Predicting Translation Difficulty on Sentence-level for MT Systems

Junyi Jessy Li and Kai Hong
University of Pennsylvania
Philadelphia, PA 19104
{ljunyi, hongkai1}@seas.upenn.edu

1 Problem Statement and Related Work

Machine Translation (MT), despite its many recent advances, still remains a hard problem. Translations produced by automatic systems suffer from problems such as information reordering, misalignments, mistranslations, etc. In this study, we propose to explore the question of how difficult a sentence is to translate automatically.

We propose to study this problem from two viewpoints. The first viewpoint relies on language dependent factors. Intuitively some properties unique to certain languages poses problems for automatic systems. For example, Chinese and Romanian are well known for their zero anaphora (Zhao and Ng, 2007; Mihăilă et al., 2011); information expressed as individual words in one language may be morphologically expressed in another (Minkov et al., 2007); reorderings often must be explicitly handled (Wang, 2007), and a single word in one language may be decomposed into several in another (Popovi et al., 2006).

On the other hand, some factors might make the sentences difficult to translate, independent of language. Length, for instance, is a natural indicator for how difficult for a human to translate a sentence (Mishra et al., 2013). We ask the question whether it is also the case for automatic systems. Other indicators we conjecture include the use of subordinations, structure, amount of re-ordering and complexities of verb phrases and noun phrases, etc.

A handful of studies have looked into predicting the difficulty of translation on language level, rather than on sentence level. There, most of the work investiage the general properties which make those language-pairs difficult to translate. Birch et al. (2008) show that the amount of reordering, morphological complexity and historical relatedness are important factors for predicting the performance of MT systems. A further study looked into 462 language pairs (in total 22 languages), where the concept of translation model entropy is introduced to capture the amount of uncertainty involved in choosing candidate translation phrases (Koehn et al., 2009).

Through this study we hope to characterize difficulties that may lie ahead for MT systems, and we aim at quantitatively estimate such difficulties by developing a set of features for a machine learning framework.

2 Datasets

We propose to use the test set from NAACL WSMT 2006 (Koehn and Monz, 2006) and ACL WSMT 2007 (Callison-Burch et al., 2007) for our study. The dataset of these two years include language pairs of English-German, English-French, English-Spanish and English-Czech (for WSMT 2007 only) in both directions. We choose these two years for the reason that they provide manual evaluation of fluency and adequency per sentence, along with the *BLEU* score. Moreover, a number of language families have been covered, including Germanic (English and German), Romance (French and Spanish) and Slavic (Czech). For year 2007, data of systems translating the same content from different languages are also available, which would provide us with better estimate of translation difficulty depending on the language. As we want to form a supervised system of predicting translation difficulty, we need to split our dataset into training, development and testing set.

If time permits, we would like to look into the problem for more languages. The WSMT workshop of later years include Hungarian, Haitian Creole and Russian, which makes it a good choice of performing further research.

3 Approach and Evaluation

3.1 Objective Function and Evaluation

The plan of this study is to investigate and collect a set of language dependent and independent factors of a sentence that are indicative of translation difficulty. One challenge of such an approach is the lack of a reliable sentence-level evaluation metric for translation hypothesis. We conjecture that with more systems, the average of automatic scores or human evaluations would provide a partial but feasible solution to this problem. Here, we regard manual evaluations as our main metric, *BLEU* scores are used as auxilliary metric. The oracle difficulty measure for a sentence is thus the average of fluency and adequency of all MT systems, where higher score indicates easier sentence-pairs. For the cases where manual evaluation is not available, we use the *BLEU* score as oracle, similar to Birch et al. (2008) and Koehn et al (2009) where they evaluate langauge-level difficulty; as well as (Nenkova and Louis, 2008) where they average over the *ROUGE* scores to predict input difficulty for automatic summarization.

To demonstrate the effectiveness of our proposed features, we will compute the Spearman Correlation between the features and the difficulty score on training set. We can also compute the goodness of fit for simple linear regression models using one feature. Then we propose two approaches of performing evaluations. Firstly, as the difficulty takes a value between 1 and 5, we can regard our problem as a regression problem, which could be evaluated through Mean-square Error. Secondly, we can also compute the pariwise accurracy, which calculates the accurracy of correctly predicting which sentence-pair is more difficult than the other.

3.2 Proposed Default and Baseline

Since sentence length is one of the most crucial indicators of translation difficulty for human (Mishra et al., 2013), we regard it as our *Default System*. As for the *Baseline System*, we propose to compute factors such as syntax complexity, subordinations and the number of noun phrases or verb phrases.

References

Alexandra Birch, Miles Osborne, and Philipp Koehn. 2008. Predicting success in machine translation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 745–754, Honolulu, Hawaii, October. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation.* Association for Computational Linguistics, Prague, Czech Republic, June.

Philipp Koehn and Christof Monz, editors. 2006. *Proceedings on the Workshop on Statistical Machine Translation.* Association for Computational Linguistics, New York City, June.

Philipp Koehn, Alexandra Birch, and Ralf Steinberger. 2009. 462 machine translation systems for europe.

Claudiu Mihăilă, Iustina Ilisei, and Diana Inkpen. 2011. Zero pronominal anaphora resolution for the romanian language. *Research Journal on Computer Science and Computer Engineering with Applications POLIBITS*, 42.

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, June. Association for Computational Linguistics.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351, Sofia, Bulgaria, August. Association for Computational Linguistics.

Ani Nenkova and Annie Louis. 2008. Can you summarize this? identifying correlates of input difficulty for multi-document summarization. In *Proceedings of ACL-08: HLT*, pages 825–833, Columbus, Ohio, June. Association for Computational Linguistics.

Maja Popovi, Daniel Stein, and Hermann Ney. 2006. Statistical machine translation of german compound words. In Tapio Salakoski, Filip Ginter, Sampo Pyysalo, and Tapio Pahikkala, editors, *Advances in Natural Language Processing*, volume 4139 of *Lecture Notes in Computer Science*, pages 616–624. Springer Berlin Heidelberg.

Chao Wang. 2007. Chinese syntactic reordering for statistical machine translation. In *In Proceedings of EMNLP*, pages 737–745.

Shanheng Zhao and Hwee Tou Ng. 2007. Identification and resolution of Chinese zero pronouns: A machine learning approach. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 541–550.