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 *lmEArd* 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.

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 # s # y # t Hl sA\q Al# sjArb fy jAgwAr Al# bRazyly lwsyAnw bwrty mkAn AyrfAyn fy Al# sb=Aq gdA Al# AhD Al# s # y # s # y # kwn Awly xTw +At +h fy EAIm sb=Aq +At AlwrmwlA
```

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.
# Morphological Analysis for Statistical Machine Translation

**Report Date:**
2004

**Report Type:**

**Dates Covered:**
00-00-2004 to 00-00-2004

**Title and Subtitle:**
Morphological Analysis for Statistical Machine Translation

**Performing Organization Name(s) and Address(es):**
IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY, 10598

**Publication Date:**

**Sponsoring/Monitoring Agency Name(s) and Address(es):**

**Availability Statement:**
Approved for public release; distribution unlimited

**Supplementary Notes:**
The original document contains color images.

**Security Classification:**
- Report: Unclassified
- Abstract: Unclassified
- This Page: Unclassified

**Limitation of Abstract:**

**Number of Pages:**
5
Sudan: alert in the red sea to face build-up of the opposition in Eritrea

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

Sudan: alert in the red sea to face build-up of the opposition in Eritrea

Figure 2. Alignment between word-segmented Arabic and English

Sudan: alert in the red sea to face build-up of the opposition in Eritrea

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 qIE +p ...

Mwskw 51-7 (Af b) - Elm An Al# qSf Al# mdfEy Al*y Ady Aly ASAb jndyyn rwssyn Avyn b# jrwH Tfyf p msA' Al# jmE p fy mTAr xAn qIE p ...

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

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 (tagA) and English (tagE).

(1) \Pr(tagE | tagA)

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(tagE | stemtagA, suffixj_tagjk)

(2) computes the translation probability of an Arabic suffix and its part-of-speech into an English part-of-speech tag in the Arabic stem tag context, stemtagA. StemtagA 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 suffixj ranges from 1 to M, M = number of distinct suffixes co-occurring with stemtagA, tagjk in suffixj_tagjk is the part-of-speech of suffixj, where k ranges from 1 to L, L = number of

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.

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 in the training corpus.

(3) \( \Pr(\text{tag}_E | \text{prefix}_{i_{tagik}} \text{, stem}_{tagA}) \)

(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, \( \text{stem}_{tagA} \). Prefix in \( \text{prefix}_{i_{tagik}} \text{, stem}_{tagA} \) may be interpreted in a manner analogous to suffix 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 \( \text{tag}_E \) given \( \text{tag}_A \) between \( \Pr(\text{tag}_E | \text{tag}_A) \) and \( \Pr(\text{tag}_E | \text{stem}_{tagA}, \text{suffix}_{j_{tagjk}}) \) for a suffix and \( \Pr(\text{tag}_E | \text{tag}_A) \) and \( \Pr(\text{tag}_E | \text{prefix}_{i_{tagik}} \text{, stem}_{tagA}) \) 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 \( \text{tag}_A \), select the top 3 most probable \( \text{tag}_E \) from \( \Pr(\text{tag}_E | \text{tag}_A) \).

**Step 2:** Partition all \( \text{prefix}_{i_{tagik}} \text{, stem}_{tagA} \) and \( \text{suffix}_{j_{tagjk}} \) into two groups in each \( \text{stem}_{tagA} \) context.

**Group I:** At least one of \( \text{tag}_{i_{tagik}} | \text{tag}_{j_{tagjk}} \) or \( \text{tag}_{i_{tagik}} | \text{tag}_{j_{tagjk}} \) occurs as one of the top 3 most probable translation pairs in \( \Pr(\text{tag}_E | \text{tag}_A) \). Prefixes and suffixes in this group are likely to have their independent counterparts in English.

**Group II:** None of \( \text{tag}_{i_{tagik}} | \text{tag}_{j_{tagjk}} \) or \( \text{tag}_{i_{tagik}} | \text{tag}_{j_{tagjk}} \) occurs as one of the top 3 most probable translation pairs in \( \Pr(\text{tag}_E | \text{tag}_A) \). 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 \( \text{prefix}_{i_{tagik}} / \text{suffix}_{j_{tagjk}} \) occurs in more than one \( \text{stem}_{tagA} \) context, and its translation probability into NULL tag is not the highest, **merge** the \( \text{prefix}_{i_{tagik}} / \text{suffix}_{j_{tagjk}} \) into its stem in the \( \text{stem}_{tagA} \) context.

Merge/deletion analysis is applied to all \( \text{prefix}_{i_{tagik}} / \text{suffix}_{j_{tagjk}} \) 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 \( \text{Al#} \) in Arabic compared with its counterpart the in English. We apply \( \text{Al#}- \) deletion to Arabic noun phrases so that only the first occurrence of \( \text{Al#} \) in a noun phrase is retained. All instances of \( \text{Al#} \) occurring before a proper noun – as in \( \text{Al#} \) qds, whose literal translation is the Jerusalem – are also deleted. Unlike the automatic induction of morphological analysis described in 3.2.1, \( \text{Al#}- \) 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

<table>
<thead>
<tr>
<th>Genre</th>
<th>3.5K</th>
<th>35K</th>
<th>350K</th>
<th>3.3M</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>1,000</td>
<td>1,000</td>
<td>9,238</td>
<td>12,002</td>
</tr>
<tr>
<td>Ummah</td>
<td>500</td>
<td>1,000</td>
<td>13,027</td>
<td>13,027</td>
</tr>
<tr>
<td>UN</td>
<td>2,000</td>
<td>33,000</td>
<td>327,735</td>
<td>3,270,200</td>
</tr>
</tbody>
</table>

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.
Table 2. Impact of morphological analysis on IBM Model 1

<table>
<thead>
<tr>
<th>corpus size</th>
<th>baseline</th>
<th>morph analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5K</td>
<td>0.10</td>
<td>0.25</td>
</tr>
<tr>
<td>35K</td>
<td>0.14</td>
<td>0.29</td>
</tr>
<tr>
<td>350K</td>
<td>0.18</td>
<td>0.31</td>
</tr>
<tr>
<td>3.3M</td>
<td>0.18</td>
<td>0.32</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>corpus size</th>
<th>baseline</th>
<th>morph analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5K</td>
<td>0.17</td>
<td>0.24</td>
</tr>
<tr>
<td>35K</td>
<td>0.24</td>
<td>0.29</td>
</tr>
<tr>
<td>350K</td>
<td>0.32</td>
<td>0.36</td>
</tr>
<tr>
<td>3.3M</td>
<td>0.36</td>
<td>0.39</td>
</tr>
</tbody>
</table>

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 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).

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


\(^5\) Our experiments indicate that addition of Al#-deletion, cf. Phrase Translation Model, does not affect the performance of IBM Model 1.