Enhancing Phrase Extraction from Word Alignments Using Morphology

Ahmed Ragab Nabhan
Department of Mathematics, Faculty of Science, Fayoum University,Egypt.
E-mail: ragab@claes.sci.eg

Ahmed Rafea
Computer Science Dept., American University in Cairo,113, Sharia Kasr El-Aini, P.O. Box 2511, 11511, Cairo, Egypt.
E-mail: rafea@aucegypt.edu

Khaled Shaalan
Informatics Institute, The British University in Dubai (BUiD), Emirates.
Email: khaled.shaalan@buid.ac.ae

Abstract

We propose a technique for effective extraction of bilingual phrases from word alignments using morphological processing. Morphological processing leads to an increase of the frequency of words in the corpus, consequently reduces Alignment Error Rate (AER). Intuitively, better word alignments enhance the quality of bilingual phrases extracted. Using alignments of a stemmed corpus for phrase extraction, instead of alignments of a raw one, shows significant improvements in translation quality, especially with small corpora.

1. Introduction

Phrase based statistical translation systems outperform classical word-based counterparts. They achieve better word ordering and are much faster during decoding. The idea is that phrases move as units during the translation process (Fox, 2002). A distortion model moves phrases into their appropriate positions in the target sentences while the phrase translation model takes care of local word ordering.

Starting from the basic idea, a variety of techniques has been proposed to build phrase based translation systems. The techniques can be differentiated based on the way phrase correspondences are learnt from parallel corpora. One of these methods is based on phrase extraction from word aligned sentence pairs (Och, 2001; Koehn, 2003). A word alignment tool is used to determine the best alignment for each pair of sentences (called the viterbi alignment). Then, bilingual phrases correspondences are determined through a phrase extractor routine. As the alignment error rate decreases, the quality of extracted phrases increases as a result.

Phrase based translation can benefit from morphological processing of both sides of the parallel corpus. Morphological analysis is very useful in increasing the data to parameter ratio and hence it decreases alignment error rate (Dejean, 2003).

In this paper, we present a technique for better phrase extraction from word aligned corpora using morphological processing of source and target sentences. The experiments

conducted on Arabic-English corpus show a remarkable improvement in translation quality.

The paper is structured as follows. In section 2, we discuss the problem of word alignment. Section 3 presents phrase based statistical translation background. In section 4, we present a technique for effective extraction of bilingual phrases from word alignments. Section 5 shows experiments and results. Section 6 presents conclusions and future work.

2. Word Alignments

Statistical Machine Translation is all about alignments. A word alignment, in general, is a mapping from source sentence words to target sentence words (Brown et al, 1993). Alignments are the basis on which translation models are built. For two sentences that are translations to each other, there is more than one possible alignment. Each alignment is given a probability value that represents how sure we are about the alignment. Word alignments are very useful in estimating translation, fertility, and word ordering frequencies for the classic word-based (fertility-based) translation models.

Alignments are represented as a vector of integers of length equal to number of words in the target sentence. Each number of the alignment vector is the position of the source word that generates that target word. Unaligned words are linked to a virtual NULL word and take a value of zero in the alignment vector. For example, suppose we have the two source and target sentence:

T: وزير صيني و برازيلي يفوزان بجائزة الأمم المتحدة للبيئة
S: A Chinese minister and a Brazilian win UN environmental prize

This pair may have the alignment vector:

A= {3, 2, 4, 6, 7, 10, 8, 9}

The first target word 'وزير' is attached to the third source word 'minister'. The alignment may be represented visually by a matrix as in figure 1.

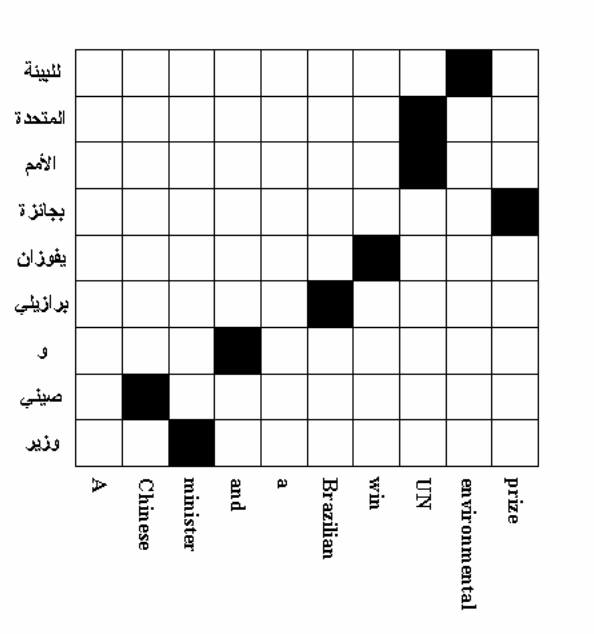

Figure.1 A word alignment example

Theoretically, an alignment is a hidden variable that training algorithms find out during the parameter estimation process. The probability of an alignment a given a source sentence e and its target translation f can be expressed as follows:

$$P(a \mid e, f) = \frac{P(a, f \mid e)}{p(f \mid e)} \quad (1)$$

Summing over ***a*** for both sides yields:

$$\sum_a P(a \mid e, f) = \sum_a \frac{P(a, f \mid e)}{P(f \mid e)} \quad (2)$$

The left hand side sums to 1 and P (f | e) can be factored out, and finally we get the translation model equation:

$$P(f \mid e) = \sum_a P(a, f \mid e) \quad (3)$$

Hence, the translation model probability P (f | e) is the sum of all probabilities of producing a French string ***f*** and an alignment **a** given an English string ***e***.

In order to estimate the probability P (a, f | e), Brown et al. (1993), introduced a series of five statistical models each of which contributes to the calculation of p (a, f | e).

3. Phrase Based Statistical Translation

The main deficiency of the IBM translation model is the distortion (permutation) component. One cannot justify that ordering of words in target language merely adheres to some statistical distribution that forces word order to be in some form or another, ignoring the syntactic regime that controls word ordering. The same words may be permuted in many ways reflecting different semantics. A word at the beginning of the source sentence may map to a word in the last position in the target language. The word ordering is conditioned only on word classes and positions in the source sentence (Yamada and Knight, 2002). The existence of a language model as a judge to penalize bad word orderings does not help much. The outcome is a high word ordering error rate and low performance, especially in longer sentences. The existence of many ordering possibilities increases as the length of target sentence increases, and consequently, decoding time increases too.

An interesting fact about the translation process is that often words in source sentence constitute phrases that are translated as units into target sentences (Brown el al, 1993). That is, phrases in one language tend to cohere (stay together) during translation. Fox (2002) studied the phrasal cohesion phenomena across two languages and conclude that there is a large amount of regularity of phrasal cohesion.

As a response to the phrasal cohesion phenomenon, many researchers introduced models for phrase-based statistical machine translation. In general, identifying phrases within sentence pairs can be done in two ways. One method is to establish phrase correspondences directly in the parallel corpus (Marcu and Wong, 2002). Alternatively, bilingual phrase correspondences can be learned from a word aligned corpus. In this paper we use the second alternative.

One of the outstanding models of phrase based statistical translation is the alignment template models, which can be viewed as a phrase translation system (Och et al, 1999). Bilingual phrases are derived from IBM Model 4 word alignments. Koehn et al. (2003) introduced a phrase-based model and a beam search decoder for phrase based machine translation. Their proposed system records high performance with the BLEU metric, by a fairly simple heuristic for learning phrase correspondences from word alignments, and then scoring phrase translation probabilities using bilingual phrase frequencies. Actually one would say that the way BLEU rates translations using phrase to phrase matching inspires the phrase-based machine translation.

The general model of PSMT is a noisy channel model. Using Bayes, we search for a source sentence e which maximizes the probability P (e | f), where f is the target sentence we want to translate:

$$\arg\max_e P(e \mid f) = \arg\max_e P(f \mid e).P(e) \quad (4)$$

This equation modularizes the system into a translation model $P(f \mid e)$ and a language model $P(e)$. The language model weights the

final string of words, while the phrase translation model takes care of local word ordering.

In the decoding phase, the target sentence is segmented into a number of phrase segments. The possible segmentations of target sentence are equally likely. The foreign phrases are translated into source phrases using the phrase translation model. Finally, the translated phrases may be reordered according to a distortion model. For more explanation on the phrase based translation system, the user is encouraged to review Koehn (2003).

In the proposed work, we used GIZA++ to word-align the parallel corpus. GIZA++ (Och, 2001) has been cited as a robust translation model training toolkit for building word-based translation models. GIZA++ generates word alignments as a byproduct of the training process. There have been a number of heuristics for using the word alignments to extract phrase-to-phrase translations. Some heuristics exist to overcome one deficiency of Model 4 word alignment model that the fertility model is asymmetric, i.e. source words can be aligned to more than one target words, but target words cannot be aligned to more than one source word. To handle the problem, GIZA++ is used to generate word alignments for both directions; source to target and target to source, and the two word alignments are intersected to get a high precision word alignments. The word alignment intersection points are expanded by candidate neighboring candidates to increase the number of phrases extracted. This expansion can be done using different heuristics (Och, 2002; Koehn, 2003), according to the definition of neighborhood (block or diagonal).

4. Effective Phrase Extraction Using Morphology

Word alignments are the basis of the phrase based translation model we present in this article. Intuitively, good word alignments yields better phrase based translation model. The lower the Alignment Error Rate (AER) the better. In general, increasing the learning curve of the model by adding more bilingual sentence pairs to the corpus improves the accuracy of word alignments. However, one can increase the accuracy of alignments by boosting the training algorithm using a variety of techniques other than using more training data. Some techniques make use of available linguistic resources such as dictionaries and stemmers. Indeed, one can improve word alignments and translation models quality without using additional training data or linguistic resources. Harrington (2003) applied bagging and boosting to the problem of word alignments and showed that word alignments can be improved without linguistic resources. Nabhan and Rafea (2004) optimize the training process by adjusting the parameters of GIZA++ to produce better translation model and better word alignments. In the presented work, we used stemmers to enhance the word alignments.

There are many facts that make the word alignment task a difficult one. One of such facts is the sparseness of training data. Many word forms, especially for morphologically rich languages, are absent in the training corpora. Another problem is related to morphology, when we have a morphologically rich language sentence paired with a less rich language one. In this case, morphologically rich sentences tend to be shorter than their translation, which makes the alignment tasks more difficult. For example, in our experiments, we found that the average sentence length of the Arabic side of the corpus was 30 words, while the average length of the English sentences was 37 words. That shows that the English sentence is longer by a factor of

1.23 than the Arabic sentence, which makes the word alignment ambiguous for word alignment models. A third fact about language pairs is the structure differences which cause problem in both word alignment and translation tasks.

To face the sparse data problem, and to achieve a degree of symmetrization between language pairs, statistical machine translation was in need for morphological and syntactic information. Morphological analysis has been identified as a good means for increasing word alignment accuracy and translation performance, even in the early days of SMT. Brown et al. (1993) suggested the use of morphology to increase the effectiveness of SMT systems. They also suggested using bilingual dictionaries. Nießen (2002) proposed incorporating morpho-syntactic information (lemma–tag representation) into SMT process and introduced hierarchical lexicon models including baseform and POS information. He also proposes reordering operations that help SMT by harmonizing word order between source and target languages. Ueffing and Ney, (2003) addressed the problem of producing correct inflected form in a richer morphology target language. They introduced transformations to the poor morphology source language based on POS information. Dejean et al. (2003) improved IBM Model 4 alignments using stemmers for HLT-NAACL 2003 Workshop shared task. They showed that word stemming reduced the parameter space by reducing vocabulary size. Sereewattana (2003) used segmentation to "effectively reduce the number of unknown words and singletons in the corpora which helps improve the translation model". Young-Suk Lee (2004) proposed word segmentation into prefix-stem-suffix sequence and using part of speech tagging to achieve a syntactic and morphological symmetry between source and target languages. Corston-Oliver and Gamon (2004) measured the impact of normalizing inflectional morphology on German-English statistical word alignment. Popovic et al. (2005) suggested augmenting a small parallel text with morpho-syntactic language resources. They suggested different processing steps for a morphologically rich language.

The main assumption that motivates this work is that the accuracy of extracted phrases can be increased if we enhance word alignments that GIZA++ produce. Generally speaking, we want to augment the corpus using morphological processing to increase the data-to-parameter ratio and to decrease noise (rare events in the form of rare word forms). A number of researchers reported work on using stemming and morphological analysis in the SMT framework to get better results. For example,

In the presented work, we used word stemming of parallel corpus to increase the accuracy of extracted phrases. If we extract bilingual phrase from the word alignments of the stemmed corpus, then the extracted phrases will be a meaningless sequence of word lemmas and many linguistic features will be lost. To preserve word forms, we used the alignment vectors of the stemmed corpus with the original un-stemmed vocabulary during phrase extraction.

For example, consider the following pair of sentences:

T: الأمين العام للأمم المتحدة يزور القاهرة
S: united nations secretary general visits Cairo

The alignment vector for this pair which GIZA++ generate is: {3,4,0,1,0,6}. The system was unable to align 'للأمم' to "nations" and 'يزور' to 'visits', as outlined in figure 2. A side effect

of the EM training algorithm is that NULL word acts as a garbage collector for unknown (rare) words. Assigning both target words to the NULL word increases the probability of the alignment more than if it aligned 'visits' to 'للأمم' or 'يزور'. For more explanation on the behavior of EM algorithm in aligning too many words to NULL, the reader may refer to Moore (2004).

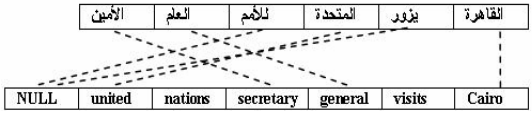

Figure 2: A noisy alignment

We can reduce alignment error rate using morphological processing. It is known that stemming reduces the number of unknown words. For example, different word forms of the noun 'أمة' such as 'أمم', 'الأمم', 'الأمة', … etc, are reduce to one word form which is the stem 'أمة'. Morphological processing helps the EM algorithm to detect more reliable alignments during training. The more accurate alignment produced after using morphological processing is outlined below.

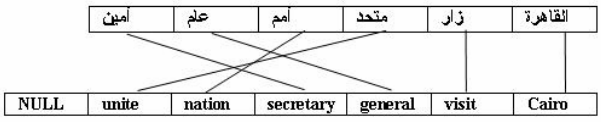

Figure 3: A more accurate alignment

During the phrase extraction process, we use the more accurate alignment of the stemmed sentences together with the original sentence pair (the original word forms), instead of collecting phrases from the stemmed sentence pair. As we will see in the next section, experiments show that this simple method increases the quality of the bilingual phrases extracted in terms of lexical weighting of the extracted phrases and increases the number of phrases extracted too.

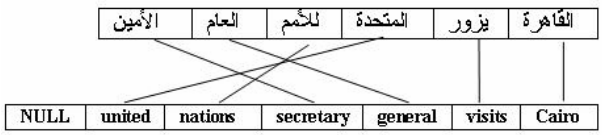

Figure 4: Using the accurate word alignments with original word forms of the sentences.

5. Experiments and Results

We used a set of public resources available on the web. The Pharaoh decoder developed by Philip Keohn (2003) was used a baseline for the experiments. We developed our tools to build the phrase translation model following the steps shown in the Pharaoh decoder user manual. We used public linguistic morphological analyzers and stemmers. The Buckwalter morphological analyzer available at LDC was used to process the Arabic side of the training corpus. We used the Porter stemming for processing the English side of the corpus. We used the LDC Arabic-English parallel corpus for conducting our experiments. We used a corpus of about 52k sentence pairs. Table 1 and Table 2 show different statistics about training and test corpora. Different preprocessing steps were taken to prepare the data, removing punctuation marks and lower-casing English words.

We ran experiments for different corpora sizes to ensure that morphological processing is effective even when increasing the size of the training data. We have sets of 20k, 30k, 40k and 52k sentence pairs, with an increase of the size of about 10k sentence. We used the translation model extracted from each corpus to translate a fixed set of 500 Arabic sentences. We used the BLEU metric (Papeneni, 2001) as a performance measure.

For each corpus size, we run two experiments. In the first experiment, we measure the quality

of the phrase translation model built without using any morphological processing steps. In the second experiment, we test the effect of morphological processing on the translation quality. We ran the Buckwalter morphological analyzer for stemming the Arabic corpus. The Porter stemmer was used for processing the English corpus. We ran GIZA++ over the parallel stemmed corpus. We used the alignments of the stemmed corpus to collect phrases from the original raw corpus as explained by the technique. We used the resultant translation model to translate the 500 Arabic sentences. Figure 5 shows BLEU score (Vertical) for each corpus (Horizontal).

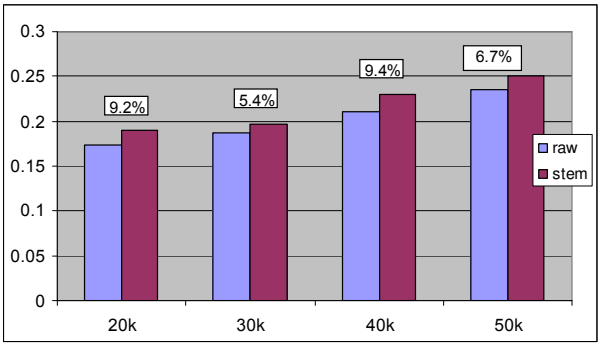

Figure 5: Experimental Results

As shown in figure 5 and tables 1 and 2, the use of morphological processing has a plenty of benefits. The parameter space of the training corpus was reduced because many word forms were reduced to its stem form. This appears clearly when comparing the vocabulary size of the raw and stemmed corpus. This in term increases the data-to-parameter ration and leads to a better parameter estimation and a reduced alignment error rate. The phrase extraction process was more effective in terms of the correctness of bilingual phrases and the number of extracted phrases increases too. By visually inspecting the phrase model before and after morphological processing, we believe that both coverage and correctness of the phrase table increases due to the processing. The net effect is a better translation quality and a less out of vocabulary rate for the test data.

Table1: Training Corpus statistics

	Raw	Stemmed
Number of Arabic tokens	1.5 M	
Number of English tokens	1.9 M	
Number of Arabic words	90 K	25 K
Number of English words	38 K	27 K
Arabic Average sentence length	30 words	
English Average sentence length	37 words	
Number of extracted phrases	1.25 M	1.6 M

Table2: Test Corpus statistics

Corpus size	500
Number of Arabic tokens	15 K
Number of English tokens	18 K
Arabic Average sentence length	31 words
English Average sentence length	37 words

6. Conclusion

We presented a simple but effective heuristic for better extraction of translation phrase pairs from a word aligned corpora. For a morphologically rich language such as Arabic, morphological processing yields significant improvements in terms of higher translation quality and lower alignment error rate. Although we did not measure the alignment error rate quantitatively, we believe that the error rate decreases, by visual inspection of alignments. The method we present yields an average translation improvement of 7.2%.

In future work, we are going to study the effect of using part of speech during the parameter estimation process. Some researchers added an extra part of speech parameter for IBM Model 2 and reported better results. We may make use of syntactic information, although some researchers report poor performance when constraining the phrase extraction process to

produce syntactically correct phrases. We still believe that available linguistic resources can be of benefit to statistical machine translation.

7. Acknowledgement

This work is supported in part by a Collaboration Project on Statistical Machine Translation, between Information Science Institute, University of Southern California, and Computer Science Department, American University in Cairo. The project is funded by US-Egypt Science Board.

8. References

Ahmed Ragab Nabhan & Ahmed Rafea: 2004, 'Tuning Statistical Machine Translation Parameters', International Conference on Computational Intelligence, Istanbul, Turkey, 2004: 181-184

Brian Harrington: 2003, 'Bagging and Boosting for Word Alignment', http://www.harringtonweb.com/brh/misc/, (2005).

Franz Josef Och: 2002, 'Statistical Machine Translation: From Single-Word Models to Alignment Templates', http://www-mgi.informatik.rwth-aachen.de/Kolloquium/pastOberseminar.html

Franz Josef Och: 2000, GIZA++ Readme File. http://wasserstoff.informatik.rwth-aachen.de/Colleagues/och/software/GIZA++.html, (2003).

Heidi Fox: 2002, 'Phrasal Cohesion and Statistical Machine Translation', in Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.

Herve Dejean, Eric Gaussier, Cyril Goutte & Kenji Yamada: 2003, 'Reducing Parameter Space for Word Alignment', HLT-NAACL 2003 Workshop, pp. 23-26

Kenji Yamada & Kevin Knight: 2002, 'A Decoder for Syntax-based Statistical MT', Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 303-310.

Kevin Knight: 1999, 'A Statistical MT Tutorial Workbook', www.clsp.jhu.edu/ws99/projects/mt/mt-workbook.htm, (2004).

Kishore Papeneni, Salim Roukos & Todd Ward: 2001, 'BLEU: A Method for Automatic Evaluation of Machine Translation', IBM Research Report, RC22176

Maja Popovi´c, David Vilar, Hermann Ney, Slobodan Jovi˘ci´c & Zoran ˘ Sari´: 2005, ' Augmenting a Small Parallel Text with Morpho-syntactic Language Resources for Serbian-English Statistical Machine Translation', Proceedings of theWorkshop on Building and Using Parallel Texts, 2005, University of Michigan

Nicola Ueffing & Hermann Ney: 2003, ' Using POS information for statistical machine translation into morphologically rich languages', Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, pp. 347 – 354, ISBN: 1-333-56789-0.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer: 1993, 'The Mathematics of Statistical Machine Translation: Parameter Estimation' http://www.clsp.jhu.edu/ws99/projects/mt/ibmpaper.ps, (2003).

Philipp Koehn: 2003, ‘A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models’, Technical Manual of the Pharaoh decoder.

Philipp Koehn, Franz Josef Och & Daniel Marcu: 2003, ‘Statistical Phrase-Based Translation’, In Proceedings of the Human Language Technology Conference (HLT), pp. 127–133

Simon Corston-Oliver & Michael Gamon: 2004, ‘Normalizing German and English inflectional morphology to improve statistical word alignment’, The 6th Conference of the Association for Machine Translation in the Americas.

Sonja Nießen: 2002, 'Improving Statistical Machine Translation using Morpho-syntactic Information', PHD dissertation.

Yaser Al-Onaizan, JanCurin, Micheal Jahr, Kevin Knight, John Lafferty, Dan Melamed, Franz Josef Och, David Purdy, Noah A. Smith & David Yarowsky: 1999, ‘Statistical Machine Translation Final Report’, www.clsp.jhu.edu/ws99/projects/mt, (2004).

Young-Suk Lee: 2004, ' Morphological Analysis for Statistical Machine Translation', proceedings of the HLT-NAACL 2004 conference, pp. 57-60.