Determining Clause Boundary and Structure for Machine Translation of Complex Sentences

Abheet Sharma[1] and Alok Debnath[2]

[1] abheet.sharma@research.iiit.ac.in
[2] alok.debnath@research.iiit.ac.in

Abstract. Companies which use MT tools all face a common problem: whenever the source sentence is too complicated (indicator of which is the presence of multiple clauses) the target sentence generated is very poor in quality. Thus a translator has to manually do these types of translations, which is an expensive operation. In this project, we aim to identify the sentences with multiple clauses, identify these clauses and mark their boundaries. The output will be sentences where the clauses are highlighted, so the post editor can choose which clauses to break down. We also have to suggest a strategy to break the clauses into sentences with no more than two clauses. We use a dependency based approach for this task.

Keywords: Preprocessing for Machine Translation · Dependency Parsing

1 Introduction

Machine translation is known to be an NLP-complete problem. Therefore, the machine translation pipeline is equipped with multiple preprocessing tools, in order to provide all the analyses required for the translation between two language pairs in most rule-based and statistical pipelines. In this project, we aim to use one such preprocessing tool in order to enhance clause identification and clause boundary detection for machine translation of complex sentences. Dependency parsing is a form of sentence parsing that determines the roles of the words in a sentence in relation to other words regardless of their syntactic arrangement. With the evolution of universal dependencies, the syntactico-semantic relations between words can be classified multi-lingually, which is very useful for machine translation. By identifying the relations between the main verb and coordinate/subordinate verbs, we can easily classify the clauses, identify the possible relations between the verbs in the clauses and determine the most optimal clause boundary in that sentence based on word order.

2 Review of Literature

The review of literature in this project will be divided into three parts: 1) English-Hindi MT, 2) clause boundary detection and pipelines 3) universal dependencies and dependency parsing.

2.1 Machine Translation for English-Hindi

Preprocessing is an important task or machine translation. Angla-Hindi [1] provides post translation aids by determining sections of the text that have not been well translated. Hybrid machine translation tools such as [2] use a tree-like structure for simplification of the semantic structure in order to ease machine translation, a task strongly required for machine translation of complex sentences. In the vein of syntactic parsing for machine translation, Arabic to English MT have performed both morphological as well as syntactic preprocessing tools [3], [4], [5], which are useful to determine the sentence structure and reordering needed in cases of complex sentences. Similar approaches have been tried from Hindi to English by [6], while [7] use UNL in order to represent the relation of different lexical items, for an interlingua based MT system. For more data rich language pairs, statistical methods exist for sentence simplification before translation [8].

2.2 Clause Boundary Detection

Another important task in preprocessing for machine translation that has not been researched in the context of downstream tasks is clause boundary detection. [9] provides a survey of predicting clause boundaries, while [10] is a rule based method for clause boundary detection. The latter method is a pipeline that uses phrase structure trees in order to determine the clauses. Our method is more task-oriented and therefore, the primary requirement is not word order based, which is why clause boundary detection, while following the former paper, will use universal dependencies (see next section) for clause boundary detection.

2.3 Universal Dependencies

Our approach uses universal dependencies because they capture common syntactic and semantic relations between words multi-lingually. [11] is the Universal Dependencies Project(v1) which is a multilingual treebank collection. We will parse these trees and find relations between the verbs in the primary and in the clause to detect clause boundaries. Since we require a semantic-annotated treebank rather than syntactic(since syntax between languages is different but the semantic relations will be similar between languages), we use [12] which provides the required semantic-annotated treebanks. [13] will help us to perform multilingual universal dependency parsing on raw text. [14] is an enhanced English universal dependency representation which helps us to get more direct relations between words (i.e, it mitigates the effect of syntax on surface-structure dependency trees). From the Hindi perspective, we use [15] which extends universal dependencies to Indian languages by converting Pninian Krakas and Dependencies to universal dependencies and create Hindi Dependency Treebanks.

3 Hypothesis

Given a sentence and its dependency parse, it is possible to:

– Identify complex sentences and detect the boundaries of clauses, and;
– Determine the optimal translation strategy given universal dependencies

References

1. Sinha, R. M. K., and A. Jain. "AnglaHindi: an English to Hindi machine-aided translation system." MT Summit IX, New Orleans, USA (2003): 494-497.
2. Narayan, Shashi, and Claire Gardent. "Hybrid simplification using deep semantics and machine translation." Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2014.
3. El Isbihani, Anas, et al. "Morpho-syntactic Arabic preprocessing for Arabic-to-English statistical machine translation." Proceedings of the Workshop on Statistical Machine Translation. Association for Computational Linguistics, 2006.
4. Habash, Nizar. "Syntactic preprocessing for statistical machine translation." Proceedings of the 11th MT Summit 10 (2007).
5. Habash, Nizar, and Fatiha Sadat. "Arabic preprocessing schemes for statistical machine translation." Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006.
6. Rao, Durgesh, et al. "A practical framework for syntactic transfer of compound-complex sentences for English-Hindi machine translation." Proceedings of KBCS. Vol. 2000. 2000.
7. Dave, Shachi, Jignashu Parikh, and Pushpak Bhattacharyya. "Interlingua-based EnglishHindi machine translation and language divergence." Machine Translation 16.4 (2001): 251-304.
8. Specia, Lucia. "Translating from complex to simplified sentences." International Conference on Computational Processing of the Portuguese Language. Springer, Berlin, Heidelberg, 2010.
9. Sharma, Sanjeev Kumar. "Clause Boundary Identification for Different Languages: A Survey." International Journal of Computer Applications Information Technology 8.2 (2016): 152.
10. Sacaleanu, Bogdan, Alice Marascu, and Charles Jochim. "Rule-based syntactic approach to claim boundary detection in complex sentences." U.S. Patent No. 9,652,450. 16 May 2017.
11. Nivre, Joakim, et al. "Universal Dependencies v1: A Multilingual Treebank Collection." LREC. 2016.
12. White, Aaron Steven, et al. "Universal decompositional semantics on universal dependencies." Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. 2016.
13. Zeman, Daniel, et al. "CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies." Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies (2018): 1-21.
14. Schuster, Sebastian, and Christopher D. Manning. "Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks." LREC. 2016.
15. Tandon, Juhi, et al. "Conversion from paninian karakas to universal dependencies for hindi dependency treebank." Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016). 2016.