Intelligent Translation Memory Matching and Retrieval Metric Exploiting Linguistic Technology

Rohit Gupta University of Wolverhampton
Hanna Bechara University of Wolverhampton
Constantin Orasan University of Wolverhampton

ABSTRACT

Translation Memories (TM) help translators in their task by retrieving previously translated sentences and editing fuzzy matches when no exact match is found by the system. Current TM systems use simple edit-distance or some variation of it, which largely relies on the surface form of the sentences and does not necessarily reflect the semantic similarity of segments as judged by humans. In this paper, we propose an intelligent metric to compute the fuzzy match score, which is inspired by similarity and entailment techniques developed in Natural Language Processing.

1. Introduction

Most of the Translation Memories research has been carried out in the industry. The focus of this research has been on providing a good graphical user interface to the translators, developing different filters to handle different file formats (e.g. pdf, xml, txt, html, word, xliff, subtitle etc), and project management features. Apart from this, current TMs are also equipped with some tools like terminology managers and plugins to support machine translation from MT service providers. One of the core features of a TM system is retrieving previously translated similar segments for post-editing in order to avoid translation from scratch when an exact match is not available. However, this retrieving process is still limited to edit-distance based measures. Although, these measures provide a strong baseline, they are not sufficient to capture the semantic similarity between the segments as judged by humans. This results in uneven post-editing time required for the same fuzzy match scored segments and non-retrieval of semantically similar segments. In this paper, we propose an intelligent metric to compute the fuzzy match score, which is based on SemEval Task 1 semantic similarity and textual entailment system (Gupta et al., 2014).

2. Our Approach

The system described in (Gupta et al., 2014) calculates the similarity and entailment between a pair of sentences. This system was adapted to measure the similarity between two TM segments. Given the amount of calculation involved in the task, we kept only those features which can be quickly calculated and proved the most useful for the original system. The system

uses features based on surface form, parts of speech information, lemma, typed dependency parsing, named entities, paraphrasing, machine translation evaluation, and corpus pattern analysis (Hanks, 2013). Stanford CoreNLP3 toolkit (Manning et al., 2014) provides lemma, parts of speech (POS), named entities, and dependencies relations of words in each sentence. We used the PPDB paraphrase database (Ganitkevitch et al., 2013) to identify paraphrases.

After extracting these features, we employed a support vector machine (SVM) in order to build a regression model to predict semantic similarity. The training dataset for the SVM is a set of 4934 parallel sentences of the SICK dataset (Marelli et al., 2014) annotated with similarity scores by humans. The SVM used an RBF kernel with C = 8 and γ = 0.125. More details about the method can be found in (Gupta et al., 2014).

The trained SVM system works as a similarity calculator between any pair of sentences provided that the same feature values are available for this pair of sentences.

3. Experiments and Results

We carried out evaluations on two different sets. The test sets were generated by a random selection of segments from DGT-TM corpora (Steinberger et al., 2012). We used English as source and French as target. The target side (French) of the input was considered as a reference for evaluation. We used the word based edit-distance measure implemented by OmegaT[1] as a baseline. The statistics for our test sets is given in the Table 1 below:

	Test-1 (# segments)	Test-2 (# segments)
Input	500	2500
TM	5000	10000

Table 1: Test sets statistics

We performed both a manual and automatic evaluation. For our automatic evaluation, we used the machine translation evaluation metrics METEOR (Denkowski and Lavie, 2014) and BLEU (Papineni et al., 2002). For each input segment, we retrieved the most similar sentence (and their proposed translation into French) as indicated by the baseline and our similarity metric. Table 2 presents the results of automatic evaluation when having a threshold of 70% over the edit-distance. BLEU-ED-70 represents BLEU score using edit distance, BLEU-SS- 70 represents BLEU score using our approach, METEOR-ED-70 represents METEOR score using edit distance, and METEOR-SS-70 represents METEOR score using our approach. The proposed method yields better results for Test-1 but not for Test-2.

	Test-1	Test-2
BLEU-ED-70	77.32	81.34
BLEU-SS-70	81.61	77.14
METEOR-ED-70	91.5	87.35
METEOR-SS-70	92.6	84.55

Table 2: Results automatic evaluation

[1] http://www.omegat.org

To gain a deeper understanding of our system's performance, we also performed a manual evaluation on Test-2. We considered the source side (English) of the segments for this evaluation. A native speaker of English performed the manual evaluation. Three different options were given to the evaluator: Semantic similarity is better; Edit-distance is better; or both are similar. When keeping the 70% threshold and ignoring exact matches, we retrieved 266 different fuzzy matched segments. In these 266 segments, 258 segments were tagged as similar, for 6 segments, edit-distance retrieved better and for 2, our semantic similarity approach retrieved better. Some of the examples from Test-2 are given in Table 3. Example 1 shows our approach (SS) performed better, while examples 2 and 3 show edit-distance (ED) performed better.

The initial results, as stated earlier, show comparable results. Although our approach does not perform better overall, there are several factors, which should be taken into consideration.

1	Input	For the purposes of this Regulation :
	ED	For the purpose of this demonstration :
	SS	For the purposes of this Regulation the following definitions shall apply:
2	Input	This Decision shall enter into force on the date of its publication in the Official Journal of the European Union .
	ED	This Decision shall enter into force on the day of its publication in the Official Journal of the European Union .
	SS	This Decision shall enter into force on the date of its adoption .
3	Input	The Commission sought and verified all information deemed necessary for the determination of dumping .
	ED	The Commission sought and verified all the information deemed necessary for the purposes of the review .
	SS	The Commission sought and verified all the information provided by interested parties and deemed necessary for the determination of dumping , resulting injury and Union interest -

Table 3: Examples from Test-2

The genre of the training set and test set were very different. The SICK dataset consists of simple sentences extracted mostly from image captions while DGT-TM corpus has much larger and complex sentences from mainly legal domain. The average words per segment for TM is 27.9 and for input is 32.54 for test set, whereas for SICK training dataset average words per sentence is only 9.63.

4. Conclusion and Future Work

In this paper we suggested an initial approach to employ a semantic similarity system in a TM framework. Our initial experiment shows some positive indication in this direction. We are in a stage of improving and speeding up our system and extend our experiment to a similar training and test set. In the future, we would also like to develop a human annotated corpus of the same domain to get the better training model. Other similarity calculation techniques involving less computation will also be explored.

References

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Juri Ganitkevitch, Van Durme Benjamin, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia.

Rohit Gupta, Hanna Bе́chara, Ismail El Maarouf, and Constantin Ora ̆san. 2014. UoW: NLP techniques developed at the University of Wolverhampton for Semantic Similarity and Textual Entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*.

Patrick Hanks. 2013. *Lexical Analysis: Norms and Exploitations*. Mit Press.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC 2014*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*, pages 311–318.

Ralf Steinberger, Andreas Eisele, Szymon Klocek, Spyridon Pilos, and Patrick Schlu ̈ter. 2012. DGT-TM: A freely available Translation Memory in 22 languages. *LREC*, pages 454–459.