Solution for Word Reordering English-Turkish SMT

Emrah UĞURLU, Hatem HADDAD and Mohammad Shukri SALMAN
Mevlana University, Konya/Turkey
eren_ugurlu@outlook.com, hhatem@mevlana.edu.tr, mssalman@mevlana.edu.tr

ABSTRACT

The process of automatically translating sentences by examining a number of human-produced translation samples is called Statistical Machine Translation. To help with word alignment of statistically translated sentences between languages having significantly different word orders, word reordering is required. In this paper, we outline the characteristics of the Turkish language and its challenges for SMT. We propose to reorder translated Turkish sentences based on linguistic knowledge. For this issue, we propose using morphological analysis and syntactic rules.

KEYWORDS


1 INTRODUCTION

Machine Translation (MT) is automatically translating from a source language into a target language by utilizing computers [1]. An approach to MT, which is characterized by the use of machine learning methods, is called Statistical Machine Translation (SMT) [2]. The idea of automatic translation dates back to the invention the modern digital computer. It was Warren Weaver, who in 1949 suggested that the problem can be addressed by using statistical methods and ideas of information theory. SMT algorithms automatically learn how to translate, by examining many samples of human-produced translations [3]. SMT has gained much popularity due to the automation involved in both developing and maintaining translation systems. In addition to being faster and more cost efficient than older translation technologies, they can also be improved easily. However, despite all their great advantages, SMT systems have some limitations. One of these limitations is that SMT may not provide good results between languages having significantly different word orders, such as English-Turkish pair [4]. Main reason for this phenomena is the misalignment at the end of the alignment process [4]. Word alignment, which is aligning the words in parallel sentences from two different languages, is a crucial phase for SMT. Figure-1 represents an example of a Turkish and an English parallel sentences. It worths to emphasize that the number of words in the two sentences are different and words with equivalent meanings do not appear in the same order in the two sentences. For instance, the word “onların” occurring in the position number 5 has an equivalent “them” in the position number 8.

Figure 1. Word alignments between Turkish-English. Reordering is mostly affected by syntactic structure of the target language. For instance, the typical sentence order in English sentences is subject-verb-object (SVO), whereas in Turkish it is very flexible. For this reason, the alignment between typologically different languages may become complicated due to different word orders and morphological structures [5]. After translations, words may not be in correct places, reducing the translation performance (BLEU Scores).
This paper is organized as follows: in Section 2, we shortly review Turkish language characteristics. In Section 3, we describe the challenges of Turkish SMT. Related work and discussion are presented in Section 4, and our proposed approach for word reordering English-Turkish SMT results presented in Section 5. We conclude our study in Section 6 with an outlook to future work.

2 TURKISH LANGUAGE CHARACTERISTICS

Turkish, which is an Ural-Altaic language, has agglutinative word structures with productive inflectional and derivational suffixes linked to a root morpheme or to other morphemes [5]. For example, the word şampiyonlardan, has the following morphological structure:

\[ \text{şampiyonlardan} \rightarrow \text{şampiyon} + \text{ı} + \text{dan} \]

\[ \begin{array}{c}
1 \rightarrow \text{From} \\
2 \rightarrow \text{champion} \\
3 \rightarrow \text{S} \\
\end{array} \]

Figure 2. Morphological analysis of “şampiyonlardan”.

On the other hand, a Turkish word can change its meaning and its grammatical tag by taking inflectional and derivational suffixes. The following words created by adding suffixes to the root word, <break> “kır”, is as follows:

- <break> kır → verb
  - <be broken> kır+ıl → passive verb
  - <fragile> kır+ıl+gan → adjective
  - <fugility> kır+ıl+gan+lık → noun
  - <fragilities> kır+ıl+gan+lık+lar → plural noun

Along with its agglutinative structure, more than one word can be combined to form a single noun, with a totally different meaning. For example, hanım + eli → hanimeli:

- <lady> hanım + <hand> el + <accusative> -i hanimeli< it is a flower>

Also, more than one word can form a noun phrase. For example:


In addition to the above, almost all morphemes have systematic variations depending on vowels and boundary consonants. For instance, the English morpheme <from> corresponds to the following four Turkish morphemes: -den, -dan, -ten and -tan; as shown below:

- <from the picture> Resimden → vowel harmony
- <from the package> Paketten → consonant harmony
- <from the car> Arabadan → vowel harmony
- <from the street> Sokaktan → consonant and vowel harmony

The dominant constituent order in Turkish is Subject–Object–Verb. However all six constituent orders are possible with six different emphasis.

- Ahmet Ayşe’yi gördü . (Ahmet saw Ayşe.)
- Ayşe’yi Ahmet gördü . (It was Ahmet who saw Ayşe.)
- Gördü Ahmet Ayşe’yı. (Ahmet saw Ayşe (but was not really supposed to see her.))
- Gördü Ayşe’yi Ahmet. (Ahmet saw Ayşe (and I was expecting that.))
- Ahmet gördü Ayşe’yı. (It was Ahmet who saw Ayşe (but someone else could also have seen her.))
- Ayşe’yi gördü Ahmet. (Ahmet saw Ayşe (but he could have seen someone else.))

All the Turkish features presented in this section can affect the performance of Turkish SMT.

3 CHALLENGES OF TURKISH SMT

As mentioned previously, the alignment between typologically different languages may become very complicated due to different constituent orders and morphological structures [5]. English and Turkish can be considered rather distant languages because, English has very limited morphology and fundamentally
obeys to fixed SVO constituent order, whereas Turkish has a very flexible (but SOV dominant) constituent order. What's more, Turkish is an agglutinative language with not only a very rich but also a productive derivational and inflectional morphology.

On the English side, a Turkish word has to align with a complete phrase, and sometimes these phrases on the English side could be discontinuous [4, 8]. Figure 3 shows the translation of 1 word in Turkish to 13 words in English. It can be seen that, the alignments are not straightforward.

In Table-1, we compare a SMT output versus human translation. Even though, SMT facilitates the translation by automation, words are not in correct places after translation. This reduces the translation performance.

The original source English sentence is the first row. The second row is its human translation into Turkish. The third row is the SMT output of the first row. It can be seen that, some words are in correct places (“onun”, “programı”); whereas some words are not (“az”, “dört”, “yüzde”). Also, the English sentence contains some phrases, such as, “her program”, “Less than four percent”. The SMT output of the first phrase will be already aligned. However, the latter one needs special attention. Because four tokens will be translated into 3 tokens with different alignments, as shown below:

<table>
<thead>
<tr>
<th>SMT Output</th>
<th>EN: Less than four percent of the folks that went through her program actually go back to jail.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUR: Onun programından geçen kimselerin yüzde dördünden azı aslında hapishaneye geri döner.</td>
<td></td>
</tr>
</tbody>
</table>

A SMT needs to overcome some challenges:
- Finding the correct translation depends on syntax information. However, statistical systems are based only on statistics. Therefore, translating with correct meaning is a challenge. For example, “millet” in third row should be replaced with “kimse” as it is in second row.
- Since a Turkish word can change its meaning and grammatical tag, depending on its suffixes correctly, to analyze the root and the morphemes is crucial for SMT. Therefore, a morphological analysis is required.
- Turkish grammar brings up some challenges because of allomorphs, voicing and other grammatical operators. Morphological analysis can produce false results, because of vowel and consonant harmony. For example, the root “hapis” takes a suffix “-e” and becomes “hapse”.
  <jail> “hapis” + <to> -e → <to jail> “hapse”
- Constituent order differences between source and target languages makes it harder to align. For example, the constituent order is SVO in English:
  Subject: “Less than four percent of the folks that went through her program”
  Verb: “go”
  Object: “back to jail”.
However, the constituent order for Turkish is SOV:
  Subject: “onun programından geçen kimselerin yüzde dördünden azı”
  Object: “hapishaneye”
  Verb: “geri döner”

Table 1. Comparison of SMT Output versus Human Translation.
Table-2 and Figure 4 provide some of the challenges.

The first line in Table-2 is the original source English sentence. The second line is its SMT output. It is a translation based on statistics. However, the meaning of some words are not translated properly.

### Table 2. Example Showing Challenges in Turkish SMT.

<table>
<thead>
<tr>
<th>No.</th>
<th>Original English</th>
<th>Turkish Equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Brenda Palms-Farber was hired to help ex-convicts reenter society and keep them from going back.</td>
<td>Brenda Palms-Farber kiralandı yardım etmek eski suçlular yeniden girmek toplum ve saklamak onları den gitmek geri.</td>
</tr>
<tr>
<td>2</td>
<td>Brenda Palms-Farber was hired to help ex-convicts reenter society and keep them from going back.</td>
<td>Brenda Palms-Farber tutuldu Yardım etmek eski suçlular yeniden katılmaya topluma ve alıkoyma onları den dönmek geri.</td>
</tr>
<tr>
<td>3</td>
<td>Brenda Palms-Farber was hired to help ex-convicts reenter society and keep them from going back.</td>
<td>Brenda Palms-Farber eski suçluların topluma yeniden katılmalarına yardım etmek ve onları geri dönmekten alıkoymak üzere tutuldu.</td>
</tr>
</tbody>
</table>

To improve the translation quality, four words, which are emphasized as bold, are the correct meanings, given in third line. However, the third line alignment is incorrect. Therefore, a reordering is required. The fourth line is the human translation of the first line.

SMT process is mostly based on statistics, and it is not possible to always detect the correct meaning without using syntax information; therefore it is likely to be mistranslated for some words, such as, “kiralandı” vs “tutuldu”, “girmeye” vs “katılmaya”, “saklama” vs “alıkoyma” and “gİmek” vs “dönmek”. Even if the words' meanings are properly chosen, they should be placed into their correct alignments. For instance, the predicate “tutuldu” in third row should be placed at the end of the sentence.

# 4 RELATED WORKS AND DISCUSSION

Brown et al. [6] first brought up the concept of statistical model for generating English sentences and for translating them into French. They defined analysis and synthesis components to be used in a statistical translation system. Many of the components did not produce high translation accuracy. However, following the same approach in terms of the analysis-transfer synthesis concept described in [6], they improved the results [7]. A number of linguistic transformations are applied by them. The transfer component operated from a string of French morphemes to a string of English morphemes. The analysis component recanted the source English sentence into a first level intermediate form, the transfer component took this first level intermediate form into another intermediate form. The target language was more compatible with this second intermediate form. Then, the translation from source sentence into target language is constructed from this new intermediate form by the synthesis component. As a result, they obtained 21% improvement in translation performance.

El-Kahlout et al. [8] proposed as a future work to exploit sub-lexical structures, based on two observations: Turkish word would have to align with a complete English sentence and sometimes these English sentences could be discontinuous. Later, Öflazer et al. [9] followed their proposal of exploiting morphology, and developed a process consisting of segmenting Turkish words into lexical morphemes and tagging the English side using TreeTagger [10]. They were inspired from research works in [11,12] that previously addressed the usage of morphology in SMT from or into morphologically richer languages. For example, Niessen et al. [11] used morphological
decomposition to improve alignment quality. They claimed that by benefiting from interdependencies of related inflected forms, bilingual training data can be better exploited. By the combination of their suggested methods, an improvement of the translation results was presented on two German-English collections. Minkov et al. [12] used morphological post-processing on the output side using structural information and information from the source side, to improve SMT quality. They utilized a set of morphological and syntactic knowledge sources from both Russian and Arabic sentences in a probabilistic model, and evaluated their contribution in generating both source and target sentences. They concluded that a significant boost in BLEU scores can be achieved by utilizing a representation between unsegmented word forms and morphologically fully segmented forms [9].

Mermer et al. [13] studied supervised segmentation comprising morphological analysis, disambiguation and manually-crafted rules. Later they presented unsupervised Turkish morphological segmentation (TMS) for statistical machine translation (SMT) [14], which had not been addressed in previous works. They compared unsupervised segmentation against two baselines: no segmentation and their previous study in [13] with supervised results. They concluded that while unsupervised segmentation improves translation BLEU scores (0.5140 for word-based baseline, 0.5455 for unsupervised segmentation and 0.5946 for supervised segmentation) over the word-based baseline; it did not reach the performance of task-optimized supervised segmentation.

El-Kahlout et al. studied reordering in English side to make it appear more Turkish like [5]. They used morphological segmentation and disassembled the Turkish words into their lexical morphemes. On the English side, they do not use derivational morphology. They tagged the English side using TreeTagger [10], which provides a part-of-speech (POS) tag for each word. Their study was based on the idea that in Turkish, most surface distinctions are the results of word-internal phenomena like vowel harmony. By using lexical morphemes instead of surface morphemes, different phonological forms corresponded to the same set of words/tags in English, after translations. For example, the two Turkish surface morphemes “+ler” and “+lar” have different pronunciations. They both indicate plurality and both are translated to “+s” on English side. In their research they considered four cases, each of which has different morphological representations on the Turkish side. In all cases, BLEU was measured for word-based representation. The four cases in [10] are as the following:

- Baseline: Full word representation of both Turkish and English sentences. For instance, the form <of his table> “masasının” would be used on the Turkish side and “table+s” on the English side.

- Full Morphological Segmentation: For the previous example, three tokens (“masa+sı+nın”) on Turkish side and two tokens (“table+s”) on English side.

- Root+Morphemes Segmentation: For English words, it is the same as it is in case two. For Turkish sentences, a single morpheme group is used, yielding two tokens (“masa + sınnı”).

- Selective Morphological Segmentation: Certain Turkish morphemes cannot be aligned with anything on the English side. For example, Turkish “sıl” does not have a corresponding unit on the English side. Therefore, morphemes with unalignment percentage over 80% are considered as part of the root. So, the same word is represented as two tokens (“masası+nn”). English words are represented as in the second case above.
The best BLEU results are obtained with selective morphological segmentation (22.81) resulting 46.8% improvement compared to the baseline (15.53).

Cakmak et al. focused on English-Turkish language pairs [4]. Based on the observation that Turkish is a highly agglutinative language, whereas English has a very poor morphology when compared to this language; one Turkish word is usually aligned with several English words. In their study, they evaluated a Giza++ system by stripping the words down to their stems and suffixes. The usage of morphological units in the training stage reduced the alignment error rate by 40%, relatively. Also, a new test corpus of 300 manually aligned sentences is released together with their study.

The previous researchers have not studied reordering on Turkish side because of the complexity and agglutinative nature of Turkish characteristics. Some of them evaluated reordering on English side, as mentioned above. However, this approach contains some drawbacks. First of all, some tokens are missing on English or Turkish side. For instance, <Ayşe's book> “Ayşe'nin kitabı” can be aligned as follows:
1: 1 ; → “Ayşe”-“Ayşe”
2: 2 ; → “'s”-“+nin”
3: 3 ; → “book”-“kitap”
4: - ; → Ø-“+ı”

The last token on Turkish side does not have a corresponding token on English side. For instance, <The most beautiful house> “En güzel ev” can be aligned as follows:
1: - ; → “The”-Ø
2: 1 ; → “most”-“en”
3: 2 ; → “beautiful”-“güzel”
4 : 3 ; → “house”-“ev”

The first token on English side does not have a corresponding token on Turkish side.
As a second drawback, we can argue that making English sentences Turkish-like is also a sophisticated process. El-Kahlout et al. mentioned those challenges in [5]. One of those challenges is the usage of the preposition “of”, which brings up special difficulties. For example, in noun phrases like “United States of America”, the “of” does not correspond to a genitive morpheme on the Turkish equivalent for “America”. Its translation is not “Amerika’nın Birleşik Devletleri”; its correct translation should be “Amerika Birleşik Devletleri”. What is more, noun phrases located on both sides of “of” have to be extracted and interchanged. Bracketing errors for more than one noun phrases makes the situation more cumbersome. For example, the complex noun phrase “The election of the president of United States of America”, has a nested structure like “The election of [the president of [United States of America]]”. Furthermore, for all prepositions, preceding tags should also be checked and the first step of extraction procedure should be finding the patterns.

According to the state of the art, the main problem for English-Turkish SMT is the word order. Indeed, Table 3 shows our preliminary results using a gold standard test collection which consists of 300 manually aligned English-Turkish sentences [4].

Table 3. BLEU Scores.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-gram</td>
</tr>
<tr>
<td>EN→TR</td>
<td>39.83</td>
</tr>
<tr>
<td>TR→EN</td>
<td>52.06</td>
</tr>
</tbody>
</table>

The collection characteristics are presented in Table 4.

Table 4. Collection Description.

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>TUR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of Sentences</td>
<td>300</td>
<td>300</td>
</tr>
<tr>
<td>Total number of Terms</td>
<td>4711</td>
<td>3528</td>
</tr>
<tr>
<td>Average Number of Terms by Sentences</td>
<td>15.71</td>
<td>11.76</td>
</tr>
</tbody>
</table>
Comparing 1-gram and 2-gram results, we can observe that BLEU score decreases significantly. The BLEU scores of 3-gram and 4-gram are very weak. It is also worth noting that Turkish-English SMT performances are much more better than English-Turkish SMT performances.

5 PROPOSED APPROACH FOR REORDERING

To correctly reorder the words, our proposed approach is based on syntactic rules. To reorder the words on Turkish side, we need to obtain their POS tags and classify the words according to their POS tags. For this purpose, a POS tagger is used.

Figure-5 explains how reordering is carried out. After SMT output in Turkish is obtained, we tag each word. To reorder the unordered words in Turkish, we use syntactic rules.

1-Punctuation marks, such as comma and semicolon, are important signs to indicate a completeness in terms of meaning. Therefore, using these punctuation marks, sentences are divided into segments and reordering is done within each segment. Therefore, the first thing is splitting sentences into segments.  
2-Once the segments are determined, predicates are placed at the end of each segment, to provide SOV constituent order.

3-There can be a confusion between bare verbs and nouns, therefore a further analysis is required to confirm that they are predicates. For example, Turkish word “at” has two meanings: <throw> and <horse>. Only the first one is a verb, while the second one is a noun.

4-If there are no predicates (or cannot be obtained by POS tagger), the order of the sentence is not changed and left as it is.

Table-5 demonstrates how reordering is done according to Figure-5. The word “harcamaktadırlar” is recognized as predicate and placed at the end based on Rule 2.

6 CONCLUSION

The performance of alignment strongly depends on POS tagger, because special cases need to be handled. For instance, Turkish word “acacağım” has multiple meanings and as result, multiple tags. It could be either “I will open”, which is tagged as predicate or “my bottle opener”, which is tagged as noun.
such cases, to improve tagging we need to use some syntactic information. Benefiting from the English side has the potential to improve the performance of reordering. As a future work, we plan to implement our approach and to apply the proposed rules on the gold standard test collection.

REFERENCES
