Semantics-based pretranslation for SMT using fuzzy matches

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

SSST, June 2015

Introduction

We propose a method for extending our current fuzzy matching framework:

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

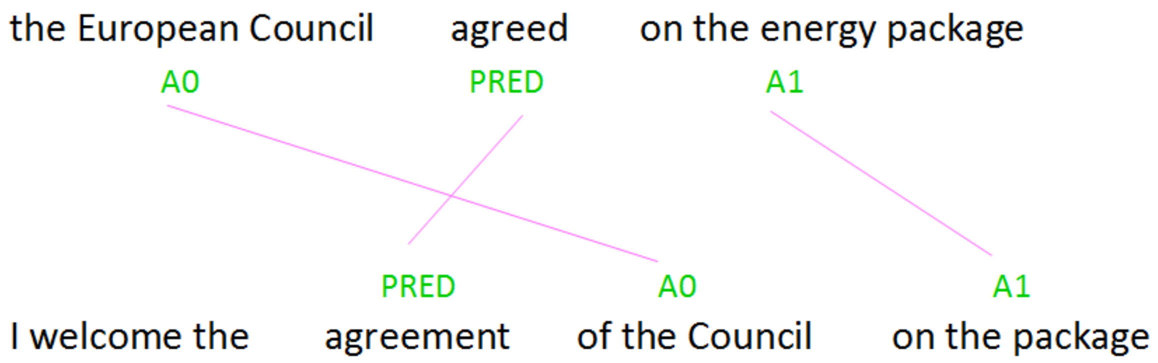
→ Partially implemented and tested, for English-Dutch

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
- Correlation of fuzzy metrics with evaluation metric

Semantics-based fuzzy matching

- Lexical semantics: METEOR
- Semantic roles:
  - ➢MEANT
  - ➢SR metrics of Asiya toolkit



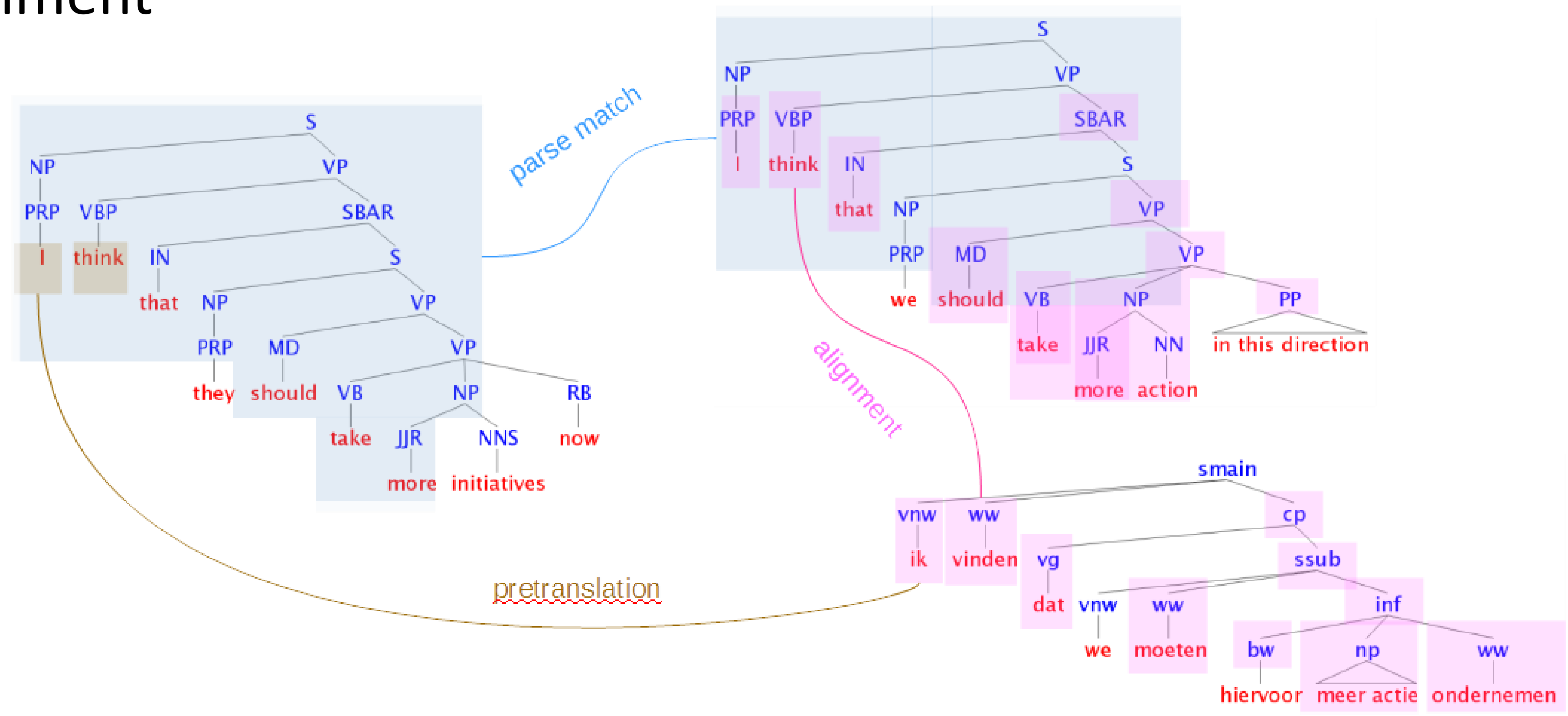
the European Council agreed on the energy package

I welcome the agreement of the Council on the package

Integration of fuzzy matches with SMT

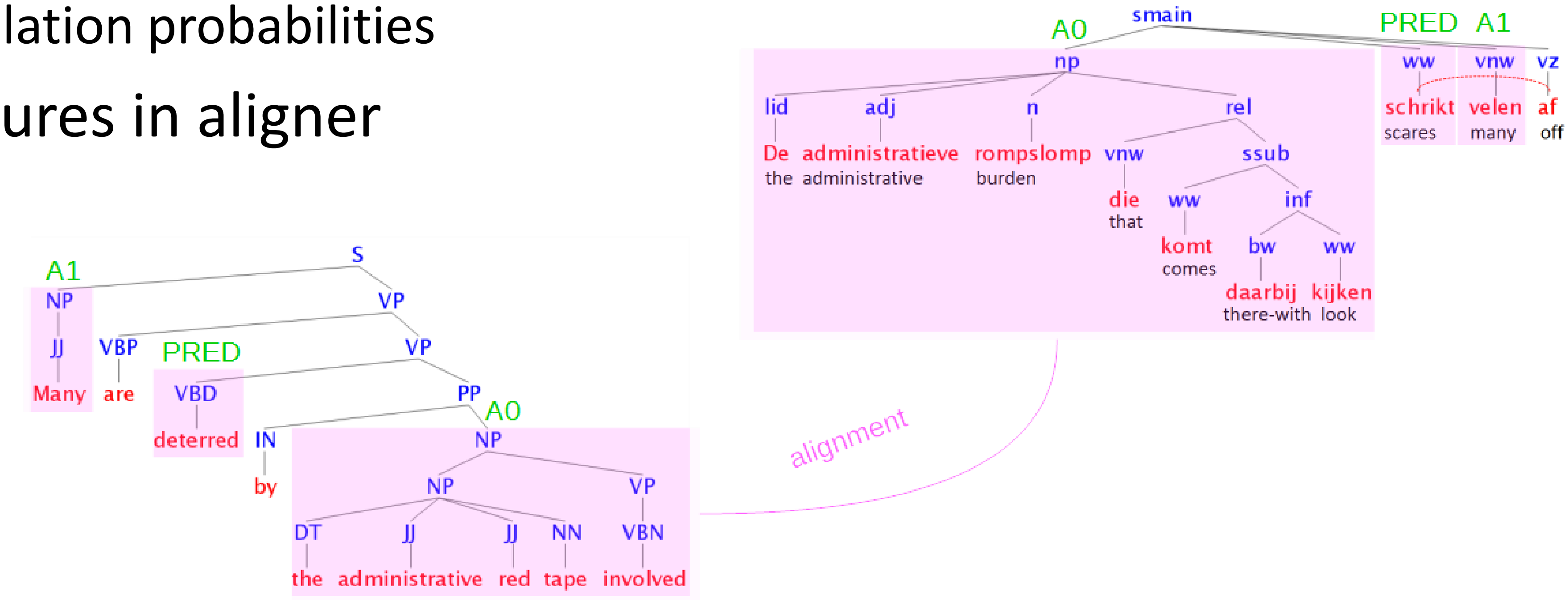
- Word alignment (consistently aligned parts)
- Parse tree alignment

→ XML markup



Semantic tree alignment

- Diverging syntactic structures
- Roleset alignment:
  - ➢PropBank/NomBank labels
  - ➢Lexical translation probabilities
- Semantic features in aligner



SRL systems

- English: LTH (Johansson and Nugues 2008) for PropBank/NomBank
- Dutch: system trained on crosslingual projections English-Dutch