Detecting And Exploiting Semantic Overlap (DAESO)

Emiel Krahmer
Tilburg University
Daeso is …

• … a Stevin project that will start on 1 October 2006 and will run for 3 years.

• … a follow up to the Imogen - IMIX project.

• … a collaboration between:
  - Tilburg University (Erwin Marsi, Emiel Krahmer),
  - University of Amsterdam (Maarten de Rijke),
  - Antwerp University (Walter Daelemans) and
  - TextKernel (Jakub Zavrel).
Semantic Overlap

- Steve Irwin, the TV host known as the "Crocodile Hunter," has died after being stung by a stingray off Australia's north coast. [www.cnn.com]

- Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef [www.time.com]
Daeso in brief

- Development of a parallel monolingual corpus.

- Based on this, development of tools for detecting semantic overlap.

- Evaluate the benefits of exploiting semantic overlap in practical applications.
Corpus

• Build a monolingual parallel corpus for Dutch [1M words].

• Text genres:
  - Parallel, recent translations
  - News stories describing same event
  - Answers from QA engines [IMIX]

• Manually aligned and classified.
Detecting Semantic Overlap

- **Alignment** of sentences on the basis of dependency trees (Alpino).

- **Classification** of semantic relations such as paraphrases, specifies, generalizes, etc. [Also relevant for recognizing textual entailments, e.g., RTE2, Marsi et al. 2006]

- **Fusion** of overlapping sentences for text-to-text generation.
Exploiting Semantic Overlap

- **Multi Document Summarization** [“beyond extraction”]: better, more informative summaries?

- **Information Extraction**: improved recall?

- **Question Answering**: more complete and more accurate answers?
More information

Contact: Emiel Krahmer
e.j.krahmer@uvt.nl