Segmentation for English-to-Arabic Statistical Machine Translation

Ibrahim Badr **Rabih Zbib** **James Glass**
Computer Science and Artificial Intelligence Lab
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{iab02, rabih, glass}@csail.mit.edu

Abstract

In this paper, we report on a set of initial results for English-to-Arabic Statistical Machine Translation (SMT). We show that morphological decomposition of the Arabic source is beneficial, especially for smaller-size corpora, and investigate different recombination techniques. We also report on the use of Factored Translation Models for English-to-Arabic translation.

1 Introduction

Arabic has a complex morphology compared to English. Words are inflected for gender, number, and sometimes grammatical case, and various clitics can attach to word stems. An Arabic corpus will therefore have more surface forms than an English corpus of the same size, and will also be more sparsely populated. These factors adversely affect the performance of Arabic↔English Statistical Machine Translation (SMT). In prior work (Lee, 2004; Habash and Sadat, 2006), it has been shown that morphological segmentation of the Arabic source benefits the performance of Arabic-to-English SMT. The use of similar techniques for English-to-Arabic SMT requires recombination of the target side into valid surface forms, which is not a trivial task.

In this paper, we present an initial set of experiments on English-to-Arabic SMT. We report results from two domains: text news, trained on a large corpus, and spoken travel conversation, trained on a significantly smaller corpus. We show that segmenting the Arabic target in training and decoding improves performance. We propose various schemes for recombining the segmented Arabic, and compare their effect on translation. We also report on applying Factored Translation Models (Koehn and Hoang, 2007) for English-to-Arabic translation.

2 Previous Work

The only previous work on English-to-Arabic SMT that we are aware of is by Sarikaya and Deng (2007). It uses shallow segmentation, and does not make use of contextual information. The emphasis of that work is on using Joint Morphological-Lexical Language Models to rerank the output.

Most of the related work, though, is on Arabic-to-English SMT. Lee (2004) uses a trigram language model to segment Arabic words. She then proceeds to deleting or merging some of the segmented morphemes in order to make the segmented Arabic source align better with the English target. Habash and Sadat (2006) use the Arabic morphological analyzer MADA (Habash and Rambow, 2005) to segment the Arabic source; they propose various segmentation schemes. Both works show that the improvements obtained from segmentation decrease as the corpus size increases. As will be shown later, we observe the same trend, which is due to the fact that the model becomes less sparse with more training data.

There has been work on translating from English to other morphologically complex languages. Koehn and Hoang (2007) present Factored Translation Models as an extension to phrase-based statistical machine translation models. Factored models allow the integration of additional morphological fea-

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 153–156,
Columbus, Ohio, USA, June 2008.

tures, such as POS, gender, number, etc. at the word level on both source and target sides. The tighter integration of such features was claimed to allow more explicit modeling of the morphology, and is better than using pre-processing and post-processing techniques. Factored Models demonstrate improvements when used to translate English to German or Czech.

3 Arabic Segmentation and Recombination

As mentioned in Section 1, Arabic has a relatively rich morphology. In addition to being inflected for gender, number, voice and case, words attach to various clitics for conjunction (*w+ 'and'*)[1], the definite article (*Al+ 'the'*), prepositions (e.g. *b+ 'by/with', l+ 'for', k+ 'as'*), possessive pronouns and object pronouns (e.g. *+ny 'me/my', +hm 'their/them'*). For example, the verbal form *wsnsAEdhm* and the nominal form *wbsyAratnA* can be decomposed as follows:

(1) a. w+ s+ n+ sAEd +hm
and+ will+ we+ help +them

b. w+ b+ syAr +At +nA
and+ with+ car +PL +our

Also, Arabic is usually written without the diacritics that denote the short vowels, and different sources write a few characters inconsistently. These issues create word-level ambiguity.

3.1 Arabic Pre-processing

Due to the word-level ambiguity mentioned above, but more generally, because a certain string of characters can, in principle, be either an affixed morpheme or part of the base word, morphological decomposition requires both word-level linguistic information and context analysis; simple pattern matching is not sufficient to detect affixed morphemes. To perform pre-translation morphological decomposition of the Arabic source, we use the morphological analyzer MADA. MADA uses SVM-based classifiers for features (such as POS, number and gender, etc.) to choose among the different analyses of a given word in context.

We first normalize the Arabic by changing final *'Y'* to *'y'* and the various forms of *Alif hamza* to bare *Alif*. We also remove diacritics wherever they occur. We then apply one of two morphological decomposition schemes before aligning the training data:

1. **S1**: Decliticization by splitting off each conjunction clitic, particle, definite article and pronominal clitic separately. Note that plural and subject pronoun morphemes are not split.

2. **S2**: Same as S1, except that the split clitics are glued into one prefix and one suffix, such that any given word is split into at most three parts: *prefix+ stem +suffix.*

For example the word *wlAwlAdh* (*'and for his kids'*) is segmented to *w+ l+ AwlAd +P:3MS* according to S1, and to *wl+ AwlAd +P:3MS* according to S2.

3.2 Arabic Post-processing

As mentioned above, both training and decoding use segmented Arabic. The final output of the decoder must therefore be recombined into a surface form. This proves to be a non-trivial challenge for a number of reasons:

1. Morpho-phonological Rules: For example, the feminine marker *'p'* at the end of a word changes to *'t'* when a suffix is attached to the word. So *syArp +P:1S* recombines to *syArty ('my car')*

2. Letter Ambiguity: The character *'Y'* (*Alf mqSwrp*) is normalized to *'y'*. In the recombination step we need to be able to decide whether a final *'y'* was originally a *'Y'*. For example, *mdy +P:3MS* recombines to *mdAh 'its extent'*, since the *'y'* is actually a *Y*; but *fy +P:3MS* recombines to *fyh 'in it'*.

3. Word Ambiguity: In some cases, a word can recombine into 2 grammatically correct forms. One example is the optional insertion of *nwn AlwqAyp* (*protective 'n'*), so the segmented word *lkn +O:1S* can recombine to either *lknny* or *lkny*, both grammatically correct.

To address these issues, we propose two recombination techniques:

1. **R**: Recombination rules defined manually. To resolve word ambiguity we pick the grammatical form that appears more frequently in the

[1] In this paper, Arabic text is written using Buckwalter transliteration

training data. To resolve letter ambiguity we use a unigram language model trained on data where the character *'Y'* had not been normalized. We decide on the non-normalized from of the *'y'* by comparing the unigram probability of the word with *'y'* to its probability with *'Y'*.

2. **T**: Uses a table derived from the training set that maps the segmented form of the word to its original form. If a segmented word has more than one original form, one of them is picked at random. The table is useful in recombining words that are split erroneously. For example, *qrDAy*, a proper noun, gets incorrectly segmented to *qrDAn +P:1S* which makes its recombination without the table difficult.

3.3 Factored Models

For the Factored Translation Models experiment, the factors on the English side are the POS tags and the surface word. On the Arabic side, we use the surface word, the stem and the POS tag concatenated to the segmented clitics. For example, for the word *wlAwlAdh ('and for his kids')*, the factored words are *AwlAd* and *w+l+N+P:3MS*. We use two language models: a trigram for surface words and a 7-gram for the POS+clitic factor. We also use a generation model to generate the surface form from the stem and POS+clitic, a translation table from POS to POS+clitics and from the English surface word to the Arabic stem. If the Arabic surface word cannot be generated from the stem and POS+clitic, we back off to translating it from the English surface word.

4 Experiments

The English source is aligned to the segmented Arabic target using GIZA++ (Och and Ney, 2000), and the decoding is done using the phrase-based SMT system MOSES (MOSES, 2007). We use a maximum phrase length of 15 to account for the increase in length of the segmented Arabic. Tuning is done using Och's algorithm (Och, 2003) to optimize weights for the distortion model, language model, phrase translation model and word penalty over the BLEU metric (Papineni et al., 2001). For our baseline system the tuning reference was non-segmented Arabic. For the segmented Arabic experiments we experiment with 2 tuning schemes: **T1** uses segmented Arabic for reference, and **T2** tunes on non-segmented Arabic. The Factored Translation Models experiments uses the MOSES system.

Scheme	Training Set	Tuning Set
Baseline	34.6%	36.8%
R	4.04%	4.65%
T	N/A	22.1%
T + R	N/A	1.9%

Table 1: Recombination Results. Percentage of sentences with mis-combined words.

4.1 Data Used

We experiment with two domains: text news and spoken dialogue from the travel domain. For the news training data we used corpora from LDC[2]. After filtering out sentences that were too long to be processed by GIZA ($>$ 85 words) and duplicate sentences, we randomly picked 2000 development sentences for tuning and 2000 sentences for testing. In addition to training on the full set of 3 million words, we also experimented with subsets of 1.6 million and 600K words. For the language model, we used 20 million words from the LDC Arabic Gigaword corpus plus 3 million words from the training data. After experimenting with different language model orders, we used 4-grams for the baseline system and 6-grams for the segmented Arabic. The English source is downcased and the punctuations are separated. The average sentence length is 33 for English, 25 for non-segmented Arabic and 36 for segmented Arabic.

For the spoken language domain, we use the IWSLT 2007 Arabic-English (Fordyce, 2007) corpus which consists of a 200,000 word training set, a 500 sentence tuning set and a 500 sentence test set. We use the Arabic side of the training data to train the language model and use trigrams for the baseline system and a 4-grams for segmented Arabic. The average sentence length is 9 for English, 8 for Arabic, and 10 for segmented Arabic.

[2]Since most of the data was originally intended for Arabic-to-English translation our test and tuning sets have only one reference

4.2 Recombination Results

To test the different recombination schemes described in Section 3.2, we run these schemes on the training and development sets of the news data, and calculate the percentage of sentences with recombination errors (Note that, on average, there is one mis-combined word per mis-combined sentence). The scores are presented in Table 1. The baseline approach consists of gluing the prefix and suffix without processing the stem. **T + R** means that the words seen in the training set were recombined using scheme **T** and the remainder were recombined using scheme **R**. In the remaining experiments we use the scheme **T + R**.

4.3 Translation Results

The 1-reference BLEU score results for the news corpus are presented in Table 2; those for IWSLT are in Table 3. We first note that the scores are generally lower than those of comparable Arabic-to-English systems. This is expected, since only one reference was used to evaluate translation quality and since translating to a more morphologically complex language is a more difficult task, where there is a higher chance of translating word inflections incorrectly. For the news corpus, the segmentation of Arabic helps but the gain diminishes as the training data size increases, since the model becomes less sparse. This is consistent with the larger gain obtained from segmentation for IWSLT. The segmentation scheme **S2** performs slightly better than **S1**. The tuning scheme **T2** performs better for the news corpus, while **T1** is better for the IWSLT corpus. It is worth noting that tuning without segmentation hurts the score for IWSLT, possibly because of the small size of the training data. Factored models perform better than our approach with the large training corpus, although at a significantly higher cost in terms of time and required resources.

	Large	**Medium**	**Small**
Training Size	3M	1.6M	0.6M
Baseline	26.44	20.51	17.93
S1 + T1 tuning	26.46	21.94	20.59
S1 + T2 tuning	26.81	21.93	20.87
S2 + T1 tuning	26.86	21.99	20.44
S2 + T2 tuning	27.02	22.21	20.98
Factored Models + tuning	27.30	21.55	19.80

Table 2: BLEU (1-reference) scores for the News data.

	No Tuning	**T1**	**T2**
Baseline	26.39	24.67	
S1	29.07	29.82	
S2	29.11	30.10	28.94

Table 3: BLEU (1-reference) scores for the IWSLT data.

5 Conclusion

In this paper, we showed that making the Arabic match better to the English through segmentation, or by using additional translation model factors that model grammatical information is beneficial, especially for smaller domains. We also presented several methods for recombining the segmented Arabic target. Our results suggest that more sophisticated techniques, such as syntactic reordering, should be attempted.

Acknowledgments

We would like to thank Ali Mohammad, Michael Collins and Stephanie Seneff for their valuable comments.

References

Cameron S. Fordyce 2007. *Overview of the 2007 IWSLT Evaluation Campaign* . In Proc. of IWSLT 2007.

Nizar Habash and Owen Rambow, 2005. *Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop*. In Proc. of ACL.

Nizar Habash and Fatiha Sadat, 2006. *Arabic Preprocessing Schemes for Statistical Machine Translation*. In Proc. of HLT.

Philipp Koehn and Hieu Hoang, 2007. *Factored Translation Models*. In Proc. of EMNLP/CNLL.

Young-Suk Lee, 2004. *Morphological Analysis for Statistical Machine Translation*. In Proc. of EMNLP.

MOSES, 2007. *A Factored Phrase-based Beam-search Decoder for Machine Translation*. URL: http://www.statmt.org/moses/.

Franz Och, 2003. *Minimum Error Rate Training in Statistical Machine Translation*. In Proc. of ACL.

Franz Och and Hermann Ney, 2000. *Improved Statistical Alignment Models*. In Proc. of ACL.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, 2001. *Bleu: a Method for Automatic Evaluation of Machine Translation*. In Proc. of ACL.

Ruhi Sarikaya and Yonggang Deng 2007. *Joint Morphological-Lexical Language Modeling for Machine Translation*. In Proc. of NAACL HLT.