Universiteit van Tilburg

_textkernel

DAESO Detecting and Exploiting Semantic Overlap

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

Universiteit Antwerpen

Universiteit van Amsterdam

PROJECT OVERVIEW

BACKGROUND

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.

DAESO CORE

Corpus and tools

- Developing software for automatic alignment and semantic relation labeling
- Building a 1M word parallel monolingual treebank for Dutch with aligned syntactic nodes

Evaluation of tools

- Multi Document Summarization (UA)
- Question Answering (UvA)
- Information Extraction (Textkernel)

DAESO TOOLS

- Automatic Sentence Aligner : classify to what extent two sentences are semantically similar
- Automatic Phrase Aligner/Labeler : classify and label semantic relations between phrases
- Sentence Fuser : merge sentences carrying the same information into one sentence
- Sentence Compressor : make sentence shorter by removing less important information

REFERENCES

Erwin Marsi and Emiel Krahmer, Annotating a parallel monolingual treebank with semantic similarity relations . In: The Sixth International Workshop on Treebanks and Linguistic Theories (TLT'07), Bergen, Norway, December 7-8, 2007

Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Celebi, A., Drabek, E., Lam, W., and Liu, D. and Qi, H., Saggion, H., Teufel, S., Topper, M., Winkel, A., The MEAD Multidocument Summarizer, 2003, available at: http://www.summarization.com/mead/

MUSA project, demo available at: http://www.cnts.ua.ac.be/cgi-bin/anja/musa

Bosma, W. Discourse -oriented Summarization, PhD. Thesis, University of Twente,2008

Daeso website: http://uvt.daeso.nl
Contact: e.j.krahmer@uvt.nl / e.c.marsi@uvl.nl

FOCUS ON ANTWERP TEAM

Work in Progress

MAIN TASKS ANTWERP

A. Evaluation in Summarization

B. Sentence compression

C. Evaluation in Coreference resolution

MEAD

Public available Toolkit for Automatic Summarization en Evaluation

Basic method: compute for every sentence in the documents an importance weight. Sort sentences on their importance. Start creating a summary by adding the sentence with highest weight. Take the next important sentence and measure the similarity with the sentence that are already in the summary. If little overlap, then add sentence to summary. Repeat until maximum summary size is reached.

EVAL CORPUS DUTCH

Topic-based Multi Document Summarization

- Similar to DUC 2005-2006-2007
- 38 clusters of 5-10 news articles, 50% Daeso, 50% DCOI
- Each cluster is summarized by 5 annotators (100w and 250w summaries)
- Each sentence is also judged (3 values) for sentence extraction tasks

MEAD in DAESO

- Dutch Version of MEAD: the baseline system
- Mead + sentence compression: *compression as post-processing: add more sentences to summary*
- Mead + Sentence Aligner: *use alignment info as additional sentence weight in determining sentence importance. Sentence that is aligned partly to other sentence gets higher weight (follow-up sentence) . Sentence that is aligned completely, will recieve a low weight (redundant).*
- Mead + Phrase Aligner: *Aligner replaces the module that computes similarities between sentences to eliminate redundancy*

SENTENCE COMPRESSION

We adapt the MUSA system for Dutch and we add a paraphrase module trained on Daeso data: replace phrases by shorter semantically similar variants.

Basics of the system: Sentences are shallow parsed (pos, lemma, chunks). Rules determine which words/phases are candidates to be selected. Each candidate gets an importance weight and the least important candidates are removed.

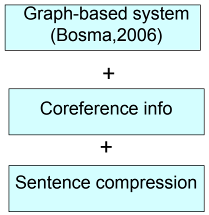
EXAMPLE

A Canadian couple on Monday stunned the Canadian Red Cross by handing over a donation of five million Canadian dollars (4.1 million US) for the tsunami relief effort in South Asia

A [4(12.44) Canadian] couple [3(12.09) on Monday] stunned the Canadian Red Cross by handing over a donation [1(11.510) of five million [5(12.44) Canadian] dollars] (4.1 million US) [6(13.46) for the tsunami relief effort] [2(11.65) in South Asia] .

A Canadian couple on Monday stunned the Canadian Red Cross by handing over a donation (4.1 million US) for the tsunami relief effort .

TAC 2008



Nodes represent sentences, arcs represent cosine similarity between sentences.

Two sentences get an extra connection in the graph when they have a coreferential relation.

(MUSA,2004) post-processing step to fit more information in the summary.