# A Syntactic and Lexical-Based Discourse Segmenter

**Milan Tofiloski** — School of Computing Science
**Julian Brooke** — Department of Linguistics
**Maite Taboada** — Department of Linguistics
Simon Fraser University

## Introduction

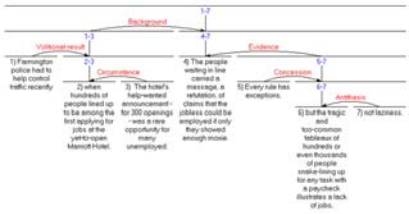
**SLSeg** – Syntactic and Lexical Segmenter
Discourse segmentation for discourse parsing
- Finding elementary discourse units (EDUs)
  - Breaking text into sentences
  - Breaking sentences into clauses

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) texts

Segmentation principles based on syntactic and lexical information

## 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)




Previous work
- SPADE, sentence-level parser which performs segmentation
- Subba and Di Eugenio (2007)
- Thanh et al. (2004)

## 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 segment candidates
- Clauses and sentences
- Coordinated clauses (not coordinated VPs)
- Adjunct clauses (finite or non-finite)
- Non-restrictive relative clauses marked by commas
- All discourse segments must contain a verb

NOT discourse segments
- Clausal complements
- Complements of attributive and cognitive verbs
- Restrictive relative clauses

## Examples

1. While looking over a Scottsdale, Ariz., model house one day,
2. my wife was amused by the real-estate agents,
3. who engaged her in the "you talk kinda funny" conversation.

1. Adjunct clause
2. Main clause
3. Non-restrictive relative clause

1. Definitely one we will buy on DVD
2. to be able to watch later snuggled on the couch

1. Main clause
2. Purpose non-finite adjunct clause

1. The thing that caught my attention was the fact that these fantasy novels were marketed to kids in the UK, but to adults in North America.

1. Main clause with two embedded clauses, neither one a discourse segment

## 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
   - Stop phrases
   - Discourse cue phrases
   - Word-level part-of-speech tags
5. Wrong boundaries removed
   - Discourse markers that resemble sentences (*if you will*)
6. Segmentation within parentheticals as well

## Data

Nine human-segmented 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
  - Number of boundaries in agreement with gold standard
- Recall
  - Number of correct boundaries divided by number of boundaries in gold standard

SLSeg compared to
- SPADE (Soricut and Marcu 2003)
- SUNDANCE parser (Riloff and Phillips 2004)
  - Would a general-purpose parser suffice for our purposes?
- Baseline
  - Segmentation after S, SBAR, SQ, SINV, SBARQ

## 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

**SPADE output**

1. Luckily we bought the extended protected plans from Lowe's,
2. so we are waiting
3. for whirlpool to decide
4. if they want to do the costly repair
5. or provide us with a new machine

3. Object clause
4. Object clause
5. Coordination within an embedded (object) clause

**SLSeg output**

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

2. No need for further segmentation; no discourse relations within the segment

## Results

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

| | Epinions | | | Treebank | | | Original RST | | | Combined Total | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| System | P | R | F | P | R | F | P | R | F | P | R | F |
| Baseline | .22 | .70 | .33 | .27 | .89 | .41 | .26 | .90 | .41 | .25 | **.80** | .38 |
| SPADE (coarse) | .59 | .66 | .63 | .63 | **1.0** | .77 | .64 | .76 | .69 | .61 | .79 | .69 |
| SPADE (original) | .36 | .67 | .46 | .37 | **1.0** | .54 | .38 | .76 | .50 | .37 | .77 | .50 |
| Sundance | .54 | .56 | .55 | .53 | .67 | .59 | .71 | .47 | .57 | .56 | .58 | .57 |
| SLSeg (Charniak) | **.97** | .66 | **.79** | **.89** | .86 | **.87** | **.94** | .76 | **.84** | **.93** | .74 | **.83** |
| SLSeg (Stanford) | .82 | **.74** | .77 | .82 | .86 | .84 | .88 | .71 | .79 | .83 | .77 | .80 |

Table 1: Comparison of segmenters

## Contribution

SLSeg – Conservative 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

All data and software available:
- http://www.sfu.ca/~mtaboada/research/SLSeg.html

## References and Acknowledgements

Lynn Carlson, Daniel Marcu and Mary E. Okurowski.2002. *RST Discourse Treebank. Philadelphia, PA: Linguistic Data Consortium.*
Eugene Charniak. 2000. A Maximum-Entropy Inspired Parser. *Proc. of NAACL, pp. 132–139. Seattle, WA.*
Dan Klein and Christopher D. Manning. 2003. Fast Exact Inference with a Factored Model for Natural Language Parsing. *Advances in NIPS 15 (NIPS 2002), Cambridge, MA: MIT Press, pp. 3–10.*
William C. Mann and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text, 8:243–281.*
Daniel Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, MA.*
Ellen Riloff and William Phillips. 2004. *An Introduction to the Sundance and AutoSlog Systems. University of Utah Technical Report #UUCS-04-015.*
Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. *Proc. of HLT-NAACL, pp. 149–156. Edmonton, Canada.*
Rajen Subba and Barbara Di Eugenio. 2007. Automatic Discourse Segmentation Using Neural Networks. *Proc. of the 11th Workshop on the Semantics and Pragmatics of Dialogue, pp. 189–190. Rovereto, Italy.*
Huong Le Thanh, Geetha Abeysinghe, and Christian Huyck. 2004. Automated Discourse Segmentation by Syntactic Information and Cue Phrases. *Proc. of IASTED. Innsbruck, Austria.*

**Acknowledgements**
This work was supported by an NSERC Discovery Grant (261104-2008) to Maite Taboada. We thank Angela Cooper and Morgan Mameni for their help with the reliability study.