NAIST-IS-MT1251126

Master's Thesis

Improving Word Alignment for Statistical Machine Translation by Prediction of Unalignable Words

Frances Pikyu Yung

March 13, 2014

Department of Information Systems
Graduate School of Information Science
Nara Institute of Science and Technology

A Master's Thesis
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Master of ENGINEERING

Frances Pikyu Yung

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Satoshi Nakamura	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Kevin Duh	(Co-supervisor)

Improving Word Alignment for Statistical Machine Translation by Prediction of Unalignable Words*

Frances Pikyu Yung

Abstract

Word alignment is a common first step in a Statistical Machine Translation (SMT) system, yet professional human translators do not translate word-for-word but sense-for-sense. The discrepancy results in unalignable words. Based on the contrast of the language pair, certain words can be predicted to be additional to the translation to fulfill senses without a correlated word in the source text. This study analyzes the distribution and linguistic properties of unalignable words based on a hand-aligned parallel corpus of Chinese and English. Based on the findings, a method is proposed to predict and pre-remove unalignable words prior to automatic word alignment, so as to improve the alignment accuracy. The method improves word alignment accuracy of GIZA^{++} when evaluated against manual annotation, as well as the end-to-end SMT results.

Keywords:

statistical machine translation, word alignment, linguistic knowledge

*Master's Thesis, Department of Information Systems, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-MT1251126, March 13, 2014.

対応のない単語の予測による統計的機械翻訳のためのワードアライメントの改善*

フランセス　ピキュー　ヤング

内容梗概

ワードアラインメント（対訳単語の整列）は、一般な統計的機械翻訳システムの最初の手順であるが、人間の翻訳プロは単語単位ではなく、概念単位で翻訳する。その不一致さによってアライン不可能（対訳単語が訳文に見つからない）な単語が発生する。一方、ターゲット言語の概念を果たすために訳文に追加された、原文に対訳語のない単語は、言語の間の差異に基づいて予測することが出来る。本研究は、人手でワードアラインメントを取った中国語ー英語の対訳コーパスにおけるアライン不可能な単語の分布や言語学的な特性を分析し、自動ワードアラインメントを行う前にアライン不可能な単語を予測して外し、アラインメント精度を高める方法を提案した。人手のアラインメントに対して評価した結果、GIZA^{++}による自動ワードアラインメントの精度が改善された。統計的機械翻訳システムの最終翻訳結果も改善された。

キーワード

機械翻訳　ワードアラインメント　言語学知識

*奈良先端科学技術大学院大学 情報科学研究科 情報システム学専攻 修士論文, NAIST-IS-MT1251126, 2014 年 3 月 13 日.

Acknowledgments

First of all, I would like to thank Matsumoto-sensei, Nakamura-sensei, Shimbo-sensei and Kevin-sensei for their kind instructions and encouragement, especially Kevin-sensei for his frequent thoughtful discussions and bitter sweet surprises. I had a great time in NAIST surrounded by interesting lab mates of the Computational Linguistic Lab. and numerous friends from the 'international community'.

Also, I would like to thank MEXT for offering me scholarship, and the government of Ikoma city, especially all staff of Pure Nursery, Shikanodai Elementary School and Shikanodai Gakudo, for their fabulous childcare and education. They made my study in Japan possible.

Most importantly, I must thank by lovely kids Yoshi and Jiro for tolerating my speed cooking, irritation before deadlines, and occasional drunken nights. I love you two most!

Lastly, I ought to thank my distant partner, T, for doing nothing in particular.

Contents

Acknowledgments **v**

1 Introduction **1**
1.1 Background . . . 2
1.1.1 Statistical machine translation . . . 2
1.1.2 The IBM Models . . . 3
1.1.3 Automatic Evaluation . . . 4
1.2 Related works . . . 5
1.2.1 Empty categories in translation . . . 5
1.2.2 Link deletion method . . . 6

2 Method **7**
2.1 Motivation . . . 7
2.2 Proposed method . . . 9

3 Corpus Analysis **13**
3.1 The ORACLE dataset . . . 13
3.2 Property of unalignable words . . . 15
3.2.1 Distribution among word types . . . 16
3.2.2 Distribution among part-of-speech . . . 17

4 Experiments **21**
4.1 Removing hand annotated unalignable words . . . 21
4.1.1 Settings . . . 22
4.1.2 Results . . . 23
4.2 Automatic prediction of unalignable words . . . 24
4.2.1 Features . . . 24
4.2.2 Settings . . . 26
4.2.3 Results . . . 27

4.3 Removing predicted unalignable words from REAL data 29
4.3.1 The REAL dataset 29
4.3.2 Settings 30
4.3.3 Results 31

5 Discussion **33**
5.1 Data size V.S. performance 33
5.1.1 Size of classifier training data 33
5.1.2 Size of MT training data 34
5.2 Output analysis 37

6 Conclusion **39**

7 Future work **41**

Bibliography **43**

List of Figures

2.1 Visualization of the effect of removing and restoring *unalignable* words 11

5.1 Accuracy of unalignable word classifiers and BLEU scores (before tuning) of hierarchical MT systems built on the REAL dataset v.s. various portion of ORACLE training data used for classifier training. The BLEU score of the baseline system was 17.02. 34

5.2 BLEU scores over REAL test set. Size is in number of sentences. *'+ Unalignment'* refers to PBMT systems built with automatic unalignable word prediction. 35

5.3 Precision, recall and F1 of GIZA++ output on ORACLE training slice . *'+ Unalignment'* refers to PBMT systems built with automatic unalignable word prediction 36

List of Tables

3.1 No. of core and unalignable words in the hand aligned ORACLE corpus 16
3.2 Coverage of *unalignable* tokens by top unalignable word types 17
3.3 Top 5 POS categories of Chinese and English unalignable words . . . 18

4.1 Word alignment accuracies of GIZA^{++}and evaluation scores of MT systems of the ORACLE dataset. The best score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05. 23
4.2 Accuracy of the unalignable word classifier and the effect on the resulting GIZA^{++} alignment and MT performance by removing the predicted unalignable words beforehand (on ORACLE data). Classification based on a list of top 30 token-POS pairs most frequently annotated as unalignable is shown for comparison. B, T and M stand for the MT evaluation scores of BLEU, TER and METEOR respectively. The highest score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline (no preprocessing) below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05. 28
4.3 Evaluation scores of MT systems built on the REAL dataset with and without pre-deduction of unalignable words. Evaluation is made against 4 references. The highest score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05. . 31

5.1 Examples of translations exclusively found in the top **15** lexical translation produced by the Baseline or Proposed (default GIZA^{++}) systems. 37
5.2 Translation examples from baseline and *unalignment* MT systems . . 38

Chapter 1

Introduction

In a typical statistical machine translation (SMT) system, translation rules are extracted from texts and their translation based on the correspondence between words of the two languages. Automatic word alignment is thus the first step of the SMT pipeline and is commonly performed by the IBM Model 4 [2], as implemented by GIZA^{++}. On the other hand, word-level correspondence is not guaranteed since the translation corpus has not been produced in a bottom-up manner by word alignment. Due to this contradiction, there are words that are unalignable across languages, even by hand, especially between distant language pairs.

Led by the contradiction in human translation and word alignment based machine translation, this study offers the following hypotheses: (1) There is not a distinct list of always unalignable words across two languages since a word can be part of different sense units. (2) Given the target language for translation, it is possible to predict source words that cannot be aligned with any words in the target language. (3) Removing unalignable words by prediction as a preprocess helps improve an SMT system.

To evaluate the validity of the hypotheses, the cross-lingual lexical contrast is examined by an analysis of the properties of alignable and unalignable words according to a hand-aligned Chinese-English corpus. Next, useful features are identified for the classification of alignable and unalignable words and a supervised classifier is built to predict unalignable words from the training corpus. Using a preprocessing technique, automatic prediction of unalignable words is incorporated to an SMT system. The experiment results show that the method improves both the accuracy of the automatic word alignment by IBM Model 4, as well as the overall performance of the SMT sys-

tem.

The rest of this chapter explains the background of the work by an overview of statistical machine translation, followed by an introduction of related works. Chapter 2 explains the motivation behind this work and a proposal to improve word alignment based on the motivation. Chapter 3 presents the findings of an analysis on manual word alignments concerning unalignable words. Chapter 4 describes a series of experiments using annotated and unannotated data. The experimental results are discussed in Chapter 5. Finally, a conclusion will be drawn in the Chapter 6 and future research directions will be described in Chapter 7.

1.1 Background

1.1.1 Statistical machine translation

A statistical machine translation system basically consists of a translation model and a language model. The translation model produces translation rules to render source language units to target language units, while the language model generates fluent target texts according to context. The two models are combined by the noisy-channel model, where **f** is a foreign source sentence and **e** is the English translation:

$$\begin{aligned} argmax_e p(\mathbf{e}|\mathbf{f}) &= argmax_e \frac{p(\mathbf{f}|\mathbf{e})p(\mathbf{e})}{p(\mathbf{f})} \\ &= argmax_e p(\mathbf{f}|\mathbf{e})p(\mathbf{e}) \end{aligned} \tag{1.1}$$

The earliest translation model is a word-based model, which formulates the translation of a sentence by splitting the sentence into words and outputting the target word(s) with the highest lexical translation probability for each source word as the translation. Estimation of the lexical translation probability is based on the statistics of a bilingual corpus, which is aligned to the sentence level. However, word-level correspondence between the source and target texts is necessary to derive the lexical translation probability distribution. The mapping of each source word to the corresponding target

word(s) in the bilingual training corpus is known as word alignment.

Later translation models extract translation rules of higher linguistic levels, such as multi-word phrases and syntactical trees, yet the rules are still based on word level correspondences. Thus, although word alignment is the by-product of word-based SMT model, it is also the basis of higher translation models. Nonetheless, the effect of automatic word alignment on the overall MT result is not straightforward. Fraser and Marcu [9] reports that measuring word alignment quality by weighted F-measure best correlates with the overall MT outputs.

1.1.2 The IBM Models

The earliest SMT models, the IBM Models [2], are generative models that generate the translation of a sentence along with its probability estimation based solely on the lexical translation probability of each word. Word alignment is induced from bilingual sentence pairs using the expectation maximization (EM) algorithm. Starting from IBM Model 1, where word order is not considered, higher IBM Models include components to model the structure of alignment more specifically: word positions in Model 2, fertility and word order distortion in Model 3, relative distortion in Model 4, and deficiency in Model 5.

Fertility of a source word refers to the number of target words the source word can produce. A NULL word token is added to each source and target sentence such that unaligned words are represented by alignments with the NULL token. To capture many-to-many links, word alignments in both translation directions, i.e. from source to target and from target to source, have to be combined. Various symmetrization heuristics can be used, ranging from the most conservative intersection and the greediest union [36]. The *grow-diag-final-and* heuristic is a widely-used standard between the two extremes.

The IBM Models are word-based SMT models, yet the word alignment algorithm used remains to be the basis of later SMT models up till present. GIZA^{++} [35] is a freely available implementation of the IBM Model 4 that is widely used to produce word alignments at the very first step of an SMT system.

1.1.3 Automatic Evaluation

Since manual evaluation is time and labour consuming, machine translation is largely evaluated automatically by subjective metrics, which are calculated based on the how well the machine translation matches with the reference translation made by human translators.

The most widely used automatic evaluation metric is BLEU (BiLingual Evaluation Understudy)[37], which is defined by equation 1.2.

$$\text{BLEU} = \textit{brevity-penalty} \cdot \exp\left(\sum_{n=1}^{N} w_n \log \textit{precision}_n\right)$$

$$\textit{brevity-penality} = \min\left(1 - \frac{\textit{reference corpus length}}{\textit{candidate corpus length}}, 0\right) \tag{1.2}$$

where $\textit{precision}_n$ is the precision of matched n-grams; N is the maximum order of n-gram to be matched; w_n is the weight for each n-gram precision; and *brevity-penality* is a function to penalize short sentences.

A candidate translation that results with a higher BLEU score is closer to the reference translation, which suggests that the candidate system outperforms the others. However, the difference in scores can also result from random variation. To test the hypothesis that a higher score is due to higher translation performance, statistical significance tests, expressed as 'p-values', are taken.

Other common evaluation metrics include METEOR [1], WER (Word Error Rate) [41], TER (Translation Edit Rate) [40], GTM (General Text Matcher) [44, 32] NIST (National Institute of Standards and Technology) [7], and RIBES (Rank-based Intuitive Bilingual Evaluation Score) [11], etc. In this work, the machine translation results are evaluated by BLEU, METERO and TER, and results with over 95% significance (i.e. p-values below 0.05) are highlighted.

1.2 Related works

Syntactic information for automatic word alignments has been investigated in a number of studies [4, 5, 6, 10, 20, 46]. On lexical level, function words are assumed to be language specific and are treated separately with content words [33, 39]. While function words are defined in previous works based on wordlists or frequencies, this study defines unalignable based on translation context.

In particular, this study shares a similar notion with the following works in that some words are lexicalized to surface forms only on one side of the language pairs (Subsection 1.2.1) and likely invalid links are predictable and can be trimmed to improve word alignment (Subsection 1.2.2).

1.2.1 Empty categories in translation

Empty categories are syntactic categories not associated to a surface word. Chung and Gildea [5] observes the problem that certain empty categories in the source language have to be translated to surface words in the target language. For example, pronouns that are dropped in Chinese and Korean have to be translated explicitly in English. Since alignment to the NULL token is not specific enough to generate target words from source empty categories [5], this work proposes a preprocessing step to insert empty pronoun tokens to which explicit pronouns in the target language are aligned.

The method was first proved useful by insertion of empty categories based on the hand annotated Chinese Treebank and Korean Treebank. Training on the Chinese Treebank, empty categories are then automatically predicted by various methods, including pattern matching [12], CRF [14] using simple context features and parsing with grammar modified with empty category annotation [5]. Experiments show that prediction of dropped pronouns by CRF results in the highest classification accuracy and the largest improvement on the overall phrase-based MT (PBMT) performance.

Comparatively, the unalignable words tackled in this work is not constrained by any particular syntactical categories. Focus is put on whether a word has a counterpart

in the other language irrespective of the underlying syntactical categories. The unalignable words are predicted by SVM classification using cross-lingual features of a sentence pair.

1.2.2 Link deletion method

GIZA^{++} alignments of the two language directions can be symmetrized by various heuristics. Symmetrization by union intuitively produces the highest recall in word alignment with low precision. Fossum et al. [8] thus proposes a *Link Deletion* algorithm that delete unlikely links from GIZA^{++} union outputs.

A scorer to evaluate the quality of a particular word alignment is trained by perceptron based on syntactic features of the extracted string-to-tree rules, lexical translation features and structural features of the alignment. Higher scores are given to alignment more similar to a test set of hand annotated alignment. Links in a GIZA^{++} union are iteratively deleted one by one and each modified alignment is automatically scored. The alignment with the highest score is returned as output. Improvement in word alignment accuracy as well as the final PBMT output over GIZA^{++} union alignments is recorded according to the experiment results.

In this study, the focus is put on mistaken links of words that actually should not be aligned. Instead of post-processing the output of GIZA^{++}, a pre-processing step is used to remove unalignable words before running GIZA^{++} to avoid incorrect links between alignable and unalignable words. Improvement over the GIZA^{++} alignment symmetrized by the standard *grow-diag-final-and* heuristics is shown in the experiments.

Chapter 2

Method

Towards improving SMT of distant language pairs, a method is proposed in this study to facilitate automatic word alignment. In view of the fact that the majority of SMT systems employ the IBM Model 4 in the word alignment step, the method is formulated as a preprocessing step that modifies input to be fed into GIZA^{++}. This chapter explains the motivation of the method and the algorithmic details.

2.1 Motivation

Building translation rules from word-level alignment assumes that words are generally alignable across languages, where unalignable words are of the minority. Closer languages, such as members of the Indo-European language family, have shared etymology to certain extent. Therefore, the assumption holds most of the time. In the case of distant language, however, it is naive to assume that words of the source language usually have counterparts in the lexicon of the target language. Since the process of lexicalization depends on culture and distant languages have evolved separately in different background, 1-to-1 mapping of lexical semantics is unlikely.

A parallel corpus consists of texts in the source language and their translation by human translators. When professional translators translate, they do not build up sentences from word-for-word translation. Major translation theories, such as the concept of *dynamic equivalence* [34], advocates *sense-for-sense* translation over *word-for-word translation*. The source text is to be translated such that the *'sense'* perceived

by the source text readers is the same as that perceived by the target text readers. Sentences are not broken down to units of words nor phrases, but units of *senses*[1], or concepts. The preference of *sense-for-sense* translation over *word-for-word* translation intuitively suggests that word chunks of the same sense are more likely to be aligned comparing with individual words. Word alignment across bilingual sentences is thus a task never meant to be done and is not straightforward.

During translation, words have to be added or omitted to achieve equivalence between the corresponding sense units. These words are not lexically dependent to any words in the source text, but is 'aligned' to concepts above word-level. For example, the Chinese term *'tai yang'* literally means *'sun'*, yet functionally its equivalence should be *'the sun'* in English. However, there is not any concepts of definite articles in Chinese, nor is it incorporated in the morphology of *'tai yang'*, so additional meaning is conveyed by the term *'the sun'*. Metaphorically speaking, translations of *'tai-yang'* to *'sun'* and *'the sun'* are like 1-to-0.8 and 1-to-1.2 alignments respectively, realized as 1-to-1 and 1-to-2 word alignments. The function word *'the'* is added to satisfy the grammatical need of English for a determiner and the semantics of *'tai yang'* provides knowledge to choose 'the' as the proper determiner, whereas the core alignment is between *'tai-yang'* and *'sun'*.

Nonetheless, the an unalignable word in one translation is not always additional in any translation. For example, while *'the'* is added when translating *'tai-yang'* to *'the sun'*, it is alignable in other contexts such as in the term *'the man'* for *'na nanren'* , where *'the'* can be the translation of the Chinese demonstrative pronoun *'na'*, literally means *'that'*. Looking at the English text alone, it can be deduced that the *'the'* in *'the sun'* is likely to be an added word, while the *'the'* in *'the man'* probably comes from a Chinese source word.

On the other hand, although it is commonly assumed that *unalignable words* are not necessarily *functional words*, content words can also be added to the translation when it is necessary to convey the *'sense'* of the source language. For example, in the translation of proper nouns, it is common to insert nouns that demonstrate the entity the name refers to, such as the translation of *'Chicago Bulls'* to *'Zhijiage Gongniu*

[1]Note that the term 'sense' here does not refer to one of the multiple meanings of a polysemous word, as when it is used in a word sense disambiguation task.

Dui', literally *'Chicago Bulls Team'*. This is because the *'sense'* evoked by the term *'Chicago Bulls'* in the minds of English readers is not the same as the *'sense'* evoked by the term *'Zhijiage Gongniu'* alone in the minds of Chinese readers.

Words within a many-to-many word alignment can thus be distinguished as *core* words and *unalignable* words. To guide the study, a *core* word is defined as a translated word with a corresponding word in the source, even the semantics of the word pair has limited overlap. An *additional* word, on the other hand, is a word without a corresponding word in the source language, but is derived from a concept above word level. Unalignability is defined given a language pair, since the lexical discrepancy is a contrast between 2 languages. Comparing with the above Chinese-English example, the English word *'the'* has French equivalents *'le'*, *'la'* or *'les'*. Thus, it is not an *unalignable* word between French and English translation.

From the above observation, it can be hypothesized that unalignable words do not form a universal list but can be deduced from the linguistic context based on the contrast of the given language pair. This study thus investigates methods to incorporate this notion in SMT systems.

2.2 Proposed method

This study separates the deduction of unalignable words from the automatic word alignment process and predict unalignable words using external linguistic features based on the contrast between the two languages. The goal is to improve the quality of automatic word alignment, which is the first step for most SMT systems. Chinese and English are chosen as the translation language pairs since unalignable words are abundant between distant languages and specific word-aligned data is available for this language pair.

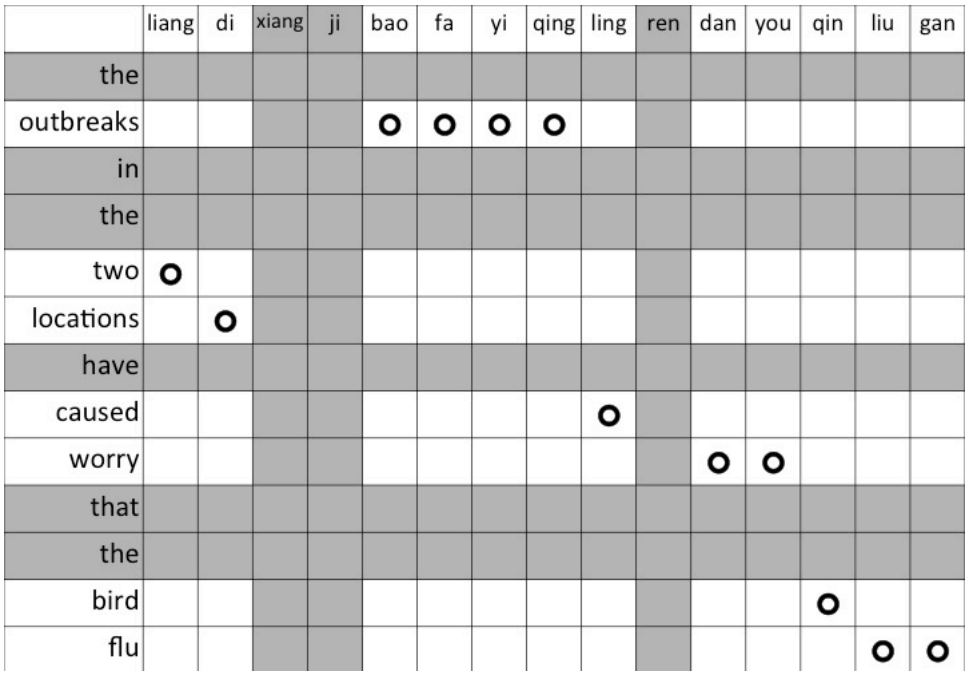
The gist of the method is to predict and remove likely unalignable words from both sides of the parallel training data before feeding it to automatic word alignment by GIZA^{++}. With the unalignable words removed, each sentence pair is left with a smaller number of possible links and more accurate word alignment can thus be inferred by GIZA^{++} and excessive alignment of rare words with the 'NULL' token can

be prevented. The actual processing steps are as follows and visualized in Figure 2.1.

1. Train a supervised classifier by relating cross-linguistic features with alignability.
2. Predict words that are likely to be unalignable using the classifier
3. Remove these words from the training corpus
4. Run GIZA++
5. Symmetrize the word alignments in both directions using heuristics
6. Restore the removed words to the original position and mark them as unalignable words
7. Continue translation rule extraction and the rest of the pipeline.

To investigate useful linguistic features that distinguish unalignable words from core ones, this study examines in depth the words that are not aligned by human annotators in a manually aligned Chinese-English corpus. Next, a classifier is built using the manually aligned data and the applicability of the above method is evaluated based on manual annotation and automatic prediction of unaligned words.

Blind the predicted *unalignable words* and run GIZA^{++} to align the remaining words



Link all *unalignable words* to the NULL token and extract translation phrases with and without the *unalignablel words*

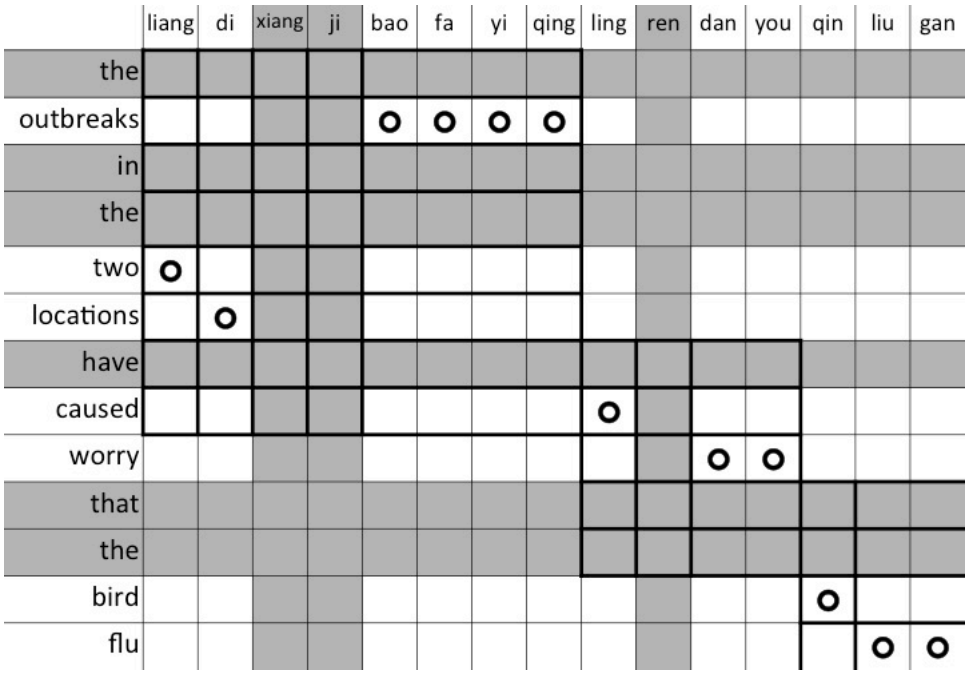


Figure 2.1: Visualization of the effect of removing and restoring *unalignable* words

Chapter 3

Corpus Analysis

The key question in this research is *'What kind of words are unalignable?'*. Before testing the applicability of our proposed method, an analysis is carried out to examine linguistic and contextual features that generalize unalignable words. The data used is a Chinese-to-English parallel corpus which is not only hand aligned to the word-level corpus but also annotated with markups explaining the nature of the alignment. This chapter first introduces the annotation criteria of this specific corpus, followed by an analysis of unalignable words based on this data, which is named the ORACLE dataset.

3.1 The ORACLE dataset

The ORACLE dataset is made up of the following data released by the Linguistic Data Consortium.

- GALE Chinese-English Word Alignment and Tagging Training Part 1 – Newswire and Web (LDC2012T16)[17]
- GALE Chinese-English Word Alignment and Tagging Training Part 2 – Newswire (LDC2012T20)[18]
- GALE Chinese-English Word Alignment and Tagging Training Part 3 – Web (LDC2012T24)[19]

In total, the data consists of about 13000 Chinese sentences and their English translation, making up to 390K English words. The source texts were collected from Chinese web data and the English translation is made based on the Chinese-English translation guidelines for the Global Autonomous Language Exploitation (GALE) project [38]. This parallel corpus is enriched with two levels of annotation: (1) manual word-level alignment and (2) classification of the nature of word alignments represented by predefined tags. Detailed guidelines are drafted for annotators to follow for each of the annotation task. An actual annotation example is shown below,

- (Chinese)
  wo yi qian ceng shuo guo , wo men zhi shao xu yao wu sheng yi he 。

- (English)
  I previously said , we need at least 5 wins and a tie .

- (Annotated alignment)
  12,13-6(SEM) 18-14(FUN) 10,11-7,8(SEM) 14-9(FUN) 15-10(SEM) 1-1(FUN) 2,3-2(SEM) 8,9-5(FUN) 7-4(FUN) 16-12(FUN) 4[TEN],5,6[TEN]-3(GIS) 17-13(SEM) -11[COO](NTR)

where the numbers in the alignment stand for positions of aligned words in the source and target sentences. For example, '12,13-6' means the *'12th and 13th Chinese words are aligned' to the 6th word in English'*; -11 means *'the 11th English word is not aligned to any Chinese words'*. Words in round brackets are link tags and words in square brackets are tags of the unaligned words. Specifically, (SEM) stands for 'semantic link', (FUN) for 'functional link', (GIS) for 'grammatically inferred link', (NTR) for 'not translated link', [TEN] for 'tense marker', and [COO] for 'context obligatory markers'. Annotators are trained to fully understand the definitions of these tags.

The alignment task is based on the *minimum match* principle [15], by which the source words are aligned with the target words by hand as fine-grained as possible, even for many-to-many links. For example, *'Happy New Year'* forms a many-to-many link with the Chinese counterpart *'Xin-nian kuaile'*, literally *'new year happy'*, yet it

can be further broken down to individual links of *'Xin'* for *'new'*, *'nian'* for *'year'*, and *'kuai-le'* for *'happy'*. To achieve this, the Chinese words are segmented to one character per word, since the word segmentation preprocess may clash with the *minimum match* principle and thus unalignable characters can be included in a Chinese word segment. For instance, *'Xin-nian'* is likely to be segmented as one unit.

In the second annotation task, the word alignments are further tagged with categories, which are defined based on the nature of the alignment. For example, FUNCTIONAL LINK for links between function words and SEMANTIC LINKS for links between content words. Unmatched words, which does not correspond to any particular word in the other language, are tagged with defined categories that state the functions of these words in the alignment. Under the annotation scheme of this corpus, these unmatched words are either aligned to 'NULL', or aligned to the corresponding word of their dependency heads, if occur, as an 'attachment'[15]. For example, *the* is annotated as 'DETERMINER', and attached to the core link between *tai yang* and *sun*.

The *attachment* principle is adopted to handle the abundant unaligned words between Chinese and English, However, the *attachment* links are complicated inference from a wide context and actually degrades the generalization of translation rules in an SMT system[45]. This work focuses on the *attached words* instead. A many-to-many link can be viewed as as a pair of aligned sense units and the attached words as unalignable words, since they are not core to the many-to-many alignment and are, by definition of the annotation scheme, added or omitted to translate cross-lingual thought.

3.2 Property of unalignable words

Using ORACLE dataset, an analysis is carried out on the unalignable words in the parallel data. All the unalignable words (*attached words* as well as other non-matched words that are not attached to any links are tagged with 20 categories based on the *reason* of the word addition/omission. Since the key purpose of the analysis is to find out what kind of words are unalignable, a distinction is not made among different categories, but only of being tagged or not, i.e. being unalignable or not.

3.2.1 Distribution among word types

Analyzing the hand aligned corpus, it is found that there are not any words that are always annotated as unalignable. Instead, the words are either always annotated as core words or annotated as either core or unalignable words. Table 3.1 reveals that although the number of 'always core' word types are 7 times that of 'sometimes unalignable' word types, the former contributes to only 17% of the total number of word tokens in the corpus. On the other hand, 17%, as well, of the total word instances are annotated as unalignable, while most of the tokens (66%) belong to 'sometimes unalignable' word types, yet annotated as core instances.

	No. of word types	**No. of unalignable tokens**	**No. of core tokens**	**Total no. of tokens**
always core word types	25320	/	147,373 (17%)	856,867 (100%)
sometimes unalignable word types	3581	146,693 (17%)	562,801 (66%)	

Table 3.1: No. of core and unalignable words in the hand aligned ORACLE corpus

Some examples of potentially unalignable words are shown below.

1. Chi: *yi* *ge* *di fang*
   one (measure word) place
   Eng: *one* *place*

2. Chi: *ge ren*
   personal
   Eng: *personal*

3. Chi: *ming tian* *zhong wu*
   (tomorrow) (midday)
   Eng: *tomorrow* *at* *midday*

4. Chi: *zai* *jia*
   at/in/on home
   Eng: *at* *home*

In example 1, *'ge'* severs as a measure word that is exclusive in Chinese, but in example 2, it is part of the multiword unit *'ge-ren'* for *'personal'*. Similarly, prepositions, such as *'at'*, can either be omitted or translated depending on context. It is thus impossible to construct a distinct list of unalignable words, although it is commonly believed that function words, which is a close set, are always not aligned across languages.

Nonetheless, among the 17% word types of potentially unalignable words, the unalignable instances are by no means evenly distributed. Table 3.2 shows that the top 50 most frequent unalignable word types cover 68% and 88% of all the Chinese and English unalignable word instances respectively, and the top 100 cover 78% and 94%. Word type is thus an important clue to predict whether a token is unalignable or not.

Word types most frequently annotated as unalignable	**No. of unalignable tokens (coverage)**	
	Chinese	English
Top 50	34,987 (68%)	83,905 (88%)
Top 100	40,121 (78%)	89,609 (94%)

Table 3.2: Coverage of *unalignable* tokens by top unalignable word types

3.2.2 Distribution among part-of-speech

Intuitively, function words and other words with part-of-speech (POS) defined only in one of the languages are unlikely to be core. This subsection thus examines if the unalignable words come from a certain POS categories. This is based on automatically POS tagging of the ORACLE data and the tagger used is the Standford Tagger [42]. As explained in the previous section, the Chinese side of the parallel corpus is in characters but not word segments, yet automatic taggers run only on segmented Chinese text. Therefore, the text was first segmented using Standford Segmenter [43] before tagging, and then assigned the Chinese Treebank POS tag to each character of the word segment, such that each character can be matched with the character-based alignment

Chinese POS	**No. of unalignable tokens**	**% of unalignable tokens**
DEG	7411	97%
NN	6138	4%
AD	6068	17%
DEC	5572	97%
VV	4950	6%
English POS	**No. of unalignable tokens**	**% of unalignable tokens**
DT	27715	75%
IN	19303	47%
PRP	5780	56%
TO	5407	62%
CC	4145	36%

Table 3.3: Top 5 POS categories of Chinese and English unalignable words

annotation.

It is found that the unalignable words found in the manual alignment include all POS categories of either language. Table 3.3 lists the top 5 POS categories that most unalignable words belong to and the percentage they are annotated as unalignable. Some POS categories, such as English determiners, are indeed mostly unalignable, yet they can also come from a Chinese source word depending on context, such as the *'the'* example illustrated in Section 2.1.

Note also that many Chinese unalignable words are nouns (NN) and verbs (VV). Clearly it is not appropriate indiscriminately consider all nouns as unalignable. Below are some examples of content words added to the Chinese translation.

1. Chi: *can jia* *hui jian* *huo dong*
   participate meeting activity
   Eng: *participate* *in the meeting*

2. Chi: *hui yi* *de* *yuan man* *ju xing*
   meeting 's successful take place
   Eng: *success* *of* *the meeting*

(adapted from [16])

English verbs and adjectives are often nominalized to abstract nouns (such as *'meeting'* from *'meet'*, or *'success'* from *'succeed'*), but such derivation is rare in Chinese morphology. Since POS is not morphologically marked in Chinese, *'meeting'* and*'meet'* are the same word. To reduce the processing ambiguity and produce more natural translation, extra content words are added to mark the nominalization of abstract concepts. For example, *'hui jian'* is originally *'to meet'*. Adding *'huo dong'*(activity) transforms it to a noun phrase (example 1), similar to the the addition of *'ju sing'*(take place) to the adjective *'yuan man'* (example 2). These additional words are not lexically dependent but are inferred from the context, and thus do not align to any source words.

Non-grammatical words are added to fill the gap between the lexical semantics of the source and target language in order to achieve sense equivalence. In naturally occurring translation, such addition is a very common technique to facilitate the understanding of the target readers. However, in translation prepared for SMT training, such addition is constrained and is allowed only if it is needed for grammaticality or fluency [38]. Therefore, the content words in the ORACLE data that are annotated as unalignable are indeed necessary addition to the translation.

To summarize, most of the words (83%) can be aligned to another word in the opposite language. There is not a list of *'always unalignable'* words nor *'always unalignable'* word categories. However, a small number of word types or categories are much more frequently classified as unalignable than others. Word types, POS and lexical contexts provide cues to differentiate unalignable words from core words. These findings will be used for automatic classification of unalignable words described in Section 4.2.

Chapter 4

Experiments

The analysis of the ORACLE data has already proved the first claim of this study that unalignable words between two languages do not come from a distinct list, such as a list of function words. Nonetheless, it is also observed that there is certain extent of regularity in the distinction between core and unalignable words. This supports the second claim that unalignable words are predictable from their linguistic contexts.

This chapter describes experiments to evaluate the remaining hypothesizes and the applicability of our proposed method. The questions to be answered are:

1. Does removing gold standard unalignable words beforehand improve automatic word alignment by GIZA^{++} and the overall MT outputs?
2. How accurate can automatic unalignable word prediction achieve based on linguistic features?
3. Does removing automatically classified unalignable words in realistic data improve automatic word alignment by GIZA^{++} and the overall MT outputs?

Experiments to answer these questions and the corresponding results are illustrated in the following three subsections respectively.

4.1 Removing hand annotated unalignable words

Gold standard unalignable words are annotated in the ORACLE dataset. In this experiment, the proposed method is tested by removing the gold unalignable words

before running GIZA^{++} on the relatively small ORACLE data. The ORACLE dataset is also hand aligned, thus the automatic word alignments produced by GIZA^{++} can be evaluated by comparing with the gold hand alignment. Since the gold alignments are bidirectional with many-to-many links, the GIZA^{++} alignments are symmetrized before evaluation.

4.1.1 Settings

As an implementation of IBM Model 4, the GIZA^{++} tool adds a 'NULL' token to each sentence to capture unalignable words, but the words fed in are supposed to be all alignable. Alignments to the 'NULL' token are suppressed by setting the *EM probability for empty words*to 0 (for the HMM model) and setting the *deficient distortion for empty word option to 2* (for IBM models 3 and 4). This run of GIZA^{++} is referred to as 'adjusted GIZA^{++}'. Another set of alignments are produced with 'default GIZA^{++}', which takes the standard parameter values passed from MOSES, for comparison. As a reference, the number of NULL alignments drops from 2671 to 172 for Chinese-to-English direction, and from 4389 to 699 for English-to-Chinese direction.

The alignments in both directions are symmetrized by the standard *grow-diag-final-and* heuristics. The links of words annotated as unalignable are not counted as positive links, although most of them are attached to their head words and included in many-to-many alignments in the manual annotation. Following [9], the word alignments are evaluated by weighted F measure since it correlates with the performance of the resulting MT systems most. α is set to 0.5 for Chinese-English alignments. The results are shown in Table 4.1.

Next, the removed words are restored to their original positions as unalignable words. Standard phrase-based and hierarchical MT systems are built using these word alignments using MOSES [13]. The reordering model was set as *mslr-bidirectional-fe* and the language model is a 4-gram model trained on the same dataset. 81% of the data was used for training, 9% for MERT tuning and 10% for testing. For comparison, a baseline is built under the same settings except the preprocess - word alignments are induced from all words in the training corpus.

4.1.2 Results

Table 4.1 lists the precisions, recalls and F1 measures of the automatic GIZA^{++} and the automatic evaluation scores of the resulting MT systems before and after tuning, respectively. It is shown that higher accuracy of alignment is achieved by removing unalignable words beforehand. Limiting NULL alignments by GIZA^{++}increases the F1 measures slightly. Higher MT scores are produced by the proposed methods in all cases - for both MT systems and 3 evaluation schemes. Improvements of about 0.4 BLEU point and 0.9 BLEU point are recorded for the resulting phrase-based and hierarchical MT systems respectively before tuning. This supports our hypothesis that removing gold unalignable words helps improve word alignment and the resulting SMT.

	Word alignment accuracy		**Phrase-based MT**			**Hierarchical MT**		
				before tuning	after tuning		before tuning	after tuning
Baseline	P	.711	BLEU	11.38	12.29	BLEU	10.32	11.05
	R	.488	TER	**70.86**	73.63	TER	75.91	74.91
	F1	.579	METEOR	21.76	22.33	METEOR	21.08	20.09
Proposed	P	**.809**	BLEU	11.73^{+}	**12.47**	BLEU	**11.21**$^{+}$	10.46^{-}
(default	R	.501	TER	71.84^{-}	73.83	TER	75.16^{+}	70.71^{+}
GIZA^{++})	F1	.619	METEOR	22.02^{+}	**22.57**$^{+}$	METEOR	**22.06**$^{+}$	20.75^{-}
Proposed	P	.802	BLEU	**11.81**$^{+}$	12.42	BLEU	11.03^{+}	**11.30**
(adjusted	R	**.509**	TER	71.39^{-}	**73.57**	TER	**74.70**$^{+}$	**70.36**$^{+}$
GIZA^{++})	F1	**.623**	METEOR	**22.08**$^{+}$	22.54^{+}	METEOR	22.00^{+}	**21.54**$^{+}$

Table 4.1: Word alignment accuracies of GIZA^{++}and evaluation scores of MT systems of the ORACLE dataset. The best score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05.

4.2 Automatic prediction of unalignable words

In order to scale the method to unannotated large bitext, it is necessary to automatically predict words that are likely to be unalignable. This is a challenging task: according to the manual annotation, unalignable words do not come from a distinct list and most of the word types are ambiguously unalignable or core depending on context. The prediction task can be thus formulated as a classification problem.

A supervised classifier is built using the ORACLE dataset, where unalignable words are tagged and core words are not. Features are selected based on the analysis on the distribution of the unalignable words in the ORACLE data. The predictions made by the classifier are evaluated against the hand annotated unalignable words. Next, the predicted unalignable words are removed and restored according to the proposed method and the effect on the resulting MT systems is assessed. This section describes the features used to train the classifier and the performance of the MT systems thus built on the ORACLE dataset.

4.2.1 Features

The second hypothesis of this study is that unalignable words can be predicted given a pair of source and target languages. The corpus analysis in Chapter 3 has shed some light on the determination of cues useful for the prediction, such as word types, POS and lexical context. On the other hand, the word alignment task links source words to the target words, thus whether a word is alignable or not depends on the translated sentence as well. Therefore, cross-lingual features are also used.

The actual features used to train the classifier include below:

- **Lexical context:** Unigrams in window sizes of 1, 3, 5, 7 are used to capture lexical context. Normalization is applied to letter case, numerals (Arabic numerals in Chinese encodings are normalized to ASCII), and certain symbols. The window size that produces the highest classification accuracy is chosen for the final classifier.
- **Part-of-speech:** The POS tags of each token, which are used in the previous analysis, are used as features. Similarly, different window sizes are tried and

the one that produces the highest classification accuracy is chosen for the final classifier.

- **Top token-POS pairs:** This feature is defined by whether the token in question and its POS tag is within the top n frequent token-POS pairs annotated as unalignable. 4 features are defined with $n = 10, 30, 50, 100$. Since the top frequent unalignable words covers most of the counts as shown in the previous analysis, being in the top n list is a strong positive features.

- **Number of likely unalignable words per sentence:** It is hypothesized that the translator will not add too many tokens to the translation and delete too many from the source sentence. In the ORACLE data, 68% sentences have more than 2 unalignable words while 17% have more than 10 unalignable words. The number of likely unalignable words in the sentence is approximated by the count of words within the top 100 token-POS pairs annotated as unalignable. By this approximation, there are at most 9 'unalignable words' per sentence. 5 features are defined with count of 'unalignable words' being over 0, 2, 4, 6 and 8 respectively.

- **Sentence length:** Longer sentences are more likely to contain unalignable words than shorter sentences. Sentence lengths in intervals of 20 words are grouped as 1 feature.

- **Source-target sentence length ratio:** As an extension to the above feature, the source sentence is likely to contain more unalignable words if it is aligned with a shorter target sentence. Note that since the Chinese sentence lengths are count by characters, they are generally longer than the English sentences, at most 4 times longer.

- **Presence of alignment candidate:** This is a negative feature defined by whether there is an alignment candidate in the target sentence for the source word in question, or vice versa. The candidates are extracted from the top n frequent words aligned to a particular word according to the manual alignments of the ORACLE data. 5 features are defined with $n = 5, 10, 20, 50, 100$ and one 'without limit', such that a more possible candidate will be detected by more features.

These features are used to train a supervised classifier on the ORACLE dataset. Detail configuration will be described in the next subsection.

4.2.2 Settings

An SVM classifier is trained to predict whether a word in the parallel text is core or unalignable. Let (e_1^J, f_1^K) be a pair of sentences with lengths J and K. For each word in (e_1^J, f_1^K) that belongs to a predefined list of potentially unalignable words, a word type and language specific classifier is ran. In this experiment, the list of potentially unalignable words is defined by word types that are annotated as unalignable in the ORACLE data for at least once. Word types that are never annotated as unalignable are classified as core, irrespective of the decision of the classifier.

The classifiers are trained using the LIBSVM [3] toolkit with a linear kernel. The margin tradeoff was not tuned on any development set but kept at default. 90% of the word instances is used for training and 10% is used for testing. For example, the *'at'-classifier* classifies whether a particular *'at'* in the English text is an *unalignable* instance or not, and 90% of the *'at'* instances are used for training while the remaining 10% are used for testing.

To avoid data sparseness, a separate classifier is built only for each of the top 100 word types with the highest count of unalignable words per language according to the hand annotated data. An additional classifier is built for all the remaining words in each language. A separated set of classifiers is built for each language direction. As a result, 202 classifiers are built in total, which are combined by an overall classifier that picks the corresponding classifier out of the 202 by rule for each token. Experimenting features using various window sizes, unigram window size of 3 in combination with POS window size of 7 works best and is chosen for the MT experiments. The overall accuracy of the unalignable word prediction is shown in the first column of Table 4.2, which is computed by the count of correctly classified tokens over the count of all tokens in the test set.

The more fine grained the classifiers are, the more specifically can the weights be tuned, yet fewer training instances will be available. Classifiers of other levels of granularity were trained and compared. For example, a more coarse-grained set of classifiers is defined by training a separate classifier for each of the top 50 words. On the contrary, recalling that the unalignable words in the ORACLE data are hand annotated with categories (refer to Section 3.1), a more fine-grained set of classifiers is defined by building a classifier for each category of unalignable words' . For instance,

the *'ge'-measure-word-classifier* classifies whether a particular *'ge'* in the Chinese text is an *additional* 'measure word' or not. However, the setting of '1 classifier per top 100 word types' and irrespective of the annotated categories works best.

Following the proposed steps, predicted unalignable words are held out before GIZA^{++} is run and restored afterwards. MT systems are built with the same default settings used in the previous experiment, where 81% of the data was used for training, 9% for MERT tuning and 10% for testing. Note that the MT data split is based on sentence while that for the classifier is based on word type, so the two splits are not exactly overlapping. Nonetheless, tokens in the MT test set are possibly included in the training test of the unalignable word classifiers.

For comparison, MT systems are also built by classifying all instances of the top 30 token-POS combinations as unalignable. According to the annotation, 72% of the manually annotated unalignable words belong to this top 30 list, yet among all word instances belonging to this list, only 20% are unalignable instances. As in Section 4.1, the accuracy of automatic word alignment on the MT training set is evaluated against the gold annotation and the MT outputs are evaluated by automatic metrics.

4.2.3 Results

Table 4.2 shows the accuracy of the unalignable word classification and the resulting performance of GIZA^{++} and the MT systems. Automatic prediction of unalignable words achieves an accuracy of about 92%. As shown in Table 3.1, only 17% of the words are unalignable thus 83% accuracy can be achieved by classifying none of the words as unalignable, which is about 10% lower than that of the proposed classifier. In fact, classifying by the top 30 token-POS list achieves the same level of accuracy

Significant improvement of the hierarchical MT system over the baseline (no preprocessing) is seen using the proposed method with automatic prediction, yet little improvement or degradation is seen on the phrase-based MT system. On the other hand, comparing with Table 4.1, the resulting GIZA^{++} accuracy after the preprocess based on automatic prediction is lower than the baseline. Preprocessing based on the gold alignment produces the best MT and alignment results, as shown in Table 4.1. This suggests that the classifier accuracy is positively related to GIZA^{++} accuracy as

well as MT performance.

	unalignable word classification accuracy	**GIZA^{++} word alignment accuracy**			**Phrase-based MT**		**Hierarchical MT**	
					before tuning	after tuning	before tuning	after tuning
Removing top 30 unalignable token-POS pairs	Accuracy .830	P R F1	.400 .394 .397	B T M	9.87^{-} 79.19^{-} 20.62^{-}	9.99^{-} **71.95**$^{-}$ 20.68^{-}	9.05^{+} 81.14^{+} 20.29^{+}	9.51^{-} 81.65^{-} 20.71^{-}
Removing unalignable words by classifier (default GIZA^{++})	Accuracy .921	P R F1	.503 .459 .480	B T M	11.53 72.32^{-} 21.82	12.12 74.33^{-} **22.18**	10.80^{+} 76.10 **21.79**$^{+}$	**11.08** 73.61^{+} **21.62**$^{+}$
Removing unalignable words by classifier (adjusted GIZA^{++})	Accuracy .921	P R F1	**.512** **.465** **.480**	B T M	**11.55** **71.95**$^{-}$ **21.90**	**12.20** **73.90** 22.08^{-}	**10.85**$^{+}$ **75.57** 21.78^{+}	**11.08** **73.33**$^{+}$ 21.53^{+}

Table 4.2: Accuracy of the unalignable word classifier and the effect on the resulting GIZA^{++} alignment and MT performance by removing the predicted unalignable words beforehand (on ORACLE data). Classification based on a list of top 30 token-POS pairs most frequently annotated as unalignable is shown for comparison. B, T and M stand for the MT evaluation scores of BLEU, TER and METEOR respectively. The highest score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline (no preprocessing) below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05.

The result of experiment 4.2 supports the second hypothesis that unalignable words are predictable by linguistic and context features, since the predictions are 92% correct. It also moderately supports the third hypothesis that pre-removing unalignable words based on prediction is beneficial to SMT, although training on a relatively small amount of data. The third hypothesis will be further tested in the experiment described in

Section 4.3.

4.3 Removing predicted unalignable words from REAL data

The experiments of Section 4.2 illustrate gains in the MT results using unalignable word predict, where the SMT systems are trained on the relatively small-scaled ORACLE dataset. However, usual SMT training requires a huge amount of parallel data, typically hundreds of thousands at least. Therefore, the proposed method is applicable to SMT only if it can be scaled to large realistic data. This section describes the experiments conducted to test the scalability of the proposed method, started by an introduction of the data used.

4.3.1 The REAL dataset

The parallel text used for MT training in this section is named the REAL data, which is a large-scale unannotated translation corpus selected from the NIST Open MT task 2008. The development and test sets are the same as those of the Open MT task, and the training data includes the following corpora:

- Hong Kong Parallel Text (LDC2004T08) [21]
- Chinese News Translation Text Part 1 (LDC2005T06) [23]
- Chinese English News Magazine Parallel Text (LDC2005T10) [22]
- GALE Phase 1 Chinese Broadcast News Parallel Text - Part 1 (LDC2007T23) [24]
- GALE Phase 1 Chinese Blog Parallel Text (LDC2008T06) [25]
- GALE Phase 1 Chinese Broadcast News Parallel Text - Part 2 (LDC2008T08) [26]
- GALE Phase 1 Chinese Broadcast News Parallel Text - Part 3 (LDC2008T18) [27]

- GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 1 (LDC2009T02) [28]
- GALE Phase 1 Chinese Broadcast Conversation Parallel Text - Part 2 (LDC2009T06) [29]
- GALE Phase 1 Chinese Newsgroup Parallel Text - Part 1 (LDC2009T15) [30]
- GALE Phase 1 Chinese Newsgroup Parallel Text - Part 2 (LDC2010T03) [31]

These make up to 34M English words and 1.1M sentences, which is two orders of magnitude larger than the ORACLE data[1].

4.3.2 Settings

The REAL dataset is tagged and normalized in the same way as the ORACLE set for feature extractions. Each token is then classified to 'unalignable' or not by the trained classifier and the predicted unalignable words are removed before GIZA^{++} alignments following the proposed method. Without available hand annotation, the classification accuracy and GIZA^{++} alignment accuracy are not evaluated. Phrase-based and hierarchical translation systems are built after restoring the unalignable words to their original positions. The language model is trained on the REAL training data as well and the weights are tuned by MERT. A baseline system is built with the same settings except the preprocess.

In the classification experiment on the ORACLE data in Section 4.2, 1 classifier is built for each of the top 100 word types with highest count of unalignable word and 1 classifier for the rest of the words, where word types never annotated as unalignable are classified as core by rule. Comparatively, the REAL dataset has a much larger vocabulary size. To avoid missing necessary links, it is preferable to be conservative when removing words. Therefore, the scope of classification on the REAL data is adjusted to running the classifier only on the top 100 unalignable words and classify the rest as core by rule. Evaluated using the ORACLE data, this constraint lowers the classification accuracy slightly from 0.921(refer to Table 4.2) to 0.919.

[1]Note that since not all the OpenMT data are accessible, e.g. FBIS, our results are not directly comparable to other systems in the evaluation campaign.

4.3.3 Results

Table 4.3 shows the performance of the resulting MT systems. Comparing the tuned and untuned scores respectively, the performance of the phrase-based system ranges from improvement of 0.32 BLEU points to degradation of 0.64 BLEU points using the proposed method. On the other hand, that of the hierarchical system shows improvement from 0.62 to 0.96 BLEU points. Again, suppressing NULL alignments by GIZA^{++} after removing predicted unalignable words further improves the MT results.

The results of the experiment support the third claim that removing unalignable words by automatic prediction helps improve SMT. Generally, more improvement is observed on the hierarchical MT systems. This suggests that our method is most effective for MT systems that depend critically on high-quality alignments during rule extraction, such as Hierarchical MT and Syntax-based MT.

	Phrase-based MT			**Hierarchical MT**		
		before tuning	after tuning		before tuning	after tuning
Baseline	BLEU 4	18.20	**19.67**	BLEU 4	17.02	16.75
	TER 4	**63.37**	**64.44**	TER 4	68.00	**63.31**
	METEOR 4	22.88	23.79	METEOR 4	22.92	21.98
Proposed	BLEU 4	18.48	19.29	BLEU 4	17.53^{+}	16.30^{-}
(default	TER 4	63.78^{-}	64.55	TER 4	67.81	63.79^{-}
GIZA^{++})	METEOR 4	23.08^{+}	**23.99**	METEOR 4	**23.46**$^{+}$	22.03
Proposed	BLEU 4	**18.55**	19.04^{-}	BLEU 4	**17.64**$^{+}$	**17.71**$^{+}$
(adjusted	TER 4	63.76^{-}	65.14^{-}	TER 4	**67.57**	63.63
GIZA^{++})	METEOR 4	**23.17**$^{+}$	23.36^{-}	METEOR 4	23.42^{+}	**22.92**$^{+}$

Table 4.3: Evaluation scores of MT systems built on the REAL dataset with and without pre-deduction of unalignable words. Evaluation is made against 4 references. The highest score for each MT system is bolded. $^{+}$ indicates improved scores with p-value between baseline below 0.05. $^{-}$ indicates degraded scores with p-value between baseline below 0.05.

Chapter 5

Discussion

The experiment in Sec 4.3 showed that the proposed method can improve SMT on realistic data. In this chapter, further experiments are performed to identify the conditions in which the proposed method works well. In particular, the questions are: (1) How much labeled classifier data is necessary? (2) How does the translation table using the proposed method differ from the usual translation table generated by the MT system?

5.1 Data size V.S. performance

5.1.1 Size of classifier training data

To evaluate how much annotation is necessary for accurate enough unalignable word prediction, experiments in Section 4.2 and Section 4.3 are repeated using different proportions of the ORACLE data. It is noteworthy that a lexical translation lexicon is necessary when extracting the 'alignment candidate' feature. In the repeated experiment, the translation lexicon was based on the manual word alignment of the whole ORACLE data. Nonetheless, the translation lexicon could also be acquired automatically from an external source (e.g. GIZA++ on bitext), which is much cheaper than annotation of word alignment.

Figure 5.1 shows that while training the classifier using only 1% of the ORACLE data is comparable with classification using a wordlist, training by 5% of the data considerably raises the accuracy, which levels off with further training data. Similar trends are shown in the hierarchical MT systems thus built. As shown in Figure 5.1,

improved evaluation scores over the baseline are achieved using over 5% of the classification training data, which is a feasible annotation effort of about 650 sentences. The improvement becomes significant (P-values below 0.05) using 20% or more of the data.

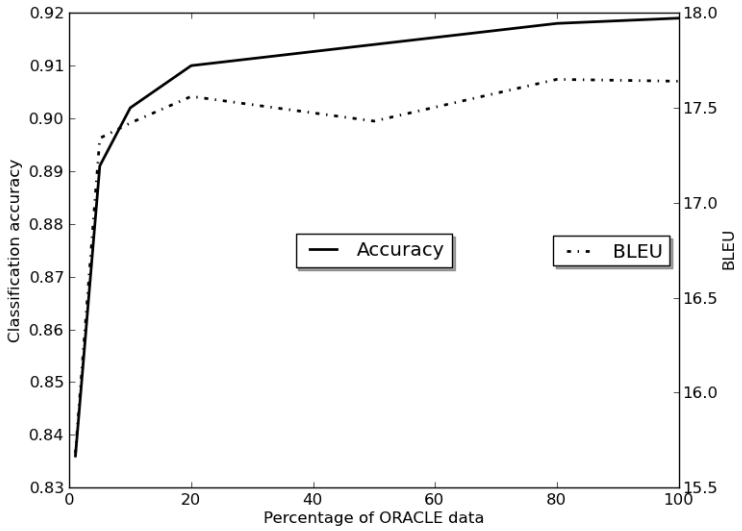


Figure 5.1: Accuracy of unalignable word classifiers and BLEU scores (before tuning) of hierarchical MT systems built on the REAL dataset v.s. various portion of ORACLE training data used for classifier training. The BLEU score of the baseline system was 17.02.

5.1.2 Size of MT training data

Further experiments are conducted to evaluate the effectiveness of removing unalignable words when using various sizes of MT training data. The hypothesis is that the preference of alignment between core words can be learnt by GIZA^{++} given large data, and artificial removal of the additional words disturbs the learning in turn. The MT experiments in Section 4.3 are repeated using the same set of classifiers, trained with the whole ORACLE data, but applied on MT training corpora of different sizes, sliced from the REAL data.

In order to evaluate the relation between the quality of the word alignments using the proposed method and MT training size, the ORACLE data is included as a small portion to be combined with the REAL training data, and evaluate the accuracy of the GIZA++ outputs against manual word alignments of the ORACLE data. Figure 5.3 shows the accuracies of the GIZA++ alignments across different sizes of REAL training data and it is obvious that *unalignment* produces higher precision and recall irrespective of the training size. These results impliy that the removing unalignable words steadily improves MT systems independent on training sizes.

On the other hand, the difference in BLEU score across MT training corpora of different sizes is shown in Figure 5.2. It is observed that the MT results of both cases improve with training data size and there is not a distinct intersection indicating the corpus size from which word alignment no longer benefits from additional word removal.

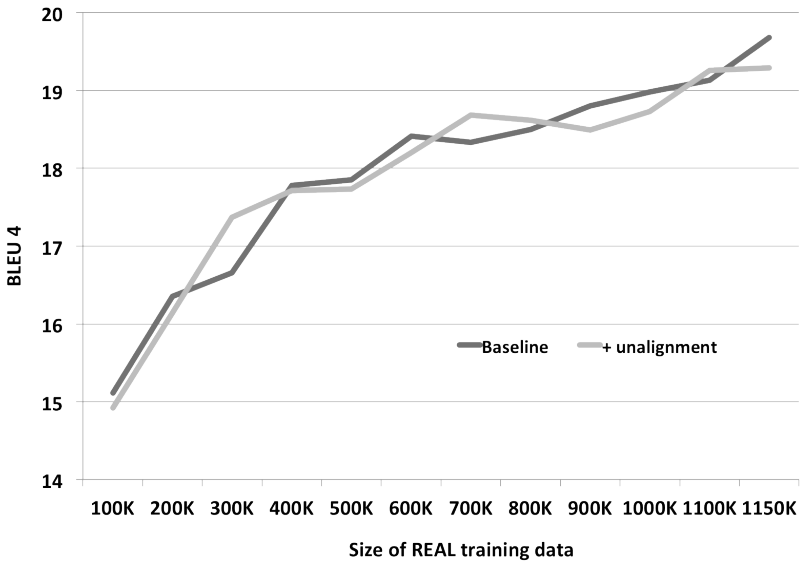


Figure 5.2: BLEU scores over REAL test set. Size is in number of sentences. *'+ Unalignment'* refers to PBMT systems built with automatic unalignable word prediction.

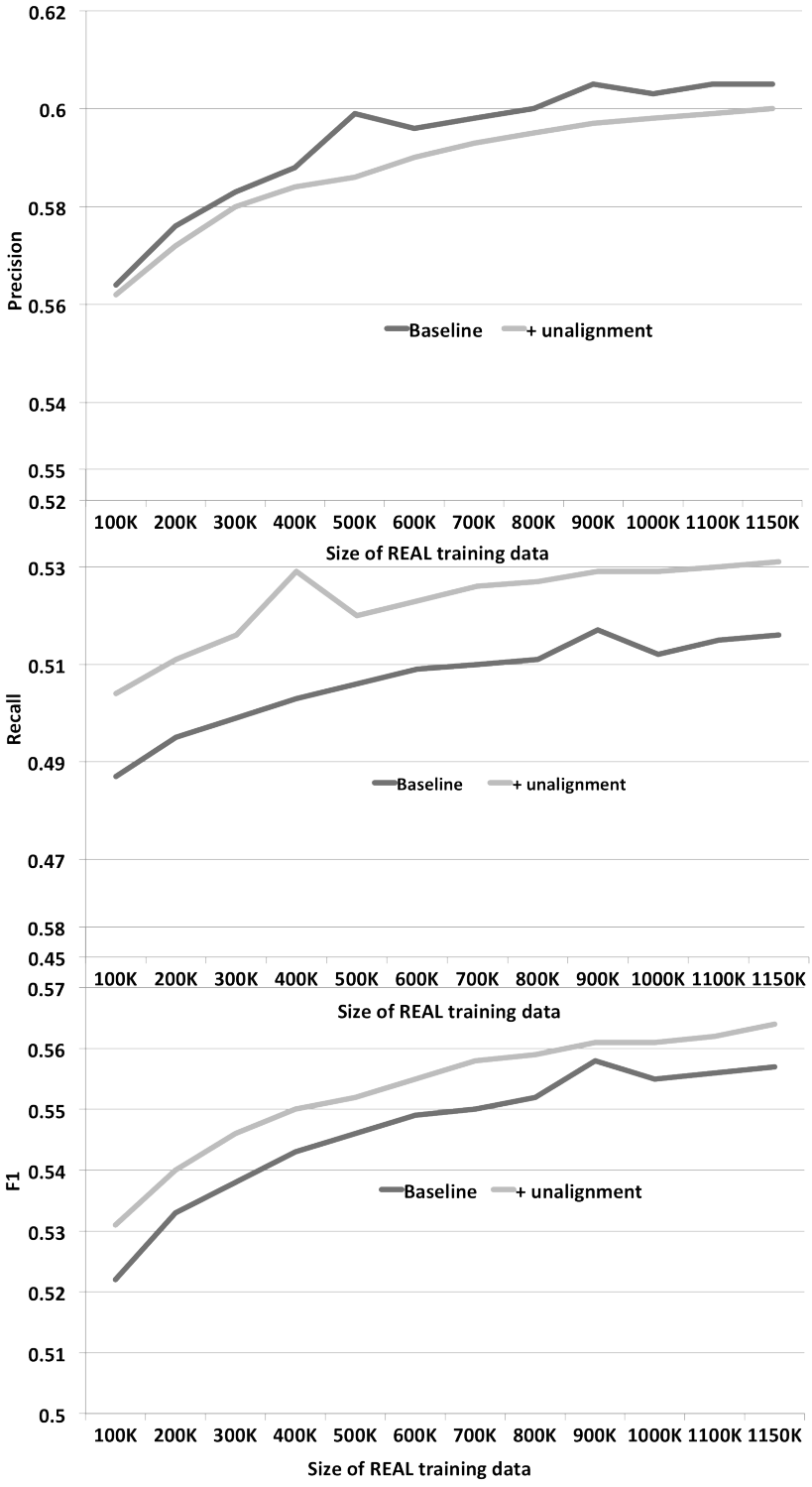


Figure 5.3: Precision, recall and F1 of GIZA++ output on ORACLE training slice . *'+ Unalignment'* refers to PBMT systems built with automatic unalignable word prediction

5.2 Output analysis

The difference between the actual alignments produced by the baseline and the proposed system are compared according to the lexical translation tables generated by MOSES. It is found that the size of the translation table produced by the proposed system exceeds that of the baseline to a large extent. There are 537K links exclusively found in the proposed system and 363K links exclusively in the baseline system. One of the possibilities is that the baseline system frequently align rare words to ought-to-be-unaligned words, which occur in high frequency, yet detailed human evaluation is required to prove this hypothesis.

Chinese word	gloss	English lexical translation	
		Baseline only	Propose only
tao	escape	from, away	surrender, fugitives
xie	bring	him	bringing
xing	form	and	model
dan	but	it, the, they	yet, nevertheless, though
she	society	our, the	agency, communities
pa	scare	that, are, be, will	fears, worried, may, frightening, terrible

Table 5.1: Examples of translations exclusively found in the top **15** lexical translation produced by the Baseline or Proposed (default GIZA^{++}) systems.

Table 5.1 shows the difference in the resulting lexical translation tables for some rare Chinese words. In particular, the top 15 English translations of each Chinese word are collected and the translations that are exclusively in one table but not the other are listed. Note that the baseline system tends to incorrectly align these infrequent Chinese words to function words (a phenomenon known as garbage collection). In contrast, the translations of the proposed system are more reasonable.

The motivation of pre-removing unalignable words is to prevent over-constriaining translation rules which require the absolute presence of unalignable words. Examining some sentences of the output translation to see if this purpose is achieved, cases are identified where Chinese *additional* words are correctly identified and removed, such as *'shu'* (Table 5.2, example 1) and *'dou'* (example 2). Yet there are also cases where commonly unaligned English *additional* words are omitted wrongly, such as *'is'* and *"'s"* (example 2).

1.Chi:	*ban*	*shu*	*yunanzhe*	
	half	number	disaster victim	
BL:	*half*	*of the number*	*of disaster victims*	
Unalign:	*half*		*of the disaster victims*	
Ref:	*half*		*of the victims*	

2.Chi:	*wode*	*yixie*	*pengyou*	*dou*	*shi*	*yixie*	*mingxing*	*de*	*fensi*
	my	some	friend	(discourse marker)	is	some	stars	's	fans
BL:	*some*	*of my*	*friends*		*are*	*some*	*of*	*the stars*	*fans*
Unalign:	*some*	*of my*	*friends*		_	*some*	_	*stars*	*fans*
Ref:	*some*	*of my*	*friends*		*are*	*fans*	*of*	*some*	*superstars*

Table 5.2: Translation examples from baseline and *unalignment* MT systems

Chapter 6

Conclusion

This study investigates the phenomenon of unalignable words in automatic word alignments for statistical machine translation. In view of the difference in lexicalization across languages, unalignable words in an aligned sentence pair are predicted according to their linguistic contexts. An easy-to-implement technique that requires minimal modification to the standard MT training pipeline is proposed to preprocess the training data based on the classification.

Three hypotheses are made concerning unalignable words in translation: (1) Unalignable words are not a closed list. (2) Unalignable words in a parallel text are predictable. (3) Removing unalignable words before automatic word alignment is beneficial for SMT. Analyses and experiments are conducted to prove these hypotheses. In Section 3, statistics of human annotated word alignments reveals that 83% tokens belong to potentially unalignable word types but only 17% tokens in the corpus are actually unalignable. Section 4.2 shows that supervised classification of core and unalignable words using cross-lingual lexical features achieves 92% accuracy. MT experiments in Sections 4.1 and 4.3 show that removing unalignable words by the proposed method based on gold standard unalignable words in small annotated data and automatic prediction of unalignable words in large realistic data improves the overall hierarchical SMT systems by 0.5 and 0.6 BLEU points respectively compared to systems using standard GIZA^{++}. The above results conclude that the claims of the study are valid.

Chapter 7

Future work

The proposed method removes unalignable words from the training corpus and returns them after automatic alignment of the core words and before extraction of phrase-based or hierarchical translation rules. This results in different unconditioned versions of translation rules containing various number of unalignable words. However, whether to contain an unalignable word or not is unlikely to be a random binary choice. A translation rule that includes a particular unalignable word should be distinguished from one excluding it based on more fine-grained conditions.

A larger view of this work is that the unalignable words are actually aligned to a structure above the lexical level. For example, when the source syntactic tree contains an empty pronoun, an unalignable pronoun is to be added to the translation. Above sentence level, discourse relations are often implicit in Chinese but explicitly marked in English. Unalginable discourse connective is thus to be added when translating Chinese to English.

Further investigation is needed to model the addition of unalignable words to the translation or omission from the source text. Other directions in future work include further exploration of features for unalignable word classification and direct incorporation of the distinctive treatment of core and unalignable words to the translation model.

Bibliography

[1] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, 2005.

[2] P. F. Brown, S. A. D. Pietra, V. J. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2), 1993.

[3] C.-C. Chang and C.-J. Lin. Libsvm : a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(27), 2011.

[4] C. Cherry and D. Lin. Soft syntactic constraints for word alignment through discriminative training. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2006.

[5] T. Chung and D. Gildea. Effects of empty categories on machine translation. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2010.

[6] J. DeNero and D. Klein. Tailoring word alignments to syntactic machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.

[7] G. Doddington. Automatic evaluation of translation quality using n-gram co-occurence statistics. *Proceedings of the International Conference on Human Language Technology*, 2002.

[8] V. Fossum, K. Knight, and S. Abney. Using syntax to improve word alignment precision for syntax-based machine translation. *Proceedings of the Workshop on Statistical Machine Translation*, 2008.

[9] A. Fraser and D. Marcu. Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, 33(3), 2007.

[10] U. Hermjakob. Improved word alignment with statistics and linguistic heuristics. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2009.

[11] H. Isozaki, T. Hirao, K. Duh, K. Sudoh, and H. Tsukada. Automatic evaluation of translation quality for distant language pairs. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2010.

[12] M. Johnson. A simple pattern-matching algorithm for recovering empty nodes and their antecedents. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning*, 2001.

[15] X. Li, N. Ge, S. Grimes, S. M. Strassel, and K. Maeda. Enriching word alignment with linguistic tags. *Proceedings of International Conference on Language Resources and Evaluation*, 2010.

[16] X. Li, N. Ge, and S. Strassel. *Tagging Guidelines for Chinese-English Word Alignment*. Linguistic Data Consortium, version 1.0 edition, 2009.

[17] X. Li, S. Grimes, and S. Strassel. Gale chinese-english word alignment and tagging training part 1 – newswire and web. Philadelphia: Linguistic Data Consortium, Web download files, 2012.

[18] X. Li, S. Grimes, and S. Strassel. Gale chinese-english word alignment and tagging training part 2 – newswire. Philadelphia: Linguistic Data Consortium, Web download files, 2012.

[19] X. Li, S. Grimes, and S. Strassel. Gale chinese-english word alignment and tagging training part 3 – web. Philadelphia: Linguistic Data Consortium, Web download files, 2012.

[20] P. Liang, A. Bouchard-Cote, D. Klein, and B. Taskar. An end-to-end discriminative approach to machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and International Conference on Computational Linguistics*, 2006.

[21] X. Ma. Hong kong parallel text. Philadelphia: Linguistic Data Consortium, Web download files, 2004.

[22] X. Ma. Chinese english news magazine parallel text. Philadelphia: Linguistic Data Consortium, Web download files, 2005.

[23] X. Ma. Chinese news translation text part 1. Philadelphia: Linguistic Data Consortium, Web download files, 2005.

[24] X. Ma. Gale phase 1 chinese broadcast news parallel text - part 1. Philadelphia: Linguistic Data Consortium, Web download files, 2007.

[25] X. Ma. Gale phase 1 chinese blog parallel text. Philadelphia: Linguistic Data Consortium, Web download files, 2008.

[26] X. Ma. Gale phase 1 chinese broadcast news parallel text - part 2. Philadelphia: Linguistic Data Consortium, Web download files, 2008.

[27] X. Ma. Gale phase 1 chinese broadcast news parallel text - part 3. Philadelphia: Linguistic Data Consortium, Web download files Philadelphia: Linguistic Data Consortium, Web download files, 2008.

[28] X. Ma. Gale phase 1 chinese broadcast conversation parallel text - part 1. Philadelphia: Linguistic Data Consortium, Web download files, 2009.

[29] X. Ma. Gale phase 1 chinese broadcast conversation parallel text - part 2. Philadelphia: Linguistic Data Consortium, Web download files, 2009.

[30] X. Ma and S. Strassel. Gale phase 1 chinese newsgroup parallel text - part 1. Philadelphia: Linguistic Data Consortium, Web download files, 2009.

[31] X. Ma and S. Strassel. Gale phase 1 chinese newsgroup parallel text - part 2. Philadelphia: Linguistic Data Consortium, Web download files, 2009.

[32] I. D. Melamed, R. Green, and J. P. Turian. Precision and recall of machine translation. *Proceedings of the North American Chapter of the Association for Computational Linguistics*, 2003.

[33] T. Nakazawa and S. Kurohashi. Alignment by bilingual generation and monolingual derivation. *Proceedings of the International Conference on Computational Linguistics*, 2012.

[34] E. A. Nida. *Toward a Science of Translating: with Special Reference to Principles and Procedures Involved in Bible Translating*. BRILL, 1964.

[35] F. J. Och and H. Ney. Improved statistical alignment models. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2000.

[36] F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1), 2003.

[37] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2002.

[38] Philadelphia: Linguistic Data Consortium. *GALE Program: Chinese to English Translation Guidelines*, version 2.9 edition, 2010.

[39] H. Setiawan, C. Dyer, and P. Resnik. Discriminative word alignment with a function word reordering model. *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, 2010.

[40] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, 2006.

[41] C. Tillmann, S. Vogel, H. Ney, and A. Zubiaga. A dp-based search using monotone alignments in statistical translation. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1997.

[42] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2003.

[43] H. Tseng, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. Optimizing chinese word segmentation for machine translation performance. *Proceedings of the Workshop on Statistical Machine Translation*, 2008.

[44] J. P. Turian, L. Shen, and I. D. Melamed. Evaluation of machine translation and its evaluation. *Proceedings of MT Summit*, 2003.

[45] J. Xu and J. Chen. How much can we gain from supervised word alignment? *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2011.

[46] H. Zhang and D. Gildea. Stochastic lexicalized inversion transduction grammar for alignment. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2005.