WeScience Corpus

Gisle Ytrestøl

July 16, 2009

Why Another Treebank?

- Not that many around (Penn Treebank, Redwoods Treebank)
- Encyclopedic texts (Wikipedia)
- *ACL-IJCNLP 2009 Workshop on: The Peoples Web Meets NLP*

`http://www.ukp.tu-darmstadt.de/acl-ijcnlp-2009-workshop/`

ACL workshop papers (examples)

Full papers:

- *Construction of Disambiguated Folksonomy Ontologies Using Wikipedia* (Noriko Tomuro and Andriy Shepitsen)
- *Named Entity Recognition in Wikipedia* (Dominic Balasuriya, Nicky Ringland, Joel Nothman, Tara Murphy and James R. Curran)

Short papers:

- *Automatic Content-based Categorization of Wikipedia Articles* (Zeno Gantner and Lars Schmidt-Thieme)
- *Evaluating a Statistical CCG Parser on Wikipedia* (Matthew Honnibal, James Curran and Joel Nothman)

WeScience

- All articles from a Wikipedia snapshot from 2008-Jul-27 15:43:45
- Automatic selection of articles
- 100 articles, 270,000 tokens, 14,000 sentences/lines

WeScience

- Redwoods grammar-based annotation approach (HPSG)
- Treebanking is carried out by Dan Flickinger (CSLI, Stanford)
- Used for training the oracle for the incremental deterministic HPSG parser (PhD project)

Domain-specific Selection

- Impartial selection of most relevant articles within NLP from Wikipedia
- Seed articles: Articles classified under the category *Computational Linguistics* and its sub-categories
- Simple link analysis: Wikipedia articles with the most cross-references from seed articles assumed most important

Restrictions

- Article must be of a certain length (more than 2000 characters in Wikipedia Markup Syntax)
- Articles concerning a year has been removed (only none-automatic interference in the article selection)
- A few somewhat off-topic articles are included (e.g. *USA* and *Microsoft Windows*).

How it was done

- Offline Wikipedia reader on a local machine
- The entire automatic selection is carried out by a number of Python scripts
- Reproducible given the same Wikipedia snapshot
- Documented in the Technical Summary

`http://wiki.delph-in.net/moin/WeScience`

Desired format

- Same format as the Redwoods Treebank — textual, line-oriented identifier prefix for each sentence
- Allows for concatination (as oppose to XML format)
- Stripped for unwanted section/markup from the source text
- Aim to preserve all markup that eventually may be important for linguistic analysis
- Eliminate markup which is linguistically irrelevant (e.g. meta information)

Wiki Markup Syntax

- Markup language that facilitates on-line rendering (as HTML) for display in a web browser
- Wikipedia guidelines aim to keep the architecture and design as consistent as possible
- These guidelines are not always followed

Markup that contributes to the linguistic content

Hyperlinks and item in bulleted list:

(1) [10120240] |* Design of [[parser]]s or [[phrase chunking|chunkers]] for [[natural language]]s

Italic:

(2) [10621290] |For example, in the following example, ''one'' can stand in for ''new car''.

Unwanted markup and content

- Entire sections: (Footnotes, references, bibliography, etc)
- Metadata like picture insertion, comments in the text etc.
- These parts of the source file are removed by regular expressions in Python scripts

Unwanted markup and content

- Comments:

The total initial investment raised for the new company <! − − **when is 'eventually'? 1998? 2008?** −− > *amounted to almost* . . . (from the *Google* Article)

- Sections:

==Bibliography==
* Christopher D. Manning, Hinrich Schütze, "Foundations of Statistical Natural Language Processing", MIT Press: 1999. ISBN 0-262-13360-1.

Sentence Segmentation

- All linebreaks removed from Wiki Source Markup
- Used `tokenizer 1.0` (`http://www.cis.uni-muenchen.de/~wastl/misc/`) to insert new sentence boundaries
- Various regular expressions were used to improve the performance of the sentence segmentation stage
- Reoccurring challenges for the Sentence Segmentation:
  - bulleted lists without periods in end of sentences
  - large mathematical formulas in the middle of sentences

The WeScience Format

- Same format as the Redwoods Treebank
- Each sentence starts with an identifier—an eight-number digit

Placeholder	Article Number	Sentence number	Decimal
1	065	049	0

Table: Identifier for sentence 10650490.

- *[10650490] |These rules can be formally expressed with [[attribute grammar]]s.*
- The decimal is initially set to 0. If there is further need to manually split the sentence, the decimal of the following new sentence(s) will be incremented.

Numbers and statistics

- WeScience corpus is divided into 16 section, with at the most 1,000 sentence in each section
- An article is never split between two sections
- Average number of tokens in each sentence is 17.9

ERG Parsing Result

- Basic parsing coverage of the corpus with the ERG reached 86 percent
- Expected to improve when the ERG grammar is fitted to the domain
- Average parse times per sentence (to produce up to 500 analyses for treebanking) of just below five seconds
- For more information on the WeScience Treebank, go to `http://www.delph-in.net/wescience/`