Semantics-based pretranslation for SMT using fuzzy matches

Tom Vanallemeersch, Vincent Vandeghinste
Centre for Computational Linguistics, University of Leuven, Belgium

We propose a method for extending our current fuzzy matching framework with semantic information and SMT pretranslation functionality:

- Use of fuzzy metrics based on lexical semantics and on semantic roles from PropBank/NomBank
- Integration of fuzzy matches with SMT by pretranslating matching parts using word alignment or parse tree alignment
- Use of semantic roles during parse tree alignment

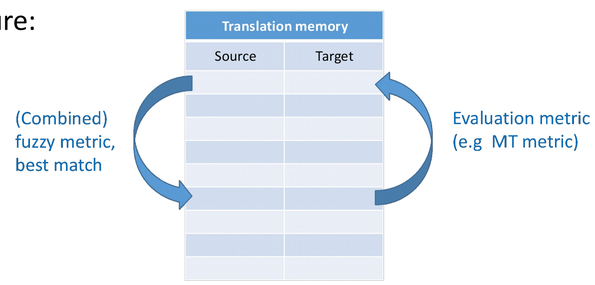
The method has been partially implemented and tested, for English-Dutch.

Fuzzy matching framework

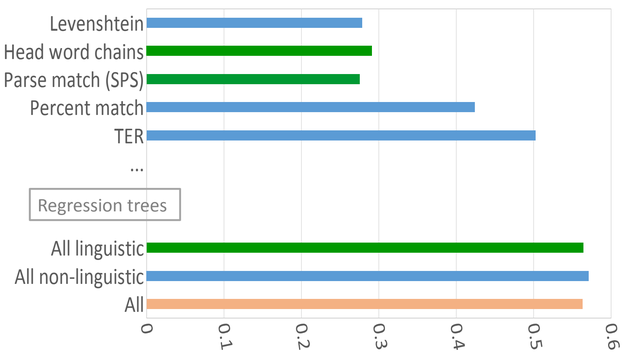
Origin: general-purpose similarity metrics, metrics for MT evaluation, …

Type: linguistically (un)aware metrics
combined metrics: regression trees with match scores as features
→ predict usability of translation of match

Evaluation procedure:



Correlation of fuzzy metrics with TER (Vanallemeersch and Vandeghinste 2015):

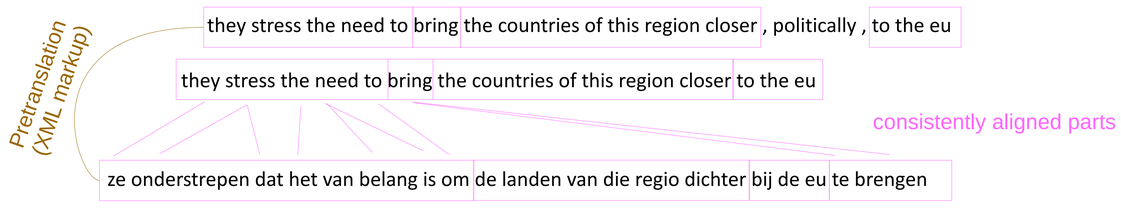


To do:
manual evaluation by translators

Integration of fuzzy matches with SMT

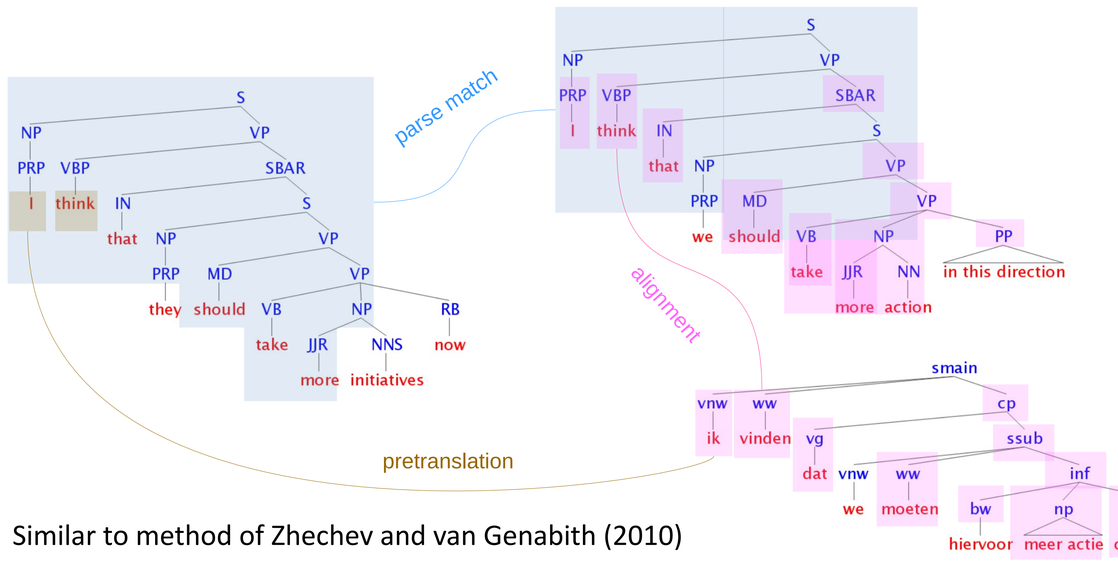
We are testing two types of alignment for determining matching parts for pretranslation.
→ Alternatively, the translation of matching parts may be used for target-language edit hints (Esplà-Gomis et al. in press).

Word alignment



Similar to method of Koehn and Senellart (2010)

Parse tree alignment



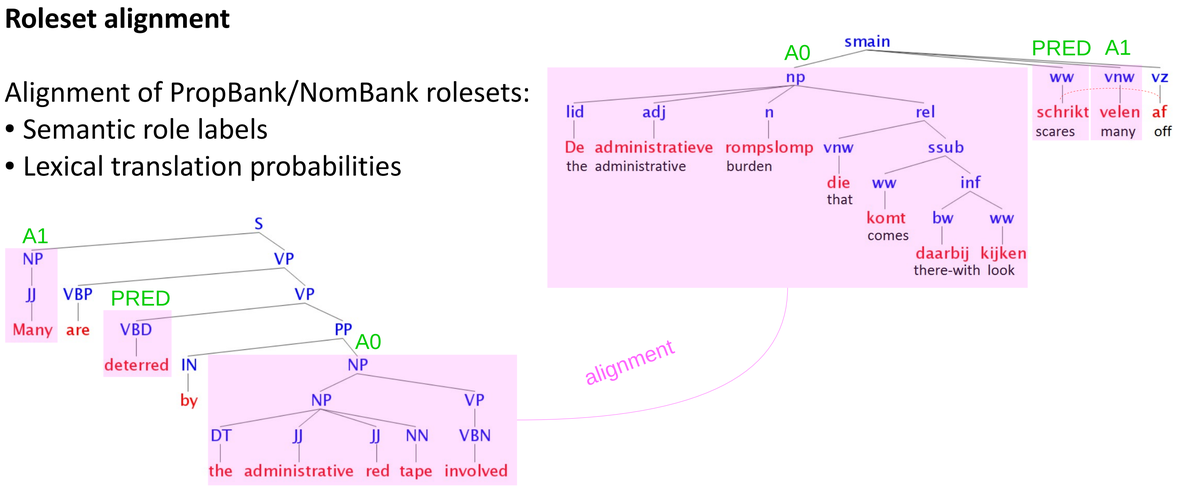
Similar to method of Zhechev and van Genabith (2010)

Semantic tree alignment

We are testing two alignment strategies which make use of semantic roles in order to approach the problem of diverging syntactic structures.

Roleset alignment

Alignment of PropBank/NomBank rolesets:
- Semantic role labels
- Lexical translation probabilities



Similar to method of Wu and Palmer (2011)

Semantic features in aligner

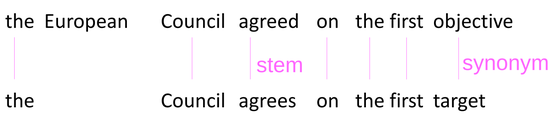
Role labels as features in discriminative tree aligner Lingua-Align (Tiedemann 2010)
→ Preliminary tests: improvement of F-score by 3% when abstracting from word alignment

Semantics-based fuzzy matching

We are testing fuzzy metrics which use two types of semantics. These metrics can also be applied for evaluating the translation of a fuzzy match.

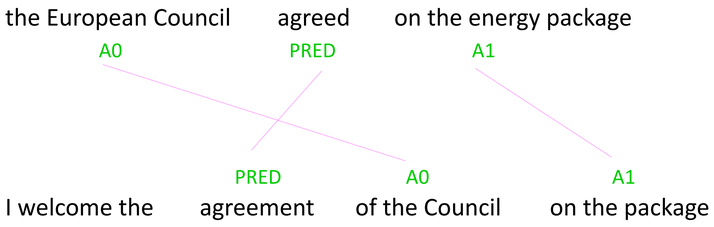
Lexical semantics

METEOR
→ Synonyms from Dutch part of EuroWordNet
→ Paraphrases from English-Dutch phrase pairs: Parex
(Denkowski and Lavie 2010; Bannard and Callison-Burch 2005)



Semantic roles

MEANT (Lo and Wu 2011)
SR metrics in Asiya toolkit (Giménez and Márquez 2010)



➡ Levenshtein:

the European Council agreed on the energy package
I welcome the agreement of the Council on the package

SRL systems

English: System of LTH (Johansson and Nugues 2008) for PropBank/NomBank
Swirl (Asiya toolkit)
Dutch : System trained on crosslingual projections from English to Dutch (Vanallemeersch 2012)

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

References

Bannard and Callison-Burch (2005) Paraphrasing with Bilingual Parallel Corpora. ACL Proceedings.

Denkowski and Lavie (2010) METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages. Proceedings ACL Workshop on SMT and MetricsMATR.

Esplà-Gomis et al. (in press) Using Machine Translation to Provide Target-language Edit Hints in Computer Aided Translation Based on Translation Memories. Journal of Artificial Intelligence Research.

Giménez and Márquez (2010) Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation. Prague Bulletin of Mathematical Linguistics.

Johansson and Nugues (2008) Dependency-based Semantic Role Labeling of PropBank. EMNLP Proceedings.

Koehn and Senellart (2010) Convergence of Translation Memory and Statistical Machine Translation. Proceedings AMTA Workshop on MT Research and the Translation Industry.

Lo and Wu (2011) MEANT: An Inexpensive, High-accuracy, Semi-automatic Metric for Evaluating Translation Utility via Semantic Frames. ACL Proceedings.

Tiedemann (2010) Lingua-Align: An Experimental Toolbox for Automatic Tree-to-Tree Alignment. LREC Proceedings.

Vanalllemeersch (2012) Parser-independent Semantic Tree Alignment. Proceedings META-RESEARCH Workshop on Advanced Treebanking, in conjunction with LREC.

Vanallemeersch and Vandeghinste (2015) Assessing Linguistically Aware Fuzzy Matching in Translation Memories. EAMT Proceedings.

Wu and Palmer (2011) Semantic Mapping using Automatic Word Alignment and Semantic Role Labeling. Proceedings ACL 5th Workshop on SSST.

Zhechev and van Genabith (2010) Maximising TM Performance through Sub-Tree Alignment and SMT. AMTA Proceedings.