Morphological Analysis for Statistical Machine Translation

Young-Suk Lee
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
Email: ysuklee@us.ibm.com

Abstract

We present a novel morphological analysis technique which induces a morphological and syntactic symmetry between two languages with highly asymmetrical morphological structures to improve statistical machine translation qualities. The technique pre-supposes fine-grained segmentation of a word in the morphologically rich language into the sequence of *prefix(es)-stem-suffix(es)* and part-of-speech tagging of the parallel corpus. The algorithm identifies morphemes to be *merged* or *deleted* in the morphologically rich language to induce the desired morphological and syntactic symmetry. The technique improves Arabic-to-English translation qualities significantly when applied to IBM Model 1 and Phrase Translation Models trained on the training corpus size ranging from 3,500 to 3.3 million sentence pairs.

1. Introduction

Translation of two languages with highly different morphological structures as exemplified by Arabic and English poses a challenge to successful implementation of statistical machine translation models (Brown et al. 1993). Rarely occurring inflected forms of a stem in Arabic often do not accurately translate due to the frequency imbalance with the corresponding translation word in English. So called a word (separated by a white space) in Arabic often corresponds to more than one independent word in English, posing a technical problem to the source channel models. In the English-Arabic sentence alignment shown in Figure 1, Arabic word *AlAHmr* (written in Buckwalter transliteration) is aligned to two English words 'the red', and *llmEArDp* to three English words 'of the opposition.' In this paper, we present a technique to induce a morphological and syntactic symmetry between two languages with different morphological structures for statistical translation quality improvement.

The technique is implemented as a two-step morphological processing for word-based translation models. We first apply word segmentation to Arabic, segmenting a word into *prefix(es)-stem-suffix(es)*. Arabic-English sentence alignment after Arabic word segmentation is illustrated in Figure 2, where one Arabic morpheme is aligned to one or zero English word. We then apply the proposed technique to the word segmented Arabic corpus to identify prefixes/suffixes to be *merged* into their stems or *deleted* to induce a symmetrical morphological structure. Arabic-English sentence alignment after Arabic morphological analysis is shown in Figure 3, where the suffix *p* is merged into their stems *mwAjh* and *mEArd*. For phrase translation models, we apply additional morphological analysis induced from noun phrase parsing of Arabic to accomplish a syntactic as well as morphological symmetry between the two languages.

2. Word Segmentation

We pre-suppose segmentation of a word into *prefix(es)-stem-suffix(es)*, as described in (Lee et al. 2003) The category prefix and suffix encompasses function words such as conjunction markers, prepositions, pronouns, determiners and all inflectional morphemes of the language. If a word token contains more than one prefix and/or suffix, we posit multiple prefixes/suffixes per stem. A sample word segmented Arabic text is given below, where prefixes are marked with #, and suffixes with +.

w# s# y# Hl sA}q Al# tjArb fy jAgwAr Al# brAzyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sbAq gdA Al# AHd Al*y **s# y#** kwn Awly xTw **+At +h** fy EAlm sbAq +At AlfwrmwlA

3. Morphological Analysis

Morphological analysis identifies functional morphemes to be merged into meaning-bearing stems or to be deleted. In Arabic, functional morphemes typically belong to prefixes or suffixes.

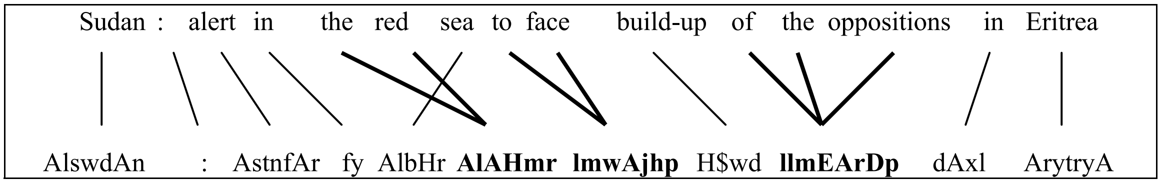

Figure 1. Word alignment between Arabic and English without Arabic morphological processing

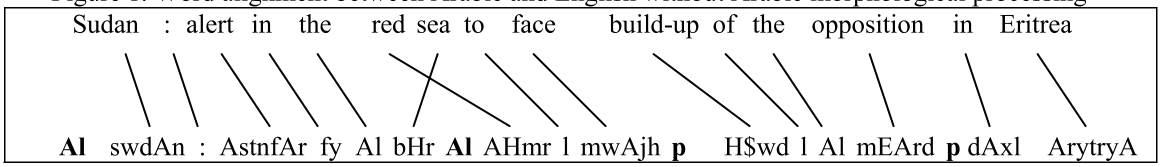

Figure 2. Alignment between word-segmented Arabic and English

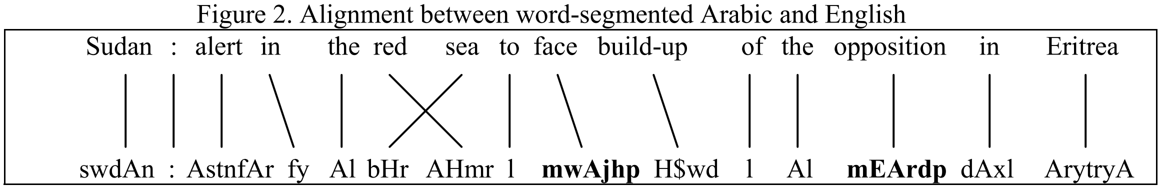

Figure 3. Alignment between morphologically analyzed Arabic and English

Sample Arabic texts before and after morphological analysis is shown below.

Mwskw 51-7 (Af b) - Elm An Al# qSf Al# mdfEy Al*y Ady Aly **ASAb** **+p** **jndy** **+yn** **rwsy** **+yn** **Avn** **+yn** b# jrwH **Tfyf** **+p** q*A}f Al# **jmE** **+p** fy mTAr xAn **qlE** **+p** …
Mwskw 51-7 (Af b) - Elm An Al# qSf Al# mdfEy Al*y Ady Aly **ASAbp** **jndyyn** **rwsyyn** **Avnyn** b# jrwH **Tfyfp** msA' Al# **jmEp** fy mTAr xAn **qlEp** …

In the morphologically analyzed Arabic (bottom), the feminine singular suffix *+p* and the masculine plural suffix *+yn* are merged into the preceding stems analogous to singular/plural noun distinction in English, e.g. *girl* vs. *girls*.

3.1 Method

We apply part-speech tagging to a symbol tokenized and word segmented Arabic and symbol-tokenized English parallel corpus. We then viterbi-align the part-of-speech tagged parallel corpus, using translation parameters obtained via Model 1 training of word segmented Arabic and symbol-tokenized English, to derive the conditional probability of an English part-of-speech tag given the combination of an Arabic prefix and its part-of-speech or an Arabic suffix and its part-of-speech.[1]

[1] We have used an Arabic part-of-speech tagger with around 120 tags, and an English part-of-speech tagger with around 55 tags.

3.2 Algorithm

The algorithm utilizes two sets of translation probabilities to determine merge/deletion analysis of a morpheme. We obtain tag-to-tag translation probabilities according to (1), which identifies the most probable part-of-speech correspondences between Arabic (tag_A) and English (tag_E).

(1) $\Pr(tag_E \mid tag_A)$

We also obtain translation probabilities of an English part-of-speech tag given each Arabic prefix/suffix and its part-of-speech according to (2) and (3):

(2) $\Pr(tag_E \mid stemtag_A, suffix_j\_tag_{jk})$

(2) computes the translation probability of an Arabic suffix and its part-of-speech into an English part-of-speech in the Arabic stem tag context, $stemtag_A$. $Stemtag_A$ is one of the major stem parts-of-speech with which the specified prefix or suffix co-occurs, i.e. ADV, ADJ, NOUN, NOUN_PROP, VERB_IMPERFECT, VERB_PERFECT.[2] J in $suffix_j$ ranges from 1 to M, M = number of distinct suffixes co-occurring with $stemtag_A$. tag_{jk} in $suffix_j\_tag_{jk}$ is the part-of-speech of $suffix_j$, where k ranges from 1 to L, L = number of

[2] All Arabic part-of-speech tags are adopted from LDC-distributed Arabic Treebank and English tags are adopted from Penn Treebank.

distinct tags assigned to the *suffix*$_j$ in the training corpus.

(3) $\Pr(tag_E \mid prefix_i\_tag_{ik}, stemtag_A)$

(3) computes the translation probability of an Arabic prefix and its part-of-speech into an English part-of-speech in the Arabic stem tag context, $stemtag_A$. $Prefix_i$ and tag_{ik} in $prefix_i\_tag_{ik}$ may be interpreted in a manner analogous to $suffix_j$ and tag_{jk} of $suffix_j\_tag_{jk}$ in (2).

3.2.1 IBM Model 1

The algorithm for word-based translation model, e.g. IBM Model 1, implements the idea that if a morpheme in one language is robustly translated into a distinct part-of-speech in the other language, the morpheme is very likely to have its independent counterpart in the other language. Therefore, a robust overlap of tag_E given tag_A between $\Pr(tag_E|tag_A)$ and $\Pr(tag_E|stemtag_A, suffix_j\_tag_{jk})$ for a suffix and $\Pr(tag_E|tag_A)$ and $\Pr(tag_E|prefix_i\_tag_{ik}, stemtag_A)$ for a prefix is a positive indicator that the Arabic prefix/suffix has an independent counterpart in English. If the overlap is weak or doesn't exist, the prefix/suffix is unlikely to have an independent counterpart and is subject to merge/deletion analysis.[3]

Step 1: For each tag_A, select the top 3 most probable tag_E from $\Pr(tag_E|tag_A)$.
Step 2: Partition all $prefix_i\_tag_{ik}$ and $suffix_j\_tag_{jk}$ into two groups in each $stemtag_A$ context.
Group I: At least one of '$tag_E|tag_{ik}$' or '$tag_E|tag_{jk}$' occurs as one of the top 3 most probable translation pairs in $Pr(tag_E|tag_A)$. Prefixes and suffixes in this group are likely to have their independent counterparts in English.
Group II: None of '$tag_E|tag_{ik}$' or '$tag_E|tag_{jk}$' occurs as one of the top 3 most probable translation pairs in $\Pr(tag_E|tag_A)$. Prefixes and suffixes in this group are unlikely to have their independent counterparts in English.
Step 3: Determine the merge/deletion analysis of the prefixes/suffixes in Group II as follows: If $\text{prefix}_i\_\text{tag}_{ik}/\text{suffix}_j\_\text{tag}_{jk}$ occurs in more than one stemtag_A context, and its translation probability into NULL tag is the highest, ***delete*** the $\text{prefix}_i\_\text{tag}_{ik}/\text{suffix}_j\_\text{tag}_{jk}$ in the $stemtag_A$ context. If $\text{prefix}_i\_\text{tag}_{ik}/\text{suffix}_j\_\text{tag}_{jk}$ occurs in more than one stemtag_A context, and its translation probability into NULL tag is not the highest, ***merge*** the $\text{prefix}_i\_\text{tag}_{ik}/\text{suffix}_j\_\text{tag}_{jk}$ into its stem in the stemtag_A context.

Merge/deletion analysis is applied to all $\text{prefix}_i\_\text{tag}_{ik}/\text{suffix}_j\_\text{tag}_{jk}$ occurring in the appropriate stem tag contexts in the training corpus (for translation model training) and a new input text (for decoding).

3.2.2 Phrase Translation Model

For phrase translation models (Och and Ney 2002), we induce additional merge/deletion analysis on the basis of base noun phrase parsing of Arabic. One major asymmetry between Arabic and English is caused by more frequent use of the determiner *Al#* in Arabic compared with its counterpart *the* in English. We apply *Al#*-deletion to Arabic noun phrases so that only the first occurrence of *Al#* in a noun phrase is retained. All instances of *Al#* occurring before a proper noun – as in *Al# qds*, whose literal translation is *the Jerusalem* – are also deleted. Unlike the automatic induction of morphological analysis described in 3.2.1, *Al#*-deletion analysis is manually induced.

4. Performance Evaluations

System performances are evaluated on LDC-distributed Multiple Translation Arabic Part I consisting of 1,043 segments derived from AFP and Xinhua newswires. Translation qualities are measured by uncased BLEU (Papineni et al. 2002) with 4 reference translations, *sysids: ahb, ahc, ahd, ahe*.

Systems are developed from 4 different sizes of training corpora, 3.5K, 35K, 350K and 3.3M sentence pairs, as in Table 1. The number in each cell indicates the number of sentence pairs in each genre (newswires, ummah, UN corpus).[4]

Genre	3.5K	35K	350K	3.3M
News	1,000	1,000	9,238	12,002
Ummah	500	1,000	13,027	13,027
UN	2,000	33,000	327,735	3,270,200

Table 1. Training Corpora Specifications

4.1 IBM Model 1

Impact of morphological analysis on IBM Model 1 is shown in Table 2.

[3] We assume that only one tag is assigned to one morpheme or word, i.e. no combination tag of the form DET+NOUN, etc.

[4] We have used the same language model for all evaluations.

corpus size	baseline	morph analysis
3.5K	0.10	0.25
35K	0.14	0.29
350K	0.18	0.31
3.3M	0.18	0.32

Table 2. Impact of morphological analysis on IBM Model 1

Baseline performances are obtained by Model 1 training and decoding without any segmentation or morphological analysis on Arabic. BLEU scores under 'morph analysis' is obtained by Model 1 training on Arabic morphologically analyzed and English symbol-tokenized parallel corpus and Model 1 decoding on the Arabic morphologically analyzed input text.[5]

4.2 Phrase Translation Model

Impact of Arabic morphological analysis on a phrase translation model with monotone decoding (Tillmann 2003), is shown in Table 3.

corpus size	baseline	morph analysis
3.5K	0.17	0.24
35K	0.24	0.29
350K	0.32	0.36
3.3M	0.36	0.39

Table 3. Impact of morphological analysis on Phrase Translation Model

BLEU scores under baseline and morph analysis are obtained in a manner analogous to Model 1 except that the morphological analysis for the phrase translation model is a combination of the automatically induced analysis for Model 1 plus the manually induced *Al#*-deletion in 3.2.2. The scores with only automatically induced morphological analysis are 0.21, 0.25, 0.33 and 0.36 for 3.5K, 35K, 350K and 3.3M sentence pair training corpora, respectively.

5. Related Work

Automatic induction of the desired linguistic knowledge from a word/morpheme-aligned parallel corpus is analogous to (Yarowsky et al. 2001). Word segmentation and merge/deletion analysis in morphology is similar to parsing and insertion operation in syntax by (Yamada and Knight 2001). Symmetrization of linguistic structures can also be found in (Niessen and Ney 2000).

[5] Our experiments indicate that addition of *Al#*-deletion, cf. Phrase Translation Model, does not affect the performance of IBM Model 1.

Acknowledgements

This work was partially supported by the Defense Advanced Research Projects Agency and monitored by SPAWAR under contract No. N66001-99-2-8916. The views and findings contained in this material are those of the authors and do not necessarily reflect the position of policy of the Government and no official endorsement should be inferred. We would like to acknowledge Salim Roukos and Kishore Papineni for technical discussions.

6. References

Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. 1993. The mathematics of statistical machine translation: Parameter Estimation. *Computational Linguistics, 19(2)*: 263—311.

Lee, Y-S., Papineni, K., Roukos, S., Emam, O., Hassan, H. 2003. Language Model Based Arabic Word Segmentation. In *Proceedings of the 41st Annual Meeting of the ACL*. Pages 399—406. Sapporo, Japan.

Niessen, S., Ney, H. 2000. Improving SMT quality with morpho-syntactic analysis. In *Proceedings of 20th International Conference on Computational Linguistics*. Saarbrucken, Germany.

Och, F. J., Ney. H. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the ACL*. Pages 295—302. Philadelphia. PA.

Papineni, K., Roukos, S., Ward, R., Zhu W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting of the ACL*. Pages 311—318. Philadelphia, PA.

Tillmann, Christoph 2003. A Projection Extension Algorithm for Statistical Machine Translation. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Pages 1—8. Sapporo, Japan.

Yamada, K. and Knight, K. 2001. A Syntax-Based Statistical Translation Model. In *Proceedings of the 39th Conference of the ACL.* Pages 523—530. Toulouse, France.

Yarowsky, D., G. Ngai and R. Wicentowski 2001. Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of HLT 2001* (ISBN: 1-55860-786-2).