CMP 670 - Statistical Natural Language Processing

Project Proposal

Student number : N18147271
Name : Mustafa
Last Name : KAYA

Project Name : English-Turkish Translation Program

1. Task

In this study will be done English-Turkish translation program. It will translate English words and sentences according to their Turkish equivalents.

2. Data:

A dataset containing 473035 data will be used in the study. In the data set, different length English sentences and Turkish counterparts are available.

3. Related Work

The TU-Language project sponsored by the NATO Science for Stability Programme was started in 1994 to establish computational foundations for the natural language processing research on the Turkish language with the collaboration of the Computer Engineering Department of Middle East Technical University, the Computer Science Department of Bilkent University and Halici Computing, Inc. The project attempts to perform extensive research on Turkish which will eventually lead to the development of an English to Turkish machine translation system, Turkish language tutorial system, a Turkish dictionary and other software tools to be used in further research (Turhan, 1997). Turkish is an agglutinative and morphologically rich language with a free constituent order. Although statistical NLP research on Turkish has taken significant steps in recent years, much remains to be done. Especially for the annotated corpora, Turkish is still behind similar languages such as Czech, Finnish, or Hungarian (Yıldız et al.,2014). For example, EuroParl corpus (Koehn, 2002), one of the biggest parallel corpora in statistical machine translation, contains 22 languages (but not Turkish). Although there exist some recent works to produce parallel corpora for Turkish-English pair, the produced corpus is only applicable for phrase-based training (Yeniterzi and Oflazer, 2010; El-Kahlout, 2009).

Bisazza and Federico (2009) have applied morphological segmentation in Turkish-to-English statistical machine translation and found that it provides nontrivial BLEU score improvements. In the context of translation from English toTurkish, Durgar-El Kahlout and Oflazer (2010) have explored different

representational units of the lexical morphemes and found that selectively splitting morphemes on the target side provided nontrivial improvement in the BLEU score. Also, Durgar-El Kahlout and Oflazer(2006) tested five method to translate from English to Turkish. They obtained these BLEU scores.

Exp.	System	BLEU
1	Baseline	0.0752
2	Morph. Concatenation.	0.0281
3	Selective Morph. Concat.	0.0424
4	Morph. Grouping and Concat.	0.0644
5	Morph. Grouping + (3)	**0.0913**

The results in paper indicate that with the standard models for SMT, we are still quite far from even identifying the correct root words in the translations into Turkish, let alone getting the morphemes and their sequences right. Although some of this may be due to the (relatively) small amount of parallel texts they used, it may also be the case that splitting the sentences into morphemes can play havoc with the alignment process by significantly increasing the number of tokens per sentence especially when such tokens align to tokens on the other side that is quite far away.

In my study, i will implement the work done before. However, i will use different dataset. This study will be applied under the guidance of previous academic studies and the success level of the program will be measured.

4. Method

In this task, i will use Deep Learning methods. Recently, these methods have been shown to perform very well on various NLP tasks such as language modeling, POS tagging, named entity recognition, sentiment analysis and paraphrase detection, among others. The most attractive quality of these techniques is that they can perform well without any external hand-designed resources or time intensive feature engineering. Despite these advantages, many researchers in NLP are not familiar with these methods (Socher and Manning, 2013).

5. Schedule

Schedule	Task
28.03.2019 – 11.04.2019	Research more work about translation method
11.04.2019 – 25.04.2019	Writing code
25.04.2019 – 09.05.2019	Testing and review
09.05.2019 – 23.05.2019	Prepare presentations

REFERENCES

Bisazza. A., and Federico, M., 2009. Morphological Pre-Processing for Turkish to English Statistical Machine Translation. International Workshop on Spoken Language Translation (IWSLT), Tokyo, Japan.

Durgar-El-Kahlout, İ. and and Oflazer, K., 2010. Exploiting Morphology and Local Word Reordering in English-to-Turkish Phrase-Based Statistical Machine Translation. IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 18, NO. 6.

Durgar-EL-KAHLOUT İ., 2009. A Prototype English-Turkish Statistical Machine Translation System. Doctor of Philosophy, Sabancı University.

Durgar-El Kahlout, İ. and Oflazer, K., 2006. Initial Explorations in English to Turkish Statistical Machine Translation. Proceedings of the Workshop on Statistical Machine Translation, pages 7–14, New York City, June 2006. 2006 Association for Computational Linguistics

Koehn, P., 2002. Europarl: A Multilingual Corpus for Evaluation of Machine Translation.

Socher, R. and Manning, C., 2013. Deep Learning for Natural Language Processing.

Turhan, Ç., K., 1997. An English to Turkish Machine Translation System Using Structural Mapping, ANLC '97 Proceedings of the fifth conference on Applied natural language processing, pages 320-323.

Yeniterzi, R. and Oflazer, K., 2010. Syntax-tomorphology mapping in factored phrase-based statistical machine translation from english to turkish. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, ACL ’10, pages 454–464, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yıldız, O., T., 2014. Constructing a Turkish-English Parallel TreeBank. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), Baltimore, Maryland, USA, June 23-25 2014. c 2014 Association for Computational Linguistics, pages 112–117.