Assessing the Relative Reading level of Sentence Pairs for Text Simplification

Sowmya Vajjala and Detmar Meurers
LEAD Graduate School and Department of Linguistics, University of Tübingen

In a Nutshell

Our Questions

- Can readability models be **effective at the sentence level**?
- Can we **evaluate simplification** using readability models?

Practical Relevance

- identify targets for text simplification
- evaluate simplification approaches

Approach

- Build a **document-level model** and evaluate it using a standard dataset.
- Apply the model to **sentences** using normal and simplified Wikipedia.

Corpora

1. (a) Document-level model: WeeBit corpus (Vajjala & Meurers, 2012)
   (b) Standard dataset for testing: Common Core Standards corpus
2. Sentence level: Wiki (**N**)-SimpleWiki(**S**) sentence aligned corpus (Zhu et al., 2010)

Modeling

- We model readability as regression (1–5).
- Evaluation of the document-level model:
  - Pearson Correlation (r) and Root Mean Square Error (RMSE)
- Evaluation of the model for sentences:
  - How well does the model identify S<N in the Wiki corpus?

Features

- Lexical Features
  - lexical richness features
    - TTR, noun variation, ...
  - POS density features
    - # nouns/# words, ...
  - traditional features and formulae
    - Flesch-Kincaid score, ...
- Syntactic Features
  - complexity features from SLA
    - # dep. clauses/clause, ...
  - other parse tree features
    - # NPs per sentence, ...
- Morphological properties of words
- Avg. # of senses per word (WordNet)
- Psycholinguistic features
  - Average Age-of-Acquisition of words (Kuperman et al., 2012)
  - word abstractness, ...

Document Level Model

- Machine learner: SMOReg (WEKA)
- **Result** (WeeBit, 10-fold cross-validation) $r = 0.92$, RMSE $= 0.53$
- **Result on Common Core Standards**

	Spearman
Our System	**0.69**
Nelson et al. (2012):	
REAP	0.54
ATOS	0.59
DRP	0.53
Lexile	0.50
Reading Maturity	**0.69**
SourceRater	**0.75**

	Pearson r
Our System	0.61
Flor et al. (2013):	
Lexical Tightness	-0.44
Flesch-Kincaid	0.49
Text Length	0.36

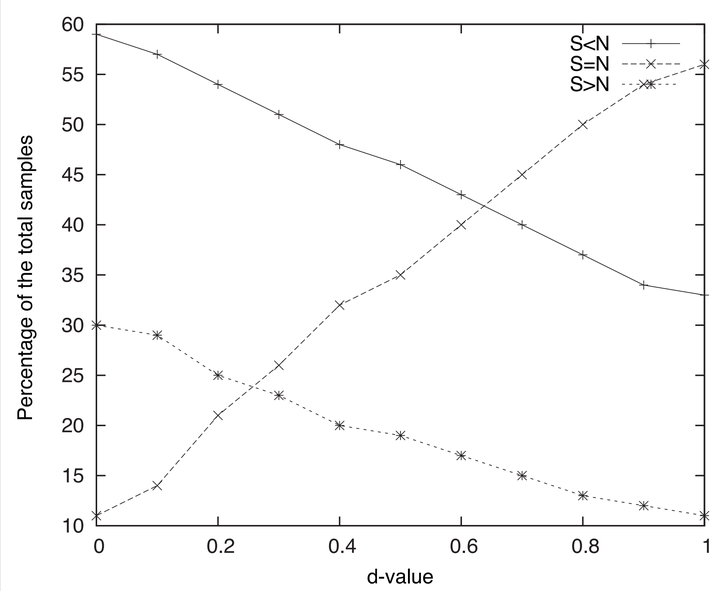
Sentence-Level Model

- We started with training a sentence-level readability model on Wiki-SimpleWiki:
  - Binary classification: simple – hard
    - 65–68% accuracy, irrespective of the training sample size
  - As regression: $r = 0.4$
- Poor performance raises two questions:
  1. Is simplification better understood as relative (instead of absolute)?
  2. Which features work best at sentence level?

Is Simplification Relative?

- How can we show this?
  - compute reading levels of normal (N) and simplified (S) sentences
  - evaluate classification using S<N, S=N and S>N
- How big must |S−N| be to interpret it as a categorical difference in reading level? → We call this the *d-value*.

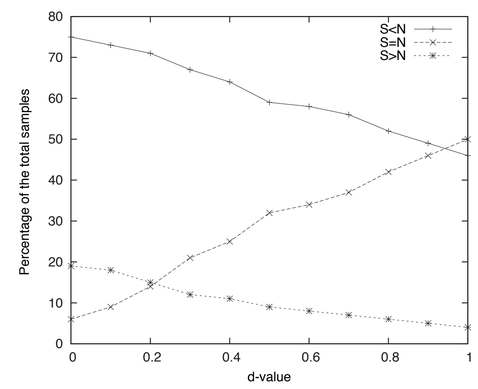
Influence of *d*



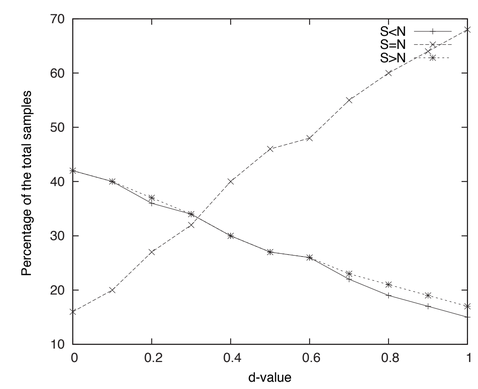
Influence of *N*

How does the reading level of the unsimplified sentence (N) affect the results?

Results for harder sentences (N >=2.5):



Results for simpler sentences (N<2.5):



⇒ It is difficult to identify simplifications for an already simple sentence.

⇒ Works well for more complex sentences.

Summary

- Relative analysis of readability better for sentence-level comparison.
- The accuracy of comparison depends on:
  1. minimum |S−N| required to identify a categorical difference (**d**).
  2. reading level of the original, unsimplified sentence (**N**).
- Our approach works well for complex sentences.

Future Work

- Explore automatic identification of transformations for simplification.
- Explore ranking or ordinal regression instead of linear regression.

References

Flor, M., B. B. Klebanov & K. M. Sheehan (2013). Lexical Tightness and Text Complexity. In *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility*.

Kuperman, V., H. Stadthagen-Gonzalez & M. Brysbaert (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4), 978–990. URL http://crr.ugent.be/archives/806.

Nelson, J., C. Perfetti, D. Liben & M. Liben (2012). *Measures of Text Difficulty: Testing their Predictive Value for Grade Levels and Student Performance*. Tech. rep., The Council of Chief State School Officers.

Vajjala, S. & D. Meurers (2012). On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In J. Tetreault, J. Burstein & C. Leacock (eds.), *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*. Montréal, Canada: Association for Computational Linguistics, pp. 163—-173. URL http://aclweb.org/anthology/W12-2019.pdf.

Zhu, Z., D. Bernhard & I. Gurevych (2010). A Monolingual Tree-based Translation Model for Sentence Simplification. In *Proceedings of The 23rd International Conference on Computational Linguistics (COLING), August 2010. Beijing, China*.