DAESO

Detecting And Exploiting Semantic Overlap

Erwin Marsi & Emiel Krahmer

daeso

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]

Semantic Overlap

- Alternative formulation:
  a linguistically-informed similarity metric

The DAESO Project

- DAESO: Detecting And Exploiting Semantic Overlap
- Follow-up to the IMOGEN/IMIX project
- Funded by Stevin programme (2nd round)
- Duration: October 2006 until October 2009
- Participants:
  - Emiel Krahmer, Erwin Marsi (C&I, UvT)
  - Walter Daelemans, Iris Hendrickx (CNTS, UA)
  - Maarten de Rijke, Erik TKS (II, UvA)
  - Jakub Zavrel (Textkernel)
- http://daeso.uvt.nl/

Goals

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 and detection of semantic overlap.

3. Evaluation of these tools in a number of NLP applications.

Overview of Corpus Material

- Targets for manual annotation #words

Book translations	125k
Autocue-subtitle pairs	125k
News headlines	24k
Press releases	225k
Answers from QA	1k

Subtotal manually annotated	500k

- Target for automatic annotation: 500k
  ====
- Target final corpus: 1M

Corpus Material: Book Translations

- Alternative translations of modern books (in Dutch) are very hard (impossible?) to find
- 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)

- Mostly converted from electronic form (MS Word, PDF) to TEI XML
  - Requires adhoc scripting
  - Character encoding issues
- Partly scanned
  - Correcting OCR errors takes a lot of time

Corpus Material: Book Translations

- *Aan boord van de H.M.S. ` Beagle ' werd ik , als natuuronderzoeker , sterk getroffen door bepaalde feiten omtrent de verspreiding van de fauna en flora van Zuid-Amerika , en de geologische relaties tussen de tegenwoordige en de vroegere bewoners van dat continent .*

- *Toen ik als natuuronderzoeker aan boord van Zijne Majesteits Beagle vertoefde , vielen mij bepaalde feiten op omtrent de verspreiding van de levende wezens van Zuid-Amerika en de geologische relaties tussen de huidige en de vroegere bewoners van genoemd werelddeel .*

Corpus Material: Autocue-subtitles

- Autocue-subtitle material from NOS Journaal
- Aligned at sentence level in the ATRANOS project by Erik TKS
- Typically subtitle is compressed version of autocue
- Not all pairs are useful, because there are a lot of (almost) identical pairs
- Filtered by means of a simple word overlap threshold

Corpus Material: News Headlines

- Mined from Google News Dutch RSS feed (>900k words) by Wauter Bosma, 2006
- News articles are clustered on their content, so headlines can be quite different
- Performed manual subclustering
- Ignored very long clusters
- Interesting task and data set to work on...

- RSS feed also provides first sentence of article, but only first n words, so usually truncated
- Another challenge: text mining of the full articles...

Corpus Material: News Headlines

Moordenaar van Joe woensdag terug
Verdachte Belgische mp3-moord voor drie maanden vast
Adam G. ondervraagd in Brussel
Verdachte Brusselse mp3–moord in België
Verdachte moord Joe in Kasteelbrakel
Adam G. morgen voor Belgische jeugdrechter
Moord op Joe - Adam G. donderdag voor jeugdrechter
Poolse verdachte mp3-moord aan België uitgeleverd
Verdachte Belgische mp3–moord voor drie maanden vast
Verdachte mp3-moord naar België
Verdachte mp3-moord aan België uitgeleverd
MP3-verdachte ontkent
Verdachte Brusselse mp3-moord in België
Verdachte mp3-moord 3 maanden vast
Verdachte mp3-moord voorlopig vast
Verdachte moord Joe overgedragen aan Belgische autoriteiten

Corpus Material: News Headlines

Moordenaar van Joe woensdag terug

Verdachte Belgische mp3-moord voor drie maanden vast

Adam G. ondervraagd in Brussel

Verdachte Brusselse mp3–moord in België

Verdachte moord Joe in Kasteelbrakel

Adam G. morgen voor Belgische jeugdrechter

Moord op Joe - Adam G. donderdag voor jeugdrechter

Poolse verdachte mp3-moord aan België uitgeleverd

Verdachte Belgische mp3–moord voor drie maanden vast

Verdachte mp3-moord naar België

Verdachte mp3-moord aan België uitgeleverd

MP3-verdachte ontkent

Verdachte Brusselse mp3-moord in België

Verdachte mp3-moord 3 maanden vast

Verdachte mp3-moord voorlopig vast

Verdachte moord Joe overgedragen aan Belgische autoriteiten

Corpus Material: Press Releases

- Press releases mined from news feeds of ANP and Novum
- Writing robust scripts for text mining from news feeds is hard
- Automatic alignment of articles
  - Within a time-window of 5 days
  - Aiming for high recall, at the expense of precision
  - Using threshold on Jaccard coeficient for word bigrams
- Followed by manual correction
- Some issues:
  - Duplicate articles or missing articles due to script restarts
  - Many-to-many alignments (revisions of press releases)

Corpus Material: Press Releases

Lijk gevonden in woning Eindhoven

(Novum)

De politie heeft maandagmiddag het lijk van een man aangetroffen in een huis aan de Schootsestraat in de wijk Strijp in Eindhoven .

De politie gaat uit van een misdrijf , meldt een woordvoerder .

De identiteit van het slachtoffer is nog onbekend .

Dode man gevonden in Eindhovense woning

EINDHOVEN (ANP)

In een woning aan de Schootsestraat in Eindhoven is maandag het stoffelijk overschot gevonden van een man .

Dat heeft de politie gemeld .

Volgens de politie is er sprake van verdachte omstandigheden .

Corpus Material: Answers from QA

- Target: alternative/similar candidate answers extracted by QA engines
- Answers on questions of the “open” type:
  - “Who invented the telephone?” vs “What are the advantages of using open source software?”
- Hard to find this type of material for Dutch
- We use answers from the IMIX QA reference corpus
  - Medical domain
  - Manually extracted from all available text material
  - For 100 Q’s
  - Many have only a single answer, or no answer at all

Corpus Material: Answers from QA

Waardoor ontstaan aften ?

1. *De oorzaak is niet bekend , maar stress lijkt een rol te spelen ; zo kan iemand aften krijgen in de eindexamenperiode .*
2. *De oorzaak is nog niet opgehelderd en er is ook nog geen afdoende behandeling van aften bekend .*
3. *Vooral als de weerstand afneemt door ziekte , stress of vermoeidheid kunnen er aften ontstaan .*
4. *Ook ontstaan er bij vrouwen meer aften rondom de menstruatie .*
5. *Levensmiddelen die aften kunnen veroorzaken , zijn bijvoorbeeld grapefruit en kiwi's .*

Tokenization

- Tokenization and end-of-sentence detection with D-COI tokenizer
- Errors manually corrected for book translations

Parsing

- All sentences are parsed with the Alpino parser
- A few are timed out
- Parsing errors are not manually corrected

Alignment at sentence level

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

- Alignment required for
  - book translations, press releases

- Procedure:
  - Automatic alignment aiming at high recall (low precision)
  - Manual correction with special annotation tool

Alignment at sentence level: book translation

- Correct alignment is almost always clear for parallel text
- Usually one-to-one, without crossing alignments
- Automatic alignment
  - Doing multiple passes helps
  - Simple word overlap measures obtains high accuracy (>95%)

Alignment at sentence level: press releases

- Correct alignment is not evident for comparable text
- Crossing alignment are the rule; many-to-many alignments no exception
- Automatic alignment
  - So far only tried shallow text features
  - Best results with tf*idf weighted cosine similarity on character n-grams
  - However, not good enough to ease manual annotation

- Manual alignment
  - Small pilot-exp on interannotator agreement shows reasonable precision but lower recall
  - Finding **all** similar sentences in two texts is a difficult task

Hitaext

- Graphical tool for manually aligning pairs of text with XML markup
- Input: two arbitrary (well-formed) XML documents
- Output: alignment between XML elements in simple format
- Supports hierarchical alignment
- Supports many-to-many alignments
- GUI in wxPython (Python + wxWidgets)
- Cross-platform (Mac OS X, GNU Linux MS Windows)
- Released as open source software

Dependency Tree Alignment

- Aim: parallel/comparable treebank
- Alignment between nodes of the Alpino dependency trees
  - Every node corresponds to a substring of the sentence
  - Node alignment is interpreted as alignment between corresponding substrings
- Alignments are labeled according to semantic similarity relations
  1. *Plain* **equals** *plain*
  2. *Go by plain* **restates** *flies*
  3. *Go to Italy* **generalizes** *drive to Italy*
  4. *Drive to Italy* **specifies** *go to Italy*
  5. *Go by plain and train* **intersects** *fly with KLM*
- Relations are mutually exclusive

Algraeph

- Graphical tool for aligning nodes from a pair of graphs
- Generalized reimplementation of Gadget
  - from Alpino trees to general graphs
  - arbitrary sets of alignment relations
  - plug-in mechanism for reading and visualizing graphs of a particular format
  - MVC architecture
- Relies on Graphviz' Dot to render graphs
- GUI in wxPython (Python + wxWidgets)
- Cross-platform (Mac OS X, GNU Linux MS Windows)
- XML input/output
- To be released as open source software

Current status

- First 500k words in electronic format, tokenized, parsed and aligned at sentence level, in XML format
- Ready for serious work on automatic sentence alignment for comparable text
- Ready to start alignment of dependency trees
- Ready for first work on multi-document summarization (abstracts) for news texts

Future work

- Development of
  - automatic alignment of dependency trees
  - sentence fusion module (alignment + merging + generation)
  - Paraphrase extraction techniques

- Application in
  - Multi-document summarization beyond extracts
  - QA (UvA)
  - IE (Textkernel)