Introduction

Discourse factors Analysis

Rhetorical Structure Theory (RST) Tree

Original

Once the deal is final, they will end up owning about 23 percent of the company.

Charter had pursued Time Warner Cable for months, but Time Warner Cable CEO Rob Marcus had consistently rejected what he called a lowball offer.

Simplified

Charter had tried to buy Time Warner Cable for months.

Time Warner Cable CEO Rob Marcus kept saying "no."

Discourse Connectives

Sentences governed by elaboration relation are more likely to be deleted.

Explanations are more likely to keep.

Sentences near root are mostly kept.

Features and Modeling

Document characteristics

• Number of tokens

• Topics

Discourse features

• Indicator of connectives

• Governing relation

Position features

• Sentence position

• Paragraph position

Semantic features

• 300D Glove Embeddings

Dataset and Evaluation

• Train: 42,264 sentences in 886 articles automatically aligned (Sent2Vec).

• Dev/Test: 450/1838 sentences from 50 manually aligned articles.

Middle school level is harder to predict.

Both sparse features and semantic information from word embeddings help.

FFNN+Gaussian Layer works best.

Corpus

• 50 article sets from the Newsela dataset.

• Each article is rewritten into 4 reading levels by professional editors.

• Sentence alignments for all level pairs.

• Inter-annotator agreement is 0.807 by Cohen’s Kappa.

Would Twain use Twitter to bemoan the deplorable state of the press?

Would Twain use Twitter to complain about the sad state of the press?

Middle School  Elementary School

17% 45% 55% 83%

17% 45% 55% 83%

Middle School  Elementary School

• Studied how sentences are deleted while a document is rewritten into lower levels.

• Crucial for document-level simplification.

• Less prior work and lack of good data.

Middle School  Elementary School

55% 45% 83% 17%

55% 45% 83% 17%

References


• Maddela, M. and Xu, W., 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In EMNLP.

Contribution

• Manually annotated corpus for document-level text simplification.

• Discourse-level factors are associated with sentence deletion.

• Discourse-level factors contribute to the challenging sentence deletion task.