) and target (PTG) phrases, along with the percentage of phrases extracted without into any gap (\neg PSG, \neg PTG). We also present the per-word average of source and target gaps (PWSG,PWTG)

3.5 The effect of Alignment Gaps in Phrase-Pair Quality

As we have seen in the previous analysis, the number of unaligned words of an alignment has a huge impact on the number of phrase pairs extracted. As we observed, the phrase extraction heuristic allows to generate phrase pairs by growing into gaps. Some of these gappy phrase pairs will be useful. They will increase coverage, and actually might be accurate phrase pairs. However, many other phrase pairs will be partially, if not completely wrong. To investigate how the quality of the extracted phrase pairs depends on the type of the underlying word alignment, and on the gaps of the phrases extracted from these alignments, a small-scale human evaluation was conducted. Several native Chinese speakers participated in this evaluation.

The procedure was the following: Each subject was presented with a set of Chinese-English phrase pairs. For each phrase-pair they judged if source and target phrase were adequate¹ translations of each other. This was done without any other contextual information (i.e. the surrounding words, or the sentence pairs from where these phrases were extracted). This was also done blindly, as the evaluators did not have any knowledge of the origin of the phrase-pairs. Furthermore, we included a noisy set, which were pairs of randomly selected source and randomly selected target phrases. Such noisy set would help us to determine how likely is to obtain a good score by just having a random pair of source and target phrases.

The phrases included are the ones generated by the alignments from the DWA with

¹By adequate, we mean that a source phrase could be used as translation of a target phrase in at least one situation, without loss of meaning

thresholds 0.1 to 0.9, the symmetric alignment and the hand-aligned data. Each set was generated by randomly selecting unique phrases generated by the alignments. We split the phrase pairs extracted from the hand-aligned data into two groups, Gaps and No-Gaps, which stand for phrase pairs generated from the hand alignment by growing into gaps, and phrase pairs generated without growing into gaps. We used this configuration because we wanted to highlight the effect of having gaps in the phrases generated from a perfect alignment. The sample sizes were of 100 for the HA-Gaps, HA-No-Gaps, symmetric alignment and noisy sets and 50 for the DWA-1 to DWA-9 sets.

After the results were collected, an analysis of covariance (ANCOVA) was done, considering the independent variable: system, the random variables: evaluator, number of gaps in source phrase and target phrases; and the dependent variable: adequacy, with non-repeated measurements. As we can see in Table 3.5, only the system (alignment) is a significant factor, i.e. the means of the evaluation by system are not equal. This, as we expected, means that there are differences in quality across systems. The interaction between system and evaluator is not significant, which means that there is no evidence that show that evaluators were biased towards any specific system, which is expected in a blind experiment. Also note that while the effect of the number of source gaps is almost significant (at $\alpha = 0.01$, there is strong evidence that suggests that there is an interaction between source and target gaps. In other words, looking at the gaps in one side of a phrase-pair may not tell us much about its quality. However, the combination of source and target gaps might be a good indicator.

Source	SS	df	MS	F	p-val
EV	0.04	2	0.02	0.15	0.8556
SYS	13.92	8	1.74	11.48	0.0000
SG	1.76	2	0.88	4.42	0.0138
TG	0.74	2	0.37	1.77	0.1745
EV*SYS	6.13	40	0.15	1.01	0.4518
SG*TG	9.24	29	0.31	2.10	0.0007
Error	113.99	752	0.15		
Total	196.66	849			

Table 3.7: ANCOVA table showing the effects in the experiment: Evaluator (EV), System (SYS), Number of Source Gaps (SG), Number of target Gaps (TG). Also two-way interactions are shown for Evaluator*System, Source Gaps*Target Gaps.

In Figure 3.5 we show the mean of the evaluation by system². As expected, random phrase-pairs perform poorly. This verifies the consistency of the judges evaluation as good scores could not have been achieved randomly. Surprisingly, the phrase pairs extracted from DWA-1 achieved the highest score. While DWA-1. phrase are nearly all gap-less, the underlying word alignment is far from perfect. Furthermore, comparing phrase pairs extracted from human word alignment, shows that a perfect word alignment does not lead to perfect phrase-pairs given the current extraction heuristic. However, the HA-no-gaps set performs better than

²The confidence intervals are merely informational. To determine statistical differences, one must perform unplanned pairwise comparisons such as Scheffé tests

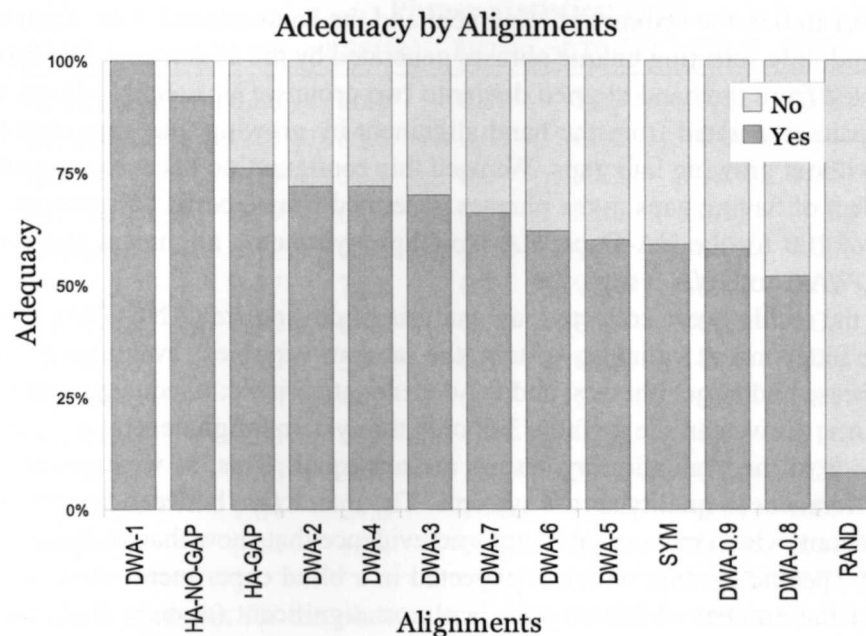


Figure 3.5: Phrase pair quality evaluation results

the HA-gaps set. This suggests that the quality of a phrase-pair extracted from a gold standard hand alignment deteriorates when it has gaps on either its source or target phrases. Overall, we do observe the tendency that less number of unaligned words in the word alignment leads to better quality of the extracted phrase pairs. In other words low precision/ high recall alignment results in fewer but higher quality phrase pairs. To balance the trade-off between higher quality phrases and coverage, we conducted a series of translation experiments where the number of unaligned words was taken as a feature. They are described in detail in Chapter 6.

3.6 Conclusions

In this chapter we studied in detail the relation between word alignment and phrase extraction. First, we analyzed word alignment according to several characteristics and compared them to hand-aligned data. We observed that there is a lot of room for improvement for our alignment models. Second, we analyzed the phrase-pairs generated by these alignments. We observed that sparser word alignments lead to a higher number of extracted phrases, which ultimately results in larger phrase tables. While these larger phrase tables contain longer phrases, many of the phrases contain unaligned words. Also, the number of unaligned words in the alignment have a large impact on the characteristics of the extracted phrase table. The unaligned words in the extracted phrase pairs follow the distribution of unaligned words in the alignment from where they were extracted. Third, a manual evaluation of phrase pair quality showed that the

more unaligned words (gaps) result in a lower human perceived quality.

Chapter 4

The Effects of Alignments on the Phrase-Based Translation Model

In the previous chapter, we observed how several alignment density characteristics (link density, source and target gaps) influence the phrase-extraction process. We performed an exploratory study that allowed us to get a sense of the importance of certain features, specially the distribution of unaligned words. We detected a strong influence of alignment gaps in the number, length and quality of extracted phrase-pairs.

In this chapter, we revisit the problem with a different focus. First, we study the translation model also known as phrase-table (as opposed to only phrase-pairs). The main difference is that in a translation model, each phrase-pair is unique and has a series of translation probabilities assigned.

Second, we make use of more variables in this analysis. In addition to alignment density, we, we include certain distortion variables, such as the proposed by (Lambert et al., 2010). For the phrase-table, we also measure the entropy of the translation model features.

Finally, we go past the exploratory study by building regression models to help us predict the most important phrase-table variables using alignment characteristics as input. We test the regression models' robustness against unseen data.

This chapter is an enhancement of the study presented before, in Chapter 3 and is tightly integrated to the analysis developed later in Chapter 5. There, we address how machine translation quality is related to several characteristics of the translation model. The remainder of this chapter is organized as follows: first we introduce the translation models and define the translation probabilities. Then, we define the setup of our experiment. Next, we perform a correlation analysis of the alignment variables and phrase-table variables to develop a sense of the relationships between the variables. Afterwards we build more fine-grained regression models and test their robustness to predict unseen data. We wrap up with a discussion of our findings.

4.1 The Phrase Translation Model

In the previous chapter we observed how we could extract phrase-pairs from an alignment. The following step in the training procedure after phrase-extraction is known as scoring. Its

objective is to create a translation model, otherwise referred as a phrase-table. This process consists in calculating different translation probabilities for each phrase-pair.

In Figure 4.1 we show a snippet of a phrase table. In addition to the English and Spanish phrases, we observe the (optional) alignment information, as well as four translation probabilities which we introduce below.

```

" familia " que figura |||
" family " |||
(0) (1) (1,2) () (2) ||| (0) (1,2) (2,4) |||
0.0140845 2.02353e-07 1 0.0163321 2.718

" familia " que se encuentra dentro |||
" family " that falls within |||
(0) (1) (2) (3) (4) (4) (5) ||| (0) (1) (2) (3) (4,5) (6) |||
1 2.03553e-06 1 1.18114e-05 2.718

" familia " que se encuentra |||
" family " that falls |||
(0) (1) (2) (3) (4) (4) ||| (0) (1) (2) (3) (4,5) |||
0.5 9.75054e-06 1 1.90139e-05 2.718

```

Figure 4.1: Snippet from a phrase table including alignment information. For display purposes each entry of the phrase table appears broken into four lines. The first and second lines correspond to the phrasal translations (phrase-pairs). The third line corresponds to the alignment information (source to target and target to source). The fourth line corresponds to the feature values for each phrase-pair.

4.1.1 Translation and Lexical Probabilities

Traditionally, there are two types of probabilities that are embedded in a phrase-table. We present them in 4.1.

ID	Variable	Description
PT1	$p(f e)$	Inverse phrase-translation probability
PT2	$p_{lex}(f e)$	Inverse lexical probability
PT3	$p(e f)$	Direct phrase-translation probability
PT4	$p_{lex}(e f)$	Direct lexical probability

Table 4.1: The four different translation probabilities included in a phrase-based translation model

On one hand we have the phrase-translation probabilities $p(f|e)$ and $p(e|f)$, which are the maximum likelihood estimates (MLE) of the phrase-pair likelihoods based on a frequency (counts) approach.

Thus the direct probability $p(e|f)$ of a foreign phrase f of being translated into and English phrase e is determined by:

$$p(e|f) = \frac{\text{count}(e, f)}{\text{count}(f)} \quad (4.1)$$

where $\text{count}(e, f)$ represents the number of times that phrase e and phrase f appear in the same phrase-pair, while $\text{count}(f)$ represents the number of times the foreign phrase f appeared in any phrase-pair. The similar criterion is used for the inverse $p(f|e)$ phrase translation probability.

Lexical translation probabilities p_{lex} (also known as lexical smoothing) measure the translation probability in a per-word basis. Since they are calculated over the whole vocabulary using alignment information, they tend to be smoother than phrase-translation probabilities. For more details on how these probabilities are calculated, please refer to Appendix B.

4.2 Experimental Setup

In this chapter, we analyze different alignments and the phrase-tables extracted from them. In addition to the alignment quality and density variables we discussed in the previous chapter, in this chapter we included distortion and dimension variables. In Table 4.2, we present each variable used in this study along with a brief description. For a thorough discussion of how each variable is calculated, please refer to Appendix B.

In the following part, we present the data we used for our study. First, we introduce the datasets and sampling techniques used. Later, we introduce the different alignment procedures employed.

4.2.1 Data and Sampling

In this study, we evaluate alignment quality and structure as characteristics of the alignments and measure their implications in translation model estimation. To measure quality, we require to use human labeled data. Unfortunately the availability of human labeled data is limited. Moreover, to make reliable estimations of the phrase-translation probabilities in a translation model we require a considerable amount of word aligned training data. Hence, we resorted to the sampling-with-replacement (bootstrapping) technique which is commonly used for obtaining smooth distributions when sampling data is scarce. For instance, this technique has already been applied for estimating confidence intervals for machine translation quality (Zhang and Vogel, 2004; Guzman and Garrido, 2008).

The initial sampling pools were populated with the datasets displayed in 4.3. In our experiments, we randomly sampled with replacement 1000 word alignments for each of the aligners. We obtained 30 of such subsamples for our Spanish-English training set and 10 subsamples of the same size for each of the aligners in our Spanish-English, Arabic-English and Chinese-English test sets.

	Alignment variables		Phrase Table variables	
	var id	description	var id	description
Alignment Quality	F	F-Score		
	P	Precision		
	R	Recall		
Alignment Density	ASG	A. Source Gaps	PSG	P. Source Gaps
	ATG	A. Target Gaps	PTG	P. Target Gaps
	ALK	A. Link Dens	PLK	P. Link Dens
Alignment Dimension	ASL	A. Source len	PSL	P. Source len
	ATL	A. Target len	PTL	P. Target len
Alignment Distortion	ACR	A. Crossings	PCR	P. Crossings
	ADT	A. Rel. Distort.	PDT	P. Rel. Distort.
	ADG	A. Diagonality	PDG	P. Diagonality
Translation Model Features			PT1	Avg entropy $p(f e)$
			PT2	Avg entropy $p_{lex}(f e)$
			PT3	Avg entropy $p(e f)$
			PT4	Avg entropy $p_{lex}(e f)$
Phrase Table entries			PSU	srcUnique
			PTU	tgtUnique
			PNE	Entries

Table 4.2: Alignment and phrase-table variables considered in this study

Language pair	Source data	Size
Spanish-English	EPPS test	400
Chinese-English	GALEP3 test	2000
Arabic-English	MT03 test	2552

Table 4.3: Hand alignment datasets

Systems

To have more diversity in our samples, we different aligners for each of our language pairs. On the discriminative side, we used the discriminative aligner by (Niehues and Vogel, 2008) and filtered to relatively balanced density-thresholds (0.4,0.5,0.6,0.7).

On the generative side, we used the GIZA(Och, 2000) source-to-target and target-to-source aligners symmetrized using the following heuristics: grow-diag, grow-diag-final, and grow-diag-final-and. Not only these heuristics are a standard (grow-diag-final is the default for the Moses (Koehn et al., 2007) decoder), but also the three of them provide a nice differentiation in terms of precision-recall balance. For instance, grow-diag-final yields high-recall, denser alignments, grow-diag provides high-precision sparse alignments and grow-diag-final-and produces a more balanced balanced alignment.

In summary we obtained 210(30x7) subsamples of Spanish-English alignments for the estimation of our regression models and 70(10x7) subsamples for each of the Spanish-English, Arabic-English and Chinese-English test sets.

Phrase-Tables

For each of the alignment samples, we obtained their corresponding phrase-tables using the phrase-extraction and phrase-scoring algorithms readily provided in the Moses package (Koehn et al., 2007). Furthermore, we measured the alignment variables of the samples and their corresponding phrase-tables and used their subsample average as input data.

The overall setup of our experiment is depicted in Figure 4.2.

4.3 Correlation Analysis

The first step in our analysis was to determine which were the variables that are the most correlated among each other. While paired correlations give us an idea of ‘relatedness’ between variables, a full correlation matrix shows us the big picture. The main objective is to see which variables relate with each other before performing regression. This helps to alleviate some problems that can be caused by multicollinearity.

To simplify this analysis, we clustered the variables by similarities. In Figure 4.3 we observe a correlation map of the alignment and phrase-table variables grouped by similarity (positive correlation) to allow an easier analysis. The procedure to generate such map is to simply run a k-nearest-neighbor clustering algorithm.

For an easier interpretation, we annotated the major clusters and labeled from 1 to 6. Below, we explain each one of them.

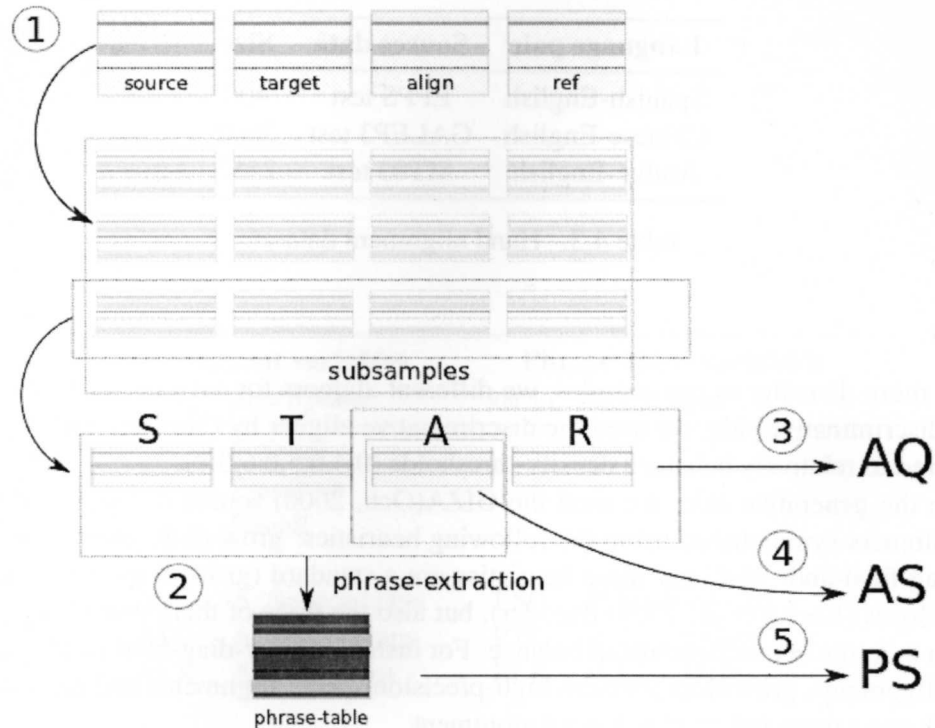


Figure 4.2: In step 1, we resample the original source and target sentences along with the computer and human generated alignments. Then we extract a phrase-table for each of these subsamples. We then evaluate the alignment quality (AQ), alignment structure (AS) and phrase-table alignment structure (PS).

C1 Diagonality, phrase-distortion and phrase-length

In the first cluster, we observe a clear correlation between the alignment diagonality ADG , the average source phrase length PSL and to a lesser extent the average target length PTL . The interpretation is straightforward. The more ‘diagonal’ or monotone our alignments are, the less trouble the phrase-extraction algorithm will have extracting longer phrases. Additionally we observe a relationship between the relative alignment distortion at the phrase-level PDT and the alignment diagonality ADG . This is surprising given that alignment diagonality and relative distortion are negatively correlated. This means that the more alignment distortion we have in our alignment, the less alignment distortion we can expect in our phrase-table PDT . The explanation for this is that the more distortion we have in our alignment, the phrase-extraction will find more restrictions and will in turn extract only phrases in regions of the alignment where there is lower distortion. Thus the resulting phrases will have lower relative distortion. Cluster $C1.b$ indicates that there is a strong relationship between $C1$ and $C4$ which also consists of alignment distortion variables (alignment relative distortion ADT and alignment crossings ACR) which indicates that alignments that have high diagonality also have lower values of distortion. In addition, alignment density clusters $C2, C6$, are to a lesser extent, also related to this cluster, which means that alignments that have high density also trend to have more distortion.

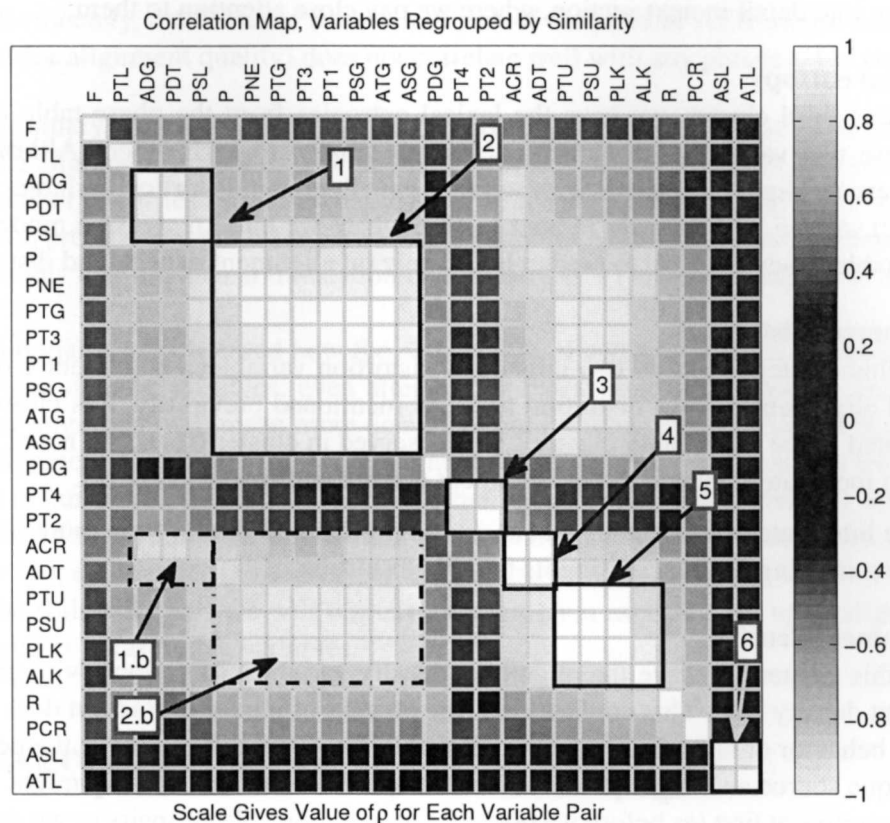


Figure 4.3: Correlation map for different variables. In this figure, the lighter colored squares indicate a strong correlation (either positive or negative) of the two variables in the respective row and column. The darker colors indicate a lack of relationship between the variables. Major clusters are labeled from 1 to 6.

C2 Sparsity

In this second cluster, we observe a large number of variables. Many of them represent alignment sparsity, as measured by source and target gaps in both the alignment and phrases (PSG, PTG, ASG, ATG). This reiterates the observations we made previously: the sparsity of an alignment remains even after phrase-extraction. Otherwise said, the phrase extraction algorithm does not modify alignment sparsity.

Additionally, in this cluster we have the alignment quality metric precision P which confirms the intuition sparser alignments trend to be more precise (this is especially true for discriminative alignments, as we observed in 3).

In this cluster, we also have three important translation model variables: the number of entries in a phrase-table PNE and the direct and inverse phrase-translation entropies ($PT3$ and $PT1$). On one hand, we confirm what we had observed previously in (Guzman et al., 2009), where we observed that sparser alignments yield a higher number of phrase-pairs. On the other hand, we observe that sparser alignments produce less determined, more ambiguous phrase-translation probabilities. We will discuss these relationships

more into detail in next section, where we pay close attention to them.

C3 Lexical entropy

In this third cluster, we have the lexical entropies from the phase-table (PT2, PT4). These two variables are almost exclusively related to each other. Although we can observe a slight correlation with some of the alignment distortion variables (ADG, ACR, ADT). As we will discover in next section where we build regression models for these variables, they are hard to predict based only on alignment density and distortion.

C4 Alignment distortion

In this cluster, we have two alignment distortion variables: alignment crossings ACR and alignment relative distortion ADT. As mentioned previously, this cluster is highly related to the diagonality cluster C1 as observed in cluster C1 . b. As noted before, it is also moderately related to the alignment density and sparsity clusters.

The interpretation for this is that as our alignment gets denser, it is more likely to have more crossings and increasing its relative distortion.

C5 Alignment density

In this cluster, we have the alignment density variables ALK along with phrase alignment density PLK. This confirms that the process of phrase-extraction does not modify the behavior of alignment density. Related to these variables we have the percentage of unique source and target phrases (PSU, PTU). While this relationship might seem counterintuitive at first (as before we observed more unique phrase-pairs as we decreased the alignment density), this relationship is explained in detail in the next section when we observe that alignment gaps are responsible for such behavior.

In this cluster, we could also include the alignment quality metric recall R. This, again is related to the fact that the more links we have in our alignment, the more likely we are to cover good links in the alignment reference.

Finally, remark one more time, how there is a strong relationship between the density variables and the sparsity variables (gaps) which measure different ends of the same spectrum. This is visible from cluster C2 . b.

C6 Sentence lengths

This last cluster brings together the average source and target lengths of the aligned sentences (the dimensions of the alignment matrix ASL, ATL). We can observe that these two variables do not relate to any other variable. Remark that they do not influence in any measure the phrase lengths PSL and PTL. This is an artifact of the extraction heuristic being limited to a certain max-phrase-length (in this case 7). Additionally, phrases are limited by link-related restrictions more than to the total length of the original sentences.

Alignment quality and distortion

While precision P and recall R are related to density variables, which is understandable given their nature, the F-metric F is only moderately related to alignment distortion. This means that the more monotonic our alignment is, the better the alignment quality.

Unfortunately, the F-metric (the counterpart of the popular AER metric used as a standard for alignment quality) does not correlate well with any phrase-table characteristic.

In summary, we have a high degree of correlation between phrase-table variables and many of the alignment variables, specially the alignment density and ones. For some variables like the number of entries in a phrase table and average phrase source length, we would expect to have good prediction models for phrase-table variables based on alignment information. For other variables like the lexical translation entropies PT_2 , PT_4 , we can expect to have weaker models.

Additionally, we observed how the distribution of certain variables remains unaffected for alignment density variables, even after phrase extraction. More density and alignment gaps in our alignment will mean more density in our phrases. The inverse can be said for the relative alignment distortion ADT . More alignment distortion ADT will result in less alignment distortion in our phrase tables PDT . For the other variables, this relationship is not kept. The phrase-extraction modifies the distribution of the source and target lengths (ASL and ATL vs. PSL and PTL), diagonality (ADG vs. PDG) and number of crossings (ACR vs PCR).

In the following part we will discuss the regression models built upon alignment variables to predict phrase-table variables.

4.4 Regression Models

Up to now, we have corroborated many of the intuitions we developed in Chapter 3. For instance, we have observed that alignment density and sparsity have a large contribution to how our phrase-tables are built. We have also developed an intuition on how alignment distortion affects other translation model variables. In this section we go beyond an exploratory study. We use multivariate regression as a tool to predict different phrase-table variables using the alignment characteristics as predictors. As we will discuss in Chapter 5 some of the phrase-table variables that we have selected for this part also show a strong influence the translation performance.

Furthermore, we use unseen data to measure the generalization of the regression models built. In addition to Spanish-English data, we also incorporate Arabic-English and Chinese-English alignments.

For a simpler analysis, we have divided this section into three parts. In the first part we discuss the effects of alignment variables into the size and percentage of unique entries in a phrase-table. Second, we discuss the effects of alignments in the average source and target phrase length. Last, we observe the relationship between the alignment density and the entropy of the phrase-table translation probabilities.

4.4.1 Entries in the Phrase-Table

The size of a phrase-table is an important factor for machine translation. The more entries we have available give us more diversity of translation options. Unfortunately it is often found that many of these entries consist of pure noise. Rare phrase pairs (which are seen just a few times during training) fall into that category. Another way of measuring the diversity of a

phrase-table consist on the percentage of unique source or target phrases. Large phrase-tables with low percentage of unique phrases will have more smooth probability distributions.

In Table 4.4 we show the standardized coefficients of the regression models using the most significant predictors. We present the R^2 results for the training (estimation) sample, as well as the unseen test sets for Spanish-English, Arabic-English and Chinese-English.

Variable	Regression coefficients				Determination coefficients (R^2)			
	ASG	ATG	ALK	ADG	train	test-es	test-ar	test-ch
PNE		0.510	-0.420	0.103	0.973	0.973	0.966	0.893
PSU		-0.976			0.953	0.959	0.923	0.983
PTU	-0.974				0.950	0.953	0.974	0.928

Table 4.4: Regression results for entries in a phrase-table

First, remark how only four variables are enough to predict the phrase-table sizes and percentage of unique variables. With these four variables we are able to determine at least 96% of the variance in the Spanish-English test set, a minimum of 90% of the variance for the Arabic-English test set and 87% of the Chinese-Test set. Below we present the analysis for each individual phrase-table variable.

Number of entries in a phrase-table

We observe that the total number of phrases depends on the number of target gaps in a positive way. The more gaps we have in our alignment, the more entries in our phrase table. In a similar way, the link density has a negative effect. Thus, the more dense our alignment, the more extraction restrictions will be presented, which in turn will reduce the number of phrase-pairs that can be extracted. Finally, diagonality plays a positive role. Thus, the more monotone our alignment is, the more phrases we will be able to extract. This phenomenon is depicted in Figure 4.5.

Unique phrases

In the correlation analysis of the previous section, we observed that the percentage of unique phrases had a positive correlation with alignment link density. However, using regression, we found that they are more dependent upon the alignment gaps. Furthermore, the percentage of unique phrase for source and target have a symmetrical dependency upon the number of gaps. For the source unique, we observe that the number of target gaps have a negative effect. This is because having more target gaps allows the same source phrase to grows into target gaps without violating any restrictions. Thus as the number of extracted phrases increases, the number of unique source phrases remains the same. This lowers the percentage of unique source phrases. The same effect occurs with the target gaps.

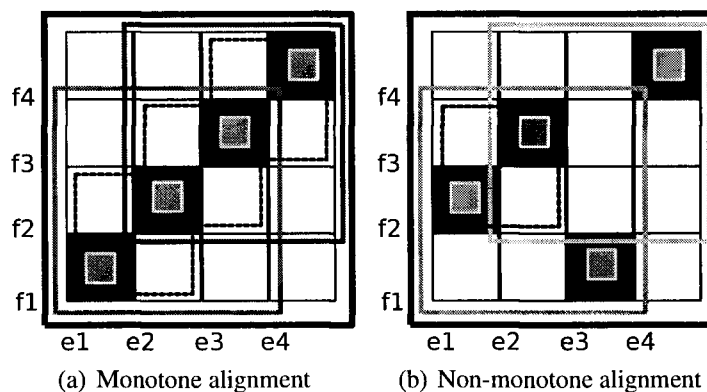


Figure 4.4: Comparison of number of generated phrase-pairs by monotone and non-monotone alignments. In the first case (a), the diagonality of this alignment is 1.0 and we have ten generated phrase pairs. Four are of length 1x1, three of length 2x2 and two of length 3x3 and one of length 4x4. On the second case (b) we have a diagonality of 0.5 while the number of generated phrases is only 8 (4+1+2+1).

4.4.2 Length of Phrases

Average phrase length is an important characteristic of a translation model. Having longer source phrases allows us to translate an input sentence using longer (thus fewer) phrases. This has the benefit that local reorderings (i.e. word swapping) are encapsulated inside the phrase. Additionally, using fewer phrases to translate give us a phrase-translation boost. In other words, by using fewer phrases to cover the input sentence, the cost of translation in terms of reordering and phrase-counts can be reduced.

The target side phrase-length is not as important as the source counterpart. During decoding, translation hypothesis length is taken care of by a word count penalty, which as its name suggests, penalizes each word in our translation (to prevent too long translations). Therefore, the role of the average target side phrase-length is (as we corroborate in next chapter) not as relevant.

In Table 4.5, we present the results for our regression models. For source phrase-length PSL we observe that for the Spanish-English samples, about 85% of its variance can be explained with two variables: the link density ALK and the alignment diagonality ADG .

Variables	Regression coefficients		Determination coefficients (R^2)			
	ALK	ADG	train	test-es	test-ar	test-ch
PSL	-0.389	0.627	0.854	0.840	0.925	0.785
PTL	-0.432	0.385	0.546	0.541	0.906	0.676

Table 4.5: Regression coefficients for phrase-table phrase length

Furthermore, we observe that for both variables (PSL and PTL) the contribution of link density is negative while the contribution of the diagonality is positive. The explanation for

this phenomenon is direct. Having more link density increases the restrictions for phrase-extraction. Thus, we will be able to extract fewer long phrases because restrictions make harder to extract them.

On the other hand, having more monotone/diagonal alignments will allow us to extract longer phrases because monotone alignments present fewer restrictions. This is exemplified in Figure 4.4.

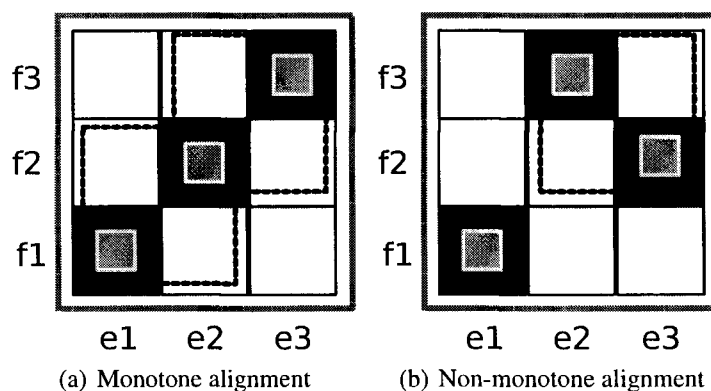


Figure 4.5: Phrase length and monotonicity. Having more monotone allow for longer phrases. On the left, the diagonality is 1.0 while the average source length is 1.66. In the second case, the diagonality is 0.5, while the average source length is 1.6

Finally, remark that the target-phrase length prediction model achieves lower determination. While this is not the best model we found, it is the simpler one. Other models excel in predicting PTL for Spanish-English, but fail to predict for the Arabic-English or Chinese-English and vice-versa. Therefore, we show the one that best represented every sample.

4.4.3 Translation Entropies

The probabilities in a phrase-table determine in a large measure which phrase-pairs will be selected during decoding. Low probabilities increase the translation cost and reduce the likelihood of a phrase-pair to make it to the final translation. For our study, we used the average entropy as a measure of the uncertainty of the model.

Measuring entropy

In Information Theory (Cover and Thomas, 1991; Manning and Schütze, 1999), the entropy of a random variable is measure of uncertainty. When the entropy is large, the uncertainty about the value of a random value is large. The entropy of a discrete random variable is defined by:

$$H(X) = - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (4.2)$$

Entropy is always positive since $0 \leq p(x) \leq 1$ and it is assumed that $0 \log 0 = 0$. Entropy is a concave function with a maximum at the uniform probability value (i.e. $p(x) =$

$1/|\mathcal{X}|; \forall x \in \mathcal{X}$). Thus, there is more entropy where all events are equally likely to happen (more uncertainty).

When dealing with conditional distributions it is customary to measure the conditional entropy instead.

$$H(Y|X) = - \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = - \sum_{x \in \mathcal{X}} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \quad (4.3)$$

We can think of conditional entropy as the weighted average of the entropy of $p(y|x)$ over all events $x \in \mathcal{X}$. In our case, we use the phrase-translation probabilities in place of $p(y|x)$ but also assume a uniform distribution over x (i.e. $p(x) = 1/|\mathcal{X}|$), which is equivalent to say that instead of using a weighted average, we simply use the arithmetic average. Thus, the average entropy we measure is:

$$\bar{H}(F|E) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \sum_{f \in \mathcal{F}} p(f|e) \log p(f|e) \quad (4.4)$$

In other words we are calculating entropy of $p(f|e)$ for each of the English phrases e and averaging it across the space of all English phrases \mathcal{E} .

We perform this operation for each of the four translation probabilities presented in Table 4.1.

Regression models

We used the alignment characteristics to predict the average entropy of the translation model probabilities. In Table 4.6, we present the most significant predictors along with the R^2 for each model on unseen data.

Variable	Regression coefficients			Determination coefficients (R^2)			
	ASG	ATG	ADG	train	test-es	test-ar	test-ch
Phrase translation entropies							
PT1	0.978			0.957	0.960	0.992	0.917
PT3		0.975		0.951	0.962	0.992	0.963
Lexical translation entropies							
PT2			-0.693	0.480	0.430	0.907	0.584
PT4			-0.580	0.336	0.266	0.505	0.604

Table 4.6: Regression coefficients for translation model entropies

Phrase-translation entropies

Similarly to what we observed in the correlation analysis, both phrase-translation entropies $PT1, PT3$ have a strong influence from the alignment gaps. This behavior is symmetrical to our observation in the percentage of unique phrases. For instance, $PT1$ has a strong positive influence from the source gaps ATG in the same fashion that for

the percentage of unique target phrases PTU receives a negative influence. Roughly speaking, our interpretation of these results is that the more gaps we have on source alignment, we will have more unique source phrases while the number of target gaps stays the same, lowering the percentage of unique target phrases. For the entropy of $p(f|e)$ ($PT1$) this means that we will have more foreign f entries per English phrases e . This in turn will increase the entropy as the probability mass will be split into more events, increasing uncertainty. The reverse effect would happen with the entropy $p(e|f)$ ($PT3$).

Lexical translation entropies

As anticipated from the correlation analysis, lexical translation entropies are more difficult to predict than their phrase-translation counterparts. From the results in 4.6 we observe only modest prediction results. We found that the most predictive variable was the diagonality of the alignment. While including other variables in the prediction models certainly increased the R^2 , the resulting models were instable (due to collinearity, we observed standardized coefficients > 1). Thus, we keep more modest yet sensible ones. The coefficients indicate a negative influence from the diagonality for both $PT2$ and $PT4$. Thus having more diagonality decreases the entropy in both variables.

4.5 Conclusions

In this chapter, we have revisited the relationship between alignments and phrase-tables. From the results, we corroborated that density variables (gaps, links) have high correlation with many phrase-table characteristics. Additionally, we discovered that the alignment distortion also plays an active role in determining certain phrase-table variables. Furthermore, we observed that the phrase-extraction algorithm plays well with alignment density making its distribution to carry to phrases. The case is different for alignment dimensions and alignment distortion. We observed that phrase-extraction changes drastically the expected behavior. Lastly, we observed that alignment quality as prescribed by the F-metric has little relationship with other phrase-table variables, suggesting that phrase-extraction exploits alignment structure rather than alignment quality.

Doing a regression analysis, we corroborated these intuitions once again but tested them with unseen data in different language pairs. For instance, we confirmed many intuitions regarding phrase-table size and phrase-length and alignment density. Additionally, we observed that alignment diagonality also plays an important role for these models such as average phrase-length. By testing our models in unseen Chinese-English and Arabic-English data, we observed that many of the predictions hold well for different language pairs. This highlights the fact that phrase-extraction does not depend of linguistic features but rather in the structure of the alignment itself.

In this chapter, we also explored the relationship between the uncertainty in our translation features and the alignments used to compute them. Here, we discovered that the entropy in the phrase-translation probability can be predicted very accurately using alignment gaps. On the other hand, the lexical scores are more difficult to predict based simply on these alignment features. While diagonality seems to predict about 30% in the entropy of these translation features, the model is not complete. Using the knowledge gained in this chapter

and the models obtained in the next chapter, we will be able to get the big picture about the effect of alignments in machine translation.

Chapter 5

Predicting Translation Quality

Improving translation performance is a major goal for researchers in the MT area. In recent years, the development of translation competitions (GALE, NIST, WMT) have highlighted the importance of achieving better translation results. Thus, new techniques, translation features, training schemes, etc. are constantly emerging.

More often, research is focused on tweaking parameters, enhancing training sets and looking at the end-to-end translation quality metrics as a measure of performance. Nonetheless, little attention is paid to an equally important task: understanding how each of the components in the complex MT systems affect performance.

In this thesis, our focus is to understand how different parts of an MT systems interact. More specifically, we pay attention to the effect of word alignments, and how their structure affect translation models and ultimately, translation quality. In this chapter, we analyze the characteristics of the phrase-tables and translation hypotheses and how they relate to translation quality.

The correlation between characteristics of the translation model and the automatic quality metrics has previously been addressed by few researchers. For instance Lopez and Resnik (2006) makes a study of different translation model (TM) features and their impact translation quality. On the other hand Birch et al. (2008) studies different language-pair characteristics and treat them as predictors of BLEU (Papineni et al., 2002) translation quality. Others, have focused on identifying characteristics of the word alignments upon which these models have been built. Fraser and Marcu (2007) study how alignment quality (AER) is related to its translation quality cousin BLEU. Lambert et al. (2009, 2010) analyze how alignment characteristics correlate with translation quality.

In this chapter, we propose the analysis of the translation process into different stages. First, we analyze translation quality in terms of the model characteristics. Then in Chapter 4 we study how alignment characteristics impact the models.

5.1 Methodology

In this section, we describe the techniques we used to analyze translation quality and build a regression model based on our translation model translations' characteristics.

5.1.1 Key Terms

Before going into detail about our models, let us define some key terms:

phrase-table Also known as translation model, the phrase table is a collection of source phrases paired with their corresponding target phrases, and several feature values. In addition, phrase tables may include alignment information from the original alignments upon which the phrases were extracted.

In the following figure, we can observe an example of a phrase-table.

translation hypothesis Translation hypotheses are the possible translations that can be built using our models. Each of them is assigned with a *translation cost* given the different scores of such translation according to the language model, the translation model (phrase-tables), etc.

First-best hypotheses are those with the lower cost, and thus, the ones that result in the output of a SMT system as the final translation.

There are several bits of information that can accompany a translation hypothesis: target sentence, feature scores, alignment information, etc. Below we present an example of the first and second best translation hypotheses for the Spanish source sentence: *la bolsa de praga terminó con menos puntos al final de la jornada* .

```

0 ||| the stock exchange of prague ended with fewer
                                points at the end of the day |||
d: 0 -1.65846 0 0 -1.41736 0 0
lm: -70.432
tm: -6.47141 -20.0611 -3.55647 -19.6429 5.99938
w: -15 |||
-7.13615 |||
0=0 1=1,2 2=3 3=4 4=5 5=6 6=7 7=8 8=9,10 9=11 10=12 11=13 12=14 |||
0=0 1=1 2=1 3=2 4=3 5=4 6=5 7=6 8=7 9=8 10=8 11=9 12=10 13=11 14=12

```

Figure 5.1: Example a first-best hypothesis for the Spanish source sentence: *la bolsa de praga terminó con menos puntos al final de la jornada* . Each the output is splitted into different lines for visualization purposes. From top down, we have the fields: Source sentence index (0), its corresponding translation. Next, we have the values for each of the model features (distortion model d , language model lm , translation model tm , and word penalty w) all in logscale. Additionally we have the accumulated weighted model score. Finally the hypothesis alignment (source-target and target-source).

5.1.2 Measuring Variables

In our study, we measure different characteristics of the phrase tables and translation hypotheses. While some of them are shared (specially those regarding underlying alignment variable measurements), other are specific to each. Below, in Table 5.1 we present those variables

	Phrase Table variables		Translation Hypothesis variables	
	var id	description	var id	description
Phrase Table entries	PSU	srcUnique		
	PTU	tgtUnique		
	PNE	pt entries		
Alignment Density	PSG	pt-sgap	FSG	fb-sgap
	PTG	pt-tgap	FTG	avg. target gaps
	PLK	pt-link	FLK	fb-link
Alignment Dimension	PSL	pt-SL	FPSL	avg. phrase SL
	PTL	pt-TL	FPTL	fb-TL
Alignment Distortion	PCR	pt-cross	FCR	fb-cross
	PDT	pt-diag	FDT	fb-diag
	PDG	pt-dist	FDG	fb-dist
Translation Model Features	PT1	pt-tm1	FT1	avg tm-1 cost
	PT2	avg. entropy pt2	FT2	fb-tm2
	PT3	pt-tm3	FT3	avg. tm-3 cost
	PT4	avg. entropy pt4	FT4	fb-tm4
Language Model			FLM	avg. lm cost
Predictors				
Translation quality	BLEU	BLEU		
	MET	Meteor		
	TER	Translation error rate		

Table 5.1: Variables used in the translation quality prediction study

used in our prediction models, as well as our predictors (translation quality): BLEU(Papineni et al., 2002), METEOR v1.3 (Denkowski and Lavie, 2011) and TER v0.7.25 (Snover et al., 2006). We highlight in **bold** typeface those variables that resulted in significant results. For further reference about each of the variables, please refer to Annex B where a more detailed description of each variable is provided.

5.1.3 Finding a Regression Model

In order to estimate a multivariate regression model when its specification is unknown (i.e. we do not know which variables belong to the model), there are several techniques that are commonly used (Hair et al., 2010; Johnson and Wichern, 2002). First, we have a combinatorial approach, which search among all possible combination of independent variables. The best example of this kind of approach is the all-possible-subsets regression. Unfortunately this approach quickly becomes impractical as the number of possible predictors grows (we would search among $2^{|X|}$ where $|X|$ is the cardinality of our predictor set).

The other category of model specification techniques are known as sequential search methods. These methods provide an objective function for selecting variables (i.e. the subset that maximizes the prediction) while employing the smallest possible number of variables.

Examples of this category are the Forward Addition (where we start with an empty model and at each iteration the most predictive variable is added), the Backward Deletion (start with an full model, and iteratively delete the least significant variable), and the Stepwise Estimation (a mixture of Forward-Backward estimation).

More specifically, the algorithm for the Stepwise Estimation, is the following (Hair et al., 2010):

1. Start with a simple regression model of only one predictor e.g. $Y = b_0 + b_1X_1$, where X_1 is usually the independent variable with the highest correlation with the dependent variable.
2. Examine the partial correlation coefficients to find additional independent variables that explain the largest statistically significant (to a threshold level p_{enter}) portion of the unexplained variance. Add the most significant variable to the model.
3. Recompute the regression equation using the newly added variable. Look for the partial p-value (F-statistic) of the previous variables in the model. Remove the one with the lowest p-value below a threshold level p_{out} .
4. Continue the procedure, adding independent variables until none of the remaining candidates are statistically significant

While stepwise regression can be very useful during specification stage, there are several caveats that need to be addressed. For instance, multicollinearity (i.e. predictors that are highly correlated) can obfuscate the effect of one predictor in favor of the other. Then, if one of those predictors enters the model, it is very unlikely that the other will be taken into account as well. Thus, one must also assess the possibility of collinearity and its impact to the model.

Another critique that can be raised to this type of method is that consequently applied significance tests can inflate Type I error. To alleviate this problem, it is suggested to use more conservative thresholds. For our research we used $p_{enter} = 0.01$ and $p_{out} = 0.05$ as thresholds.

In addition, to test the generalization of our models and detect any capitalization of sampling error (over-fitting), we performed prediction tests with unseen data using the regression coefficients for each model.

5.2 Experimental Setup

In this section we outline the configuration of our experiments including the translation systems used, the datasets and the sampling techniques used.

5.2.1 Data and Sampling

Data Sets

For this analysis, we used a variety of different test-sets available for the Spanish-English translation task for the WMT 2011 competition¹ as well as previous years. For Chinese test

¹Data can be obtained directly from <http://www.statmt.org/wmt11/>

data, we used the sets from NIST Open MT evaluation ². For Arabic, the test sets used were extracted from GALE³ evaluation data. The statistics for these sets are summarized in Table A.2.

Sub-document Sampling

In their majority, translation metrics are designed to evaluate complete documents. BLEU for instance, rely on aggregate n-gram matches for the whole document. As such, it is deficient as the sentence level. To better appreciate the effect of translation model into translation quality, we used several sub-documents, long enough to provide accurate translation stats, but short enough to allow us to appreciate difference between different translation models. Sub-document splitting is a known technique that has been used previously for confidence interval estimation (Koehn, 2004b; Guzman and Garrido, 2008). Lengths of sub-documents vary, but it's been reported around 50 test sentences per document. In our study, we chose to use a sub-document size of 100 translation sentences to get smoother results.

Train and Test

As mentioned before, we used the Spanish-English language pair for model estimation (training). Additionally we used test sets from Spanish-English, Chinese-English, and Arabic-English language pairs. For the training part, we used three sub-documents of 100 sentences each per document-set for training. For testing, we used one sub-document of the same size from each Spanish document set. For Arabic test, we used 5 documents from each of the document sets. For Chinese, we used 2 sub-documents from each of the document sets.

The final representation of document sets in our training and test samples are summarized in Table A.2

Systems

For Spanish-English experiments, we used different translation models build upon different alignments. The aligners used for these systems were the discriminative aligner (DWA) (Niehues and Vogel, 2008) with different density thresholds (Guzman et al., 2009)(0.4,0.5,0.6,0.7) to have a variety of dense and sparse alignments. The DWA aligner was trained using hand aligned data from the EPPS(Lambert et al., 2006) dataset.

Additionally, we used the symmetrized GIZA++ alignments using the heuristics grow-diag , grow-diag-final and grow-diag-final-and.

For the Chinese and Arabic test sets, we used the grow-diag-final symmetrized alignment system. For further information about the systems' training, please refer to Appendix A.

²See <http://www.itl.nist.gov/iad/mig/tests/mt/>

³<http://projects.ldc.upenn.edu/gale/>

5.2.2 Data Measurement and Consolidation

For each of our phrase-tables, we filtered them to each of the sub-documents. I.e. from the whole phrase-table, we kept only those entries which had a match in the corresponding sub-document. In practice, the filtered phrase-table has all the information needed by the decoder at runtime. From the deployment perspective, filtered phrase-tables have the advantage of saving the decoder's footprint. Additionally, they represent the search-space available to the decoder given an input document. In this study, we are interested in the characteristics of this search space as well as the characteristics of the translation hypotheses resulting from using these filtered phrase-tables.

For that reason, we filtered each of our seven phrase-tables to each of the different sub-documents. Next, we used the filtered phrase-tables to translate their corresponding sub-documents. Additionally, we evaluated the quality of each of the translations. Finally, we formed a vector for each sub-document/phrase-table combination with the following form $[PT, FB, Q]$ where PT stands for the variables related to the phrase-table and FB for the variables regarding the first-best translation hypothesis and Q for their corresponding translation quality.

In this scenario, our total data points consisted in $7 \times 3 \times 9$ for training (Es), $7 \times 1 \times 9$ for Spanish testing, $1 \times 6 \times 4$ for Arabic testing and $1 \times 8 \times 4$ for Chinese testing.

5.3 Experimental Results

In this section we present the findings of our model building along with their interpretation and some thoughts on the actions we could follow to improve translation. In the first part, we present the general models obtained for the whole set of documents. We discuss their robustness and predictability not only for Spanish but also for Arabic and Chinese. In the second part, we present refined models adapted to 'easy' test sets and tougher 'medium' and 'hard' test sets. There, we discuss the increase of predictability obtained by better targeting our models. Also discuss the trade-offs of such refinement.

5.3.1 General Model

While the use of stepwise regression is useful for discovering models where there is no 'strong' theory about the dependency between variables, it does tend to 'capitalize' on sampling error. To avoid this, we tried several runs of the algorithm, deleting the most predictive variables from the pool, and running the algorithm again. The best results on the overall data are presented below.

BLEU

In Figure 5.2 we present the results for our model. On the top, we present the predictors present in the final model along with their standardized regression coefficients to facilitate the comparison among them. On the bottom part, we present the regression determination coefficient (R^2) for the training model (tra) at each of the iterations, after including one by one each of the most significant variables. These results represent the amount of variation that

can be explained by the regression model. Additionally, we present the determination for the unseen Spanish(es), Arabic (ar) and Chinese (ch) test sets.

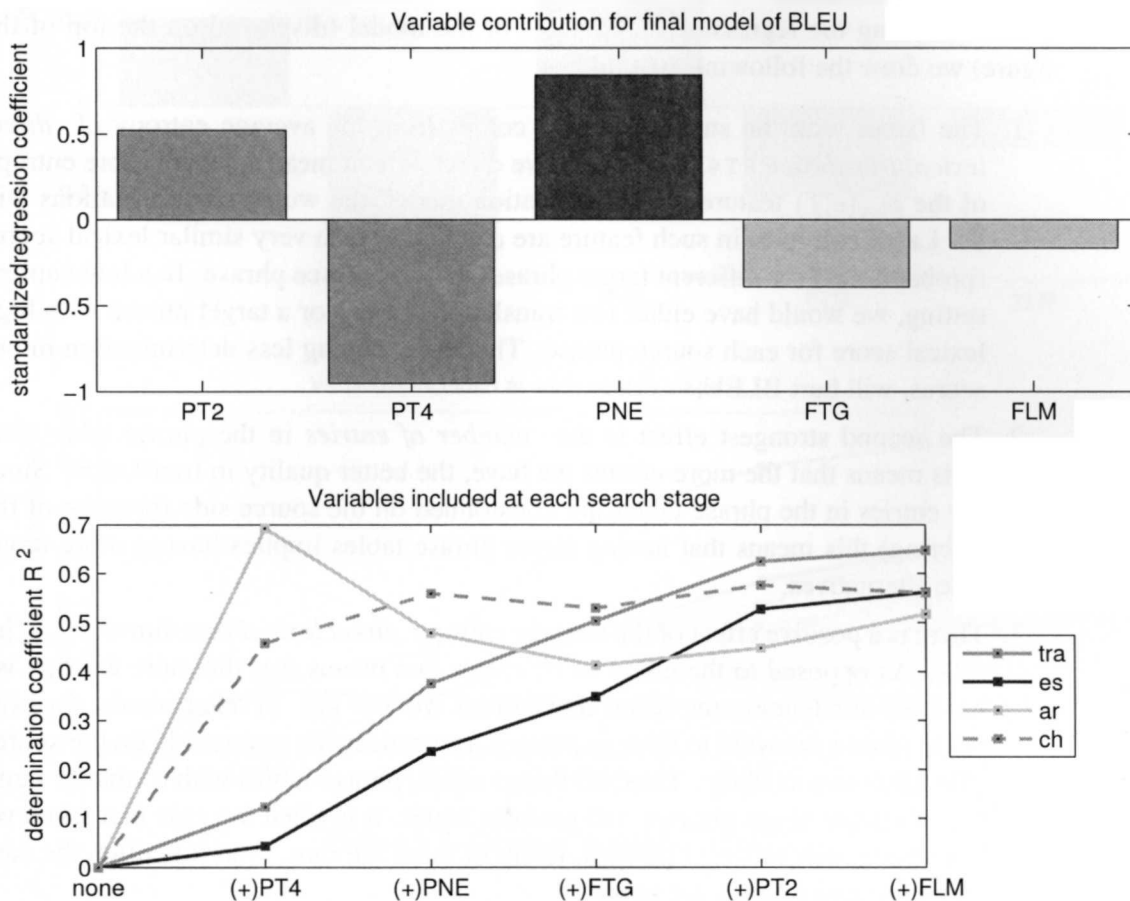


Figure 5.2: General model for BLEU

– Model Analysis

If we regard the lower part of fig. 5.2 we observe the predictive behavior of the regression model at each stage of our stepwise search. From this graph, the behavior of our model depends largely on the language pair. For instance, this model was estimated on a Spanish-English sample, and we observe that the inclusion of each variable in the model (tra) has also a positive effect on the unseen Spanish test set (es). The total variance explained by the model for this language pair is around 55%. For the Arabic-English test set (ar), the story is a bit different. The best prediction can be achieved using only the *average entropy of the direct lexical translation probability* PT4 which accounts for about 70% of the variability. From there, every variable included in the model degrades the prediction, with the exception of the *average entropy of the reverse lexical translation probability* PT2 and the *language model cost* FLM. The final model determination is around 50%. For the Chinese-English test set (ch), only two variables degrade prediction: *the average number of alignment gaps in the hypothesis* FTG and the *language*

model cost FLM. Except from that, each variable included, has an important contribution in predicting BLEU. The final determination is around 55%.

– Model Interpretation

From analyzing the regression coefficients of the model (displayed on the top of the figure) we draw the following conclusions

1. The factor with the strongest effect comes from the average entropy of *direct lexical translation* PT4. It is a negative effect, which means that the more entropy of the $p_{lex}(e|f)$ feature in our translation model, the worse our translations will be. Large entropies in such feature are associated with very similar lexical scores (probabilities) for different target phrases given a source phrase. In a low entropy setting, we would have either few translation options or a target phrase with high lexical score for each source phrase. Therefore, having less determination in our scores, will hurt BLEU.
2. The second strongest effect is the *number of entries* in the phrase table PNE. This means that the more entries we have, the better quality in translation. Since the entries in the phrase tables are constrained on the source side (because of the filtering) this means that having larger phrase tables implies having more target side alternatives.
3. There is a positive effect of the average entropy *direct lexical translation* $p_{lex}(f|e)$ PT2. As opposed to the effect of PT4, this one means that the more entropy we have for this feature, the better translations we will get. In other words, for each target phrase we want to have as many correspondences as possible on the source side, each one as likely. Thus, all things equal, phrase-tables with a smaller number of unique target phrases will perform better. It is often the case that when we have many unique target phrases, many of them are rare. And it is often the case that these rare phrases are noisy.
4. The *hypothesis target side gaps* FTG has a negative effect on the BLEU score. This means that, the more unaligned words are in the target side of the phrases we used, the lower the quality of our translations will be. This is in accordance with the study we developed in (Guzman et al., 2009) and discussed in next chapter.
5. Lastly, the *per word language model cost* FLM has a slight negative effect. This means that the more ungrammatical or *bad English* our translation is, the worse it will fare with BLEU.

METEOR

In Figure 5.3, we can observe the results of our model for the MERT metric. While there are some similarities to the BLEU regression model, the results do not generalize as well.

From the bottom of the graph, we observe a mix of results for the training and testing sets. While the variables included in the final model are the same as the ones included for BLEU model, the order in which they were added differs significantly. For Spanish-English, with exception of PT4, every variable improves the prediction of the model, reaching a final R^2 of about 40%. For Arabic-English, every variable included has a positive effect. The

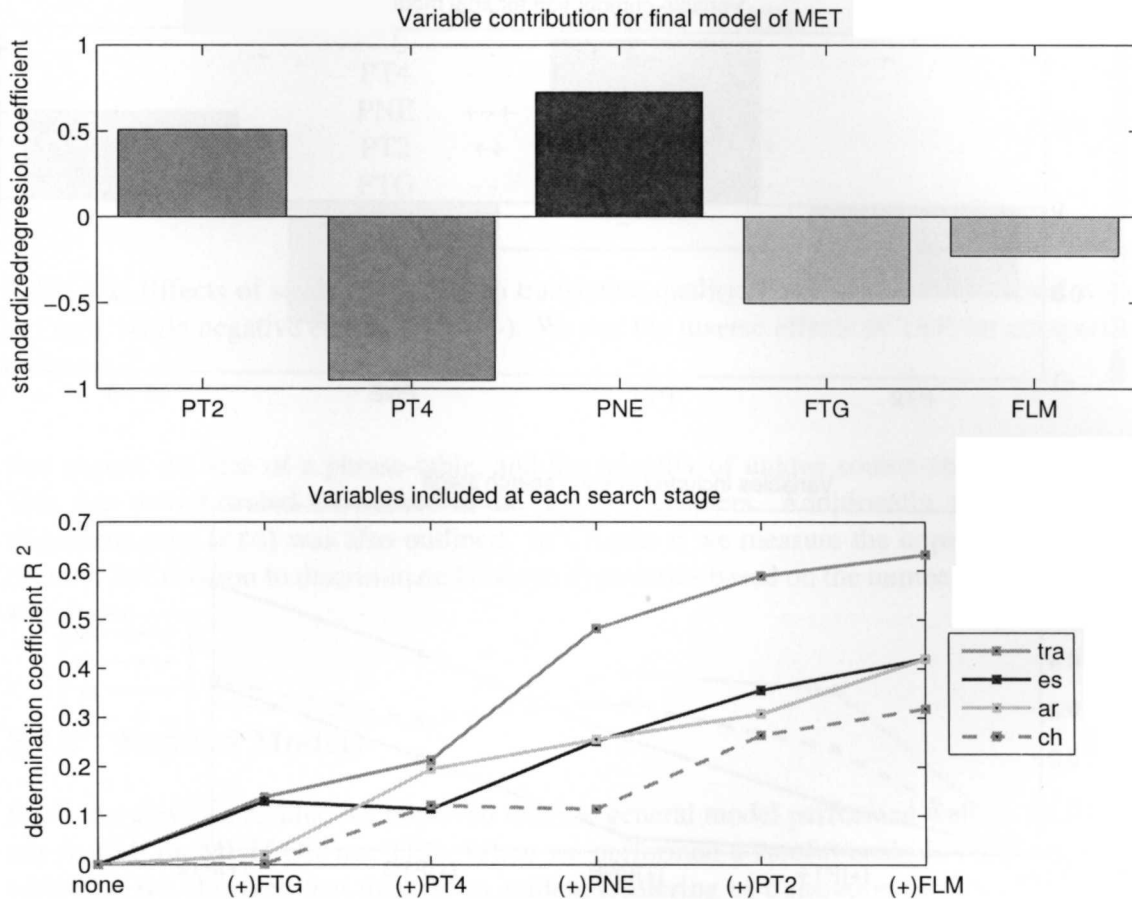


Figure 5.3: General model for METEOR

most important feature seems to be $PT4$ (as with BLEU). The final prediction lies around 40%, which is lower than for BLEU. For Chinese-English, the only three features that seem to improve predictability are $PT4$ and $PT2$ and FLM . Unfortunately, the final model only accounts for about 30% of variance, which is much lower compared to the results observed for BLEU.

Similarly to BLEU model, the direction and magnitude of the coefficients is similar. The main difference lies in the magnitude of the coefficients, but the order of importance remains the same.

TER

The model for TER is shown in Figure 5.4. It is simpler than the previous models, including only four variables. However, it reaches only about 35% of accuracy on the test sets.

From the bottom graph, we observe a consistent behavior between the Spanish-English train and test sets. Each variable adds a some prediction. Nonetheless, the last two variables FTG and $PT2$ seem to be more important for the test set. The behavior for Arabic-English and Chinese-English test sets is very similar. $PT4$, PNE and $PT2$ increase the predictability while FTG has a negative effect.

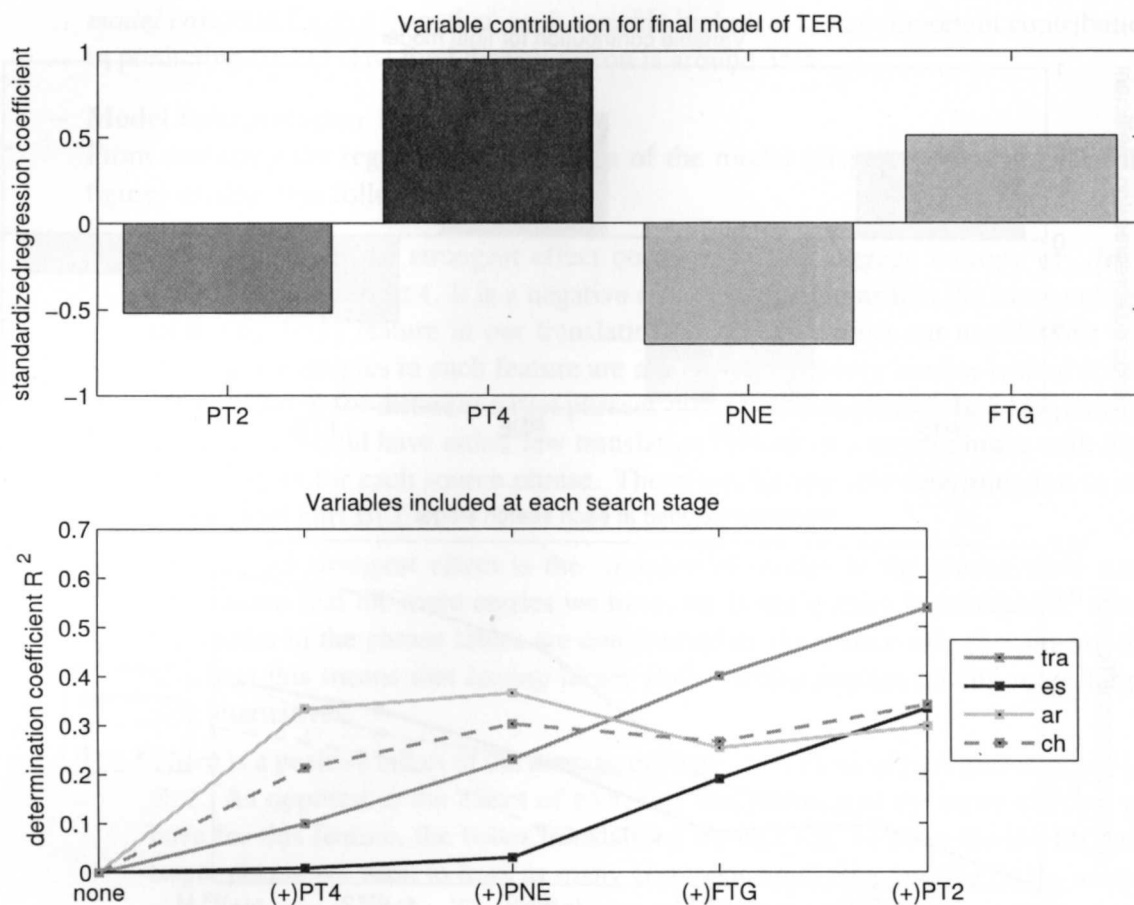


Figure 5.4: General model for TER

Since TER is a negative measure (less is more), opposed to BLEU or METEOR, the results have similar interpretation as for the previous metrics. Notice that the *language model cost* FLM does not appear to have a significant effect in predicting TER.

Discussion

In Table 5.2, we summarize the effects of the phrase-table and hypotheses in translation quality. The most important effects are PT4 and PNE. Medium strength effects are PT2 and FTG. Finally, we also have a weak effect from the FLM only for BLEU and METEOR. From these results, we can draw some conclusions. First, the most relevant effects are related to the translation model. Thus, it is important to have a well defined translation model (PT4,PT2), with many options (PNE). Additionally, it is also that our translations have good word alignment support (FTG) and are in correct English(FLM).

While these findings could be straightforward to anyone with long standing record of empirical work, this is the first study (to the best of our knowledge) that actually uses statistical methods to corroborate the intuition. Furthermore, this study showcases the relevance of word alignment into translation. As we have seen before, there are several alignment characteristics

	BLEU	METEOR	(-)TER
PT4	---	---	---
PNE	+++	+++	++
PT2	++	++	++
FTG	--	--	--
FLM	-	-	

Table 5.2: Effects of several variables in translation quality. Positive effects are noted with a (+) sign, while negative effects with a (-). We use the inverse effects of TER for comparison purposes.

that impact the size of a phrase-table, and the quantity of unique source and target entries. This was demonstrated showcased in the previous chapters. Additionally, the relevance of alignment gaps (FTG) was also outlined. In Chapter 6 we measure the impact of giving the decoder information to discriminate between hypotheses based on the number of gaps in their alignment.

5.3.2 Refined Models

From the previous results, we observed that the general model performed well on BLEU but not so well on METEOR nor TER. When we performed a careful analysis on the response variables, we observed that there is an evident clustering of translation quality according to the set documents we are dealing with. In Figure 5.5 we show the kernel density estimation (kde) of the translation metrics for the Spanish-English train and test sets.

A kde is analogous to an histogram in the sense that shows us the distribution of the variables. However it is calculated using a gaussian smoothing, revealing a continuous distribution that does not suffer from the problem of binning. Here we use it to show the multi-modality of the distributions.

From the figure we observe that while the BLEU distribution has three evident peaks, corresponding to easy, medium difficulty and hard translation tasks, the overall distribution is more symmetric than those of TER and METEOR. This explains why the general models perform better for BLEU than for the other metrics. Furthermore, observe that in both sets, METEOR and TER have roughly two sets: a medium-hard and an easy set of translation tasks. When we analyzed the document sets to which these tasks belong, we discovered a natural per document set division. Table 5.3 we present the manual labeling of the document sets.

5.3.3 Refined Models: Easy Set

In this section, we present the results of training our stepwise regression algorithm in the easy Spanish-English set, and testing on easy Spanish-English and both complete Arabic-English and Chinese-English, to test generalization.

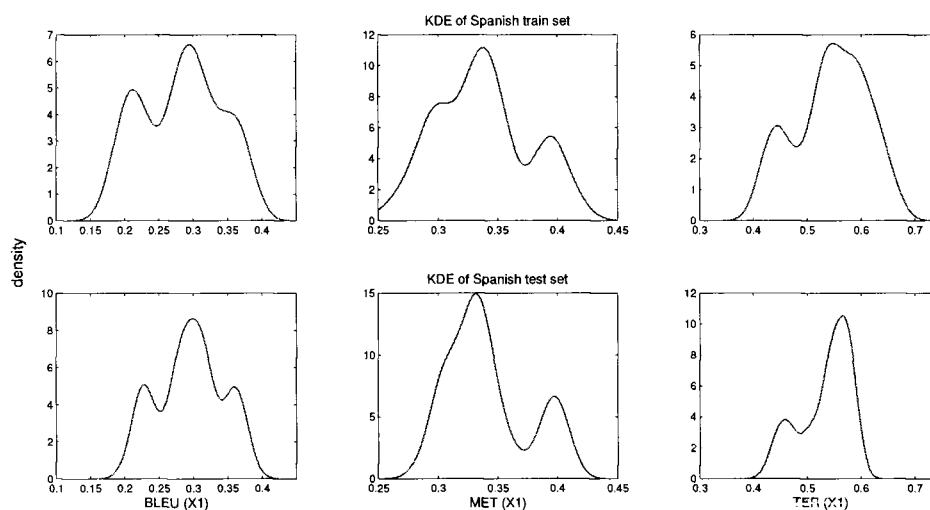


Figure 5.5: Kernel density estimation for BLEU, METEOR and TER for the Spanish-English data sets. The top row represents the distributions for the training sets, while the bottom row, the test sets. From left to right, we show the distributions for BLEU, METEOR and TER.

BLEU

For the easy set, the BLEU metric is relatively easy to predict using few variables. In fig. 5.6 we observe the results for this model.

– Model Analysis

This model is relatively simple. Two variables help us to predict very accurately the Spanish-English BLEU metric. By using only the FLM we can predict about 60% of the variability. Incorporating the FTG we get up to 75% accuracy. Performance differs for Arabic, where only FLM give us about 60% of determination but FTG does not contribute much. The story is very different for Chinese-English, for which the models does not hold well.

Spanish	classification
ac-test	medium
nc-test2007	easy
nc-test2008	easy
newstest2009	hard
newstest2010	hard
test2006	medium
test2007	medium
test2008	medium

Table 5.3: Classification of documents according to their nature

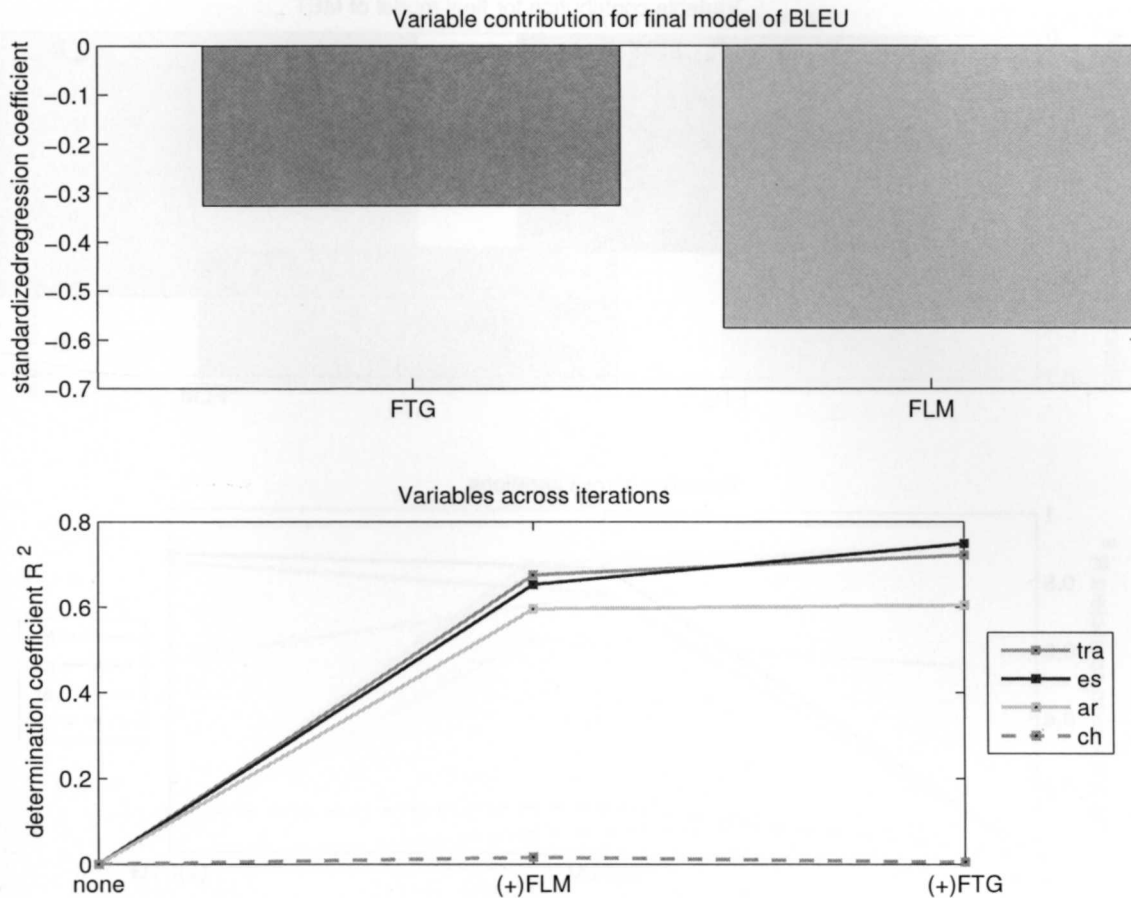


Figure 5.6: Model of BLEU for easy set

– Model Interpretation

This model explains BLEU as having a large penalty from the per word *language model cost* FLM. This means that the more fluent our translation hypotheses are, the better quality they will have according to BLEU. Additionally, we observe a negative effect from *hypothesis target gaps* FTG. This means that the more gaps we have in our translation, the worse it will be. These results are intuitive: first FTG can be regarded as an indicator of the quality of the translation model while is FLM represents the influence of the language model. These two components are the canonical components of any machine translation system. However, the relevance of FTG is noteworthy.

METEOR

The model for METEOR is very similar to the one for BLEU. However, it performs much better for Spanish-English. The results for this model can be observed in Figure 5.7.

For Spanish-English, this model reaches high accuracy using two variables. Each one increases the determination coefficient on the test set, reaching to an R^2 of 85%. For Arabic-English, only the language model FLM seem to help predict METEOR. The inclusion of FTG

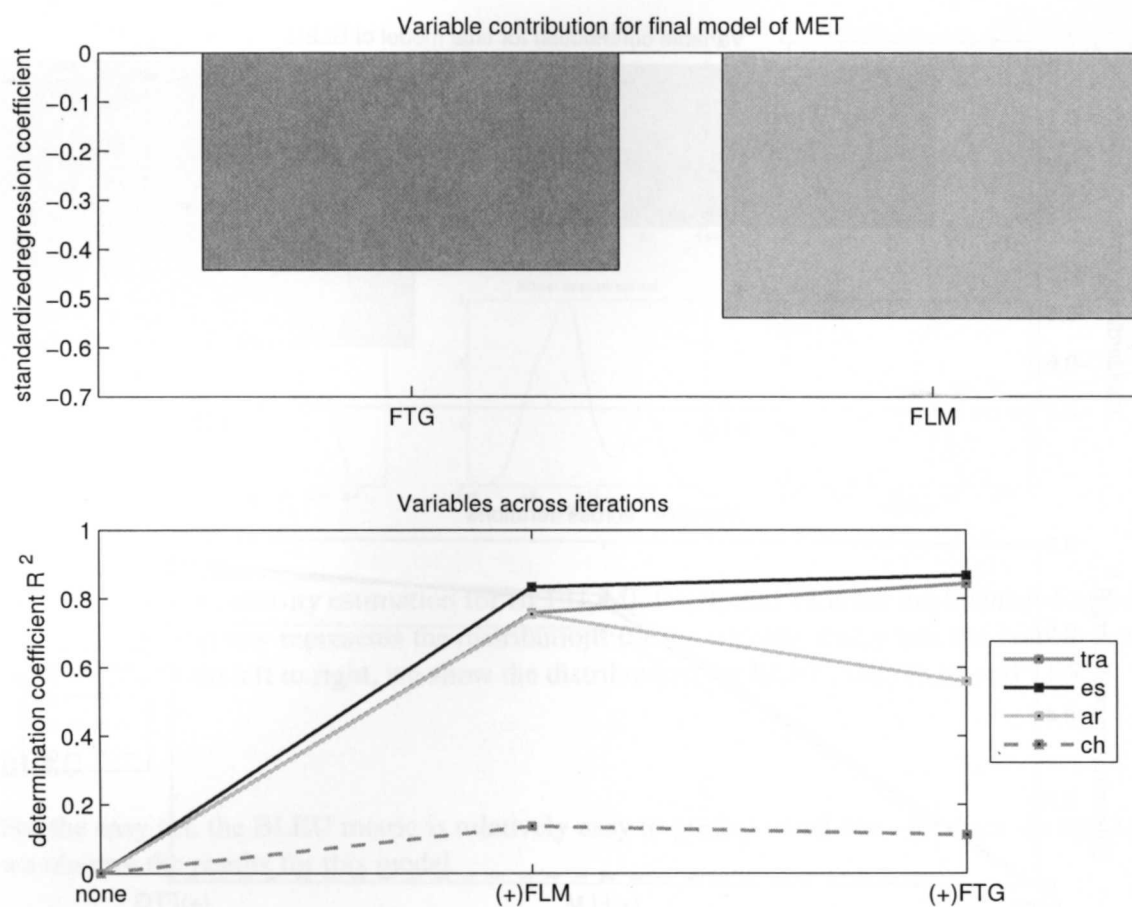


Figure 5.7: Model of METEOR for easy set

variables only hurts prediction. The final model accuracy is about 60%. For Chinese-English, the model is only able to make poor predictions, accounting for about 10% of the variance.

The direction and magnitude of the regression coefficients are very close to the coefficients for BLEU, thus receive a similar interpretation

TER

The last of the models for easy data, the TER model includes the same two variables as the previous models. In this case, since the TER variable is measured inversely, the direction of those variables is reversed. The general outline of the model is displayed in Figure 5.8.

From the figure, we observe that the behavior is very similar to the BLEU case. Only that in this case, Arabic-English reaches about 45% in accuracy while Chinese-English receive 10%. The interpretation in this case is similar. The more alignment gaps we have in our hypotheses, the more translation error we will see. On the other hand, the less fluent our English constructions are, the more penalized our translation will be.

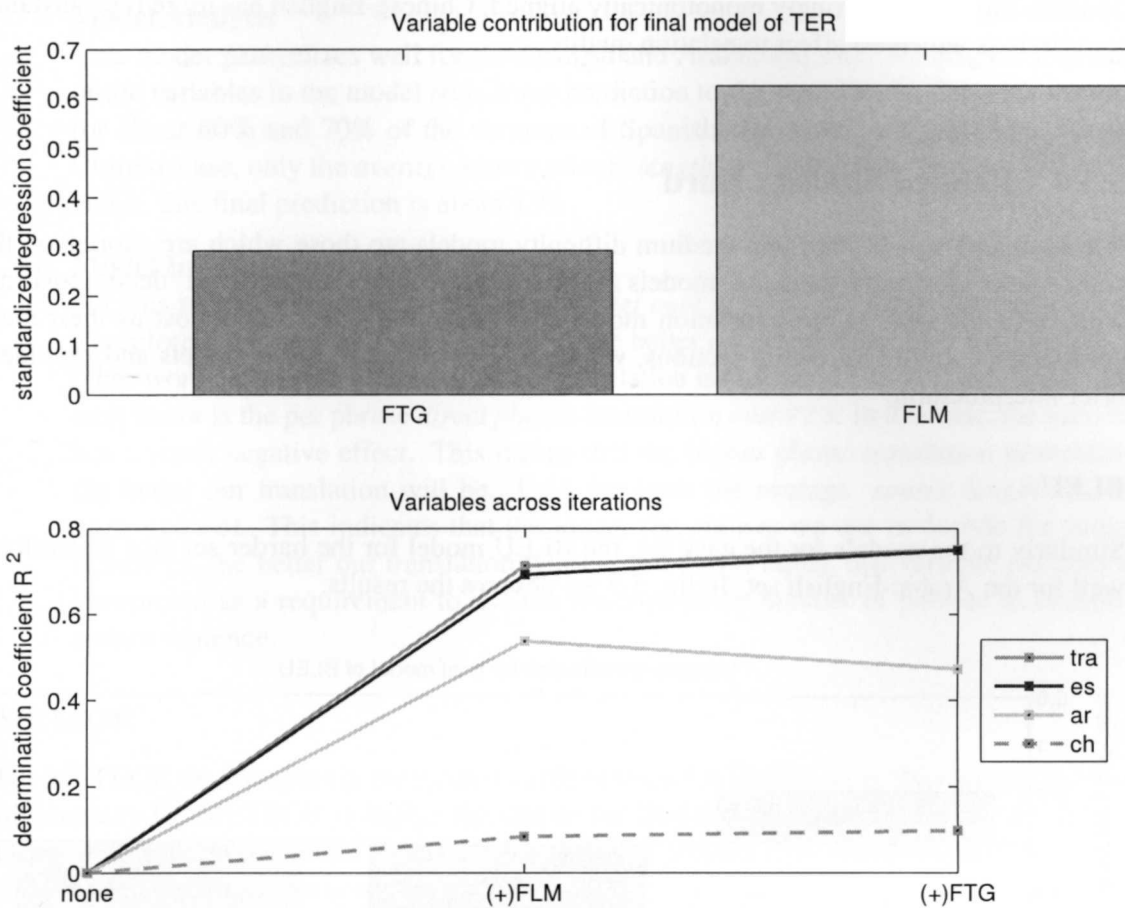


Figure 5.8: Model of TER for easy set

Summary

So far, we have observed that when dealing with an easy document, that is, one document with long source sentences, the models that better predict our translation quality variables can be summarized as in Table 5.4. The prediction levels are substantially higher than the general models.

	BLEU	METEOR	(-)TER
FLM	---	--	---
FTG	-	--	--

Table 5.4: Summary of effects for the easy document set. The (+) and (-) signs summarize the strength of the effect. (+) stands for low, (++) medium and (+++) for large.

We observe in every case, two components, the language model and the alignment gaps. Notice also that these models fit very well Spanish-English and Arabic-English. Chinese, on the other hand, performs very well. The immediate justification for this divergence could reside in the absence of any distortion component in the model. While Arabic-English and

Spanish-English are strongly monotonically aligned, Chinese-English has more long-distance re-orderings which do affect translation quality.

5.3.4 Refined Models: Hard

For Spanish-English, hard and medium difficulty models are those which are shorter on the source side. Generally speaking, models for this set have a more 'traditional' decomposition, with variables such as the translation model cost or the language model cost as their main components. In the following sections, we analyze each one of these models and provide a brief interpretation.

BLEU

Similarly to the models for the easy set, the BLEU model for the harder set also generalizes well for the Arabic-English set. In fig. 5.9 we observe the results.

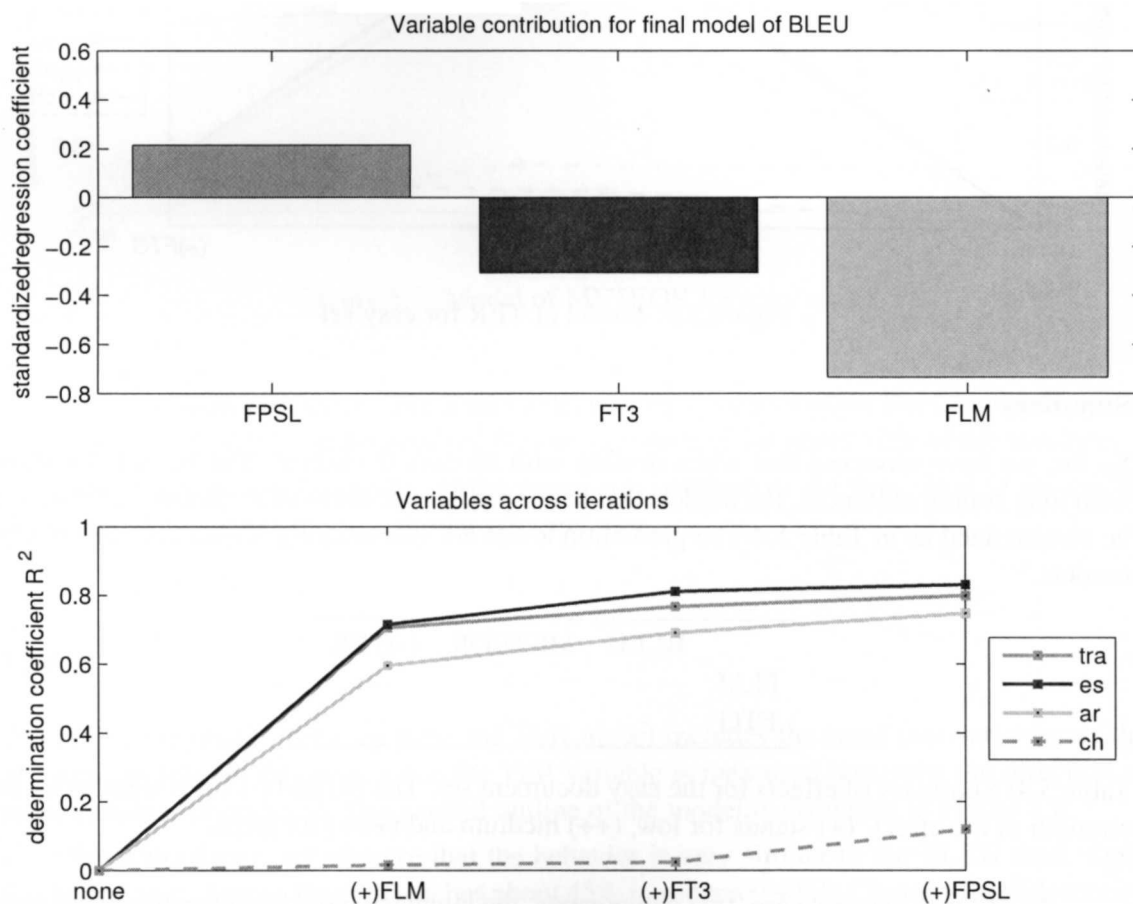


Figure 5.9: Model of BLEU for medium-hard set

– **Model Analysis**

This model generalizes well for the Spanish and Arabic test sets. We observe that each of the variables in the model adds some prediction to the sets. The final models account for about 80% and 70% of the variance of Spanish and Arabic, respectively. For the Chinese case, only the *average source phrase length* $FPSL$ seems to have any predictive value. The final prediction is about 15%.

– **Model Interpretation**

For this model, the *per word language model cost* FLM has the largest negative effect. Therefore the lower our cost per word is, the better our translation will be evaluated. In other words, the more grammatical our translation is (i.e. good English), the better. The next factor is the per phrase *direct phrase-translation cost* $FT3$. In this case, the variable has a small negative effect. This means that the higher phrase translation probability, the better our translation will be. Last, we have the average *source length of used phrases* $FPSL$. This indicates that the longer the phrases we use to decode the source sentences, the better our translation will be. All things equal, this variable can also be interpreted as a requirement to use the fewer possible number of phrases to decode a source sentence.

METEOR

For METEOR the variables in the model are the same as for BLEU. As in the easy set, the level of accuracy for METEOR is higher than those for BLEU. For Spanish-English, we observe close to 90% accuracy, while for Arabic-English it is around 75%. For Chinese-English, the total model determination is of 20%.

TER

From Figure 5.11, we observe that the model for TER is slightly different from the model for BLEU and METEOR. Instead of a long-phrase boost $FPSL$, it uses the $FT1$. Notice that this model has lower accuracy as well. For Spanish-English and it only reaches about 50% of accuracy. For Arabic-English, the most predictive feature is the Language Model, reaching for about 50%. The final model has accuracy of 45%. The Chinese-English case, as previously observed, fits poorly, with about 20%.

The interpretation of this model is similar to the previous. Notice that TER has a 'boost' from *reverse phrase translation cost*. This might seem contradictory to the $FT3$ penalty. Our take on this phenomenon is that while TER prefers low cost Foreign-to-English phrases $p(e|f)$ it actually prefers high cost in the English-to-Foreign $p(f|e)$, thus penalizing phrases that have high phrase translation probabilities. This is in accordance to what we discovered previously for the general models, where we discovered a boost for high entropies on the $PT2$ feature.

5.3.5 Summary

The models for the medium-hard document sets have some common characteristics across quality metrics. These models are more intuitive than the previous. They represent they accommodate the classical components of any SMT system. First, we see a strong contribution

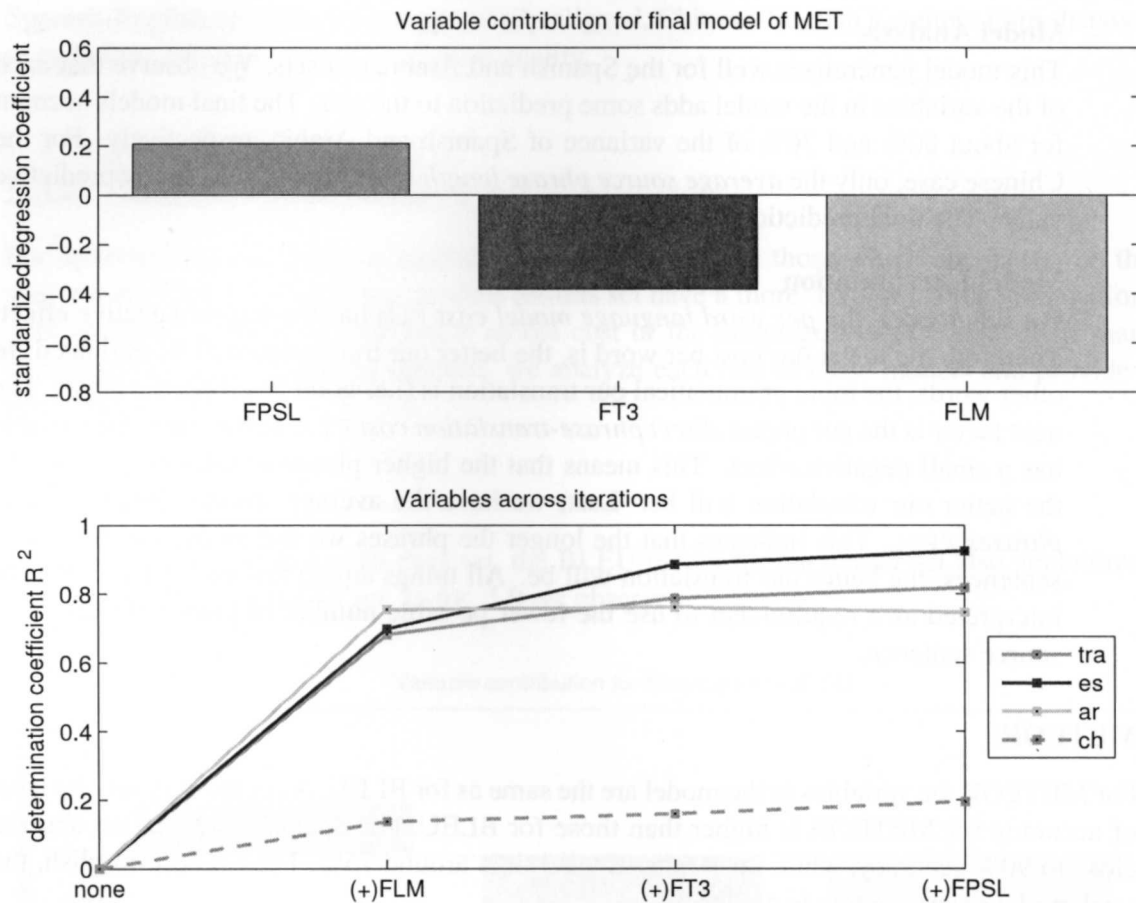


Figure 5.10: Model of METEOR for medium-hard set

from the language model. Then we observe that the phrasal translations also come into play. While in the previous models the importance of the translation model was represented by simpler features (such as FTG), in this case the phrase translation features gain more importance. Additionally, for BLEU and METEOR, we observe the length of the used phrases being important.

5.4 Conclusions

In this Chapter, we have developed several regression models that effectively allow us to predict translation quality in terms of the characteristics of the translation model (phrase-table) and translation hypotheses characteristics. In Table 5.6 we present the summary of our findings.

What we observed is that when aiming for a general model, the most important characteristics for translation quality come from the phrase table. In short, we need to have large translation models, that are well defined. We need to have good alignment support, to reduce the number of alignment gaps that go into our translation. When dealing with target models, we discovered that short sentences are harder to translate. While experience tells us

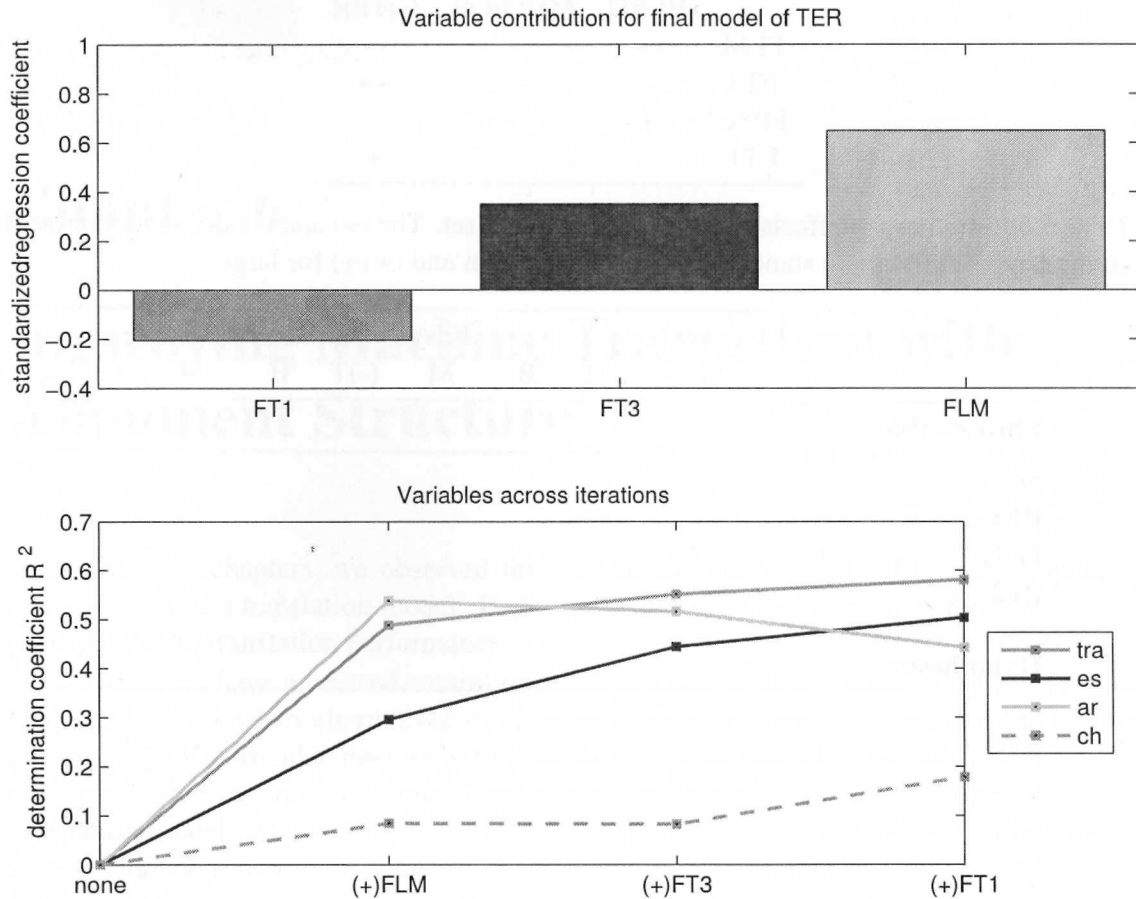


Figure 5.11: Model of TER for medium-hard set

that individual shorter sentences are easier to translate, at the aggregate level we observed the opposite. Documents whose average length is shorter, tend to fare worse. The explanation is that having longer sentences allows us to have a broader span of translation options at decoding time. Another explanation might come from the nature of the data, it could be the case that the harder documents belong to a different domain that just happen to be shorter. In any case, by targeting our models to describe more precisely those samples, we obtained a better description of translation quality.

When dealing with difficult tasks, translation quality comes determined by the cost of the language model and the direct phrase-translation feature and the length of the source side of the phrases used. In other words, when we dealt with shorter sentences, we observed that we need to use as few as possible phrases to translate, while having a fluent output that also represents well the meaning of the original sentence.

When dealing with longer translation tasks, the situation is slightly different. We need that our final translation is fluent (longer translations could degrade quickly because of re-orderings). However the translation model score does not have a significant effect. Instead, we observed that the aggregate number of alignment gaps help to discriminate between good and bad translations.

	BLEU	METEOR	(-)TER
FLM	--	--	--
FT3	-	--	--
FPSL	+	+	
FT1			+

Table 5.5: Summary of effects for the hard document set. The (+) and (-) signs summarize the strength of the effect. (+) stands for low, (++) medium and (+++) for large.

	General			Easy			Hard		
	B	M	(-)T	B	M	(-)T	B	M	(-)T
Phrase table									
<i>tm</i>									
PT4	---	---	---						
PNE	+++	+++	++						
PT2	++	++	++						
Hypotheses									
<i>tm</i>									
FTG	--	--	--	-	--	--			
FT3							-	--	--
FT1									+
FPSL							+	+	
<i>lm</i>									
FLM	-	-		---	--	---	--	--	--

Table 5.6: General summary of the regression model with the translation model and hypotheses effects.

We measured the generalization of these models in terms of how well they helped to predict translation scores for unseen data in several language pairs. We observed that the generic models make weaker predictions that hold across language pairs. On the other hand when using the targeted models, we observed that Arabic-English and Spanish-English performed very well while Chinese-English performed poorly. We attribute this to the lack of reordering as a predictor. While Arabic and Spanish align more monotonically to English, Chinese has more long-distance re-orderings. Lack of fit on unseen data was a trade-off we expected when doing target models for Spanish. Surprisingly, Arabic performed better than expected.

Finally, the results from this study set the grounds for further alignment research. We demonstrated that alignment related features impact translation quality. Thus, we need to use that information to improve alignments and translation. In the next chapter, we observe how important characteristics of the phrase table are closely related to the alignment structure and quality.

Chapter 6

Improving Machine Translation with Alignment Structure

In the previous chapters, we observed that certain alignment structural features impact the consolidation of a translation model. Furthermore, we have discussed how those effects can propagate up to translation performance.

In order to have increased control over the structure of the alignments in the final translation, we propose two alternatives to enhance the use of alignment structure for machine translation. Of particular interest is to provide means to control the number of words left unaligned, which according to our observations, determine greatly the characteristics of the translation model. Thus, the metrics we propose control that characteristic at two different stages: alignment training and decoding. On one hand, we present two new alignment quality metrics, which make use of the alignment structure the gaps to calculate the final quality score. Additionally, we introduce the alignment gap feature for decoding and its experimental translation results presented in (Guzman et al., 2009). Finally, we combine and compare both strategies and their effects in a final experiment.

The remainder of this chapter will be organized as follows: On the first section we describe in detail our proposed alignment metrics, along with the experimental translation results for alignments trained to maximize this metric. Next, we describe the new translation features used during decoding and the experimental results. Finally, we mix-and-match these enhancements and compare the results when they are used in isolation and combined.

6.1 Improving Word Alignment for Phrase-Extraction

As we have observed previously, there is a strong influence from unaligned words into how the phrase-translation model. While there is no rule-of-thumb to whether we should have sparser or denser alignments, conventional wisdom indicates that we should have alignments as close to the Human Aligned reference. In Chapter 3, we confirmed that intuition when we observed that phrases extracted from hand alignments are perceived as having better quality than phrases from automatic alignments. Additionally, we observed that human-alignment generated phrases with unaligned words are perceived as having lower quality when compared to phrases without gaps. We understand that quality and structure both play an important role

in how the final translation model will be estimated. Therefore, we propose two new alignment metrics that take into account more of the alignment structure, playing special attention to the negative space of the alignment matrix.

Traditional alignment metrics such as precision, recall, the F-measure and AER focus on quantifying the matches of true positive instances (i.e. the links in the alignment) while leaving the true negative instances out of their computation. In other words, they focus on how the positive space of an alignment matches the positive space of the reference. As shown in Chapter 3 optimizing towards maximizing these metrics leads to alignments that are less dense than the reference. Thus, optimization encourages alignments to be more precision-oriented. In some situations, more precise alignments are preferable because they result in larger phrase-tables which result handy when dealing with low coverage situations. In other situations, recall oriented alignments will be preferred because their translation models will have lower entropy.

In Chapter 3, we observed how word alignments have an impact on how the translation model is estimated. In addition, the quality of phrases is related to the alignment quality and the alignment structure. Therefore, to have better quality phrases, we need to make alignments more structurally-like to hand alignments.

There are several ways to accomplish this task. On one hand, we can devise a heuristic process (much like symmetrization heuristics) which takes into account the structural characteristics of the alignments to decide which links are added to the final alignment. Also, we can integrate more structural characteristics into a generative alignment model (like IBM) to let the algorithms discover the best alignments automatically. Another alternative is to create new alignment quality metrics that take into account not only the number of link matches but also some of the alignment structure.

We favor the last alternative because a new alignment metric can easily be coupled with different discriminative alignment frameworks. Furthermore, a new metric can be used to compare and evaluate the output of several systems, while taking the advantage of human annotated data.

In this section, we explore two new alignment metrics that take into account more of the structure of the alignment. First we make a modification to the F-metric (AER) to include the number of unaligned words into the match count. Additionally we propose the use of Balanced Accuracy (BA) as an alternative metric in which the positive instances in the learning problem (i.e. links) have equal importance as the negative instances (i.e. voids in the alignment matrix). Below we give a brief explanation of each.

– F^0 score

The traditional F1-score uses the harmonic mean of Precision and Recall to compute its score. In other words it is defined as $F = (R^{-1} + P^{-1})^{-1}$. In terms of a confusion matrix, the F-score is computed as:

$$F = \frac{tp}{2tp + fp + fn} \quad (6.1)$$

where tp are the true positives, fp are the false positives and fn are the false negatives.

Unaligned words or alignment gaps are also known as NULL-alignments. They get their name from the generative models where an explicit NULL word was considered during training to model the probability of a source word to be aligned to no word on the target side. It was also said that the source word in question had a zero fertility. Different views from the same phenomenon are depicted in Figure 6.1.

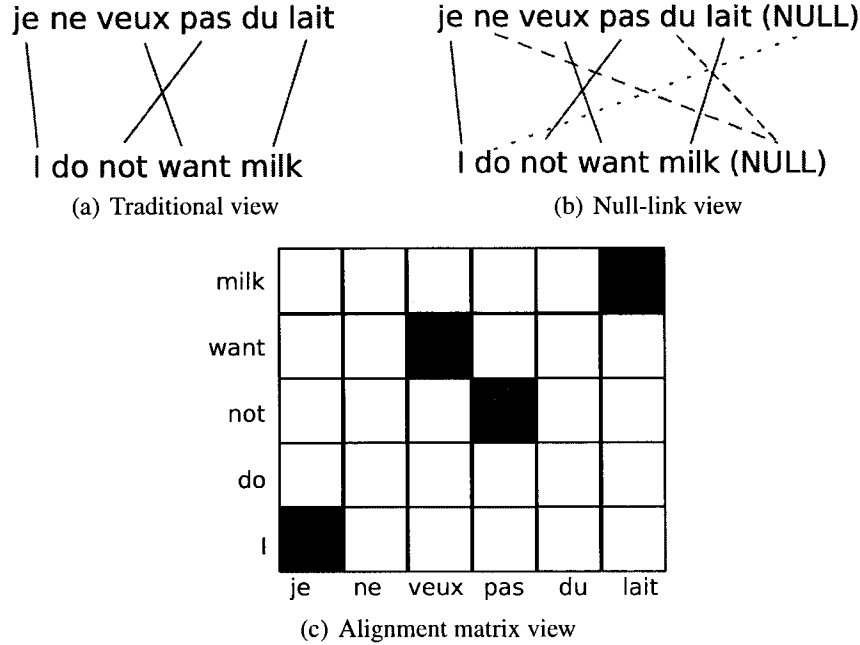


Figure 6.1: Three different representations of the alignment gaps/ NULL-links .

In the traditional link-representations, unaligned words have no link with words on the opposite side (e.g. French words *ne*, *du*). In the matrix representation of the alignment, unaligned words are represented by an empty column. In the IBM-model's view, they are represented by an explicit link towards a NULL word. Inspired by this last representation, we introduce the F^0 (F-null) metric, which uses null links in the computation of the quality score. It is defined by:

$$F^0 = \frac{tp + tp^0}{2(tp + tp^0) + fp + fp^0 + fn + fn^0} \quad (6.2)$$

Where tp^0 , fn^0 and fp^0 stand for the matches (non-matches) of null links.

By taking into account the positive and the null link matches this metric has the advantage of incorporating some of the alignment structure into the quality score.

– **Balanced accuracy**

Besides Precision and Recall (also called sensitivity), Specificity is another metric which is typically used in medical domain applications. Specificity is defined as the amount of true negatives by the amount of negatives in the gold standard. Thus it takes into account the negative space in our alignments. It is defined by:

$$sp = \frac{tn}{tn + fp} \quad (6.3)$$

While sensitive (recall) oriented alignments trend to be more dense, very specific alignments would only have the opposite constitution. To balance these two quantities in equal terms, the Balanced Accuracy is defined as the average of the two.

$$BA = \frac{1/2tp}{tp + fn} + \frac{1/2tn}{tn + fp} \quad (6.4)$$

Balanced accuracy gives equal importance to the positive matches as to the negative matches. In other words, its not only concerned of the true space, but also to the structure of the alignment defined in the negative space. As we have observed before, the sparsity of an alignment is also an important characteristic for phrase-extraction.

6.1.1 Alignment Results

We used the new alignment metrics to tune the parameters of the Discriminative Word Alignment (Niehues and Vogel, 2008) to measure the effectiveness of these metrics in discriminative training. Our goal was to determine how the resulting alignments compare the AER (F-measure) tuned alignments in terms of alignment quality and alignment structure.

Traditionally, the tuning for DWA alignment comes in two stages: first a maximum likelihood tuning of parameters is performed. Then in the second stage, the parameters (scaling factors) are fine-tuned towards a second alignment metric as this is reported to lead to the best results (Niehues and Vogel, 2008).

We performed an experiment in which our baseline was the AER metric tuning. We compared it to the F^0 and BA tuning measures based on the alignments produced for a unseen test sample.

Alignment Quality

In Table 6.1, we present the quality results for each of the alignments tuned to different metrics.

Tuning metric	Alignment Quality							
	P	R	F	P^0	R^0	F^0	SP	BA
F	85.0	62.8	72.2	79.4	61.3	69.2	99.4	81.1
F^0	85.6	62.6	72.3	79.6	61.4	69.4	99.5	81.0
BA	73.6	66.6	69.9	70.7	64.6	67.5	98.8	82.7

Table 6.1: Alignment quality metrics for each of the different alignments: F-tuned alignment (F), F^0 tuned alignment and Balanced accuracy (BA) tuned alignment. The metrics considered are precision (P), recall (R), F-measure (F), the null-link metrics (F^0 , P^0 and R^0), Balanced accuracy (BA) and specificity (SP). We present in bold font the best results for each metric.

Notice how optimizing towards different metrics yields slightly different alignments. For instance, alignments tuned towards F^0 are more precise than its counterparts in terms of both P and P^0 . However the gains are larger for P. This indicates that in order to achieve better F^0 score and have more NULL link matches, the aligner omits more true links. As

expected, this results in a slight increase of the F^0 score. Additionally it performs better on the F-score when compared to the F-tuned alignment. Finally notice how it has the highest Specificity. This means that this alignment accurately improves the rate of true negatives (non-links) matches.

The BA tuned alignments have the highest BA score. They also perform significantly better recall, both R and R^0 . However they report a significant loss in precision, which hurts the F and F^0 scores.

Alignment Structure

In Table 6.2 we present the structural differences between the alignments and the reference.

	Alignment Structure					
	ASG	ATG	ALK	ACR	ADT	ADG
Reference	0.065	0.045	1.473	2.121	3.910	0.906
F	0.082	0.047	1.099	0.744	2.792	0.917
F^0	<i>0.088</i>	<i>0.051</i>	1.088	<i>0.725</i>	<i>2.767</i>	0.917
BA	0.071	0.038	1.357	0.953	3.238	<i>0.921</i>

Table 6.2: Alignment structure metrics for each alignment. In bold-font we present each of the results which are structurally closer to the hand alignment. In slanted font, the results that are remarkably high.

In terms of structure, we observe that the BA alignment is denser than its counterparts and it is closer to the reference in several structural criteria. For instance, it has a rate of source gaps ASG closer to the reference alignment and the lowest ATG from all the alignments. Additionally its link density is closer to the link density of a hand alignment. Since it is more dense, we also observe more crossing ACR and more relative distortion ADT. Surprisingly, it has the highest diagonality overall.

As anticipated, the F^0 alignments present more alignment gaps than the rest of the alignments, and their density is slightly lower than the F-tuned alignments. They also present the less distortion overall both in terms of alignments crossings ACR and relative distortion ADT.

Summarizing, changing the tuning metric for discriminative alignments does have an effect on the quality and structure of the resulting alignments. For instance, we observed that including NULL-links in the computation of the F-score yields more precise alignments which leave more unaligned words. In terms of F and F^0 quality, they present a slight improvement when compared to the regular F-tuned alignments.

On the other hand, having alignments tuned towards Balanced Accuracy (which take into account the true negatives of the alignments) yields more recall-oriented alignments, which in terms of structure are more similar to the human generated data. As we can expect, these alignments will yield very different translation models. As predicted by our models in chapter 4 these alignments also yield very different phrase tables. Taking as a reference the F alignment which has 20 million entries, the BA alignment which is more dense, yields a shorter phrase table with only 15 million entries. On the other hand the F^0 alignments yields

a larger phrase table with 21 million entries. In the next part we will present the translation experiments.

6.1.2 Translation Experiments

As a final experiment, we used the phrase-tables from each alignment for translating unseen data. The optimization of each model was done using MERT (Och, 2003) to maximize BLEU for the news part of WMT2008 test data (nw08). For translation we used two newscast datasets corresponding to years 2007 and 2008 of the WMT newscast test data (NC07, NC08). The results of that experiment are summarized in tab. 6.3. For comparison, we included the heuristic symmetrized alignment grow-diag-final-and.

	BLEU		METEOR		TER	
	NC07	NC08	NC07	NC08	NC07	NC08
DWA-F	35.53	36.36	40.39	40.47	45.86	45.84
DWA-BA	35.15	36.03	40.00	40.13	46.47	46.36
DWA-F0	36.20	36.83	40.60	40.61	45.05	45.29
G DFA	36.14	36.75	40.42	40.37	45.11	45.36

Table 6.3: Translation results using four different models: A discriminative aligner tuned towards three different metrics (F, F^0 , BA) and a symmetrized alignment. The different evaluation metrics used were BLEU, METEOR and TER

First, remark how the initial F-tuned alignment has some trouble translating these sets of documents in comparison with the symmetrized alignment. However, including the NULL-link information in the F-metric shows significant improvements not only with respect to the baseline, but also with respect to the symmetrized heuristic. This behavior is replicated in the different metrics. While it is expected that improvements for BLEU are the largest (because the systems are optimized to maximize BLEU), observe how the DWA- F^0 alignment has consistent improvement across the different metrics.

The Balanced Accuracy alignment, which was structurally more alike to the human aligned data, has the worse results of all systems. The results from these experiments are consistent with the general predictive model that we obtained in Chapter 5. In fact, the correlation between the obtained BLEU scores and the predicted scores are of 0.88. As we observed before, having more entries in the phrase-table is an important factor in obtaining better scores. Thus, it is not surprising that the BA alignment obtained lower scores, given that its high density alignment results in fewer translation entries in the translation model.

6.1.3 Conclusions

As a result of the analysis performed in the previous chapters of this thesis, we observed that alignment structure is very important for the estimation of the translation model. To provide a mechanism to better control the structure of such alignments, in this section we introduced two new alignment quality metrics which take into account more of the alignment structure. We introduced two new alignment optimization metrics: Balanced Recall and the F-0 metric

which make use of NULL-alignment information. We observed that the F^0 -tuned alignments present improvements in terms of precision with respect to the F -tuned alignments. Also, we observed that the BA alignments are the most human-like in terms of structure. However, this similarity turns out to be of little help for machine translation.

In terms of translation quality, the clear winner is the DWA- F^0 alignment, which yields consistent improvements across metrics.

While the original concern was to reduce the number of unaligned words, tuning towards a metric that uses unaligned information yields a slightly larger number of unaligned words. However, this alignment is able to obtain better results mainly due to the fact that it results in a significantly larger phrase-table.

In the next section, we present experiments where we control the alignment gaps at decoding time via an unaligned feature.

6.2 Improving Translation Using Alignment Gap Features

In Chapter 5, we observed that there are several features from the translation model that predict well translation quality of translation hypotheses on a static scenario. We observed how the translation hypotheses with more target gaps seemed to have lower translation quality. Inspired by that fact, and to give the decoder more control over the gaps in the translation hypotheses we introduce a new decoding feature that takes into account the alignment gaps information and uses it dynamically at decoding time. As we will observe later, using such feature enables the decoder to turn an originally liability (more gaps meant less translation quality) into an asset. The target gap decoding feature (h_{ftg}) is defined as follows :

$$h_{ftg}(e_1^I, f_1^J) = J - \sum_{j \in J} \prod_{i \in I} (1 - l(i, j)) \quad (6.5)$$

Where e_1^J stands for the target phrase spanning from words 1 to J , f_1^I stands for the source word spanning from 1 to I , and $l(i, j)$ is an indicator function $\{0,1\}$ that tells us whether there is an alignment link between word i and word j . By multiplying all $(1 - l(i, j))$ for the same j we verify if the j th word is left unaligned. Thus, when calculating the feature value for a full translation hypothesis, the cost will be:

$$C_{ftg}(\bar{e}_1^K, \bar{f}_1^K) = w_{ftg} \sum_K h_{ftg}(\bar{e}_k, \bar{f}_k) \quad (6.6)$$

In the end, an optimization procedure (MERT) will set the optimal weight for w_{ftg} based on a development (tuning) translation set. In the next part, we will discuss the results of performing the optimization for these new features.

6.2.1 Tuning Weights for Quality

In this part of the analysis, we use the weights for the new features tuned by the Minimum Error Rate optimization. For comparison, we used a set of different translation models, which were generated from the alignments we had analyzed in previous chapter (DWA-4 to DWA-7, grow-diag, grow-diag-final, grow-diag-final-and).

For MERT we used the news-test for the WMT 2008. This dataset consists of News data, which is considered out-of-domain for our translation models, which were trained on Europarl, News Commentary and UN data. For further reference about the data please refer to Appendix A.

In Table 6.4, we present the weights for each of the translation models used. Since each of the optimization is performed independently (i.e. for each system and for each setting), since adding one new feature changes the search space and thus, the optimizer might modify the assigned weights for each feature. Therefore, we present the percentage of weight change corresponding to this feature, as defined by:

$$\Delta\hat{w}_{ftg} = \frac{\Delta w_{ftg}}{\sum_{w_i \in \mathcal{W}} |\Delta w_i|} \quad (6.7)$$

System	w_{ftg}	$\Delta\hat{w}_{ftg}(\%)$
DWA-4	0.107	19.80
DWA-5	0.059	8.17
DWA-6	0.092	17.68
DWA-7	0.050	11.12
GD	0.092	19.24
GDF	0.071	8.41
G DFA	0.132	21.45

Table 6.4: MERT-tuned weights for the FTG feature by system. On the central column we present the nominal weights for the decoding feature. These weights are normalized so $\sum_w |w_i| = 1$. On the right column we present the percent of tuning weight change attributed to this feature.

First of all, notice how all of the weights given to this feature are positive. This means that the feature is acting as a boost rather than a penalty. In this new scenario, it uses the gap information to discriminate good and bad hypotheses locally. In fact, decoder is using the target gap-feature (as we will observe later on fig. 6.4) as a counterweight to phrasal probability. By giving a boost to phrases-with gaps, it can then penalize harder other phrases that do not have gaps, yet have low translation probability.

Also note that a substantial amount of the weight change for these systems comes from the weight optimization for the newly added feature. For most of the systems, the percentage of weight change due to the gap-feature is less than 20%.

In the next part, we will observe how the new feature contributes to the improvement of the translation.

6.2.2 Translation Results

For this part, we used the models with the new features and the optimized weights to translate several test sets. The considered documents were both in-domain Europarl test-sets (WMT06,

WMT07, WMT08), the also limited domain Acquis corpus (AC) and the News test set (NW09, NW10, SC09).

The best scores for each document set are presented in Table 6.5. Notice that from all the tasks available, most of the best results were collected by the DWA-5 translation model the translation gap feature. And only one was obtained by the GROW-DIAG-FINAL symmetrized heuristic without gap.

Doc Set	best BLEU	System	Gap feature
AC	35.65	DWA-5	yes
SC09	26.78	DWA-5	yes
NW09	25.98	GDF	no
NW10	28.82	DWA-5	yes
WMT06	30.70	DWA-5	yes
WMT07	30.40	DWA-5	yes
WMT08	30.73	DWA-5	yes

Table 6.5: Best scores per document set obtained by system/feature combination

The individual BLEU gains scores between the translation models which use the new feature and those that do not are shown in represented in Figure 6.2, where we plot the deltas $\Delta BLEU = BLEU_{FTG} - BLEU_{bl}$ by document set.

Notice how we observe mixed results for the feature depending on the test-set. For in-domain sets, we get significant improvements. For instance, for the WMT06, we get an improvement of 1.04 blue-points for the best-case, while the worst case increases 0.08 BP, with an average of 0.38 BP gain. For the out-of domain sets the results are less optimistic. For instance, for NW10, the best-case system gets an improvement of 0.8, while the worse, presents a loss of 0.46 BP, with an average gain of 0.15BP which is barely significant. On the positive side, the general average gain is of 0.27 BP.

Analyzing the results by system (fig. 6.3) we observe an interesting pattern. For the discriminative systems, we observe that a tendency to achieve large gains by using the feature. These gains are more pronounced for moderately sparse alignments such as the DWA-5 and DWA-6. Which register average gains larger than 0.5BP. For the symmetrized alignments, the addition of the feature does not represent any significant change.

By comparing the gains by the system to the relative weight changes, we observed an interesting correlation between the gains and the shift in the weight for the inverse phrasal probability feature $p(f|e)$ as shown in fig. 6.4.

Observe how there is a clear tendency in the gains obtained by adding the new feature. Systems for which the new feature was used to counterweight the $p(f|e)$ feature, achieve better results. This was done by MERT by increasing the penalization for the $p(f|e)$ feature.

Summarizing, using the phrase-gap feature, results in a boost for phrases that carry many gaps. When combined with an increased phrase-translation penalty, we can achieve large improvements in translation quality. These improvements were more noticeable for in-domain test-sets. However, for the most part, the best-quality results were delivered by the DWA-5 system using the target gap feature.

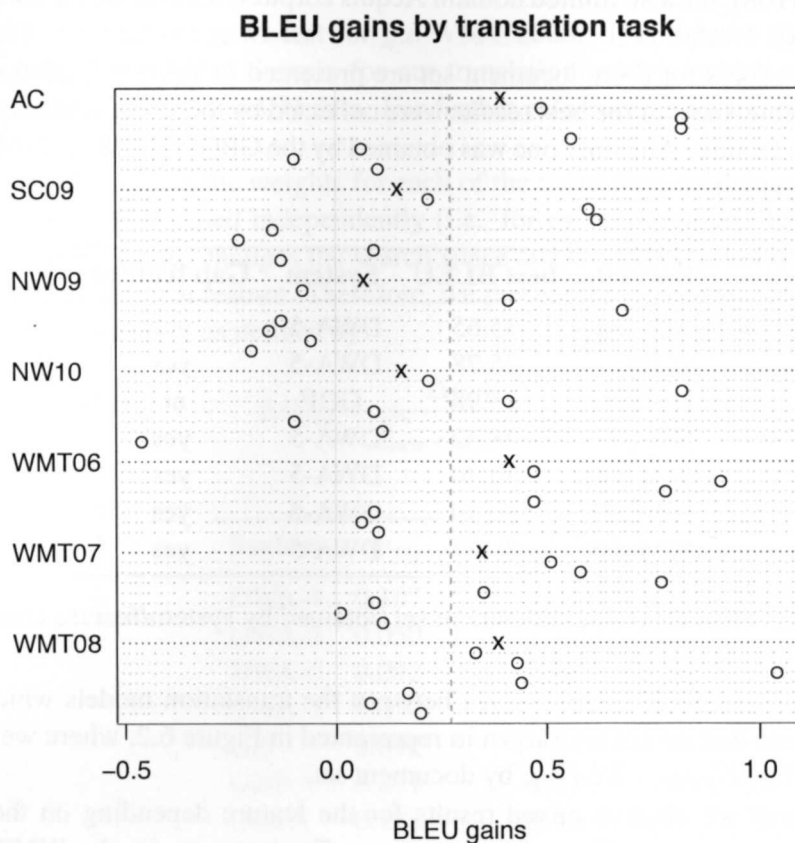


Figure 6.2: BLEU gains of systems using the target gap feature segmented by test-document. On the left axis we present the different document groups. On the horizontal axis, the gains of $BLEU \Delta BLEU = BLEU_{ftg} - BLEU_{bl}$. For each document set, the overall average gain is marked with an x . The general average is represented by the vertical dashed line.

Gap features can be more useful in situations where we have limited training data. In the next section we will present the results of experiments done in Chinese.

6.3 Translation Gaps for Limited Scenarios

In this section, we present the translation results of our experiments using limited data and a combination of two gap-features (source and target). These results were presented in (Guzman et al., 2009).

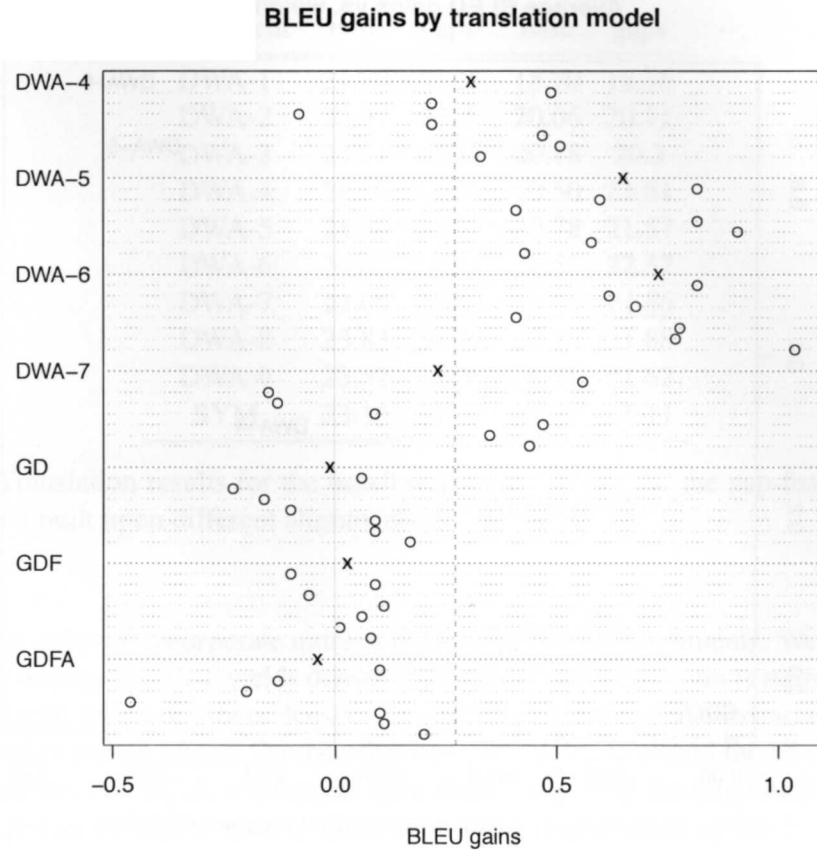


Figure 6.3: BLEU gains using the target gap feature segmented by system. On the left axis we present the different translation models used. On the horizontal axis, the gains of BLEU. For each system, their average gain is marked with an \times . The general average is represented by the vertical dashed line.

6.3.1 Setup

For this experiment, we used a training data set consisting of the GALE P3 Data¹. The data was filtered to have maximum sentence length 30. The final training set contains one million sentences. This setting is more limited than the previous experiments. The different systems that were used, were built upon the alignments from the DWA with alignment threshold $p = \{0.1..0.9\}$, and the symmetrized alignment (grow-diag-final). We use the MT05 test set for tuning, and used a subset of the development dataset of GALE07 Evaluation (DEV07) as the blind testing data. The data set consists subsets from different sources: Newswire (NW) and Weblog (Web) with 427 and 358 sentences respectively. In Table 6.6 we display the BLEU Scores for these sets. First, notice how in our baseline the best results are obtained by the sys-

¹ FOUO data (LDC2006G05), HKnews (LDC2004T08), XinhuaNews (LDC2003T05), and parallel data from GALE (LDC2008E40, LDC2007E101, LDC2007E86, LDC2007E45, LDC2006E92, LDC2006E34, LDC2006E26, and LDC2005E83).

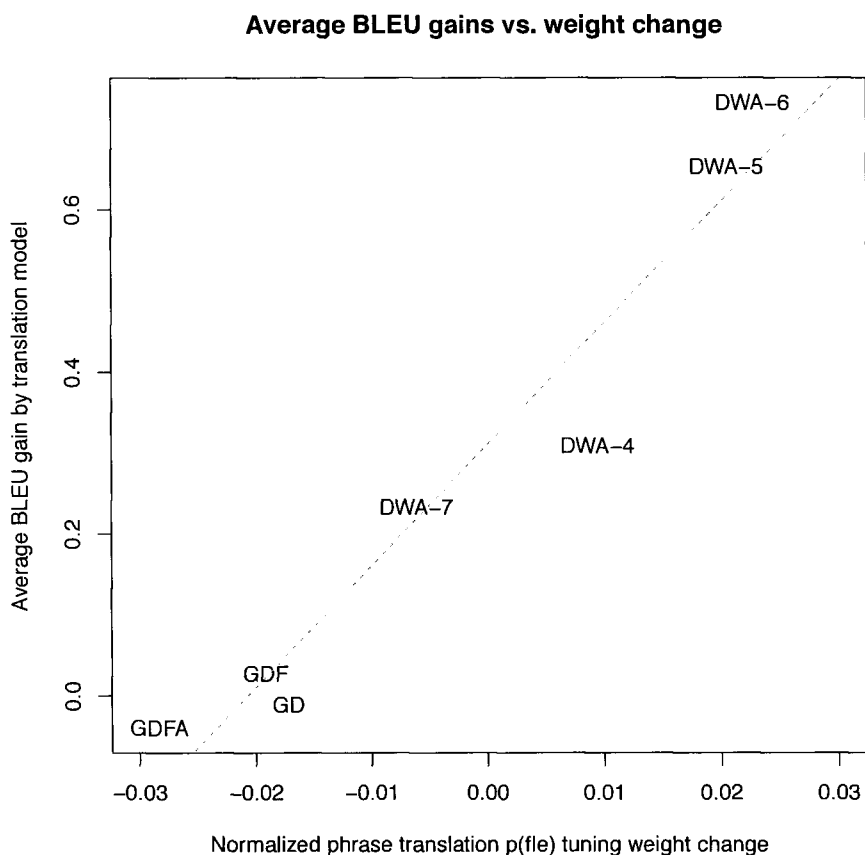


Figure 6.4: Average BLEU gains by system vs. the normalized weight change for the $p(f|e)$ feature $\Delta\hat{w}_{p(f|e)}$.

tem that previously achieved the highest AER (DWA-5). From there on, the systems trend to have lower quality as we shift the balance from precision/recall in our alignment. However, the alignments with higher recall (DWA-1) trend to perform more poorly than the high precision ones (DWA-9). This is not surprising, as this phenomenon has been observed previously. For the systems that use the number of unaligned words as a feature, we observe that the best results are found with a higher precision alignment (DWA-6). This can be explained as the result of penalizing the phrases that include a lot of gaps, which as shown before have lower human-perceived quality. The improvements of using gap features are more striking for the Web test set, where we obtain up to 2BP of improvement (for DWA-7).

6.4 Conclusions

In this chapter, we introduced two alternatives to take advantage from alignment structure for improving machine translation at the alignment training and decoding phases.

We proposed two alignment quality metrics, which by counting the negative space of the

Alignment	NW		WEB	
	base	gaps	base	gaps
DWA-1	21.20	22.25	18.70	18.76
DWA-2	22.97	23.3	20.06	20.11
DWA-3	23.11	23.35	20.18	20.3
DWA-4	24.19	24.81	20.50	21.81
DWA-5	24.56	24.72	20.78	21.57
DWA-6	24.05	24.97	20.53	22.57
DWA-7	23.05	<u>24.62</u>	19.54	<u>21.86</u>
DWA-8	23.83	24.74	20.52	21.88
DWA-9	23.32	24.26	20.26	21.62
SYM	23.15	24.18	20.22	21.11

Table 6.6: Translation results for the baseline systems (base) and the gap-feature enhanced systems (gaps) built upon different alignments.

alignment into account, incorporate more of the structure of the alignments. We observed how the Balanced Accuracy metric yields denser alignments whose structure is more similar to the hand-aligned data. However, these denser alignments inhibit the phrase-extraction process and result in compact phrase tables. On the other hand we also introduced the F^0 measure which takes into account the NULL alignments into when computing the alignment quality score. This metric results in higher quality alignments with respect both to the F and F^0 metric. Furthermore, this metric also represents an improvement to translation quality with respect to the baseline and other symmetrization heuristic. We also observed how these translation results are consistent with the predictions drawn from our general translation quality predictive model.

At the decoding stage, we proposed to incorporate alignment structure by integrating the unaligned word count as a decoding feature. This modification works well for translation models coming from discriminative alignments for which produces the best translations (except in one case). We also observed how the target gap feature interacts with the translation model phrase-translation probabilities and yields the best results when optimization weights for such phrasal probabilities are penalized more. We also observed that the gain from using this feature is dependent on the translation task and yields best results for in-domain tasks.

When using not only the target gap feature but also the source gap feature to translate data in a constrained setting, we observed a consistent improvement for different machine translation models. In the best of cases, we were able to gain 2 BLUE points which is a highly significant result. Finally, the combination of both improvements do not present particular improvements.

In conclusion, by using the alignment structural information, in particular the gaps in the alignment, we were able to turn a feature that originally appeared as a liability into an advantage. In other words, by using the knowledge gained during the development of this thesis, by identifying important alignment characteristics, controlling them and providing means to use the information to our advantage, we were able to improve machine translation.

Chapter 7

Conclusions

Improving word alignment quality has been a major focus of research in the SMT community. However, little attention was focused to understand the effects of alignment structure into the Machine Translation Pipeline. Therefore, despite the improvements reached in alignment quality, only modest improvements in translation performance were observed. In this dissertation, we studied how alignment structure impacts the translation model and how a translation model impacts translation performance. Additionally, we proposed alternatives to take into account the alignment structure and moderate its effects.

In this chapter, we summarize the body of knowledge accumulated through the development of this work. First, we revisit the observations made across different chapters. Then, we revisit our original hypothesis and research questions. Finally we wrap up with a discussion of the findings and propose future research directions.

7.1 Summary of our Findings

In this section we recapitulate the findings made throughout this thesis. From each chapter, we only extract the most important observations and consolidate them in a comprehensive list.

7.1.1 The Effects of Alignments on the Phrase-Based Translation Model

In Chapter 3, we performed a series of experiments that allowed us to compare different alignments according to their quality and structure. The observations made were the following:

- By analyzing word alignments according to several characteristics and comparing them to hand-aligned data, we observed that there is a lot of room of improvement for our alignment models. Both in terms of alignment quality and structure, our alignments are far from human generated data
- We observed that sparser word alignments lead to a higher number of extracted phrases, which ultimate results in larger phrase tables. While these larger phrase tables contain longer phrases, many of the phrases contain unaligned words.
- The number of unaligned words in the alignment have a large impact on the characteristics of the extracted phrase table.

- The unaligned words in the extracted phrase pairs follow the distribution of unaligned words in the alignment from where they were extracted.
- A manual evaluation of phrase pair quality showed that more unaligned words (gaps) result in a lower human perceived quality.
- Phrases extracted from human aligned data have better quality than computer generated word alignments.

7.1.2 The Characteristics of the Translation Model

In Chapter 4 we generated a large number of phrase-tables using different alignments. Then, we estimated regression models for different characteristics of the phrase-tables. Our most important findings were:

- We corroborated that density variables (gaps, links) have high correlation with many phrase-table characteristics, especially the number of entries and the length of the phrase-pairs.
- We discovered that the alignment distortion also plays an important role in determining certain phrase-table variables, such as the entropy of the lexical translation features.
- We observed that the phrase-extraction algorithm preserves the distribution of alignment density allowing it to carry to phrases. Thus, denser alignments yield denser phrase-pairs.
- The case is different for alignment dimensions and alignment distortion. We observed that phrase-extraction changes drastically the expected behavior of distortion variables (e.g. the more crossings in our alignment, the fewer crossing inside the phrase-pairs, because highly distorted phrases can't be extracted with the current heuristic).
- We observed that alignment quality as measured by the F-metric (AER) has little relationship with other phrase-table variables, confirming that phrase-extraction exploits alignment structure and has little to do with alignment quality.
- We discovered that the entropy in the phrase-translation probability can be predicted very accurately using alignment gaps.
- Lexical scores were more difficult to predict using the structural alignment features. While diagonality predicts about 30% of the variance in the entropy of these translation features, the model is not complete.
- By testing our models on unseen Chinese-English and Arabic-English data, we observed that many of the predictions hold well for different language pairs. This highlights the fact that phrase-extraction does not depend on linguistic features but rather on the structure of the alignment itself.

7.1.3 Predicting Translation Quality

In Chapter 5, we ran a series of experiments where we used full translation models to translate different data sets. Then, we predicted the translation quality using the information from the translation models. We discovered that:

- Predicting translation is not as easy as predicting the characteristics of a translation model. This is mainly because translation performance is dependent on the translation task. Thus, we analyzed data from a general perspective and from a translation-task targeted view.
- When aiming for a general model, the most important characteristics for translation quality come from the phrase table. In short, we need to have many translation options (from large translation models, with high inverse lexical entropy), that are well defined (with low lexical entropy). Additionally, hypotheses need to have good alignment support (with few gaps) and high fluency (with low language model cost).
- When dealing with easier translation tasks, we need that our final translation is fluent (with low language model cost), and that it has good alignment support (i.e. that the aggregate number of alignment gaps remains low).
- For difficult translation tasks we observed that using fewer, but more reliable translation options is preferable. Language fluency is also very important.
- While the generic models make weaker predictions, they hold across language pairs. On the other hand, targeted models are more dependent on the language pair. They performed very well for Spanish while presented lack-of-fit on Chinese-English.

7.1.4 Improving Translation Using Alignment Structure

Finally in Chapter 6, we introduced two alternatives to take advantage of alignment structure for improving machine translation at the alignment training and decoding levels. In summary:

- We proposed two alignment quality metrics (BA and F^0), which by taking the negative space of the alignment into account, incorporate more of the structure of the alignments.
- We observed that the Balanced Accuracy metric yields denser alignments whose structure is more similar to the hand-aligned data. However, these denser alignments inhibit the phrase-extraction process and result in phrase tables with fewer translation options.
- We concluded that having more human-alignment-like structure can be counterproductive when human alignments are very dense.
- We also introduced the F^0 measure which takes into account the NULL alignments into when computing the alignment quality score. This metric results in higher quality alignments with respect to both the F and F^0 metric and also improved translation quality. We observed that this improvement can be traced to the increase in translation options in the phrase-table coupled with other structural alignment differences.
- We also observed that our general translation quality prediction models hold well for predicting translation gains based on changes in alignment structure.
- At the decoding stage, we proposed to incorporate alignment structure by integrating the unaligned word count as a decoding feature. This modification works well for translation models coming from discriminative alignments for which produce the best translations overall.

- When using not only the target gap feature but also the source gap feature to translate data in a constrained setting, we observed a consistent improvement for different machine translation models.
- By using the alignment structural information, in particular the gaps in the alignment, we were able to turn a feature that originally appeared as a liability into an advantage. In other words, by using the knowledge gained during the development of this thesis, by identifying important alignment characteristics, controlling them and providing means to use the information to our advantage, we were able to improve machine translation. In the best of cases, we were able to gain 2 BLEU points which is a highly significant result.

7.2 Hypothesis and Research Questions Revisited

In Section 1.2 we presented our main hypothesis and dissected into three different parts. The first part was related to the effects that alignment structure has in the consolidation of our translation model. The second part was related to the influence of alignments on translation performance. The third part was related to the modulation of alignment structure in favor of translation performance. Below, we revisit each point individually.

The impact of alignment structure on the translation model

We observed that structural alignment characteristics influence greatly the characteristics of the resulting translation model. Our models explain accurately the variations in the size and lexical diversity of our translation models using only alignment structure variables as predictors. Furthermore, these models hold very well for different language pairs, which highlights the fact that phrase-extraction in its current form is very dependent on the structure of the alignments.

The impact of alignment structure of the translation model on MT performance

Translation performance depends largely on the translation task. Thus, the variables of the translation model that are more useful differ depending on the domain of application. However, we were able to obtain predictive models that explain translation performance in general moderately well. These models rely upon the characteristics of the translation model to explain differences in performance. Additionally, we also observed that precisely these characteristics are largely influenced by the structure of the alignment.

Providing means to control alignment structure will result in improvements in MT performance

We discovered that the right engineering of features can help to improve performance. We also observed that the human alignment is not the best gold-standard in terms of structure, thus having discriminative training to imitate human alignment's structure is not recommended. We observed that alignment target gaps explained some of the variance in translation performance. When we added alignment gap information to the decoder, we observed that we could achieve better estimation and improve performance

for discriminative alignments. Furthermore, in cases of limited training data, these new features helped to improve estimation and yield better translation results

In addition to our hypothesis, there were different research questions that were initially posed. Below we address them individually.

Which variables describing word alignments, translation models and translation hypotheses that we are going to include in this study?

Our study was motivated by previous research. Thus, we considered some variables that had already been used for describing alignment structure (e.g. number of links, crossings). Additionally we proposed new measures for alignment such as the alignment gaps or the alignment diagonality. We measured alignment structure at several stages of the pipeline. First, we measured word alignments directly, then we measured the alignments embedded into the phrase pairs and finally we measured the alignment structure of the phrase-pairs used for translation.

How are we going to characterize the decoding search space? Which variables are we going to use to discriminate between good and bad translation hypotheses?

In addition to the cost of the decoding features, we used the alignment's variables for the translation hypotheses. Based on those features we constructed models to predict translation quality.

How are we going to deal with undesired effects in multivariate analysis such as collinearity?

To deal with multicollinearity, we opted for the use of feature selection (stepwise) and retained those features that yielded stable models. While other options were available (regularization, dimension reduction), this choice was motivated by a preference towards model interpretability.

Which type of multivariate analysis better suits our scenario?

Also to favor interpretability, we opted for multivariate linear regression.

Which model or multivariate technique should be used to build our model?

While we tested several approaches to build complex models, we opted to use the simpler models, to favor the interpretation of the results.

How do we compare predictive models? How do we evaluate their robustness?

We employed the determination coefficient (R^2) on unseen data. Such a metric gives us an idea of how well a model fits our dependent variables.

How do we control alignment structure for Machine Translation?

As we mentioned before, we opted for two different alternatives: using discriminative training towards an alignment structure aware metric, and using decoding features that incorporate alignment structure (in form of alignment gaps) to refine its translation model.

7.3 Final Remarks

During the development of this dissertation we have discovered how different characteristics of the alignment impact Machine Translation. We discovered that while good quality alignments yielded good phrase-pairs, the consolidation of a translation model is more dependent on the alignment structure than alignment quality. We observed that human-alignments are more dense than the computer generated counterparts, which trend to be more sparse and precision-oriented. Trying to emulate human-like alignment structure resulted in poorer systems, despite that denser alignments with fewer alignment gaps have better phrase-pairs. This is due to the fact that those translation models trend to be more compact and lack translation options. On the other hand, more translation options, even if they are noisier, help to improve the quality of the translation. This is due to the fact that translation does not rely only on the translation model, but also on other factors (e.g. the language model). Lastly, when we provide the decoder with features that help it to make “more informed decisions” we observe a clear improvement in translation quality. This was specially true for the discriminative alignments which inherently leave more words unaligned. The result is more evident in low-resource settings where having larger translation lexicons represent more translation options. Using simple features to help the decoder discriminate translation hypotheses, clearly showed consistent improvements.

By performing a detailed analysis, and understanding better how phrase-based translation works, we were able to better engineer decoding features. We are confident that there are more directions to explore in terms of alignment structure. In the last section of this document, we briefly discuss them.

7.4 Future Work

In this dissertation we have observed how different alignment characteristics affect phrase-based statistical machine translation. We also observed how adding structural alignment information can improve translation quality specially for discriminative alignments. However, this is an ongoing process and there are several directions in which this research can be expanded.

Investigate the effect of coupling Alignment enhancements such as the F^0 training with the use of structural alignment decoding features such as the gap features.

We would like to explore the interaction between tuning towards an structure-aware alignment metric and structure-aware decoding features. This could lead to improvements specially for the F^0 alignment which is more sparse, with more alignment gaps (and more noise) that could be smoothed using the gap decoding features.

Use other alignment characteristics into decoding. Specially, distortion features such as the number of crossings and the diagonality of the translation hypothesis alignment.

In this dissertation we explored the inclusion of alignment gaps as a decoding feature and we observed improvements in translation quality. However, we could include more decoding features that incorporate alignment structure, such as diagonality or number of

link crossings. Optimizing log-linear weights for a higher dimensionality set of features would require some modifications of our training approach. In order to do so, we should move from a linear-descent-based optimization metric such as MERT towards one more resilient to higher dimensionality such as MIRA (Chiang et al., 2008; Watanabe et al., 2007).

Explore other regression model specification alternatives.

During our prediction model specification, we faced some issues related to higher dimensionality. We opted to use feature selection to resolve them. However, we could explore the use of mapped features (to build polynomial models) coupled with regularization. In that sense, we would be able to explore the interaction of several features and higher order terms. On the other hand, to cope with the original dimensionality problem, we could use feature dimension reduction techniques such as PCA to reduce the impact of collinearity. This would require that we make an accurate description of the principal components to ensure that the interpretation remains consistent.

Propose a complex hierarchical model that combines regression from alignment to phrase-tables and regression from phrase-tables to translation quality.

In this thesis, we provided two different prediction models that work at two separate stages. The situation arises from the fact that they necessarily have to be estimated using separate samples. Anyhow, we could couple both of their findings into a single more complex model, by using multivariate techniques such as path analysis or structural equation modeling.

Explore new optimization techniques.

We observed that the translation quality models we obtained are susceptible to MERT optimization. Specially when adding new features to decoding that change the decoder's search space. Thus it would be interesting to incorporate linear regression modeling coupled with response surface methods to propose a new optimization scheme for Machine Translation.

Appendix A

Baseline System

In this annex we describe the data used for training and testing of the Spanish-English systems used through this dissertation.

A.1 Translation Training Data

In this section we introduce the characteristics of our baseline Spanish-English translation system. We start by describing the domain of the data and its characteristics. We also outline the preprocessing steps performed to clean the data.

A.1.1 Data Sources

The source of the data is European Parliament Proceedings (EUROPARL Version 5), News commentary text, as well as the United Nations proceedings from year 2000 as provided for WMT 2010 competition. The total number of lines of data is displayed in Table A.1.

A.1.2 Preprocessing

For this data, we performed three preprocessing steps: tokenization, lowercasing and truncation. In fig. A.1 we present an illustration of the first two processes.

raw input

The Parliament''s proceedings have been handed out .

after tokenization

The Parliaments `s proceedings have been handed out .

after lowercasing

the parliaments `s proceedings have been handed out .

Figure A.1: Example of preprocessing

Tokenization splits words into simpler forms. It separates punctuation marks from words, so they are considered as single words. Thus it is expected to find longer sentences after the

procedure. However, the vocabulary size is expected to go down because morphology complexity is reduced. Lowercasing does not change to the distribution of length of the sentences, however further reduces the vocabulary. Truncation removes sentences longer than a certain limit. This is recommended due to the aligners’ memory/processing limitations. For this corpus, we removed sentences longer than 50 words using standard cleaning corpus script from Moses decoder. This script removes empty lines, removes redundant space characters, drops lines (and their corresponding lines) that are empty, lines that are too short, too long or violate the 9-1 sentence ratio limit for GIZA++.

In Table A.1, we present the statistics for RAW and Preprocessed (PP) Data

Set	RAW			PP			Change(%)		
	Lines	Tok	Voc	Lines	Tok	Voc	Lines	Tok	Voc
Spanish									
EU	1.7M	43.1M	393.1K	1.4M	35.1M	140.0K	13%	19%	64%
NC	98.6K	2.5M	123.1K	90.0K	2.3M	59.2K	9%	8%	52%
UN	6.2M	190.6M	1.4M	4.9M	129.8M	330.1K	21%	32%	76%
<i>total</i>	8.0M	236.2M	1.6M	6.4M	167.2M	387.9K	19%	29%	76%
English									
EU	1.7M	41.2M	286.7K	1.4M	33.9M	91.9K	13%	18%	68%
NC	98.6K	2.1M	101.9K	90.0K	2.0M	43.0K	9%	4%	58%
UN	6.2M	164.3M	1.3M	4.9M	115.3M	305.2K	21%	30%	76%
<i>total</i>	8.0M	207.6M	1.5M	6.4M	151.1M	346.3K	19%	27%	76%

Table A.1: Statistics for Raw and preprocessed data for Europarl (EU), News Commentary (NC) and UN training data. We present the total number of training examples (lines), number of tokens (tok) and the vocabulary size (voc).

From this table, we can observe that preprocessing indeed reduces vocabulary (unique words) size while preserving most of the content. If we look at the reduction in vocabulary size (Change column), we find that more than 50% of the vocabulary is reduced, while the total number of tokens (word instances) is reduced at most 32%. This drastic drop in number of words can be attributed mostly to the cleaning part, where we drop non-conforming sentences (and thus, some non-functional content).

A.2 Translation Test Data

In this section we provide some statistics about the data used in our translation tests. The sources for our data come from different years of the WMT competition, and each consists of samples coming from different domains:

- Europarl data
Consists of proceedings from the European Parliament. Test sets: WMT06, WMT07, WMT08.

Spanish-English data	
Dataset	Size
AC	4107
NC07	2007
NC08	2028
NW09	2525
NW10	2489
WMT06	2000
WMT07	2000
WMT08	2000

Table A.2: Datasets used for translation

- **Acquis Communautaire corpus**
Consists of European law. Test sets: AC.
- **News Commentary**
Consists of news commentary (editorials) from different sources. Test sets: NC07, NC08
- **News Wire**
Consists of excerpts of news from European newspapers (Le Monde, El País, etc). Test sets: NW10, NW09

Depending on their domain, each set can be considered to be in or out of domain. For instance WMT06 through WMT08 are considered in-domain because the translation tasks come from the Europarl Corpus. On the other hand, NW and NC data is considered out of domain. Below, we present the statistics for this data.

A.3 Hand Aligned Data

In the following table we present the statistics for the Spanish-English hand alignment data from EPPS. The original data consisted of 500 lines. But after preprocessing the total amount was reduced to 420. We then proceeded to divide that set and use 200 lines for training and 220 for testing.

	Length	EN-Words	ES-Words	Links
RAW				
EPPS-dev	100	3162	3310	3969
EPPS-test	400	12990	13665	17475
total	500	16152	16975	21444
Preprocessed				
EPPS-dev	84	2075	2082	2858
EPPS-test	336	7997	8398	11915
total	420	10072	10480	14773
Final Set				
final-dev	200	4760	4887	6779
final-test	220	5312	5593	7994
total	420	10072	10480	14773

Table A.3: Data Statistics for Spanish-English hand alignments

Appendix B

Variables

In this annex we describe in detail the variables used across this study.

B.1 Alignment Variables

Given a set of alignments \mathcal{A} where each alignment $A_k \in \mathcal{A}$ is represented by matrix of I_k source words by J_k target words, $l_k(i, j)$ represents a link between the i th source word and the j th target word, and L_k represents the total number of links in the alignment A_k , we can compute the following statistics.

B.1.1 Alignment Dimension

ASL Average number of source words

$$ASL_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} I_k \quad (\text{B.1})$$

ATL Average number of target words

$$ATL_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} J_k \quad (\text{B.2})$$

w The normalized alignment matrix width

The geometric average of the source and target lengths of an alignment k

$$W_k = \sqrt{I_k J_k} \quad (\text{B.3})$$

B.1.2 Alignment Density

ANLK Average number of links

This variable represents the average number of links.

$$ANLK_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} L_k \quad (\text{B.4})$$

ALK Average link density

This variable represents the normalized number of links per word in the alignment matrix.

$$ALK_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \sum_{l \in A_k} \frac{L_k}{W_k} \quad (\text{B.5})$$

ASLK Average number of links per source word

This variable represents the average number of links per source word.

$$ASLK_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \frac{L_k}{I_k} \quad (\text{B.6})$$

ANSG Average number of source gaps

This variable represents the average number of unaligned source words.

$$ANSG_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \sum_{i \in I_k} \prod_{j \in J_k} (1 - l(i, j)) \quad (\text{B.7})$$

ASG Average number source gaps per normalized number of words

This variable represents the average number of unaligned source words divided by the normalized length of the matrix.

$$ASG_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \frac{1}{W_k} \sum_{i \in I_k} \prod_{j \in J_k} (1 - l(i, j)) \quad (\text{B.8})$$

B.1.3 Alignment Distortion

ACR Average number of alignment crossings

This variable represents the average the number of crossings in an alignment normalized by the matrix width.

$$ACR_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \frac{\text{cross}(A_k)}{W_k} \quad (\text{B.9})$$

The number of crossings cross are calculated in the by the algorithm B.1.

ADG Average alignment diagonality

This variable represents the average diagonality of an alignment set. Diagonality is defined as the absolute correlation between the positions in the i and j positions of the links in the alignment A_k .

$$DG_{A_k} = |\rho_{i(l), j(l)}| \quad (\text{B.10})$$

This gives us a notion of the monotonicity in the positions of the alignments.

$$ADG_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} DG_{A_k} \quad (\text{B.11})$$

Algorithm B.1 Calculate the number of crossings cr in a set of links l of size L

```

 $cr \leftarrow 0$ 
sort( $l$ )
for  $k = 1 \rightarrow L$  do
  ( $i_1, j_1$ )  $\leftarrow l_k$ 
  for  $m = k + 1 \rightarrow L$  do
    ( $i_2, j_2$ )  $\leftarrow l_m$ 
    if ( $i_1 > i_2$  and  $j_1 < j_2$ ) or ( $i_1 < i_2$  and  $j_1 > j_2$ ) then
       $cr \leftarrow cr + 1$ 
    end if
  end for
end for
return  $cr$ 

```

ADT Average relative distortion

This variable represents the distortion between the i th and j th positions of the links in the alignments.

$$ADT_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{A_k \in \mathcal{A}} \sum_{l \in A_k} \frac{|i(l) - j(l)|}{W_k} \quad (\text{B.12})$$

B.1.4 Alignment Quality

P Precision

R Recall

F F measure

AER Alignment Error Rate For practical purposes $AER = 1 - F$.

BA Balanced Accuracy

B.2 Phrase-table Variables

For each phrase-pair P_k in the phrase table \mathcal{P} , we have I_k source words in the source phrase and J_k target words in the target side, $l_k(i, j)$ represents a link between the i th source word and the j th target word in the phrase-pair, and L_k represents the total number of links in the alignment embedded in P_k . Additionally, let $|\mathcal{P}_{\mathcal{I}}|$, $|\mathcal{P}_{\mathcal{J}}|$ be the total number of unique source and target phrases, respectively.

B.2.1 Alignment Variables

PSG,PTG,PLK,PDT,PDG,PCR are calculated as its alignment counterparts.

PWSG Average number of source gaps per word

This variable represents the average proportion of source unaligned words per source word.

$$PWSG_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{P_k \in \mathcal{P}} \frac{1}{I_k} \sum_{i \in I_k} \prod_{j \in J_k} (1 - l(i, j)) \quad (\text{B.13})$$

B.2.2 Phrase-table Dimension

PSL Average source phrase length

$$PSLP = \frac{1}{|\mathcal{P}|} \sum_{P_k \in \mathcal{P}} I_k \quad (\text{B.14})$$

PTL Average target phrase length

$$PTL_{\mathcal{P}} = \frac{1}{|\mathcal{P}|} \sum_{P_k \in \mathcal{P}} J_k \quad (\text{B.15})$$

B.2.3 Phrase-table Entries

PNE Total number of entries in the phrase-table

This quantity is equivalent to $|\mathcal{P}|$. Also referred as the unique number of phrase-pairs PNU.

PNI Total number of extracted phrase-pairs

This quantity is the number of non-unique phrase-pairs extracted from an alignment set.

PNU Total number of unique extracted phrase-pairs

See PNE.

B.2.4 Coverage Variables

PSU Percentage of source unique phrases

$$PSU_{\mathcal{P}} = \frac{|\mathcal{P}_{\mathcal{I}}|}{|\mathcal{P}|} \quad (\text{B.16})$$

PTU Percentage of target unique phrases

$$PSU_{\mathcal{P}} = \frac{|\mathcal{P}_{\mathcal{J}}|}{|\mathcal{P}|} \quad (\text{B.17})$$

B.2.5 Translation Entropies

PT1 Average Entropy of the inverse phrasal probability

$$PT1_{\mathcal{P}} = -\frac{1}{|\mathcal{P}_{\mathcal{I}}|} \sum_{(e_k, f_k) \in \mathcal{P}} p(f_k|e_k) \log p(f_k|e_k) \quad (\text{B.18})$$

PT2 Average Entropy of the inverse lexical probability

$$PT2_{\mathcal{P}} = -\frac{1}{|\mathcal{P}_{\mathcal{I}}|} \sum_{(e_k, f_k) \in \mathcal{P}} p_{lex}(f_k|e_k) \log p_{lex}(f_k|e_k) \quad (\text{B.19})$$

PT3 Average Entropy of the direct phrasal probability

$$PT3_{\mathcal{P}} = -\frac{1}{|\mathcal{P}_{\mathcal{J}}|} \sum_{(e_k, f_k) \in \mathcal{P}} p(e_k|f_k) \log p(e_k|f_k) \quad (\text{B.20})$$

PT4 Average Entropy of direct lexical probability

$$PT4_{\mathcal{P}} = -\frac{1}{|\mathcal{P}_{\mathcal{J}}|} \sum_{(e_k, f_k) \in \mathcal{P}} p_{lex}(e_k|f_k) \log p_{lex}(e_k|f_k) \quad (\text{B.21})$$

B.3 Translation (First-best) Variables

In this part we define the first-best hypothesis variables. For compactness, we define only the most important. For each first-best translation F_k from source sentences in document \mathcal{D} , we have I_k source words in the source phrase and J_k target words in the target side, $l_k(i, j)$ represents a link between the i th source word and the j th target word in the phrase-pair, and L_k represents the total number of links in the alignment embedded in F_k . Additionally, let $|\Phi_k|$ be the number of phrases used in translation F_k . Then:

B.3.1 Alignment Variables

FSG,FTG,FLK,FDT,FDG,FCR are calculated as its alignment counterparts

B.3.2 Cost Variables

Cost variables are calculated by computed the negative log-likelihood of the probability in question. For instance, for language models:

FLM Average language model cost

$$FLM_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{F_k \in \mathcal{D}} -\log(p_{lm}) \quad (\text{B.22})$$

For the rest of the cost variables, the procedure is the same.

B.3.3 Translation Quality

BLEU Bleu

MET Meteor

TER Translation Error Rate

Bibliography

- Ayan, N. F. and Dorr, B. J. (2006). Going Beyond AER: An Extensive Analysis of Word Alignments and Their Impact on MT. In *Proc. of the Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, pages 9–16.
- Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. of the Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72.
- Berger, A., Della Pietra, S., and Della Pietra, V. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22:39–71.
- Birch, A., Osborne, M., and Koehn, P. (2008). Predicting Success in Machine Translation. In *Proc. of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 745–754.
- Blunsom, P. and Cohn, T. (2006). Discriminative Word Alignment with Conditional Random Fields. In *Proc. of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association of Computational Linguistics (COLING-ACL)*, pages 65–72.
- Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics*, 19(2):263–311.
- Brown, P. F., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79 – 85.
- Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proc. of the Conference of the European Association for Machine Translation (EACL)*, pages 249–256.
- Chen, S. F. and Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. In *Proc. of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, pages 310–318.

- Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 263–270.
- Chiang, D., Marton, Y., and Resnik, P. (2008). Online Large-margin Training of Syntactic and Structural Translation Features. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 224–233.
- Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. Number Wiley Series in Telecommunications in Wiley Series in Telecommunications. Wiley.
- Denkowski, M. and Lavie, A. (2011). Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In *Proc. of the Workshop on Statistical Machine Translation at the Conference on Empirical Methods in Natural Language Processing*, pages 85–91.
- Fraser, A. and Marcu, D. (2007). Measuring Word Alignment Quality for Statistical Machine Translation. *Computational Linguistics*, 33:293–303.
- Gao, Q. and Vogel, S. (2008). Parallel Implementations of Word Alignment Tool. In *Proc. of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57.
- Guzman, F., Gao, Q., Niehues, J., and Vogel, S. (2011). Word Alignment Revisited. In Olive, J., Christianson, C., and McCary, J., editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, pages 164–175. Springer.
- Guzman, F., Gao, Q., and Vogel, S. (2009). Reassessment of the Role of Phrase Extraction in PBSMT. In *Proc. of the Machine Translation Summit XII*, pages 49–56.
- Guzman, F. and Garrido, L. (2008). Translation Paraphrases in Phrase-based Machine Translation. In *Proc. of the Conference on Computer Linguistics and Natural Language Processing (CICLing)*, Lecture Notes in Computer Science, pages 388–398. Springer Berlin Heidelberg.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2010). *Multiple Regression Analysis*. Prentice Hall, 7 edition.
- Hutchins, J. (2003). Machine Translation: General Overview. In *Oxford Handbook of Computational Linguistics*, chapter 27, pages 501–511. Oxford University Press.
- Ittycheriah, A. and Roukos, S. (2005). A Maximum Entropy Word Aligner for Arabic-English Machine Translation. In *Proc. of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 89–96.
- Johnson, R. A. and Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*. Prentice Hall, 5th edition.

- Jurafsky, D. and Martin, J. H. (2007). Machine Translation. In *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.*, chapter 25. Prentice Hall.
- Kobdani, H., Fraser, A., and Schütze, H. (2009). Word Alignment by Thresholded Two-Dimensional Normalization. In *Proc. of the Machine Translation Summit XII*, pages 260–267.
- Koehn, P. (2004a). Pharaoh: A Beam Search Decoder for Phrase-based Statistical Machine Translation. In *Proc. of Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 115–124.
- Koehn, P. (2004b). Statistical Significance Tests for Machine Translation Evaluation. In *Proc. of Conference on Empirical Methods for Natural Language Processing (EMNLP)*, volume 4, pages 388–395.
- Koehn, P., Hoang, H., Birch, A., Callison-burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proc. of Association of Computer Linguistics*, pages 177–180.
- Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proc. of Human Language Technology Conference of the North American Chapter of the Association of Computer Linguistics (HLT-NAACL)*, pages 127–133.
- Lambert, P., Gispert, A., Banchs, R., and Mariño, J. B. (2006). Guidelines for Word Alignment Evaluation and Manual Alignment. *Language Resources and Evaluation*, 39(4):267–285.
- Lambert, P., Ma, Y., Ozdowska, S., and Way, A. (2009). Tracking Relevant Alignment Characteristics for Machine Translation. In *Proc. of the Machine Translation Summit XII*, pages 268–275.
- Lambert, P., Petitrenaud, S., Ma, Y., and Way, A. (2010). Statistical Analysis of Alignment Characteristics for Phrase-based Machine Translation. In *Proc. of the 14th Annual conference of the European Association for Machine Translation (EAMT)*.
- Liang, P., Taskar, B., and Klein, D. (2006). Alignment by Agreement. In *Proc. of the Human Language Technology Conference of the North American Chapter of the Association of Computer Linguistics (HLT-NAACL)*, pages 104–111.
- Lopez, A. and Resnik, P. (2006). Word-based Alignment, Phrase-based Translation: What’s the Link? In *Proc. of the 7th Biennial Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 90–99.
- Manning, C. and Schütze, H. (1999). *Foundations of Natural Language Processing*. MIT Press.

- Marcu, D. and Wong, W. (2002). A Phrase-based, Joint Probability Model for Statistical Machine Translation. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, pages 133–139.
- Matusov, E., Zens, R., and Ney, H. (2004). Symmetric Word Alignments for Statistical Machine Translation. In *Proc. of the 20th International Conference on Computational Linguistics (COLING)*, pages 219–225.
- Melamed, I. D. (2000). Models of Translation Equivalence Among Words. *Computational Linguistics*, 26(2):221–249.
- Moore, R. (2005). A Discriminative Framework for Bilingual Word Alignment. In *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 81–88.
- Ney, H., Essen, U., and Kneser, R. (1994). On Structuring Probabilistic Dependencies in Stochastic Language Modelling. *Computer Speech and Language*, 8:1–38.
- Niehues, J. and Vogel, S. (2008). Discriminative Word Alignment via Alignment Matrix Modeling. *Computational Linguistics*, pages 18–25.
- Och, F. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proc. of the 41st Annual Meeting on Association for Computational Linguistics (ACL)*, pages 160–167.
- Och, F. and Ney, H. (2004). The Alignment Template Approach to Statistical Machine Translation. *Computational Linguistics*, 30(4):417–449.
- Och, F., Tillmann, C., Ney, H., and Others (1999). Improved Alignment Models for Statistical Machine Translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28.
- Och, F. J. (2000). GIZA++: Training of Statistical Translation Models. Technical report, RWTH Aachen, University of Technology.
- Och, F. J. and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 295–302.
- Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-j. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of Association for Machine Translation in the Americas (AMTA)*, pages 223 – 231.

- Somers, H. (2003). Machine Translation: Latest Developments. In *The Oxford Handbook of Computational Linguistics*. Ed. University Press. Oxford, chapter 28, pages 512–528. Oxford University Press.
- Taskar, B. and Lacoste, J. (2005). A Discriminative Matching Approach to Word Alignment. In *Proc. of Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 73–80.
- Vilar, D. and Ney, H. (2006). AER: Do We Need to “Improve” our Alignments? In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, pages 205–212.
- Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based Word Alignment in Statistical Translation. In *Proc. of the 16th Conference on Computational Linguistics (COLING)*, pages 836–841.
- Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. (2007). Online Large-Margin Training for Statistical Machine Translation. In *Proc. of EMNLP-CoNLL*, pages 764–773.
- Zhang, Y. and Vogel, S. (2004). Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. In *Proc. of International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2004)*.

Tecnológico de Monterrey, Campus Monterrey



30002007498561

<http://biblioteca.mty.itesm.mx>