

# Context-based Arabic Morphological Analysis for Machine Translation

**ThuyLinh Nguyen**

Language Technologies Institute  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
thuylinh@cs.cmu.edu

**Stephan Vogel**

Language Technologies Institute  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15213, USA  
vogel@cs.cmu.edu

## Abstract

In this paper, we present a novel morphology preprocessing technique for Arabic-English translation. We exploit the Arabic morphology-English alignment to learn a model removing nonaligned Arabic morphemes. The model is an instance of the Conditional Random Field (Lafferty et al., 2001) model; it deletes a morpheme based on the morpheme's context. We achieved around two BLEU points improvement over the original Arabic translation for both a travel-domain system trained on 20K sentence pairs and a news domain system trained on 177K sentence pairs, and showed a potential improvement for a large-scale SMT system trained on 5 million sentence pairs.

## 1 Introduction

Statistical machine translation (SMT) relies heavily on the word alignment model of the source and the target language. However, there is a mismatch between a rich morphology language (e.g Arabic, Czech) and a poor morphology language (e.g English). An Arabic source word often corresponds to several English words. Previous research has focused on attempting to apply morphological analysis to machine translation in order to reduce unknown words of highly inflected languages. Nießen and Ney (2004) represented a word as a vector of morphemes and gained improvement over word-based system for

German-English translation. Goldwater and McClosky (2005) improved Czech-English translation by applying different heuristics to increase the equivalence of Czech and English text.

Specially for Arabic-English translation, Lee (2004) used the Arabic part of speech and English parts of speech (POS) alignment probabilities to retain an Arabic affix, drop it from the corpus or merge it back to a stem. The resulting system outperformed the original Arabic system trained on 3.3 million sentence pairs corpora when using monotone decoding. However, an improvement in monotone decoding is no guarantee for an improvement over the best baseline achievable with full word forms. Our experiments showed that an SMT phrase-based translation using 4 words distance reordering could gain four BLEU points over monotone decoding. Sadat and Habash (2006) explored a wide range of Arabic word-level preprocessing and produced better translation results for a system trained on 5 million Arabic words.

What all the above methodologies do not provide is a means to disambiguate morphological analysis for machine translation based on the words' contexts. That is, for an Arabic word analysis of the form *prefix\*-stem-suffix\** a morpheme only is either always retained, always dropped off or always merged to the stem regardless of its surrounding text. In the example in Figure (1), the Arabic word “AlnAfi\*h” (“window” in English) was segmented as “Al nAfi\* ap”. The morpheme “ap” is removed so that “Al nAfi\*” aligned to “the window” of the English sentence. In the sentence “hl ldyk mqAEd bjwAr AlnAf\*h ?” (“do you have window tables ?” in English) the word “AlnAfi\*h” is also segmented as “Al nAfi\* ap”. But in this sentence, morphological preprocessing should remove both “Al” and “ap” so that only the remain-

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (<http://creativecommons.org/licenses/by-nc-sa/3.0/>). Some rights reserved.

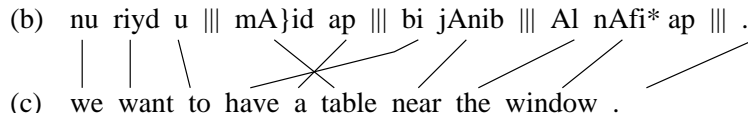
- (a) nryd mA}dh bjAnb AlnAf\*h .
- (b) nu riyd u ||| mA}id ap ||| bi jAnib ||| Al nAfi\* ap ||| .  

- (c) we want to have a table near the window .
- (d) nu riyd u mA}id ap bi jAnib Al nAfi\* .

Figure 1: (a) Romanization of original Arabic sentence, (b) Output of morphological analysis toolkit—words are separated by ‘|||’, (c) English translation and its alignment with full morphological analysis (d) Morphological analysis after removing unaligned morphemes.

ing morpheme “nAfi\*” aligned to the word “window” of the English translation. Thus an appropriate preprocessing technique should be guided by English translation and bring the word context into account.

In this paper we describe a context-based morphological analysis for Arabic-English translation that take full account morphemes alignment to English text. The preprocessing uses the Arabic morphology disambiguation in (Smith et al., 2005) for full morphological analysis and learns the removing morphemes model based on the Viterbi alignment of English to full morphological analysis. We tested the model with two training corpora of 5.2 millions Arabic words(177K sentences) in news domain and 159K Arabic words (20K sentences) in travel conversation domain and gain improvement over the original Arabic translation in both experiments. The system that trained on a subsample corpora of 5 millions sentence pairs corpora also showed one BLEU score improvement over the original Arabic system on unseen test set.

We will explain our technique in the next section and briefly review the phrase based SMT model in section 3. The experiment results will be presented in section 4.

## 2 Methodology

We first preprocess the Arabic training corpus and segment words into morpheme sequences of the form *prefix\* stem suffix\**. Stems are verbs, adjectives, nouns, pronouns, etc., carrying the content of the sentence. Prefixes and suffixes are functional morphemes such as gender and case markers, prepositions, etc. Because case makers do not exist in English, we remove case marker suffixes from the morphology output. The output of this process is a *full morphological analysis* corpus. Even after removing case markers, the token count

of the full morphology corpus still doubles the original Arabic’s word token count and is approximately 1.7 times the number of tokens of the English corpus. As stated above, using original Arabic for translation introduces more unknown words in test data and causes multiple English words to map to one Arabic word. At the morpheme level, an English word would correspond to a morpheme in the full morphology corpus but some prefixes and suffixes in the full morphology corpus may not be aligned with any English words at all. For example, the Arabic article “Al” (“the” in English) prefixes to both adjectives and nouns, while English has only one determiner in a simple noun phrase. Using the full morphological analysis corpus for translation would introduce redundant morphemes in the source side.

The goal of our morphological analysis method for machine translation is *removing nonaligned* prefixes and suffixes from the full morphology corpus using a data-driven approach. We use the word alignment output of the full morphology corpus to the English corpus to delete morphemes in a sentence. If an affix is not aligned to an English word in the word alignment output, the affix should be removed from the morphology corpus for better one-to-one alignment of source and target corpora. However, given an unseen test sentence, the English translation of the sentence is not available to remove affixes based on the word alignment output. We therefore learn a model removing non-aligned morphemes from the full morphology Arabic training corpus and its alignment to the English corpus. To obtain consistency between training corpus and test set, we applied the model to both Arabic training corpus and test set, obtaining pre-processed morphology corpora for the translation task.

In this section, we will explain in detail each steps of our preprocessing methodology:

- Apply word segmentation to the Arabic training corpus to get the full morphological analysis corpus.
- Annotate the full morphological analysis corpus based on its word alignment to the English training corpus. We tag a morpheme as “Deleted” if it should be removed from the corpus, and “Retained” otherwise.
- Learn the morphology tagger model.
- Apply the model to both Arabic training corpus and Arabic test corpus to get preprocessed corpus for translation.

## 2.1 Arabic Word Segmentation

Smith et al. (2005) applies a source-channel model to the problem of morphology disambiguation. The source model is a uniform model that defines the set of analyses. For Arabic morphology disambiguation, the source model uses the list of un-weighted word analyses generated by BAMA toolkit (Buckwalter, 2004). The channel model disambiguates the morphology alternatives. It is a log-linear combination of features, which capture the morphemes’ context including tri-gram morpheme histories, tri-gram part-of-speech histories and combinations of the two.

The BAMA toolkit and hence (Smith et al., 2005) do not specify if a morpheme is an affix or a stem in the output. Given a segmentation of an original Arabic word, we considered a morpheme  $a_i$  as a stem if its parts of speech  $p_i$  is either a noun, pronoun, verb, adjective, question, punctuation, number or abbreviation. A morpheme on the left of its word’s stem is a prefix and it is a suffix if otherwise. We removed case marker morphemes and got the full morphology corpus.

## 2.2 Annotate Morphemes

To extract the Arabic morphemes that align to English text, we use English as the source corpus and aligned to Arabic morpheme corpus using GIZA++ (Och and Ney, 2003) toolkit. The IBM3 and IBM4 (Brown et al., 1994) word alignment models select each word in the source sentence, generate fertility and a list of target words that connect to it. This generative process would constrain source words to find alignments in the target sentence. Using English as source corpus, the alignment models force English words to generate their alignments in the Arabic morphemes.

GIZA++ outputs Viterbi alignment for every sentence pair in the training corpus as depicted in (b) and (c) of Figure (1). In our experiment, only 5% of English words are not aligned to any Arabic morpheme in the Viterbi alignment. From Viterbi English-morpheme alignment output, we annotate morphemes either to be deleted or retained as follows:

- Annotate stem morphemes as “Retained”(R), in dependant of word alignment output.
- Annotate a prefix or a suffix as “Retained” (R) if it is aligned to an English word.
- Annotate a prefix or a suffix as “Deleted” (D) if it is not aligned to an English word.

Note that the model does not assume that GIZA++ outputs accurate word alignments. We lessen the impact of the GIZA++ errors by only using the word alignment output of prefix and suffix morphemes.

Furthermore, because the full morphology sentence is longer, each English word could align to a separate morpheme. Our procedure of annotating morphemes also constrains morphemes tagged as “Retained” to be aligned to English words. Thus if we remove “Deleted” morphemes from the morphology corpus, the reduced corpus and English corpus have the property of one-to-one mapping we prefer for source-target corpora in machine translation.

## 2.3 Reduced Morphology Model

The reduced morphology corpus would be the best choice of morphological analysis for machine translation. Because it is impossible to tag morphemes of a test sentence without the English reference based on Viterbi word alignment, we need to learn a morpheme tagging model. The model estimates the distributions of tagging sequences given a morphologically analysed sentence using the previous step’s annotated training data.

The task of tagging morphemes to be either “Deleted” or “Retained” belongs to the set of sequence labelling problems. The conditional random fields (CRF) (Lafferty et al., 2001) model has shown great benefits in similar applications of natural language processing such as part-of-speech tagging, noun phrase chunking (Sha and Pereira, 2003), morphology disambiguation (Smith et al., 2005). We apply the CRF model to our morpheme tagging problem.

Let  $\mathcal{A} = \{(\mathbf{A}, \mathbf{T})\}$  be the full morphology training corpus where  $\mathbf{A} = a_1|p_1 a_2|p_2 \dots a_m|p_m$  is a morphology Arabic sentence,  $a_i$  is a morpheme in the sentence and  $p_i$  is its POS;  $\mathbf{T} = t_1 t_2 \dots t_m$  is the tag sequence of  $\mathbf{A}$ , each  $t_i$  is either “Deleted” or “Retained”. The CRF model estimates parameter  $\bar{\theta}^*$  maximizing the conditional probability of the sequences of tags given the observed data:

$$\bar{\theta}^* = \underset{\bar{\theta}}{\operatorname{argmax}} \sum_{(\mathbf{A}, \mathbf{T}) \in \mathcal{A}} \tilde{p}((\mathbf{A}, \mathbf{T})) \log p(\mathbf{T}|\mathbf{A}, \bar{\theta}) \quad (1)$$

where  $\tilde{p}((\mathbf{A}, \mathbf{T}))$  is the empirical distribution of the sentence  $(\mathbf{A}, \mathbf{T})$  in the training data,  $\bar{\theta}$  are the model parameters. The model’s log conditional probability  $\log p(\mathbf{T}|\mathbf{A}, \bar{\theta})$  is the linear combination of feature weights:

$$\log p(\mathbf{T}|\mathbf{A}, \bar{\theta}) = \sum_k \theta_k f_k((\mathbf{A}_q, \mathbf{T}_q)) \quad (2)$$

The feature functions  $\{f_k\}$  are defined on any subset of the sentence  $\mathbf{A}_q \subset \mathbf{A}$  and  $\mathbf{T}_q \subset \mathbf{T}$ . CRFs can accommodate many closely related features of the input. In our morpheme tagging model, we use morpheme features, part-of-speech features and combinations of both. The features capture the local contexts of morphemes. The lexical morpheme features are the combinations of the current morpheme and up to 2 previous and 2 following morphemes. The part-of-speech features are the combinations of the current part of speech and up to 3 previous part of speeches. The part of speech, morpheme combination features capture the dependencies of current morphemes and up to its 3 previous parts of speech.

## 2.4 Preprocessed Data

Given a full morphology sentence  $\mathbf{A}$ , we use the morpheme tagging model learnt as described in the previous section to *decode*  $\mathbf{A}$  into the most probable sequence of tags  $\mathbf{T}^* = t_1 t_2 \dots t_m$ .

$$\mathbf{T}^* = \underset{\mathbf{T}}{\operatorname{argmax}} \Pr(\mathbf{T}|\mathbf{A}, \bar{\theta}^*) \quad (3)$$

If a  $t_i$  is “Deleted”, the morpheme  $a_i$  is removed from the morphology sentence  $\mathbf{A}$ . The same procedure is applied to both training Arabic corpus and test corpus to get preprocessed data for translation. We call a morphology sentence after removing “Deleted” tag a *reduced* morphology sentence.

In our experiments, we used the freely available CRF++<sup>1</sup> toolkit to train and decode with the morpheme tagging model. The CRF model smoothed the parameters by assigning them Gaussian prior distributions.

## 3 Phrase-based SMT System

We used the open source Moses (Koehn, 2007) phrase-based MT system to test the impact of the preprocessing technique on translation results. We kept the default parameter settings of Moses for translation model generation. The system used the “grow-diag-final” alignment combination heuristic. The phrase table consisted of phrase pairs up to seven words long. The system used a tri-gram language model built from SRI (Stolcke, 2002) toolkit with modified Kneser-Ney interpolation smoothing technique (Chen and Goodman, 1996). By default, the Moses decoder uses 6 tokens distance re-ordering windows.

## 4 Experiment Results

In this section we present experiment results using our Arabic morphology preprocessing technique.

### 4.1 Data Sets

We tested our morphology technique on a small data set of 20K sentence pairs and a medium size data set of 177K sentence pairs.

#### 4.1.1 BTEC Data

As small training data set we used the BTEC corpus (Takezawa et al., 2002) distributed by the International Workshop on Spoken Language Translation (IWSLT) (Eck and Hori, 2005). The corpus is a collection of conversation transcripts from the travel domain. Table 1 gives some de-

	Arabic			Eng
	Ori	Full	Reduced	
Sentences	19972			
Tokens	159K	258K	183K	183K
Types	17084	8207	8207	7298

Table 1: BTEC corpus statistics

tails for this corpus, which consists of nearly 20K sentence pairs with lower case on the English side. There is an imbalance of word types and word tokens between original Arabic and English. The

<sup>1</sup><http://crfpp.sourceforge.net/>



original Arabic sentences are on average shorter than the English sentences whereas the Arabic vocabulary is more than twice the size of the English vocabulary. The word segmentation reduced the number of word types in the corpus to be closed to English side but also increased word tokens quite substantially. By removing nonaligned morphemes, the reduced corpus is well balanced with the English corpus.

The BTEC experiments used the 2004 IWSLT Evaluation Test set as development set and 2005 IWSLT Evaluation Test set as unseen test data. Table 2 gives the details of the two test sets. Both of them had 16 reference translations per source sentence. The English side of the training corpus was used to build the language model. To optimize the parameters of the decoder, we performed minimum error rate training on IWSLT04 optimizing for the IBM-BLEU metric (Papineni et al., 2002).

#### 4.1.2 Newswire Corpora

We also tested the impact of our morphology technique on parallel corpus in the news domain. The corpora were collected from LDC’s full Arabic news translation corpora and a small portion of UN data. The details of the data are give in Table 3. The data consists of 177K sentence pairs, 5.2M words on the Arabic and 6M words on the English side.

	Arabic			Eng
	Ori	Full	Reduced	
Sentences	177035			
Tokens	5.2M	9.3M	6.2M	6.2M
Types	155K	47K	47K	68K

Table 3: Newswire corpus statistics

We used two test sets from past NIST evaluations as test data. NIST MT03 was used as development set for optimizing parameters with respect to the IBM-BLEU metric, NIST MT06 was used as unseen test set. Both test sets have 4 references per test sentence. Table 4 describes the data statistics of the two test sets. All Newswire translation experiments used the same language model estimated from 200 million words collected from the Xinhua section of the GIGA word corpus.

## 4.2 Translation Results

### 4.2.1 BTEC

We evaluated the machine translation according to the case-insensitive BLEU metric. Table 5 shows the BTEC results when translated with default Moses setting of distance-based reordering window size 6. The original Arabic word translation was the baseline of the evaluation. The second row contains translation scores using the full morphology translation. Our new technique of context-based morphological analysis is shown in the last row.

	IWSLT04	IWSLT05
Ori	58.20	54.50
Full	58.55	<b>55.87</b>
Reduced	<b>60.28</b>	<b>56.03</b>

Table 5: BTEC translations results on IBM-BLEU metrics(Case insensitive and 6 tokens distance reordering window). The boldface marks scores significantly higher than the original Arabic translation scores.

The full morphology translation performed similar to the baseline on the development set but outperformed the baseline on the unseen test set. The reduced corpus showed significant improvements over the baseline on the development set (IWSLT04) and gave an additional small improvement over the full morphology score over the unseen data (IWSLT05).

So why did the reduced morphology translation not outperform more significantly the full morphology translation on unseen set IWSLT05? To analysis this in more detail, we selected good full morphology translations and compared them with the corresponding reduced morphology translations. Figure 2 shows one of these examples. Typically, the reduced morphology translations

Source: *لوقت ام جفا ال يننا ينفسوي*  
 Ref: *i'm afraid what you are saying doesn't make any sense to me.*  
 Full: *i'm sorry i don't understand what you are saying.*  
 Reduced: *i'm sorry i don't understand what you say.*

Figure 2: An example of BTEC translation output.

are shorter than both the references and the full morphology outputs. Table 2 shows that for the IWSLT05 test set, the ratio of the average English reference sentence length and the source sen-

	IWSLT04 (Dev set)				IWSLT05 (Unseen set)			
	Arabic			English	Arabic			English
	Ori	Full	Reduced		Ori	Full	Reduced	
Sentences	500			8000	506			8096
Words	3261	5243	3732	64896	3253	5155	3713	66286
Avg Sent Length	6.52	10.48	7.46	8.11	6.43	10.19	7.34	8.18

Table 2: BTEC test set statistics

	MT03 (Dev set)				MT06 (Unseen set)			
	Arabic			English	Arabic			English
	Ori	Full	Reduced		Ori	Full	Reduced	
Sentences	663			2652	1797			7188
Words	16268	27888	18888	79163	41059	71497	48716	222750
Avg Sent Length	24.53	42.06	28.49	29.85	22.85	39.79	27.1	30.98

Table 4: Newswire test set statistics

tence length is slightly higher than the corresponding ratio for IWSLT04. Using the parameters optimised for IWSLT04 to translate IWSLT05 sentences would generate hypotheses slightly shorter than the IWSLT05 references resulting in brevity penalties in the BLEU metric. The IWSLT05 brevity penalties for original Arabic, reduced morphology and full morphology are 0.969, 0.978 and 0.988 respectively. Note that the BTEC corpus and test sets are in the travel conversation domain, the English reference sentences contain a large number of high frequency words. The full morphological analysis with additional prefixes and suffixes outputs longer translations containing high frequency words resulting in a high n-gram match and lower BLEU brevity penalty. The reduced translation method could generate translations that are comparable but do not have the same effect on BLEU metrics.

#### 4.2.2 Newswire results

Table 6 presents the translation results for the Newswire corpus. Even though morphology segmentation reduced the number of unseen words, the translation results of full morphological analysis are slightly lower than the original Arabic scores in both development set MT03 and unseen test set MT06. This is consistent with the result achieved in previous literature (Sadat and Habash, 2006). Morphology preprocessing only helps with small corpora, but the advantage decreases for larger data sets.

Our context dependent preprocessing technique

	MT03	MT06
Ori	45.55	32.09
Full	45.30	31.54
Reduced	<b>47.69</b>	<b>34.13</b>

Table 6: Newswire translation results on IBM-BLEU metrics (Case insensitive and 6 tokens distance reordering window). The boldface marks scores significantly higher than the original Arabic translation’s scores.

shows significant improvements on both development and unseen test sets. Moreover, while the advantage of morphology segmentation diminishes for the full morphology translation, we achieve an improvement of more than two BLEU points over the original Arabic translations in both development set and unseen test set.

#### 4.3 Unknown Words Reduction

A clear advantage of using morphology based translation over original word translation is the reduction in the number of untranslated words. Table 7 compares the number of unknown Arabic tokens for original Arabic translation and reduced morphology translation. In all the test sets, morphology translations reduced the number of unknown tokens by more than a factor of two.

#### 4.4 The Impact of Reordering Distance Limit

The reordering window length is determined based on the movements of the source phrases. On an average, an original Arabic word has two mor-

Reorder Window		0	2	3	4	5	6	7	8	9
IWSLT04	Ori	57.21	57.92	58.01	<b>58.31</b>	58.16	58.20	58.20	58.12	58.01
	Full	56.89	57.54	<b>58.62</b>	58.39	58.32	58.55	58.55	58.55	58.57
	Reduced	58.36	59.56	60.05	<b>60.70</b>	60.32	60.28	60.46	60.30	60.55
MT03	Ori	41.75	43.84	45.24	<b>45.61</b>	45.40	45.55	45.21	45.22	45.19
	Full		41.45	43.12	44.32	44.71	45.30	45.80	<b>45.88</b>	45.82
	Reduced	44.08	45.28	46.50	47.40	47.41	47.69	47.59	47.75	<b>47.79</b>

Table 8: The impact of reordering limits on BTEC’s development set IWSLT04 and Newswire’s development set MT03. The translation scores are IBM-BLEU metric

Test Set	Ori	Reduced
IWSLT04	242	100
IWSLT05	219	97
MT03	1463	553
MT06	3734	1342

Table 7: Unknown tokens count

phemes. The full morphology translation with a 6-word reordering window has the same impact as a 3-word reordering when translating the original Arabic. To fully benefit from word reordering, the full morphology translation requires a longer reorder distance limit. However, in current phrase based translations, reordering models are not strong enough to guide long distance source-word movements. This shows an additional advantage of the nonaligned morpheme removal technique.

We carried out experiments from monotone decoding up to 9 word distance reordering limit for the two development sets IWSLT04 and MT03. The results are given in Table 8. The BTEC data set does not benefit from a larger reordering window. Using only a 2-word reordering window the score of the original Arabic translations(57.92) was comparable to the best score (58.31) obtained by using a 4-word reordering window. On the other hand, the reordering limit showed a significant impact on Newswire data. The MT03 original Arabic translation using a 4-word re-ordering window resulted in an improvement of 4 BLEU points over monotone decoding. Large Arabic corpora usually contain data from the news domain. The decoder might not effectively reorder very long distance morphemes for these data sets. This explains why machine translation does not benefit from word-based morphological segmentation for large data sets which adequately cover the vocabu-

lary of the test set.

#### 4.5 Large Training Corpora Results

We wanted to test the impact of our preprocessing technique on a system trained on 5 million sentence pairs (128 million Arabic words). Unfortunately, the CRF++ toolkit exceeded memory limits when executed even on a 24GB server. We created smaller corpora by sub-sampling the large corpus for the source side of MT03 and MT06 test sets. The sub-sampled corpus have 500K sentence pairs and cover all source phrases of MT03 and MT06 which can be found in the large corpus. In these experiments, we used a lexical reordering model into translation model. The language model was the 5-gram SRI language model built from the whole GIGA word corpus. Table 9

	MT03	MT06
5M Ori	56.22	42.17
Sub-sample Ori	54.54	41.59
Sub-sample Full	51.47	40.84
Sub-sample Reduced	54.78	<b>43.20</b>

Table 9: Translation results of large corpora(Case insensitive, IBM-BLEU metric). The boldface marks score significantly higher than the original Arabic translation score.

presents the translation result of original Arabic system trained on the full 5M sentence pairs corpus and the three systems trained on the 500K sentence pairs sub-sampled corpus. The sub-sampled full morphology system scores degraded for both development set and unseen test set. On development set, the sub-sampled reduced morphology system score was slightly better than baseline. On the unseen test set, it significantly outperformed both the baseline on sub-sampled training data and even outperformed the system trained on the entire

5M sentence pairs.

## 5 Conclusion and Future Work

In this paper, we presented a context-dependent morphology preprocessing technique for Arabic-English translation. The model significantly outperformed the original Arabic systems on small and mid-size corpora and unseen test set on large training corpora. The model treats morphology processing task as a sequence labelling problem. Therefore, other machine learning techniques such as perceptron (Collins, 2002) could also be applied for this problem.

The paper also discussed the relation between the size of the reordering window and morphology processing. In future investigations, we plan to extend the model such that merging morphemes is included. We also intent to study the impact of phrase length and phrase extraction heuristics.

## Acknowledgement

We thank Noah Smith for useful comments and suggestions and providing us with the morphology disambiguation toolkit. We also thank Sameer Badaskar for help on editing the paper. We also thank anonymous reviewers for helpful comments. The research was supported by the GALE project.

## References

- Brown, Peter F., Stephen Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.
- Buckwalter, T. 2004. Arabic Morphological Analyzer version 2.0. LDC2004L02.
- Chen, Stanley F. and Joshua Goodman. 1996. An Empirical Study of Smoothing Techniques for Language Modeling. In *Proceedings of the ACL*, pages 310–318.
- Collins, Michael. 2002. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *Proceedings of EMNLP '02*, pages 1–8.
- Eck, M. and C. Hori. 2005. Overview of the IWSLT 2005 Evaluation Campaign. In *Proceedings of IWSLT*, pages 11–17.
- Goldwater, Sharon and David McClosky. 2005. Improving Statistical MT through Morphological Analysis. In *Proceedings of HLT/EMNLP*, pages 676–683, Vancouver, British Columbia, Canada.
- Koehn, et al. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Annual Meeting of ACL, demonstration session*.
- Lafferty, John, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of 18th ICML*, pages 282–289.
- Lee, Young S. 2004. Morphological Analysis for Statistical Machine Translation. In *HLT-NAACL 2004: Short Papers*, pages 57–60, Boston, Massachusetts, USA.
- Nießen, Sonja and Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources Using Morpho-Syntactic Information. *Computational Linguistics*, 30(2), June.
- Och, Franz Josef and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.
- Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pages 311–318, Philadelphia.
- Sadat, Fatiha and Nizar Habash. 2006. Combination of Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proceedings of the ACL*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.
- Sha, Fei and Fernando Pereira. 2003. Shallow Parsing with Conditional Random Fields. In *Proceedings of NAACL '03*, pages 134–141, Morristown, NJ, USA.
- Smith, Noah A., David A. Smith, and Roy W. Tromble. 2005. Context-Based Morphological Disambiguation with Random Fields. In *Proceedings of HLT/EMNLP*, pages 475–482, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.
- Stolcke, A. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Intl. Conf. on Spoken Language Processing*.
- Takezawa, Toshiyuki, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of LREC 2002*, pages 147–152.