Word Alignment for English-Turkish Language Pair

M. Talha Çakmak, Süleyman Acar, Günsen Eryiğit
Department of Computer Engineering, Istanbul Technical University
Istanbul, 34469, Turkey
E-mail: cakmakmeh@itu.edu.tr, acarsu@itu.edu.tr, gulsenc@itu.edu.tr

Abstract

Word alignment is an important step for machine translation systems. Although the alignment performance between grammatically similar languages is reported to be very high in many studies, the case is not the same for language pairs from different language families. In this study, we are focusing on English-Turkish language pairs. Turkish is a highly agglutinative language with a very productive and rich morphology whereas English has a very poor morphology when compared to this language. As a result of this, one Turkish word is usually aligned with several English words. The traditional models which use word-level alignment approaches generally fail in such circumstances. In this study, we evaluate a Giza++ system by splitting the words into their morphological units (stem and suffixes) and compare the model with the traditional one. For the first time, we evaluate the performance of our aligner on gold standard parallel sentences rather than in a real machine translation system. Our approach reduced the alignment error rate by 40% relative. Finally, a new test corpus of 300 manually aligned sentences is released together with this study.

Keywords: Word Alignment, Machine Translation, Turkish

1. Introduction

Word alignment is a crucial phase for statistical machine translation (MT). The aim of this process is to align the words in parallel sentences from two different languages. The impact of alignment performance on translation quality is inevitable but not understood completely. Fraser and Marcu (2007) investigate the relationship between alignment and translation performances on different language pairs: French-English, English-Arabic, Romanian-English. The alignment between typologically different languages may become very complicated due to different word orders and morphological structures. For example, Turkish word order generally obeys SOV (Subject-Object-Verb) pattern in written text whereas English word order is most of the time SVO. In addition to this, Turkish rich and agglutinative morphology creates highly inflected and derived word forms which are sometimes equivalent to a whole sentence in English.

Figure 1.A and Figure 1.B shows aligned sentences from English-French and English-Turkish language pairs. For English-Turkish, the alignment is shown both on word level (Figure 1.B) and on morphological units level (Figure 1.C). A morphological unit is either the stem of a word or some suffix affixed after a stem. The alignment complexity between typologically different languages is far away from the alignment complexity between grammatically similar languages. This observation may also be validated by looking at the results in the literature: Liang et al. (2006) reports the alignment error rate (AER) for English-French language pair as 4.9 whereas the AER for Chinese-English language pair which are topologically very different is measured as 43.4 (Deng and Byrne 2005). The result is drastically low when compared to the first one. The results for Czech-English (14.7) (Bojar and Prokopová 2006) and Inuktitut-English (31.2) (Gotti, et al. 2005) are again very low when compared to the results for typologically similar languages.

In this study, we explore the usage of morphological units rather than words in the training stage of the statistical word alignment. El-Khalout and Ofazer (2010) report that the BLEU (Papineni, et al. 2002) score of their MT system slightly worsens by using this strategy. In this study, we want to see if the reason of this degradation is really the drop in the alignment quality or not. To the best of our knowledge, this is the first study which makes an intrinsic evaluation of the alignment performance for Turkish. With this purpose, we created a gold-standard alignment test corpus of 300 manually aligned sentences. We have two motivations for our study: 1st. Similar approaches give better results for different nlp layers for Turkish: Hakkani-Tür et al. (2002) show the impact on an HMM morphological disambiguator and Eryiğit, et al. (2008) show the improvement on a statistical dependency parser. 2nd. The studies which measure the alignment performance for other agglutinative or inflectional languages report performance improvement by using smaller units than words: Bojar and Prokopová (2006) reports an AER drop from 27.1 to 14.7 for Czech-Eng. and Gotti, et al. (2005) from 45.5 to 32.1 for Inuktut-Eng.. Singh and Bandyopadhyay (2010) reports a BLEU improvement by 2.3 for Manipuri-Eng.

The paper is structured as follows: Section 2 compares the two languages in short, Section 3 gives the details of the data sets used in the experiments and Section 4 the system configuration. The different training models and evaluation criteria are given in Section 5. Section 6 gives and discusses the results and Section 7 gives the

1 On the other hand, they report an increase with a more complicated strategy: a selective morphological segmentation where some of the consistently unaligned units on the Giza++ output are combined with the stems rather than acting as separate units.
2. The Turkish-English Language Pair

Turkish is an agglutinative language and has a very rich derivational and inflectional morphology which results to an infinite number of word forms in the language. The affixes are affixed to the end of the word one after another and the part-of-speech of a word may change several times in a single surface form due to its rich derivational structure. Almost all of the function words are represented as suffixes in Turkish. As a result, a Turkish word may sometimes be equivalent to a very long English phrase. The following example is given in order to reflect the idea: The English phrase “from my homeworks” will be translated into Turkish as the word “ödevlerimden”. In this example, the Turkish stem “ödev” means “homework”, “+ler” is the plural suffix, “+im” is the 1st singular possessive suffix which means “my” on the English side and “+den” is the ablative suffix which is matched with the English function word “from”.

In the language processing applications of Turkish, the usage of the morphological units is not only necessary for improving the systems’ performances but also required for representing the syntax and semantic relations within a sentence (Eryiğit, et al. 2008).

Turkish is a free-constituent order language which most of the time obeys the SOV word order in written text. On the other side, English has a very poor morphology when compared to Turkish and its SVO word order which differs from Turkish makes the alignment more complicated (Figure 1.B&C) between these two.

3. Data

We are using a 130K parallel corpus (Tyers and Alperen 2010) for the training. The original size of the corpus was 166K. We eliminated some of the sentences which were:

a) too long for the training of our statistical aligner
b) labels and titles
c) faulty translations (both the source and target sentences are in English)
d) false translations (the translation has totally a different meaning)

In addition to this, we split the translations consisting of multiple sentences into multiple lines. At the end of this processing, we prepared 3 different size training sets (small, medium, large) of sizes 25K, 70K and 130K for using in our experiments.

For testing our models’ performances, we are using a gold standard test data which consists of 300 manually aligned English-Turkish sentences.

We collected this data from old language proficiency exams organized by ÖSYM (Student Selection and Placement Center of Turkey). These exams include questions which consist of high quality exact translations for English-Turkish. We aligned the data manually by using a manual annotation software that we have developed. The software is designed so that the user annotates the parallel sentences by aligning the morphological units.

4. System Configuration

In order to split the languages into their morphological units we are using some automatic processors:

a) A morphological analyzer and disambiguator for

2 The constituents may easily change their position in the sentence with/without changing the meaning of the sentence. The closer the constituent is to the verb, the higher is the emphasis on that constituent.

3 Maximum 101 tokens in a sentence.

4 The gold standard test data is available via internet in http://web.itu.edu.tr/gulsenc/resources.htm.
Turkish (Sak, et al. 2008) which firstly lists the possible morphological analyses for a given Turkish word and then select the most appropriate one within the current context. The below example shows the outputs of the morphological analyzer for the word “kalemi” (“pencil” or “my castle” or “my goal” according to context.) The ambiguity in the output is resolved by the morphological disambiguator in the second stage.

input: kalem
output: 1) kale[Noun]+[A3sg]+Hm[P1sg]+[Nom]
2) kalem[Noun]+[A3sg]+[Pnon]+[Nom]

The morphological analyzer also produces some tags which are not directly related to a suffix. Such as [Noun], [A3sg], [Nom] which are the parts-of-speech of the stem, the singular markup and the nominative case marker sequentially. We only take the stem and the suffixes from this analysis and discard the remaining tags. The first analysis above will be represented in the alignment as two morphological units: ‘kale’ and ‘+Hm’.

b) TreeTagger for English (Schmid 1994) which is a tool for annotating text with part-of-speech and lemma information. We again discarded the produced tags which are not related with any suffix.

input: pencils
output: pencil+NNS

For the statistical alignment, we are using Giza++ (Och and Ney 2000) which is one of the most frequently used aligners in the state of the art SMT systems. Giza++ is trained with default system configuration and the models which will be defined in the next section.

5 A3sg: 3rd singular, P1sg: 1st person possessive, Pnon: null possessive, Nom: nominative case markers
6 +Hm is the phonological representation of the suffix +m. This suffix may take the following different forms under vowel harmony: +im, +im, +um, +im, +m.
7 NNS: noun plural
8 This model is exactly the same with representation 2 of El-Kahlout and Oflazer (2010)

Results and Discussions

<table>
<thead>
<tr>
<th></th>
<th>Word-based AER</th>
<th>Unit-based AER</th>
</tr>
</thead>
<tbody>
<tr>
<td>TM1</td>
<td>Small 0.563</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Medium 0.522</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Large 0.500</td>
<td></td>
</tr>
<tr>
<td>TM2</td>
<td>Small 0.340</td>
<td>0.405</td>
</tr>
<tr>
<td></td>
<td>Medium 0.314</td>
<td>0.380</td>
</tr>
<tr>
<td></td>
<td>Large 0.298</td>
<td>0.365</td>
</tr>
</tbody>
</table>

Table 1: Alignment Error Rates for English-Turkish

As expected, we see from Table 1 that the increase on the training set size has positive effect on the performance. But the improvement between our medium data set and large data set is still substantial which gives the signals that more training data will still be valuable for a performance increase.

Our TM2 model outperforms the TM1 model drastically on all of the training set sizes; the word-based AER for large data set drops from 0.5 to 0.298. The unit-based AER is a much more strict evaluation than the word-based AER since although the words are aligned correctly if the sub-word parts are not so, the score will be lower. We may see that the word-based score for the large-data set of TM2 is 0.298 whereas the unit-based score is only 0.365 for the same experiment. But our TM2 model still outperforms TM1 when we compare its unit-based AER (0.365 for large set) with the word-based AER of TM1 (0.5 for large set).

We showed that using morphological units in the training stage increases the alignment performance significantly. This result conflicts with the previous results which give an extrinsic evaluation of a similar model: El-Kahlout and Oflazer (2010) show that the BLEU score of their SMT system decrease by using this approach. This shows that
we can’t explain the success of a machine translation system by just the improvement or decrease in the alignment performance. There are certainly other hidden factors that affect the results of a real application. Although we need to make more research to draw the picture clearly, our first impression is that, since the usage of morphological units increase the sentence length (token size), the SMT system may have difficulties in combining morphological units and constructing valid Turkish words which may be ameliorated by including more syntax information in the process. Another reason that comes in mind may be the deficiency of BLEU score measuring the translation quality (Tantuğ, et al. 2008) of agglutinative languages.

We have stated in the introduction part that similar conflicting results are reported in the literature and the direct relation between the alignment performance and translation quality is not understood clearly. Fraser and Marcu (2007) explain the relation between alignment and translation for large French-English dataset partially but they could not generate a general result.

7. Conclusion

In this work, we investigated the alignment between English-Turkish languages which have severe differences in their morphology and syntax. We showed that the usage of morphological units in the training stage of the alignment process improves the alignment performance. By making a literature survey in the field, we conclude that there is no direct relation between the alignment quality and translation scores based on the results of previous studies. As a future work, we plan to investigate the effect of more complex morphological representations and word reordering in the alignment performance.

8. References


