Annotating a parallel monolingual treebank with semantic similarity relations

Erwin Marsi & Emiel Krahmer
Dept. Of Communication & Information / ILK
Tilburg University
Outline

• Background
  – Aligned monolingual treebanks
  – Semantic similarity relations
  – Motivation/Applicaton
  – DAESO project

• Corpus
  – Text material
  – Processing
  – Annotation tools

• Current status & future work
Aligned text space: type of text material
Aligned text space: level of alignment

Semantic form
Syntactic tree
Discontinuous phrase
Continuous phrase
Text (word, sent.)

Parallel
Comparable

December 8, 2007
Aligned text space: bilingual vs. monolingual

- Semantic form
- Syntactic tree
- Discontinuous phrase
- Continuous phrase
- Text (word, sent.)

Bilingual

Monolingual

Parallel

Comparable

December 8, 2007

TLT07

Daeso
Aligned text space: aligned monolingual treebanks

Semantic form
Syntactic tree
Discontinuous phrase
Continuous phrase
Text (word, sent.)

Aligned Monolingual Treebanks

Monolingual
Bilingual
Parallel
Comparable
Semantic similarity

- **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]
Semantic similarity: $X$ equals $Y$

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

- **Steve Irwin**, the daredevil wildlife documentarian, is killed in a *stingray* attack while filming on the Great Barrier Reef.
Semantic similarity: X restates Y

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

- Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.
Semantic similarity: X generalizes Y (Y specifies X)

• Steve Irwin, the TV host known as the "Crocodile Hunter," has died after being stung by a stingray off Australia's north coast.

• Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.
Semantic similarity: X intersects Y

• Steve Irwin, the TV host known as the "Crocodile Hunter," has died after being stung by a stingray off Australia's north coast.

• Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.
Semantic similarity relations over strings

- Five semantic similarity relations defined over pairs of substrings \((X,Y)\) from a sentence pair [Marsi & Krahmer '05]:
  1. \(X\) equals \(Y\)
  2. \(X\) restates \(Y\)
  3. \(X\) specifies \(Y\)
  4. \(X\) generalizes \(Y\)
  5. \(X\) intersects \(Y\)

- Relations are mutually exclusive

- Can easily be transferred to typed alignment relations between syntactic nodes ...
Semantic similarity relations over syntactic nodes

- Every node in a syntactic tree corresponds to a substring (phrase) of the whole sentence

STR(N3) = "w2 w3"
Semantic similarity relations over syntactic nodes

- Semantic similarity relations are defined over the strings corresponding to syntactic nodes.

```
  N1
    /   \     restates
  /     \     \
N2      N3         N2'
   |      |             |
  w1    w2    w3     w1'  w2'  w3'
```

\[\text{STR}(N3) = "w2 \; w3" \text{ restates } "w1' \; w2' \; w3'" = \text{STR}(N2')\]
Motivation

• Automatically detecting semantic similarity has many potential applications:
  – Multidocument summarization
  – Combining answers in Question-Answering systems
  – Intelligent document merging
  – Detection of plagiarism
  – Information Extraction (IE)
  – Recognizing Textual Entailment (RTE)
  – Anaphora resolution
  – Ranking parse trees
  – ...
Multi-document summarization

Given a set of similar documents

1. Rank sentences within each document on importance
2. Align similar sentences in different documents
3. Create an:
   - **Extract**: keep the most important sentences from all documents, using alignments to avoid redundancy
   - **Abstract**: keep the most important sentences from all documents, using the alignments to fuse sentences into new sentences which combine their information

[Barzilay & McKeown]
Sentence fusion

Input sentences:

1. Steve Irwin, the TV host known as the "Crocodile Hunter", has died after being stung by a stingray off Australia's north coast.
2. Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.

Intersection fusion := combine the most general information shared by both sentences (without introducing redundancy)

[162\&\text{\textordfeminine}212\&\text{\textordfeminine}230\&\text{\textordfeminine}308\&\text{\textordfeminine}][\text{\textordfeminine}Marsi&\text{\textordfeminine}Krahmer\text{\textordfeminine}05]

Example:

Steve Irwin is killed in a stingray attack off Australia's north coast.
Sentence fusion

Input sentences:

1. Steve Irwin, the TV host known as the "Crocodile Hunter", has died after being stung by a stingray off Australia's north coast.

2. Steve Irwin, the daredevil wildlife documentarian, is killed in a stingray attack while filming on the Great Barrier Reef.

Union fusion := combine the most specific information from either sentence (without introducing redundancy)

[ Marsi & Krahmer '05 ]

Example:

Steve Irwin, the TV host known as the "Crocodile Hunter" and daredevil wildlife documentarian, has died after being stung by a stingray while filming on the Great Barrier Reef.
Sentence fusion

1st sentence

2nd sentence

Analysis

Merge

Generation

Fused sentence

analyzing and aligning sentences

merging sentences

generating new sentence
The DAESO Project

- **DAESO**: Detecting And Exploiting Semantic Overlap
- Funded by the Dutch Stevin programme
- Duration: October 2006 until October 2009
- Participants:
  - Emiel Krahmer, Erwin Marsi - Tilburg University
  - Walter Daelemans, Iris Hendrickx - Antwerp University
  - Maarten de Rijke, Erik TKS - University of Amsterdam
  - Jakub Zavrel - Textkernel
- [http://daeso.uvt.nl/](http://daeso.uvt.nl/)
Goals within DAESO

1. Development of a corpus of parallel/comparable Dutch text aligned at the level of sentences, phrases and words.

2. Based on this, development of tools for automatic alignment, detection of semantic overlap and sentence fusion

3. Application of these tools in a number of NLP applications.
Overview

• Background
  – Aligned monolingual treebanks
  – Semantic similarity relations
  – Motivation/Applicaton
  – DAESO project

• Corpus
  – Text material
  – Processing
  – Annotation tools

• Current status & future work
Corpus: text material

- Text material ranges from parallel to comparable Dutch text

- The composition results from a compromise between:
  - covering different text styles;
  - availability in electronic format;
  - targeted applications;
  - strong requirements regarding copyright imposed by the research funder
**Corpus: text material for manual annotation**

<table>
<thead>
<tr>
<th>Material</th>
<th>#words:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book translations</td>
<td>125k</td>
</tr>
<tr>
<td>Autocue-subtitle pairs</td>
<td>125k</td>
</tr>
<tr>
<td>News headlines</td>
<td>24k</td>
</tr>
<tr>
<td>Press releases</td>
<td>225k</td>
</tr>
<tr>
<td>Answers from QA</td>
<td>1k</td>
</tr>
<tr>
<td>Subtotal manually annotated</td>
<td>500k</td>
</tr>
</tbody>
</table>

The degree of overlap is indicated as **High** for book translations and autocue-subtitle pairs, and **Low** for news headlines, press releases, and answers from QA.
**Corpus: automatic annotation**

<table>
<thead>
<tr>
<th>Material</th>
<th>#words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Book translations</td>
<td>125k</td>
</tr>
<tr>
<td>Autocue-subtitle pairs</td>
<td>125k</td>
</tr>
<tr>
<td>News headlines</td>
<td>24k</td>
</tr>
<tr>
<td>Press releases</td>
<td>225k</td>
</tr>
<tr>
<td>Answers from QA</td>
<td>1k</td>
</tr>
</tbody>
</table>

Subtotal manually annotated: 500k

Subtotal automatically annotated: 500k

Total corpus size: 1M
Corpus material: book translations

- Ideally: alternative translations (in Dutch) of modern books
- In practice: very hard to find (impossible?)
- We used *modern* translations of
  - "Le petit prince", by Antoine de Saint-Exupéry (integral)
  - "Les Essais" by Michel de Montaigne (partial)
  - "On the origin of species" by Charles Darwin (partial)
Corpus material: autocue-subtitles

- Autocue-subtitle material from Dutch broadcast news
- Aligned at sentence level in the ATRANOS project
- Typically subtitle is compressed version of autocue
- Filtered out nearly identical pairs by means of a simple word overlap threshold
Corpus material: news headlines

- Mined from Google News Dutch RSS feed (>900k words)
- News articles are clustered on their content, so headlines can be quite different
- Hence, we performed manual subclustering
- Skipping very long clusters
Corpus material: press releases

- Press releases from news feeds of two Dutch press agencies
- Automatic alignment of articles
  - within a time-window of 5 days
  - aiming for high recall, at the expense of precision
  - using threshold on Jaccard coefficient for word bigrams
- Subsequent manual correction
Corpus material: answers from QA

- Target: alternative/similar candidate answers in output of Question Answering (QA) engines
- Not the answers on "closed" questions
  - e.g., “Who invented the telephone?”
- But answers on “open” questions
  - e.g., “What are the advantages of using open source software?”
- However, hard to find this type of material for Dutch
- We use answers from the QA reference corpus (IMIX project)
  - medical domain
  - manually extracted from all available text material
  - for 100 Q’s
  - many have only a single answer, or no answer at all
Tokenization & parsing

• Tokenization
  – includes sentence splitting
  – with D-COI tokenizer
    (Reynaert '07)
  – errors manually corrected for book translations

• Parsing
  – all sentences parsed with the Alpino parser
    (Van Noord et al)
  – parsing errors were not manually corrected
Sentence alignment

- No sentence alignment required for
  - autocue-subtitle pairs
  - answers from QA
  - news headlines

- Alignment is required for
  - book translations (= parallel text)
  - press releases (= comparable text)

- Procedure:
  1. automatic alignment aiming at high recall (low precision)
  2. manual correction with a special annotation tool
Sentence alignment: book translation

• Task is clear:
  – in the case of parallel text, correct alignment is undisputed
• Normally one-to-one, without crossing alignments
• Many algorithms and implementations [Gale & Church '93]
• Automatic alignment
  – doing multiple passes helps
  – simple word overlap measures obtains high accuracy (>95%)
Sentence alignment: press releases

• Task is not so clear-cut:
  – Correct alignment not always evident for comparable text; see examples/discussion in the paper
• Crossing alignment are the rule; many-to-many alignments no exception
• Manual alignment
  – Small pilot-exps on interannotator agreement shows reasonable precision but low recall
  – Finding all similar sentences in two texts is a difficult task for humans
Sentence alignment: press releases

• Automatic alignment
  – Algorithms for parallel text alignment do not work for comparable text [Nelken & Shieber '06]
  – Better algorithms for comparable text alignment must be developed [Barzilay & Elhadad '03, Nelken & Shieber '06]
  – So far, we only tried simple methods relying on shallow textual features
  – Best results with tf*idf weighted cosine similarity on normalized words: F-scores around 0.6
  – However, not good enough to ease manual annotation
Sentence alignment tool: Hitaext

- Graphical tool for manually aligning two marked-up texts
- Input: two arbitrary (well-formed) XML documents
- Output: alignment between XML elements in simple format
- Supports
  - alignment at different text levels
  - many-to-many alignments
- Implementation
  - GUI in wxPython (Python + wxWidgets)
  - cross-platform (Mac OS X, GNU Linux MS Windows)
  - released as open source software
Tree alignment tool: Algraephe

- Graphical tool for aligning nodes from a pair of graphs
- Generalized reimplementiation of Gadget
  - from Alpino trees to general graphs
  - arbitrary sets of alignment relations
  - XML input/output
  - plug-in mechanism for reading and visualizing graphs of a particular format
- Implementation
  - relies on Graphviz’ Dot to render graphs
  - GUI in wxPython (Python + wxWidgets)
  - cross-platform (Mac OS X, GNU Linux MS Windows)
  - to be released as open source software
Zo heb ik in de loop van mijn leven heel veel contacten gehad met heel veel serieuze personen.

Op die manier kwam ik in het leven met massa's gewichtige mensen in aanraking.
Overview

• Background
  – Aligned monolingual treebanks
  – Semantic similarity relations
  – Motivation/Applicaton
  – DAESO project

• Corpus
  – Text material
  – Processing
  – Annotation tools

• Current status & future work
Current state of affairs

• Created
  – First 500k words in electronic format, tokenized, parsed and aligned at the sentence level, in (TEI) XML format
  – GUI tools for (correcting) manual alignment

• Working on
  – automatic sentence alignment for comparable text
    • using only shallow textual features
    • Using parse trees and lexical resources as well
  – manual alignment of dependency trees
  – multi-document summarization (abstracts) for news texts
Future work

- Development of
  - automatic alignment of dependency trees
    [Marsi & Krahmer '05]
  - sentence fusion module (alignment + merging + generation)
    [Marsi & Krahmer '05]
  - Paraphrase extraction techniques

- Application in
  - Multi-document summarization beyond extracts
  - QA
  - IE
Annotating a parallel monolingual treebank with semantic similarity relations

Erwin Marsi & Emiel Krahmer
Dept. Of Communication & Information / ILK
Tilburg University