Translating Between Morphologically Rich Languages: An Arabic-to-Turkish Machine Translation System

İlknur Durgar El-Kahlout Emre Bektaş Naime Şeyma Erdem Hamza Kaya
Tübitak-Bilgem, Kocaeli, Turkey
{ilknur.durgar,emre.bektas,seyma.erdem,
hamza.kaya}@tubitak.gov.tr

Abstract

This paper introduces the work on building a machine translation system for Arabic-to-Turkish in the news domain. Our work includes collecting parallel datasets in several ways for a new and low-resource language pair, building baseline systems with state-of-the-art architectures and developing language specific algorithms for better translation. Parallel datasets are mainly collected three different ways; i) translating Arabic texts into Turkish by professional translators, ii) exploiting the web for open-source Arabic-Turkish parallel texts, iii) using back-translation. We performed preliminary experiments for Arabic-to-Turkish machine translation with neural (Marian) machine translation tools with a novel morphologically motivated vocabulary reduction method.

1 Introduction

It is a well-known fact that to develop robust systems with data-driven methods, it is crucial to have large amounts of data. If the problem needs only raw monolingual data, the solution is straightforward; crawl the web and collect the data in the specific domain. In cases of annotating the data (e.g., treebanks) or parallel data (e.g., for machine translation) collecting the needed data is a bit harder.

Even though machine translation (MT) is one of the popular topics in natural language processing, most of the existing parallel texts include English as one of the languages (e.g., Europarl (Koehn, 2005), Multi-UN (Eisele and Chen, 2010)). For the rest of the languages, generating a new language pair from scratch is tough work that needs extensive human effort and substantial funding. One way of translating languages with no parallel data is pivoting, which means one should find corpora for two language pairs such as source-to-pivot and pivot-to-target with sufficient number of sentences in the same domain and then train and maintain two MT systems. Even though we can find such corpora in the expected domain for the given languages, the error propagation is the biggest problem of pivoting as the second system will try to translate erroneous output of the previous system.

In this work, our goal is building an Arabic-Turkish machine translation on the news domain. The task is very interesting for several reasons; primarily, both the source and the target languages are morphologically rich which proves to be a quite challenging task. Our attention on this language pair has both social and political grounds. Arabic is the official language in most of the Middle East countries that Turkey has relations with. Moreover, there is a need for quick and cheap translation solutions in communicating with the increasing number of refugees in Turkish spoken areas.

The news domain is selected as it has several benefits such as the fact that at least one side of the parallel texts can be found publicly on the web (e.g. several news portals) and Arabic is written in Modern Standard Arabic format for the news domain which is common for all Arabic speakers. To collect the data, both monolingual and bilingual data on the web is exploited. Selected portion of a monolingual data is translated into Turkish by professional translators, the publicly available but out-of-domain parallel data is cleaned and used directly and, lastly, rest of the monolingual Turkish data is back-translated to train our systems. Both unsupervised and supervised morphology reduction techniques are used to reduce the vocabulary size to a fixed number and let to fit our vocabulary into a given number of tokens while training the neural machine translation (NMT) systems .

This paper is organized as follows; Section 2 gives brief information about the source and tar-

Proceedings of the Fourth Arabic Natural Language Processing Workshop, pages 158–166
Florence, Italy, August 1, 2019.

get languages. Section 3 describes the data obtaining methods, and Section 4 introduces the segmentation methods for Turkish to alleviate the morphological differences and explains generation of surface word forms as post-processing. In section 5, we talk about our experimental setup including the data sizes and morphology abstraction/separation experiments with Marian (Junczys-Dowmunt et al., 2018) NMT tool. Finally, we conclude in Section 6.

2 Arabic and Turkish

2.1 Arabic

Arabic is a member of the Central Semitic language family. It is spoken by approximately 300 million people (ranked as sixth language) and accepted as official language in 27 countries (ranked as the third language after English and French). Arabic can be classified into three categories as; Classical Arabic (the language of the Qur'an), Modern Standard Arabic (is used in written texts and formal speeches, not a native language) and Arabic dialects (spoken by locals, mostly not written). Arabic is written from right to left with distinct 28 letters with various combinations of dots above or below these shapes. There are no capital letters. Roots are mostly composed of consonants and can have different meanings with the help of the vowels and diacritics. Arabic has a very complex and sometimes inconsistent orthography [1].

Arabic has a highly complex concatenative derivational and inflectional morphology. Words can take prefixes and suffixes at the same time for tense, number, person, gender information. For an example of the concatenation processes, the Arabic word, وسينهي (gloss; and he will finish) can be decomposed as ينهي (finished), +س (he will), and +و (and).

2.2 Turkish

Turkish is a member of the Ural-Altay language family and is the most commonly spoken Turkic language by more than 90 million people. It is the official language of Turkey and Northern Cyprus. There are lots of minority groups all over the world mainly in Europe (approximately 5M speakers).

From the machine translation point of view, Turkish has interesting and challenging properties when compared to the mostly studied languages in data-driven MT research such as English, German, French and Spanish. First of all Turkish is a highly agglutinative language where words are formed by concatenating morphemes (by suffixation) with very productive inflectional and derivational processes. Turkish morpheme surface realizations are generated by several morphophonemic processes such as vowel harmony, consonant assimilation, and elisions. The morphotactics of word forms could be quite complex when multiple derivations are involved. Indeed, Turkish is one of the languages that needs special attention because of its morphological richness. An example of the Turkish morphology can be shown with the Turkish word *partisindeydi* (gloss: s/he was at his/her party), this word can be decomposed into four morphemes as *parti* (party), +*si* (her/his), +*nde* (in) and +*ydi* (s/he was).

3 Obtaining Data

The backbone of the machine translation system is a "good" data like the most of the machine learning problems. In case of MT, a parallel corpora is required. The domain of the data, the quality and the quantity directly effect the translation output. On the other hand, obtaining such data for the machine translation purpose is not that easy. There have been efforts made to obtain parallel texts for machine translation by crawling web for parallel data (Uszkoreit et al., 2010), and by using MechanicalTurk (Ambati and Vogel, 2010; Zbib et al., 2012). Even though we spent some efforts to use MTurk, it is not yet available for requesters outside USA.

We specify three different ways to obtain the Arabic-Turkish parallel corpora; i) by translating Arabic texts into Turkish by professional translators, ii) by exploiting web for open-source Arabic-Turkish parallel texts and, iii) by back-translating monolingual Arabic data by using existing machine translation systems.

3.1 Obtaining In-domain Training Data

We selected approximately 170K Arabic sentences in the news domain from LDC datasets and had them translated to Turkish by professional translators in order to obtain gold-standard training data. Even though the translators are experts, quality assurance is an important issue. We aimed to avoid low-quality translations with a few steps.

[1] http://www.nizarhabash.com/tutorials/EMNLP-2014-Diab+Habash-Tutorial.pdf

Before the translation process, we labeled each sentence to keep the parallelism in translations. This labeling is done to prevent translators not to join any two sentences or split one sentence into pieces while translating. Then, we asked each translator to translate 50 sentences. We analyzed the outputs, detected common translation errors and prepared a translation procedure for machine translation purpose. The translation procedure had rules such as;

- Every information in the source sentence should be translated into Turkish. Neither addition nor deletion of a part of a sentence was allowed.
- Translations should not have any meaning disorder or fluency problems. Constituents can be arranged due to grammar rules without changing the meaning. Phrases should be chosen as precisely as possible.
- Each sentence should be translated independently, without considering the previous context.
- Sentences in two different lines should not be combined into a single sentence or vice versa.

After the translation was completed, we employed a bilingual consultant to randomly select 5% of the sentence pairs from each document and score them according to the quality of translation. If the quality is lower than given threshold, translators re-translated each problematic document once more. After this process, if the quality was still low, we rejected the translations for this document.

We separated 1,600 sentences for development and 1,357 sentences for testing and demanded four Turkish references to be translated by four different translators. Table 1 shows the time and cost spent to generate the gold-standard translations for training and development. As seen in the table, generating a parallel corpora by human translation from Turkish to Arabic is a time and money consuming task as the number of such translators are limited [2]. Moreover, after spending a huge budget and time, the size of the corpora is not still sufficient to train a NMT system. These facts forced us to search the web for publicly available data.

[2]As the Arabic part of the corpus is licensed by LDC, the generated corpora can not be shared with any third parties

Corpus	# Sents	Cost ($)	Time
Training	160,764	202K	7 months
Development	11,828	12K	2 months

Table 1: Time and cost spent to generate gold-standard translations.

3.2 Searching Web for Publicly Available Data

We exploited the web in order to take advantage of already existing parallel Arabic-Turkish data. We obtained two subsets of parallel data with small effort but both were out-of-domain. The corpora are;

WIT: Web Inventory of Transcribed and Translated Talks (Cettolo et al., 2012) contains transcriptions of TED talks in more than hundred languages. We selected the IWSLT 2014[3] training data as it contains both Arabic-English and Turkish-English language pairs. Firstly, common talk titles are searched and then on these common talks, Arabic and Turkish sentences that have the same English translation for each talk are matched. As a result, 130K such Arabic-Turkish parallel sentences are obtained.

OpenSubtitles2018[4]: OpenSubtitles2018 (Lison and Tiedemann, 2016) is a large database of TV and movie subtitles for sixty languages. The database has Arabic-Turkish parallel texts that contains almost 28M sentences. Even though these subtitles are aligned based on time stamps, the word order differences between the languages make one-to-one sentence alignment harder. To solve this problem and obtain more reliable parallel data, the text was re-aligned by a bilingual sentence aligner (Moore, 2002). Using this method, 21M out of 28M sentences are selected.

Both WIT and OpenSubtitles2018 are out-of-domain (OOD) for the news domain MT task, and the ratio of this OOD corpora to the news domain is huge (20M to 130K). To increase the size of the news corpora, we used a well known technique, backtranslation.

3.3 Monolingual Turkish Data and Backtranslation

In recently published NMT systems, backtranslation (Sennrich et al., 2016a) is applied commonly to increase the parallel corpora if the training data

[3]https://wit3.fbk.eu/mt.php?release=2014-01
[4]http://opus.nlpl.eu/OpenSubtitles2018.php

Corpus	In-Dom.?	# Sents
Baseline (BASE)	Yes	160K
Subtitles (OOD1)	No	21M
WIT (OOD2)	No	130K
Monolingual (MONO)	Yes	3M
Test	Yes	1357
Development	Yes	1600

Table 2: Type and size of the corpora used in the experiments.

is limited. For backtranslation, two freely available monolingual Turkish news corpora CNN-Turk[5] (2.14M sentences) and Aljazeera[6] (718K sentences) are used.

Collected monolingual Turkish corpora is preprocessed to separate each sentence to a line, to remove sentences only consisting of foreign words, symbols, numbers, and blank lines, and to replace carriage returns with line feed characters. Lastly, the corpus is sorted and the duplicate sentences are removed.

After backtranslation, as automatic systems can not produce gold-standard translations for all sentences, we need to filter the translated output to obtain a "better" subset of it. We remove translations if; i) output has only one word, ii) the ratio of input/output words is more than three and, iii) any word except the Turkish stop-words repeats more than three times. After all the collection efforts, the size and the domain of the parallel corpora is shown Table 2.

4 Incorporating Linguistically Segmented Subwords

4.1 Previous Work

Incorporating morphology when working with morphologically rich languages in SMT has been addressed by several researchers for many years. (Yang and Kirchhoff, 2006) decomposed the unknown source words at the test time into morphological subwords and translated these subwords that are unknown to the decoder by using phrase-based (PB) back-off models. For Arabic, (Zollmann et al., 2006; Sadat and Habash, 2006) exploited morphology by using morphologically-analyzed and/or tagged resources. (Popovic and Ney, 2004) presented different ways of improving translation quality for inflected languages Serbian, Catalan and Spanish by using stems, suffixes and part-of-speech information. (Goldwater and McClosky, 2005) replaced Czech words with lemmas and pseudo words to obtain improvements in Czech-to-English statistical machine translation. (Minkov et al., 2007) used morphological post-processing on the target side by using structural information and information from the source side in order to improve translation quality for Russian and Arabic. (Luong et al., 2010) proposed a hybrid morpheme-word representation in the translation models of morphologically-rich languages.

The first effort for Turkish morphological segmentation, (Durgar El-Kahlout and Oflazer, 2010), used morphological analysis to separate some Turkish inflectional morphemes that have counterparts on the English side in English-to-Turkish statistical machine translation. (Bisazza and Federico, 2009) present a series of segmentation schemes to explore the optimal segmentation for statistical machine translation of Turkish. (Mermer and Akin, 2010) worked on unsupervised morphological segmentation from parallel data for the task of statistical machine translation.

With the rise of neural machine translation, fitting the whole corpora into a fixed number vocabulary has become a challenge. Despite its success over the previous SMT methods, NMT has the lack of using large vocabularies as the training/decoding complexity is directly proportional to the vocabulary size. One solution is to limit the vocabulary size to a fixed number but this is a challenging problem especially for morphologically rich languages.

A well-known and effective method to solve this problem is the Byte-pair encoding (Sennrich et al., 2016b) (BPE) which splits words into "reasonable" number of subwords to satisfy the fix vocabulary criteria. BPE is an unsupervised word segmentation method originally used as a word compression algorithm. It iteratively "merges" the most frequent character n-grams into subwords leaving no out-of-vocabulary words. BPE is totally statistical, likelihood-based word splitting method and involves no means of linguistic information. So, researchers exploit morphology once more to incorporate "linguistically" separated subword representation when translating from/to morphologically rich languages (Sánchez-Cartagena and Toral, 2016; Bradbury and Socher, 2016) with

[5]https://www.cnnturk.com/
[6]http://www.aljazeera.com.tr/

neural machine translation.

Recently, (Ataman et al., 2017) incorporate both supervised and unsupervised morphological segmentation methods for Turkish sub-word generation for Turkish-to-English NMT. They used morphological features for the suffixes in order to decrease the sparseness caused by suffix allomorphy.

4.2 Morphological Abstraction of Turkish

The productive morphology of Turkish potentially implies a very large vocabulary size: noun roots have about 100 inflected forms and verbs have much more. These numbers are much higher when derivations are allowed. For example, one can generate thousands of words from a single root even when at most two derivations are allowed. Turkish employs about 30,000 root words (about 10,000 of which are highly frequent) and about 150 distinct suffixes. As an example to the morphological variation, in our Turkish corpora, the root word *inisiyatif* (literally: initiative) occurs totally 258 times in 47 different forms where 25 of these forms are singletons. Using morphologically segmented subwords is straightforward and sufficient when Turkish is on the source side of the translation. In case of Turkish is on the target side, any process such as segmentation or abstraction must be done more carefully as in the final representation the surface word should be generated. As a result, the "best" representation have to be selected that covers the whole information for Turkish words to generate the correct surface form.

In this work, we present an abstraction method similar to our previous work (Durgar El-Kahlout and Oflazer, 2010). Our abstraction can generate back the surface form after translation easily which allows us to use this method even if Turkish is on the target side. Simply we abstracted all possible letters in the morpheme suffixes to alleviate the differences due to the morphophonemic processes such as vowel harmony, consonant assimilation, and elisions. First we apply a morphological analysis and detect the root and the morpheme of the word, and then on morpheme we replace i) vowels *a* and *e* to capital *A* (vowel harmony); ii) *i, ı, u and ü* to capital *H*; iii) *ğ* and *k* to *K* (consonant assimilation) and; iv) *t* and *d* to *D* (consonant assimilation). In order to combine the statistics and reduce the data sparseness problem, abstraction is a better choice for morpheme representation as most surface distinctions are manifestations of word-internal phenomena such as vowel harmony and morphotactics. When surface morphemes are considered by themselves as the units in BPE, statistics are fragmented.

Table 3 shows examples of Turkish words in surface form, abstracted word and the gloss in English with highlights for the common parts. As seen in table, the first and the second columns share three morphemes +mAK+DA+DHr (Write Features) but differentiate on the surface form because of the morphophonemic processes. After the abstraction, the morphemes are same as in the English case.

On top of abstraction, we also kept *root +morphemes* separated versions of the both surface and abstracted Turkish words and experimented with each scenario to understand the effect of abstraction and separation (Table 4 number (5)). In each case we also employed BPE for the vocabulary fitting.

Table 4 shows a Turkish sentence with surface form, abstraction and separation and also BPE applied on each version. Root word *inisiyatif* (literally: initiative) separated by BPE into two or three segments depending on the length of the morphemes in the surface and abstracted representations. In representation (4), we observe that BPE tends to keep first (root) segment longer than the surface case because of the abstracted morphemes. By applying separation over surface or abstraction form, the effect of BPE is lost and only the unknown/singleton words are segmented by the algorithm as in the word *IGAD* in representation (6).

4.3 Word Generation

As stated above, making abstraction and/or segmentation processes on the target side always requires much more attention then the source side. Generating the correct surface form is crucial for the end user as they do not need to be aware of the inner representations. In order to generate the correct surface form, we employed an in-house morphological generation tool which transforms the given text with words in the format of root word and abstracted morphemes, to the correct single-word form. As a first step, this generation tool has been trained by a large Turkish corpus and works by simply creating a reverse-map through morphological segmentation of the corpus. This map contains root+morpheme sequences as keys and their corresponding surface word forms as values.

Word	Abstraction	Gloss
kahrolmaktadır	kahrol+**mAKDADHr**	**s/he is** depress**ed**
şüphelenilmektedir	şüphe+lAnHl**mAkDADHr**	**s/he is** suspect**ed**
partisindeydi	parti+s**HnDAyDH**	**s/he was at his/her** party
sarayındaydı	saray+**HnDayDH**	**s/he was in her/his** palace

Table 3: Turkish abstraction examples

(1) TR: Bu ortak inisiyatif kapsamında Sudan sorununa kapsamlı bir çözüm yer alıyor , IGAD inisiyatifinde ise yalnızca güneyle sınırlı . **(2) 1+BPE:** Bu ortak inisiya@@ tif kapsamnda Sudan sorununa kapsamlı bir çözüm yer alyor , I@@ G@@ AD inisiya@@ tifi@@ nde ise yalnızca gün@@ eyle sınırlı .
(3) 1+Abst.: Bu ortak inisiyatif kapsamHnDA Sudan sorunHnA kapsamlH bir çözüm yer alHyor , IGAD inisiyatifHnDA ise yalnzca güneylA snrlH . **(4) 3+BPE:** Bu ortak inisiyat@@ if kapsamHnDA Sudan sorunHnA kapsamlH bir çözüm yer alHyor , I@@ GA@@ D inisiyat@@ ifH@@ nDA ise yalnzca gün@@ eylA snrlH .
(5) 3+Sep.: Bu ortak inisiyatif kapsam +HnDA Sudan sorun +HnA kapsam +lH bir çözüm yer al +Hyor , IGAD inisiyatif +HnDA ise yalnzca güney +lA snr +lH . **(6) 4+BPE:** Bu ortak inisiyatif kapsam +HnDA Sudan sorun +HnA kapsam +lH bir çözüm yer al +Hyor , I@@ G@@ AD inisiyatif +HnDA ise yalnzca güney +lA snr +lH . **English:** Within this joint initiative, there is a comprehensive solution to the Sudanese problem, while in the IGAD initiative it is limited to the south

Table 4: Turkish sentences after different segmentation schemes

While creating this map, disambiguation step of morphological segmentation is omitted to increase the coverage, as keeping multiple resolutions for a surface word form will increase the number of keys for the reverse-map. Then the reverse-map is sorted by the number of occurrences of segmentation in order to select the most common ones.

In our experiments, the reverse-map succeeds to recover the 92% of the abstracted words into surface forms successfully. For the rest of the words, we defined 23 hand-written rules to generate the words which works with 97% success. Defining the generation rules are not straightforward. For example the morphemes attached to the proper foreign words can be different depending on how the words are pronounced in Turkish.

5 Machine Translation Setup

All available data shown in Table 2 was tokenized, truecased (for Turkish) and the maximum sentence length were fixed to 90 for the translation model. As different segmentations of Arabic is out of our scope in this paper, we segmented Arabic prefixes and suffixes from with MADAMIRA (Pasha et al., 2014) with ATB parameter.

To produce the abstracted Turkish words, the first step is the segmentation of morphemes and then an accurate disambiguation of the morphemes within the sentence. Thus, we first pass each word through a morphological analyzer (Oflazer, 1994). The output of the analyzer contains the morphological features encoded for all possible analyses and interpretations of the word. Then we perform morphological disambiguation using morphological features (Sak et al., 2007). Once the contextually-salient morphological interpretation is selected, we process the abstraction algorithm. On top of the abstraction and segmentation processes, we also trained BPE models over the training sets, for each language disjointly.

For the neural machine translation experiments reported in this paper, comparatively new and better performing NMT architecture, Transformer (Vaswani et al., 2017) is used by Marian (Junczys-Dowmunt et al., 2018) toolkit. System is trained on a workstation housing 4 NVIDIA titan GPUs. The GPU memory parameters are set as follows; *mini-batch-fit* is checked, workspace reserved to 8000, and maxi-batch to 900. With this setup, 24k words/s training speed using all the GPUs in parallel is achieved. Transformer *-type* is employed for training. Depth of the network is set to 4, learning

rate is set to 0.0001 with no warmup, and vocabulary size is set to 40k. *Mini-batch-fit* option is enabled. Usually it took 4-5 days to converge for the experiments.

Our early stopping criteria is 20 runs without a BLEU (Papineni et al., 2002) increase. Moreover, we use Marian-decoder's beam search decoding with size 16. We ensemble two different models which resulted in the highest two BLEU scores on the development set during validation runs. We then merge the subwords back together in the hypothesis as described in 4.3.

5.1 Results

First group of experiments are performed to evaluate the effect of the data collected from different sources. As seen in Table 5, our baseline experiment is trained on the union of in-domain human translated corpora (BASE) and out-of-domain corpora WIT (OOD1) with a ratio 1:1. We did not perform with only BASE corpora as it is quite small to make sense for NMT training. On top of this experiment, we augmented corpora with approximately 2M backtranslated corpora (MONO) with a ratio almost 1:7. Even though this ratio is above the suggested (Sennrich et al., 2016a), we observed an improvement of +6 BLEU points. We argue that if the backtranslated data is preprocessed to satisfy some quality criterion as we described in Section 3.3, one can extend training corpora with much more backtranslated data. As a last experiment, we combined the Subtitles18 data (OOD2) with 21M sentences with a ratio 1:10 to the experiment (2). As a result, despite adding a huge out-of-domain, we again obtained an improvement more than +2 BLEU points. The improvement on BLEU scores seems lower than predicted when compared to the size of the data but we should be aware of that the OOD2 corpora share very limited part with news domain.

For the second group of experiments, we investigate the effect of abstraction and segmentation of Turkish. In experiment (3), we applied three different segmentation/abstraction representations. In the first representation (exp. 4), we separated root words and morphemes into two (e.g. *kahrolmaktadır* as *kahrol +maktadır*), in second representation (exp. 5), we only employed abstraction (e.g. *kahrolmaktadır* as *kahrolmAKDADHr*) and in the third representation (exp. 6), we applied both segmentation and abstraction together (e.g. *kahrolmaktadır* as *kahrol +mAKDADHr*). It is noticed that both segmentation and abstraction processes help to improve the translation. The improvement caused by segmentation is expected as supported with previous researches. The results achieved by this work show that our novel abstraction representation is a better alternative than segmentation to help BPE for Turkish. We observe almost no improvement with segmentation (some small positive change in development data) but an improvement of +0.8 BLEU with abstraction even with huge training data of 24M sentences. Similarly, combining both segmentation and abstraction in one representation does not help the system as much as abstraction does.

Corpora/System	Dev	Test
(1) BASE + OOD1	15.70	15.91
(2) 1 + MONO	21.91	21.78
(3) 2 + OOD2	22.76	24.09
(4) (3) + Separated	23.01	24.13
(5) (3) + Abstracted	23.98	**24.92**
(6) (3) + Abst.+ Sep.	24.11	24.83
Google	19.62	20.70
Yandex	10.91	11.82

Table 5: Arabic-to-Turkish MT BLEU scores due to the different traning corpora

As this work is the first attempt for Arabic-to-Turkish MT to our best knowledge, in order to compare our systems, we also translated test data with Google[7] and Yandex [8] and listed the scores in last two rows. The unique word counts (vocabulary) after each representation are shown in Table 6. It is noticed that just separation root words and morphemes drops the vocabulary more than half but as the final vocabulary is fitted to 40K this reduction does not make a significant impact on the translation. The small count increases in the abstracted representations comes from the different morphological disambiguations of the same word.

In the following example, we show both ours and Google translations of an Arabic sentence. Even both of the translations are almost perfect, there is an important difference in handling the correct tense selection (present vs. past tense). Our translation selects the more suitable tense than Google translation which is also closer to the reference.

[7] translate.google.com
[8] ceviri.yandex.com.tr

Corpora Type	# Unique Tokens
Baseline	1026957
Separated	425216
Abstracted	1027991
Abst. + Sep.	426585

Table 6: Type and size of the corpora used in the experiments.

- **Source:** الصين تحقق منجزات باهرة فى تطوير العلوم والتكنولوجيا فى فترة ٢٠٠١-٢٠٠٥
- **Morp-NMT:** Çin , 2001-2005 yıllarında bilim ve teknolojinin gelişiminde büyük başarılar elde etti .
- **Google:** Çin 2001-2005 yıllarında bilim ve teknolojinin gelişmesinde önemli baarılar elde ediyor
- **Reference:** Çin 2001-2005 yıllarında bilim ve teknolojinin gelişmesinde önemli baarılar elde ediyor
- **English:** Between 2001 and 2005, China Recording Science and Technological Innovation

6 Conclusion

This paper focused on machine translation system for a new low-resourced language pair Arabic-Turkish in news domain which is the first effort for this language pair to the best of our knowledge. We obtained standard in-domain data by human translators. As this method is both time consuming and expensive, we exploited publicly available corpora such as TED talks and subtitle translations. Later, we backtranslated monolingual Turkish news corpora. Finally, we performed experiments with all of these corpora and reported +8 BLEU increase over the baseline setup for state-of-the-art neural machine translation system Marian. On top of these experiments, we also incorporate language specific processes such as the abstraction of morphemic processes caused by vowel harmony and consonant assimilation. We showed an improvement of +0.8 BLEU points with our abstraction representation. We also run a morphological generation tool after the translation process which covers 98% words correctly. Our future work includes applying the same abstraction algorithm to Turkish while translating from/to other European languages.

Acknowledgments

We thank the anonymous reviewers for their detailed and constructive comments. This work is supported by The Scientific and Technological Research Council of Turkey (project no: 110G125)

References

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 62–65.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *CoRR*, abs/1707.09879.

Arianna Bisazza and Marcello Federico. 2009. Morphological pre-processing for Turkish to English statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 129–135.

James Bradbury and Richard Socher. 2016. Metamind neural machine translation system for wmt 2016. In *Proceedings of the First Conference on Machine Translation*, pages 264–267.

Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. Wit^3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT)*, pages 261–268, Trento, Italy.

İlknur Durgar El-Kahlout and Kemal Oflazer. 2010. Exploiting morphology and local word reordering in english-to-turkish phrase-based statistical machine translation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(6):1313–1322.

Andreas Eisele and Yu Chen. 2010. MultiUN: A multilingual corpus from United Nation Documents. In *LREC*.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 676–683, Vancouver, British Columbia, Canada.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, Andr F. T.

Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in c++. *arXiv preprint arXiv:1804.00344*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit*.

Pierre Lison and Jörg Tiedemann. 2016. Opensubtitles2016: Extracting large parallel corpora from movie and tv subtitles. In *In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 148–157, Cambridge, MA. Association for Computational Linguistics.

Coşkun Mermer and Ahmet Afşin Akin. 2010. Unsupervised search for the optimal segmentation for statistical machine translation. In *Proceedings of the ACL 2010 Student Research Workshop*, ACLstudent '10, pages 31–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Einat Minkov, Kristina Toutanova, and Hisami Suzuki. 2007. Generating complex morphology for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 128–135, Prague, Czech Republic. Association for Computational Linguistics.

Robert C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Machine Translation: From Research to Real Users*, pages 135–144, Berlin, Heidelberg. Springer Berlin Heidelberg.

Kemal Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9:137–148.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

A Pasha, Mohamed Elbadrashiny, Mona Diab, A Elkholy, Rushdi Eskandar, Nizar Habash, M Pooleery, Owen Rambow, and Ryan Roth. 2014. Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1094–1101.

Maja Popovic and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proceedings of the 4th LREC*, pages 1585–1588.

F Sadat and Nizar Habash. 2006. Combination of arabic prerocessing schemes for statistical machine translation. In *Proceedings of the COLING/ACL*, AMTA.

Haşim Sak, Tunga Güngör, and Murat Saraçlar. 2007. Morphological disambiguation of turkish text with perception algorithm. In *Proceeding of CICLING, LNCS 4394*, pages 107–118.

Víctor M. Sánchez-Cartagena and Antonio Toral. 2016. Abu-matran at wmt 2016 translation task: Deep learning, morphological segmentation and tuning on character sequences. In *WMT*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Procceedings of the 54th Annual eeting of the Association for Computational Linguistics*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. *Procceedings of the 54th Annual eeting of the Association for Computational Linguistics*, 1715-1725.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proceedings of EACL*, pages 41–48.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F. Zaidan, and Chris Callison-Burch. 2012. Machine translation of Arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 49–59.

Andreas Zollmann, Ashish Venugopal, and Stephan Vogel. 2006. Bridging the inflection morphology gap for Arabic statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 201–204, New York City, USA.