

Erwin Marsi & Emiel Krahmer (UvT) - Iris Hendrickx & Walter Daelemans (UA) - Maarten de Rijke (UvA) - Jakub Zavrel (Textkernel)

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 by:
 - ▶ Building a 1M word parallel monolingual treebank for Dutch with aligned syntactic nodes
 - ▶ On the basis of this corpus, developing software for automatic alignment and semantic relation labeling
 - ▶ Applying this software in a number of NLP task such as automatic summarization, sentence compression, QA and IR

1. TEXT SOURCES

- 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 (130k words).
- Autocue-subtitle pairs: from NOS Journaal - collected in Atranos project (125k)
- News headlines: mined from Dutch version of Google news (24k).
- QA-system output: alternative answers to medical questions from a QA evaluation corpus - collected in IMIX project (1k)
- Press releases: different reports of the same event, collected from news agencies ANP and NOVUM (125k)

TEXT PROCESSING

1. Conversion from MS Word, PDF, HTML, etc to raw text
2. XML markup (partly TEI, partly custom format)
3. Tokenization with D-COI tokenizer
4. Syntactic parsing with the Alpino parser

SEMANTIC RELATIONS

Five mutually exclusive semantic similarity relations

1. "Dementia" equals "Dementia"
2. "risk" restates "chance"
3. "daily coffee" specifies "three cups of coffee a day"
4. "three cups of coffee a day" generalizes "daily coffee"
5. "Alzheimer and Dementia" intersects "Parkinson and Dementia"

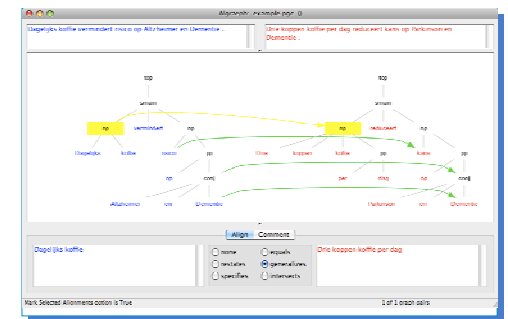
TEXT ALIGNMENT



- Text is automatically aligned down to the sentence level
- Alignment is manually corrected using the new HITAEXT tool
- The tree viewer visualizes the hierarchical structure of the XML documents
- The text viewer to visualizes text segments in context.

TREE ALIGNMENT

- Syntactic nodes are aligned and labeled with a semantic similarity relation
- Alignment is partly automatic
- Next alignment is manually corrected using the ALGRAEPH tool
- Bootstrapping: automatic alignment is further improved
- Interannotator agreement is measured



STATE OF AFFAIRS

- All corpus material has been collected, converted, marked-up, tokenized, parsed, and aligned down to the sentence level
- Tree alignment is ongoing
- Implementation of automatic alignment and multi-document summarization is ongoing
- Text and tree alignment tools are released as open source software
- IPR is settled for all text sources

References:

- Erwin Marsi, Emiel Krahmer (2007). Annotating a parallel monolingual treebank with semantic similarity relations, *International Workshop on Treebanks and Linguistic Theory*, Bergen, Norway.
- Emiel Krahmer, Erwin Marsi, Paul van Pelt (2008). Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. *ACL 2008*, Columbus, Ohio.
- Erwin Marsi, Emiel Krahmer (to appear). Detecting Semantic Overlap: A parallel monolingual treebank for Dutch (accepted for *Proceeding of CLIN 2007*).

Contact information
Erwin Marsi & Emiel Krahmer
Tilburg University
P.O. Box 90153
NL-5000 LE Eindhoven
The Netherlands

Phone : + 31 - 13 - 4663070
E: (e.i.krahmer / e.c.marsi)@uvt.nl
Daeso website: <http://daeso.uvt.nl/>

