Detecting Semantic Overlap

Annotating a parallel monolingual treebank with semantic similarity relations

Erwin Marsi & Emiel Krahmer

Communication and Cognition, Tilburg University, The Netherlands

{E.C.Marsi / E.J.Krahmer}@uvt.nl

INTRODUCTION

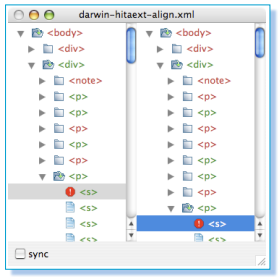
- Similar information can be expressed in many different ways
- This is an important stumbling block for applications such as question-answering, automatic summarization, information retrieval, etc.
- Resources exist on the word level (e.g., Wordnet), but are lacking for more complex phrases.
- The Stevin DAESO (Detecting and Exploiting Semantic Overlap) project intends to fill this gap.
- In the first year of the Daeso project we have collected a 1M word parallel monolingual treebank, with different text genres and different degrees of overlap (from parallel to comparable).

CORPUS MATERIALS

- Book translations: recent pairs of translations of "Le petit prince", by Antoine de Saint-Exupéry, "Les Essais" by Michel de Montaigne and "On the origin of species" by Charles Darwin.
- Autocue-subtitle pairs: from NOS Journaal (collected in Atranos project)
- News headlines: mined from Dutch version of Google news.
- QA-system output: alternative answers to medical questions from a QA evaluation corpus (collected in IMIX project)
- Press releases: different reports of the same event, collected from news agencies ANP and NOVUM.

		Manual	Available
Book translations	Darwin1	25k	154k
	Darwin 2	25k	191k
	Montaigne 1	25k	462k
	Montaigne 2	25k	~500k
	Saint-Exupéry 1	15k	15k
	Saint-Exupéry 2	15k	15k
Press releases	ANP	125k	197k
	Novum	125k	136k
QA system output		1k	1k
News headlines		24k	> 900k
Autocue-subtitles		125k	192k

CORPUS PROCESSING

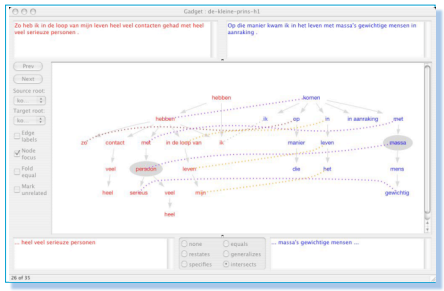


Preprocessing:

- XML TEI format (text encoding initiative)
- Tokenization with the D-Coi tokenizer for Dutch
- Dependency parsing with the Alpino parser

Alignment:

- Of paragraphs and sentences using a new alignment tool (HITAEXT) [left]. It uses a tree viewer to visualize the hierarchical structure of the XML documents and a text viewer to visualize text segments in context. Allows for 1-to-1, 1-to-n and n-to-m alignments.
- Of words and phrases using updated version of the GADGET tool [right]. Enables alignment of nodes (both leafs and intermediate ones) in dependency trees.



STATE OF AFFAIRS

- All corpus material has been collected and properly formatted.
- Linguistic preprocessing (tokenization, dependency parsing) finished.
- Sentence alignment and alignment of comparable texts (essentially) finished.
- Alignment of words and phrases within aligned sentences: ongoing.

References:

- Erwin Marsi, Emiel Krahmer and Wauter Bosma (2007). Dependency-based paraphrasing for recognizing textual entailment. In: *Proceedings of ACL-PASCAL Workshop on Textual Entailment and Paraphrasing,* Prague, pp. 83-88.
- Erwin Marsi and Emiel Krahmer (2007). Annotating a parallel monolingual treebank with semantic similarity relations, *submitted*.

Contact information

Erwin Marsi & Emiel Krahmer

Tilburg University

P.O. Box 90153

NL-5000 LE Eindhoven

The Netherlands

Phone : + 31 - 13 - 4663070

E: {e.j.krahmer / e.c.marsi}@uvt.nl

Daeso website: http://daeso.uvt.nl/