Daeso

Results of the Antwerp team

*Site visit, 10 June 2009*
Goal

Development of NLP applications for extrinsic evaluation of the Daeso corpus and Daeso tools.

-> Main focus on multi-document summarization application
Overview

1. Coreference resolution

2. Sentence compression

3. Multi-document summarization
Part 1

Coreference resolution
Coreference resolution

Can we use the Daeso corpus for coreference resolution?

Idea: pronouns or nouns that are aligned to full names/nouns can resolve coreference.

Exploratory study: count in 30 newspaper texts cases in which additional information from an aligned sentence could resolve the coreferential link. Found only two of these cases!

Conclusion: Does not work due to a lack of coverage.
Coreference resolution

- Participation in yearly automatic summarization competition (DUC/TAC) together with Wauter Bosma
- Graph-based summarization application for English
- Adding coreference information and sentence compression
- Results did not show significant improvements.
- Proceedings of Text Analysis Conference 2008
Part 2

Sentence compression
Sentence compression

Basis: sentence compression module of MUSA project at CNTS (2002-2005)

Joint work with Vincent van Asch, NEON project

Overview:
- port MUSA module to Dutch
- error analysis of compressed sentences
- paraphrases for sentence compression
Sentence compression

System description

• Sentences are shallow parsed (pos, lemma, chunks)
• Rules determine which words/phrases are candidates e.g:
  - adjectives, adverbs, rel clauses, appositions, fill words, prep phrases, bracketed parts, non-final nouns in NP
• Each candidate gets an importance weight
  \[ \text{importance weight} = \text{surprise value} + \text{rule weight} \]
• Remove least important candidate.

Universiteit Antwerpen
Original sentence:
Een toestel van Air France, met aan boord 228 mensen, onder wie 12 bemanningsleden, is maandag boven de Atlantische Oceaan voor de kust van Brazilië verdwenen.

Weighted candidates:
Een toestel [van Air France (70.13)], [met aan boord 228 mensen, onder wie 12 (90.26)] bemanningsleden, is maandag [boven de Atlantische Oceaan (70.13)] [voor de kust (70.14)] [van Brazilië (70.12)] verdwenen.

Compression (57%):
Een toestel, met aan boord 228 mensen, onder wie 12 bemanningsleden, is maandag verdwenen.
Main sources of errors:

- Multi-word expressions
- Meaningful words (*intussen, andere, slechte*)
- Word order (pp relocation)
- Shallow parser errors
  - complex prep or noun phrases
  - relative clauses
  - phrasal verbs (*denken aan*)

Universiteit Antwerpen
Sentence compression as word deletion

- Extensive corpus analysis of subtitling corpus
- Only 55% of compressed sentences are generated with a deletion model,
- 45% need more elaborate models that include word order changes and paraphrases

Published in Proceedings of the ENLG workshop, Athens, 2009
Paraphrase extraction in compression

• long-short variants
• problems: manual filtering + lack of coverage
• remedy: paraphrase patterns

In het zuiden van Irak - In zuid-Irak
In het zuiden van X - In zuid-X

• exploratory study: can we automatically extract paraphrase patterns?
• our study shows: feasible when using WWW
Future plans

- Follow up experiments for a chapter in book on state-of-the-art Natural Language Generation, Springer Publishers
Part 3

Automatic multi-document summarization
Automatic Summarization

- Multi-document summarization for Dutch
- Optimization of summarizer
- Evaluation corpus of 30 document clusters
- First experiments with integration of Daeso overlap detection
Publicly available Toolkit for Automatic Summarization and Evaluation

Basic method:
- Compute for every sentence an importance weight
- Sort sentences on their weight
- Start summary with sentence with highest weight
- Take the next important sentence
  - measure similarity with sentences already in summary
  - if little overlap, then add sentence to summary
- Repeat until maximum summary size is reached.
Evaluation Corpus

- 30 query-based document clusters
- Documents from Daeso and DCOI news data
  - Each cluster contains 5-25 relevant documents
  - Each cluster has 5 X 2 summaries (250/100 words)
  - And 5 extracts of 10 sentences.
- Funded by Antwerp BOF
- Available via INL
Optimization

- two baseline systems (random / lead)
- the effect of several features individually
- several combinations of features
- three reranker modules
- query-based versus general summarization
Features

- Position
- Length
- Centroid
- Similarity to title

Query-based summarization:
- queryWordOverlap
- queryCosine
- queryCosnoIDF
First experiments

Rerankers

- **Cosine**: compute cosine similarity score between two sentences represented as tf*idf weighted word vectors
- **MMR**: maximum margin relevance (Carbonell and Goldstein, 1998)
- **Novelty**: boost sentences following important sentence
- **Daeso**
Daeso reranker

- Similarity between a pair of sentences is estimated based on word alignments
- Word alignment uses Cornetto information *(synonyms, hyperonyms, least common subsumer)*
- Alignments as a weighted bipartite multigraph to combine the evidence
- Words are weighted with IDF
Example of Word Alignments


Exp setup

- Data set: split in 10/20 clusters for development/testing
- First: optimize automatic summarizer for Dutch news text: feature and parameter optimization
- Next: Optimized system is evaluated with the different rerankers
Exp setup

- Evaluation follows DUC 2006:
  - **Automatic evaluation with ROUGE**
    (scores system-generated summary by comparing it to 5 manually written summaries)
  - **Manual evaluation** of content and linguistic quality (average of 2 students)
## Rouge scores on test set

<table>
<thead>
<tr>
<th>Method</th>
<th>Rouge-2</th>
<th>Rouge-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand baseline</td>
<td>0.101</td>
<td>0.153</td>
</tr>
<tr>
<td>Lead baseline</td>
<td>0.139</td>
<td>0.179</td>
</tr>
<tr>
<td>Optim-cosine</td>
<td>0.152</td>
<td>0.193</td>
</tr>
<tr>
<td>Optim-mmr</td>
<td>0.149</td>
<td>0.191</td>
</tr>
<tr>
<td>Optim-daeso</td>
<td>0.150</td>
<td>0.193</td>
</tr>
</tbody>
</table>
Manual evaluation

Scores range from 5 (very good) to 1 (very poor)

<table>
<thead>
<tr>
<th></th>
<th>Gram</th>
<th>Redun</th>
<th>Refs</th>
<th>Focus</th>
<th>Struct</th>
<th>respon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>4.08</td>
<td>3.9</td>
<td>2.58</td>
<td>2.6</td>
<td>2.</td>
<td>2.25</td>
</tr>
<tr>
<td>Lead</td>
<td>3.05</td>
<td>3.6</td>
<td>3.25</td>
<td>2.88</td>
<td>2.38</td>
<td>2.4</td>
</tr>
<tr>
<td>Cosine</td>
<td>3.9</td>
<td>3.18</td>
<td>2.65</td>
<td>3.15</td>
<td>2.43</td>
<td>2.75</td>
</tr>
<tr>
<td>Mmr</td>
<td>3.98</td>
<td>3.13</td>
<td>2.55</td>
<td>3.13</td>
<td>2.38</td>
<td>2.7</td>
</tr>
<tr>
<td>Daeso</td>
<td>4.05</td>
<td>3.13</td>
<td>2.85</td>
<td>3.23</td>
<td>2.5</td>
<td>2.7</td>
</tr>
</tbody>
</table>
Discussion

- MEAD works well for Dutch
- Our first attempt did not improve upon the other rerankers

possible explanations:
- Often word overlap is enough to detect similarity
- Alignment tool makes errors
- Reranking is last step and cannot correct sentence ranking in general
Future plans

Multi-document summarization

• Re-run with improved version of Daeso reranker
  - Using tree structure
• Query-relevance estimated with Daeso aligner
• Sentence compression for summarization
• Sentence fusion for summarization
Dutch Sentence compression
http://www.cnts.ua.ac.be/cgi-bin/vincent/neondemo.py

Dutch Multi-document summarization
http://www.cnts.ua.ac.be/~iris/sumdemo.html
General conclusions(1)

• Sofar we did not show that the Daeso tools improve the presented NLP applications
  -> future plans for summarization experiments and sentence compression

• Lack of coverage of paraphrases is a problem
  - Possible solution: paraphrase pattern extraction
General conclusions(2)

- All deliverables are ready, published results
- Dutch evaluation corpus for automatic summarization
- General platforms for extrinsic evaluation are ready: multi-document summarizer, sentence compressor