A Syntactic and Lexical-Based Discourse Segmenter

Milan Toñiloski  Julian Brooke  Maite Taboada
School of Computing Science  Department of Linguistics  Department of Linguistics
Simon Fraser University

Introduction
SLSeg – Syntactic and Lexical Segmenter
Discourse segmentation for discourse parsing
• Finding: elementary discourse units (EDUs)
• Breaking test into segments
• Breaking segments into clauses
Quality EDUs critical in building quality discourse representations (Soricut and Marcu 2003)
Final goal
• Build a discourse segmenter that is robust in handling formal (newswire) and informal (online reviews)

Segmentation Principles
Capture interesting relations (Condition, Evidence, Purpose) rather than all relations
“interesting” in terms of informativeness and with a view towards applications
Applications
• Summarization
• Sentiment detection (Taboada et al. 2009, Brooke et al. 2009)

Discourse Parsing
Build a tree for a text, capturing relations across clauses and sentences
In our case, based on Rhetorical Structure Theory
• Bottom-up, with lexically-marked relations across clauses first
• Adding relations across sentences as we find them
• Text example from the RST web site (www.sfu.ca/rst)

Examples
1. While looking over a Scottsdale, Ariz., model house one day,
2. my wife was amused by the real-estate agent,
3. who engaged her in the “you talk kinda funny” conversation.
1. Adjust clause
2. Main clause
3. Non-restrictive relative clause

Implementation of SLSeg
1. Sentence segmentation with NIST’s breaksent
2. Part-of-speech tagging and syntactic parsing (Charniak parser)
3. 12 syntactic segmentation rules
4. A few lexical rules
5. Word-level part-of-speech tags
6. Wrong boundaries removed
7. Discourse markers that resemble sentences (if you will)
8. Segmentation within parentheticals as well

Data
Nine human-annotated texts
• 3 from RST literature (RST web site)
• 3 on-line product reviews (Epinions)
• 3 Wall Street Journal articles (RST Discourse Treebank)
Average length 21.2 sentences
• Longest 43 sentences
• Shortest 6 sentences
Total 191 sentences, 340 discourse segments (EDUs)

Evaluation
F-score
• Precision
• Recall
• Number of boundaries in agreement with gold standard

Results
Higher precision in combined (formal and informal texts)
Parser-independent (similar performance for both Charniak and Stanford parsers)

Contribution
SLSeg – Consensual discourse segmenter
• Higher precision compared to a statistical parser
• No significant loss in recall (high F-score)
• No training needed for a new domain, unlike statistical parsers
• SLSeg could assist in manual annotation, by providing discourse segments as starting point

Qualitative Comparison
Luckily we bought the extended protected plans from Lowe’s, so we are waiting for whirlpool to decide if they want to do the costly repair or provide us with a new machine

References and Acknowledgements

Contact: Maite Taboada – mtaboada@sfu.ca – http://www.sfu.ca/~mtaboada