Semantics-based pretranslation for SMT using fuzzy matches

Tom Vanallemeersch, Vincent Vandeghinste
Centre for Computational Linguistics
Blijde Inkomststraat 13
B-3000 Leuven, Belgium
{tom,vincent}@ccl.kuleuven.be

Abstract

Semantic knowledge has been adopted recently for SMT preprocessing, decoding and evaluation, in order to be able to compare sentences based on their meaning rather than on mere lexical and syntactic similarity. Little attention has been paid to semantic knowledge in the context of integrating fuzzy matches from a translation memory with SMT. We present work in progress which focuses on semantics-based pretranslation before decoding in SMT. This involves applying fuzzy matching metrics based on lexical semantics and semantic roles, aligning parse trees based on semantic roles, and pretranslating matching source sentence parts using aligned tree nodes.

1 Introduction

Semantic knowledge has been adopted recently for SMT preprocessing, decoding and evaluation. Using such knowledge helps for comparing sentences based on meaning rather than form, and for moving away from the assumption of lexical and syntactic similarity between source and target sentences. Little attention has been paid to semantic knowledge in the context of integrating fuzzy matches with SMT. Fuzzy matching methods were originally designed for translation memories, in which translators store their translations. They are now also being used in the context of SMT, for pretranslating parts of sentences before or during decoding. These methods pretranslate matching sentence parts through word alignment, parse node alignment and phrase tables, and use different degrees of linguistic knowledge.

As far as we know, semantic knowledge has not yet been applied for pretranslating sentence parts before decoding in SMT. Therefore, we would like to present our work in progress, which investigates, on the one hand, the use of semantic knowledge (lexical semantics and semantic roles) for improving the usability of fuzzy matches, and, on the other hand, the pretranslation of matching sentence parts using parse nodes aligned through semantic role information.

In Section 2, we provide background on fuzzy matching and on semantic knowledge in SMT, including our own previous research on fuzzy matching and tree alignment. In Section 3, we provide the methodology we are currently devising for semantics-based pretranslation. As this is work in progress, results are not yet provided. However, the discussion of our recent work on combination of fuzzy matching metrics and on semantics-based tree alignment will hint at the potential of using additional sources of linguistic information, such as lexical semantics and semantic roles, for fuzzy matching.

2 Background

The principle of fuzzy matching in a translation memory can be applied to flat sequences or to trees, and either be applied in a linguistically unaware way or involve some degree of linguistic knowledge. Fuzzy matching may be performed using classical sequence comparison metrics like Levenshtein distance (Levenshtein, 1966) or other metrics specifically designed for fuzzy matching, like the ones of Bloodgood and Strauss (2014). It may also be ap-

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 61–64,
Denver, Colorado, June 4, 2015.

plied using MT evaluation metrics like TER (Snover et al., 2006) and Meteor (Denkowski and Lavie, 2014), which were originally designed to compare MT output with one or more reference translations. In this respect, it should be noted that fuzzy matching is performed at the sub-segment level, as it determines matching parts, while MT evaluation is performed on the segment level (Callison-Burch et al., 2012). However, evaluating MT output at the sub-segment level may also be helpful, for instance to determine whether specific parts are translated better than other ones. As for the quality of fuzzy matching metrics, combined metrics appear to perform better than individual ones. For instance, Vanallemeersch and Vandeghinste (2015) combine linguistically unaware with syntactically oriented metrics using regression trees.

In recent years, there has been increasing interest in integrating fuzzy matches with SMT. An example of a linguistically unaware approach is described by Koehn and Senellart (2010), who pretranslate sentences before decoding, using the word alignment between the matching source sentence in the translation memory and its translation. Instead of using the translation of matched parts for pretranslation, the parts and their translation may also be used for enriching a phrase table, as shown by Simard and Isabelle (2009). An example of a linguistically aware integration approach is described by Zhechev and van Genabith (2010), who pretranslate sentences using the node alignment between the parse trees of the source and target sentences in the translation memory. He et al. (2011) apply linguistic knowledge on matching parts during – instead of before – decoding, for instance semantic knowledge.

As indicated above, pretranslation using fuzzy matching involves word alignment or tree alignment. The latter may be based on syntactic information in the trees, but may also involve semantic roles (Vanallemeersch, 2012). Semantic roles are increasingly being used in SMT, in various ways. For instance, Aziz et al. (2011) and Liu and Gildea (2010) annotate source sentences or parses with semantic roles before training an SMT system, while Wu and Fung (2009) compare the semantic roles in the parse tree of a translation hypothesis with the roles in the source parse tree. As regards MT evaluation using semantic roles, metrics like MEANT (Lo and Wu, 2011) have been developed.

3 Methodology

Below, we explain the methodology we are currently devising for semantic pretranslation. It consists of two steps: a fuzzy matching step which makes use of semantic knowledge (lexical semantics and semantic roles), and a pretranslation step which detects the translation of matching sentence parts through semantics-based node alignment of source and target parse trees.

3.1 Semantics-based fuzzy matching

We apply MT evaluation metrics like Meteor and MEANT to source sentences. Meteor allows for matching using synonyms and paraphrases (lexical semantics), while MEANT focuses on semantic roles. We apply a testing framework for applying metrics to sentences in the source and target language and comparing metrics (Vanallemeersch and Vandeghinste, 2015). The framework takes a leave-one-out approach: each source sentence in the translation memory is compared to all other source sentences in the memory. Given some source sentence X (with translation Y), we select the source sentence X′ in the memory which has the highest match score according to a metric, and compare its translation, Y′, to Y, the desired translation. The comparison of Y and Y′, like the comparison of source sentences, takes place using some similarity metric like TER or MEANT (which we refer to as the *target language metric*).

We compare the performance of linguistically unaware fuzzy matching metrics and syntactically oriented metrics on the one hand with semantically oriented metrics on the other hand. When comparing linguistically unaware to syntactically oriented metrics using the above framework (Vanallemeersch and Vandeghinste, 2015), we noted combined metrics have a greater ability to predict the quality of Y′, i.e. they are better at predicting how useful the target language metric will consider Y′ for translating X. Therefore, we expect that combining a semantically oriented fuzzy matching metric with other types of metrics will lead to better predictions than using the metric in isolation. We also investigate the relation between source language and target

language metrics (using the same metric in both languages may favour the source language metric over other ones). Therefore, it may be interesting to make use of human judgments of matches. However, as human evaluation is labour-intensive, and the final use of the matches lies in the integration with SMT, it may be more interesting to focus attention to the evaluation of the MT output produced after the pretranslation step described in section 3.2.

We primarily focus on the language pair English-Dutch. When applying Meteor and MEANT to English sentences, we make use of resources such as the set of English paraphrases in Meteor and the syntactic-semantic parser of Johansson and Nugues (2008), which assigns PropBank and Nombank labels (Palmer et al., 2005; Meyers et al., 2004). For Dutch, we make use of our semantic role labeler described in section 3.2 and of a Dutch paraphrase set created from English-Dutch Moses phrase tables (Koehn et al., 2003) using the *parex* tool (Denkowski and Lavie, 2010; Bannard and Callison-Burch, 2005).

3.2 Semantics-based pretranslation

Applying tree alignment in order to link nodes between source and parse trees (Zhechev and van Genabith, 2010) allows for making use of syntactic information during fuzzy matching. However, the semantic load of a sentence may be expressed in different syntactic ways, leading to possibly different syntactic structures in a source and parse tree. As an example, a source sentence may contain an structure with an active verb and its translation a structure with a passive verb, leading the semantic load to be identical, but the syntactic structure to be different. Another example is a sentence pair in which an English deverbal noun (say, *judgment*) corresponds to a Dutch verb (*beoordelen*). Therefore, we perform tree alignment based on predicates and semantic roles rather than syntactic information. To this effect, we apply semantic role labelers to source and target parses in the translation memory and align the nodes of the resulting parses using a combination of semantic information and lexical probabilities from SMT.

A Dutch semantic role labeler trained on manually annotated data which is able to identify both verbal and nominal predicates does not exist yet; the labeler used in the SoNaR project (Schuurman et al., 2010) only identifies verbal predicates. Therefore, we apply crosslingual projection from English source to Dutch target trees, parsed with Alpino (van Noord, 2006), and train a semantic role labeler for Dutch based on the target trees with projected information (Vanallemeersch, 2012). This approach for training a labeler does not require manual intervention.

After applying a fuzzy matching metric to a source sentence to be translated, we select the best match in the translation memory, and apply a procedure similar to the one of Zhechev and Van Genabith (2010): we find out the translation of the matching source parts by detecting the source nodes overlapping with these parts and retrieving the tokens dominated by the aligned target nodes. In the input to the SMT system, we mark up the source parts with the target tokens, which allows the SMT system to make use of the tokens during decoding. We evaluate the SMT output produced using semantics-based pretranslation through an MT evaluation metric such as MEANT, and compare the SMT output to the one obtained with pretranslation based on mere word alignment or on syntax-based tree alignment.

Acknowledgments

This research is funded by the Flemish government agency IWT (project 130041, SCATE). See http://www.ccl.kuleuven.be/scate.

References

Wilker Aziz, Miguel Rios, and Lucia Specia. 2011. Shallow semantic trees for SMT. *Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, Scotland, July 30–31*, pp. 316–322.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with Bilingual Parallel Corpora. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan, USA, June 25–30*, pp. 597–604.

Michael Bloodgood and Benjamin Strauss. 2014. Translation memory retrieval methods. *Proceedings of the 14th Conference of the European Association for Computational Linguistics, Gothenburg, Sweden, April 26–30*, pp. 202–210.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the Seventh Workshop on Statistical Machine Translation, Montréal, Canada, June 7–8*, pp. 10–51.

Michael Denkowski and Alon Lavie. 2010. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR, Uppsala, Sweden, July 15–16*, pp. 339–342.

Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specic Translation Evaluation for Any Target Language. *Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA, June 26–27*, pp. 376–380.

Yifan He, Yanjun Ma, Andy Way, and Josef van Genabith. 2011. Rich Linguistic Features for Translation Memory-Inspired Consistent Translation. *Proceedings of MT Summit XIII, Xiamen, China, September 19–23*, pp. 456–463.

Richard Johansson and Pierre Nugues. 2008. Dependency-based Semantic Role Labeling of PropBank. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, Waikiki, Honolulu, Hawaii, USA, October 25-27*, pp. 69–78.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-based Translation. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, Edmonton, Canada, May 27 – June 1*, pp. 48–54.

Philipp Koehn and Jean Senellart. 2010. Convergence of Translation Memory and Statistical Machine Translation. *Proceedings of AMTA Workshop on MT Research and the Translation Industry, Denver, Colorado, USA, November 4*, pp. 21–31.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. *Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China, August 23–27*, pp. 716–724.

Chi-kiu Lo and Dekai Wu. 2011. MEANT: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Portland, Oregon, USA, June 19–24*, pp. 220–229.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. Annotating Noun Argument Structure for NomBank. *Proceedings of LREC-2004, Lisbon, Portugal, May 26–28*, pp. 803–806.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.

Ineke Schuurman, Véronique Hoste, and Paola Monachesi. 2010. Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. *Proceedings of the Seventh conference on international language resources and evaluation, Valletta, Malta, May 17–23*, pp. 2471–2477.

Michel Simard and Pierre Isabelle. 2009. Phrase-based machine translation in a computer-assisted translation environment. *Proceedings of MT Summit XII, Ottawa, Ontario, Canada, August 26–30*, pp. 120–127.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. *Proceedings of the Seventh Conference of the Association for Translation in the Americas, Cambridge, Massachusetts, USA, August 8–12*, pp. 223–231.

Tom Vanallemeersch. 2012. Parser-independent Semantic Tree Alignment. *Proceedings of META-RESEARCH Workshop on Advanced Treebanking, in conjunction with LREC-2012, Istanbul, Turkey, May 22*, pp. 1–5.

Tom Vanallemeersch and Vincent Vandeghinste. 2015 [accepted]. Assessing linguistically aware fuzzy matching in translation memories. *Proceedings of the 18th Annual Conference of EAMT, Antalya, Turkey, May 11–13*.

Gertjan van Noord. 2006. At Last Parsing Is Now Operational. *Proceedings of TALN 2006, Leuven, Belgium, April 10–13*, pp. 20–42.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: a hybrid two-pass model. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Boulder, Colorado, USA*, pp. 13–16.

Ventsislav Zhechev and Josef van Genabith. 2010. Maximising TM performance through sub-tree alignment and SMT. *Proceedings of the Ninth conference of the Association for Machine Translation in the Americas, Denver, Colorado, USA, October 31 – November 4*, http://www.mt-archive.info/AMTA-2010-Zhechev.pdf.