Analyzing Context using Natural Language Parsers for Sentence Boundary Disambiguation

Andrew Tjia

hung_yao@cs.ubc.ca

Sentence Boundaries

- Terminators marking the beginning and end of sentences
  - In English, we can terminate sentences with periods (.), question marks (?), and exclamation points (!)
- First step in many NLP tasks
  - Parts of speech tagging, grammar, summarization, sentence alignment

Existing Approaches

- Hand-crafted rules
  - (Aberdeen et al., 1995)
- Satz
  - (Palmer and Hearst, 1997)
- mxTerminator
  - (Reynar and Ratnaparkhi, 1997)
- Splitta
  - (Gillick, 2009)

Existing Approaches

Method	Error Rate
Hand-crafted rules	0.9%
Satz	1.0%
mxTerminator	1.2%
Splitta	0.25%

N-gram models

- Used NLTK's nltk.model.ngram.NgramModel.
- With word level n-grams, use training and test sets, to train and test an n-gram model

ParseReduce

- Segment sentence into fragments
- Join fragments that give a better PCFG score combined than individually

- $If\ P(A.B.) \geq P(A.) \times P(B.)$:
  - Is Nonterminator
- $Else$:
  - Is Terminator

ParseReduce

Original: *The suit was filed by plaintiffs' securities lawyer Richard D. Greenfield in U.S. District Court in Philadelphia.*

Segmented:

The suit was filed by plaintiffs' securities lawyer Richard D. Greenfield in U.S.
Score = -104.303

District Court in Philadelphia.
Score = -36.6166

Combined:

The suit was filed by plaintiffs' securities lawyer Richard D. Greenfield in U.S. District Court in Philadelphia.
Score = -129.032

Stanford Parser

Input: The remainder of the debt will be exchanged for new Costa Rican bonds with a 6 1/4% interest rate.

Most likely parse tree:

```
(ROOT
  (S
    (NP
      (NP (DT The) (NN remainder))
      (PP (IN of)
        (NP (DT the) (NN debt))))
    (VP (MD will)
      (VP (VB be)
        (VP (VBN exchanged)
          (PP (IN for)
            (NP (JJ new)
              (ADJP (JJ Costa) (JJ Rican))
              (NNS bonds)))
          (PP (IN with)
            (NP (DT a)
              (NP (CD 6 1⁄4) (NN %))
              (NN interest) (NN rate))))))
    (. .)))
```

Score: -122.74738311767578

Corpora

10% Penn Treebank	
Class of SBD	*Count*
Terminators	3848
Nonterminators	2487
RST Corpus	
Class of SBD	*Count*
Terminators	6459
Nonterminators	4383
Combined (10% Penn + RST)	
Class of SBD	*Count*
Terminators	10307
Nonterminators	6870

Splitta

- (Gillick, 2009)
- Best of breed feature based supervised classifier
- 0.25% error rate trained over WSJ and Brown corpora

Results

Comparison of SBD Classifiers	
Classifier	*Overall Error Rate*
ParseReduce	1.24653%
Splitta	0.343885%
N-grams	12.0474%
Baseline	39.9953%

Examples of errors

- ParseReduce: “The information on 125 metropolitan markets is supplied by retailers such as Sears, Roebuck & Co. and K mart Corp. as well as closely held concerns such as R.H. Macy & Co. The council plans to release its regional reports monthly.<S>”

Examples of errors

- Splitta: “Takuma Yamamoto, president of Fujitsu Ltd., believes the `money worship' among young people . . .<S> caused the problem.”

Common errors

- Sentence fragments
- Ellipsis – neither Splitta nor ParseReduce handles this very well

Error Correlation

Errors on Treebank	
Type of error	*Proportion of errors*
Uncorrelated errors	97%
Correlated errors	3%
Errors on RST Corpus	
Type of error	*Proportion of errors*
Uncorrelated errors	96%
Correlated errors	4%

Conclusion

- Higher error than Splitta, but…
- Showed low correlation between errors
- Different approach which generalizes to any corpora so as long a parser exists

Future Work

- Perform grammar induction for unsupervised sentence boundary disambiguation
- Modify parser grammar for terminator/nonterminator disambiguation
- Evaluate ParseReduce against entire WSJ test corpus
- Extend technique to other parsers and other corpora, possibly in other languages

References

- Gillick, D. (2009). Sentence Boundary Detection and the Problem with the U.S. Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, (pp. 241-244). Boulder, Colorado.
- Kiss, T., & Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Disambiguation. *Computational Linguistics, 32*(4), 485-525.
- Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of English: the penn treebank. *Computational Linguistics, 19*(2), 313-330.
- Palmer, D., & Hearst, M. (1997). Adaptive Multilingual Setence Boundary Disambiguation. *Computational Linguistics, 23*(2), 241-267.
- Reynar, J., & Ratnaparkhi, A. (1997). A maximum entropy approach to identifying sentence boundaries. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, (pp. 16-19).