THE OHIO STATE UNIVERSITY

Discourse Level Factors for Sentence Deletion in Text Simplification

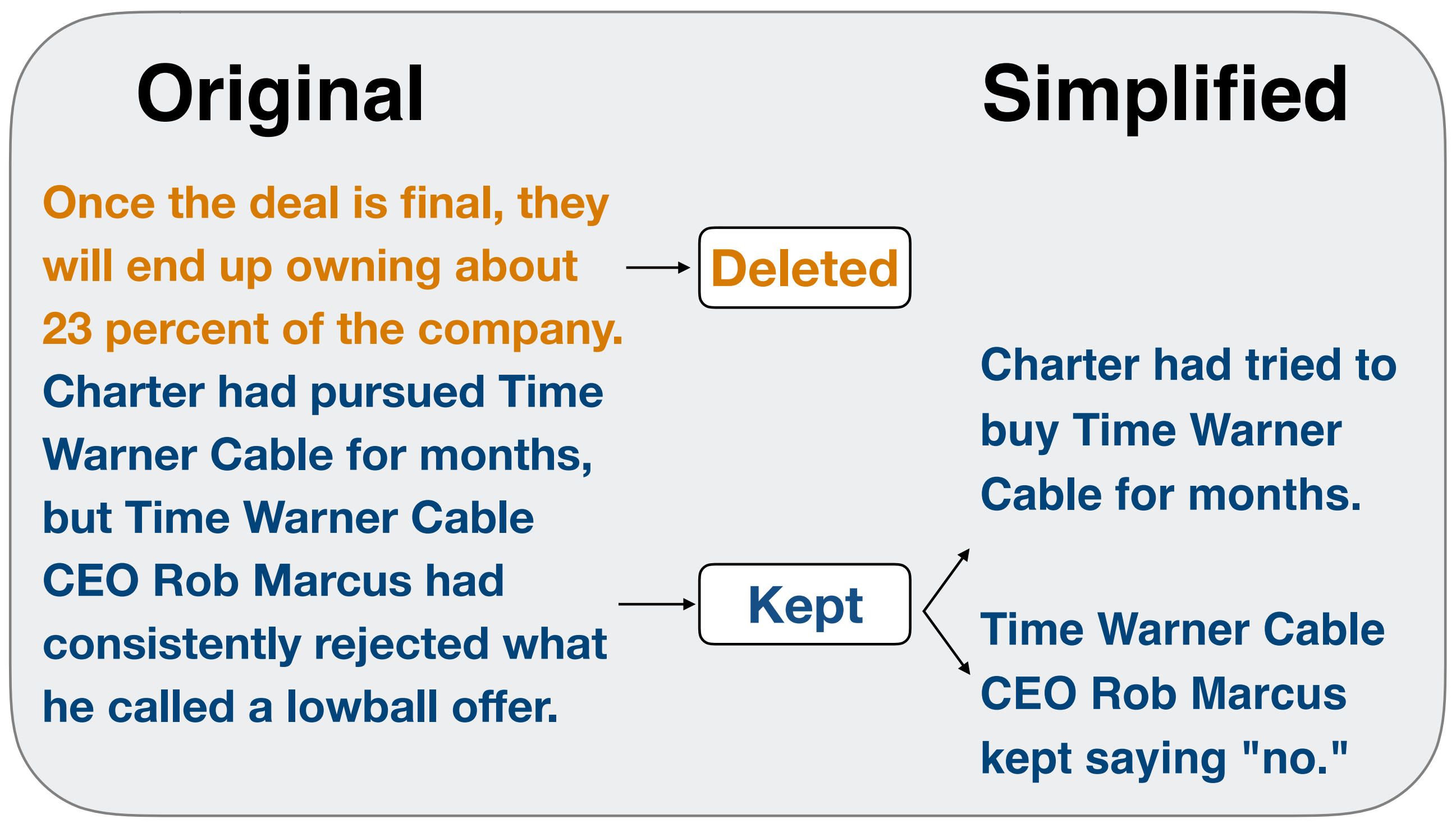
Yang Zhong[1],Chao Jiang[1], Wei Xu[1] and Junyi Jessy Li[2]

zhong.536@osu.edu, jiang.1530@osu.edu, weixu@cse.ohio-state.edu, jessy@austin.utexas.edu

[1]The Ohio State University, [2]The University of Texas at Austin

TEXAS
The University of Texas at Austin

Introduction



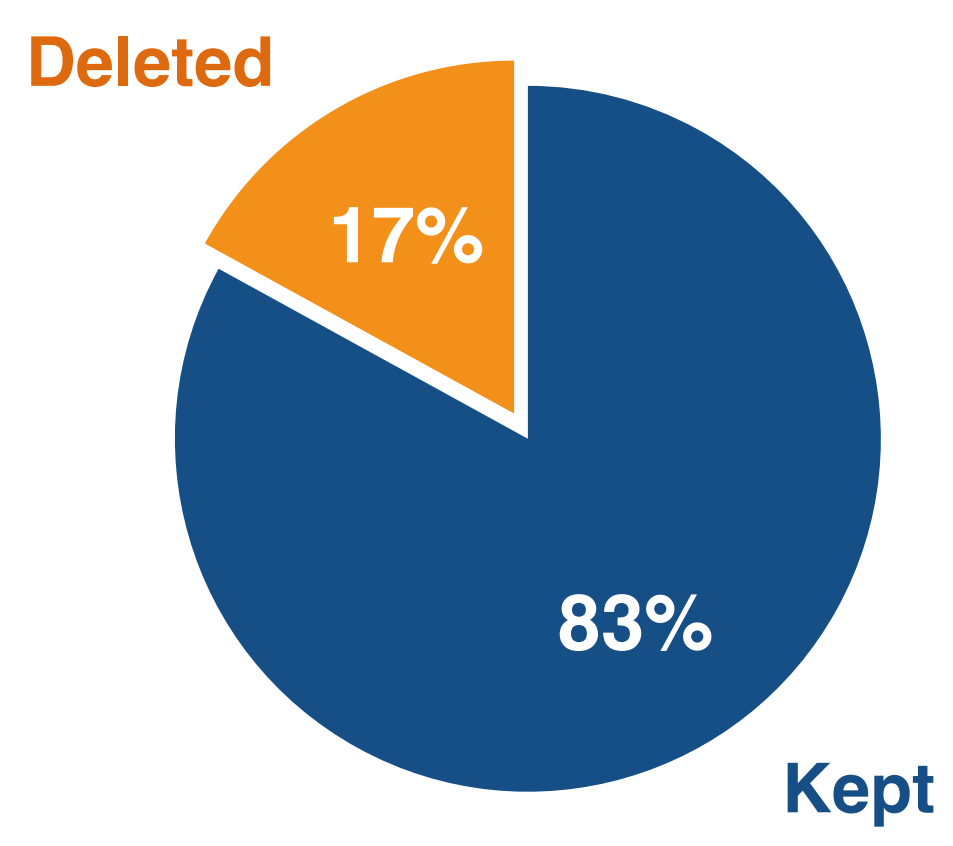
Original

Simplified

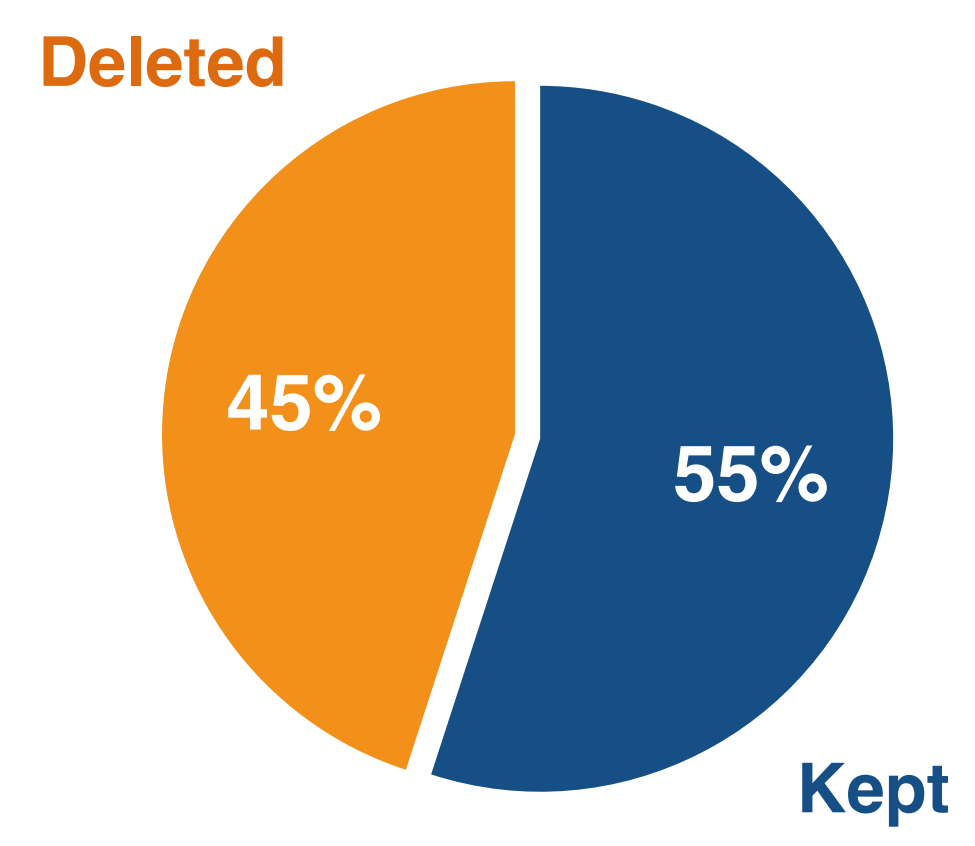
Once the deal is final, they will end up owning about 23 percent of the company. → **Deleted**

Charter had pursued Time Warner Cable for months, but Time Warner Cable CEO Rob Marcus had consistently rejected what he called a lowball offer. → **Kept** → **Charter had tried to buy Time Warner Cable for months.** / **Time Warner Cable CEO Rob Marcus kept saying "no."**

- 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

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.

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

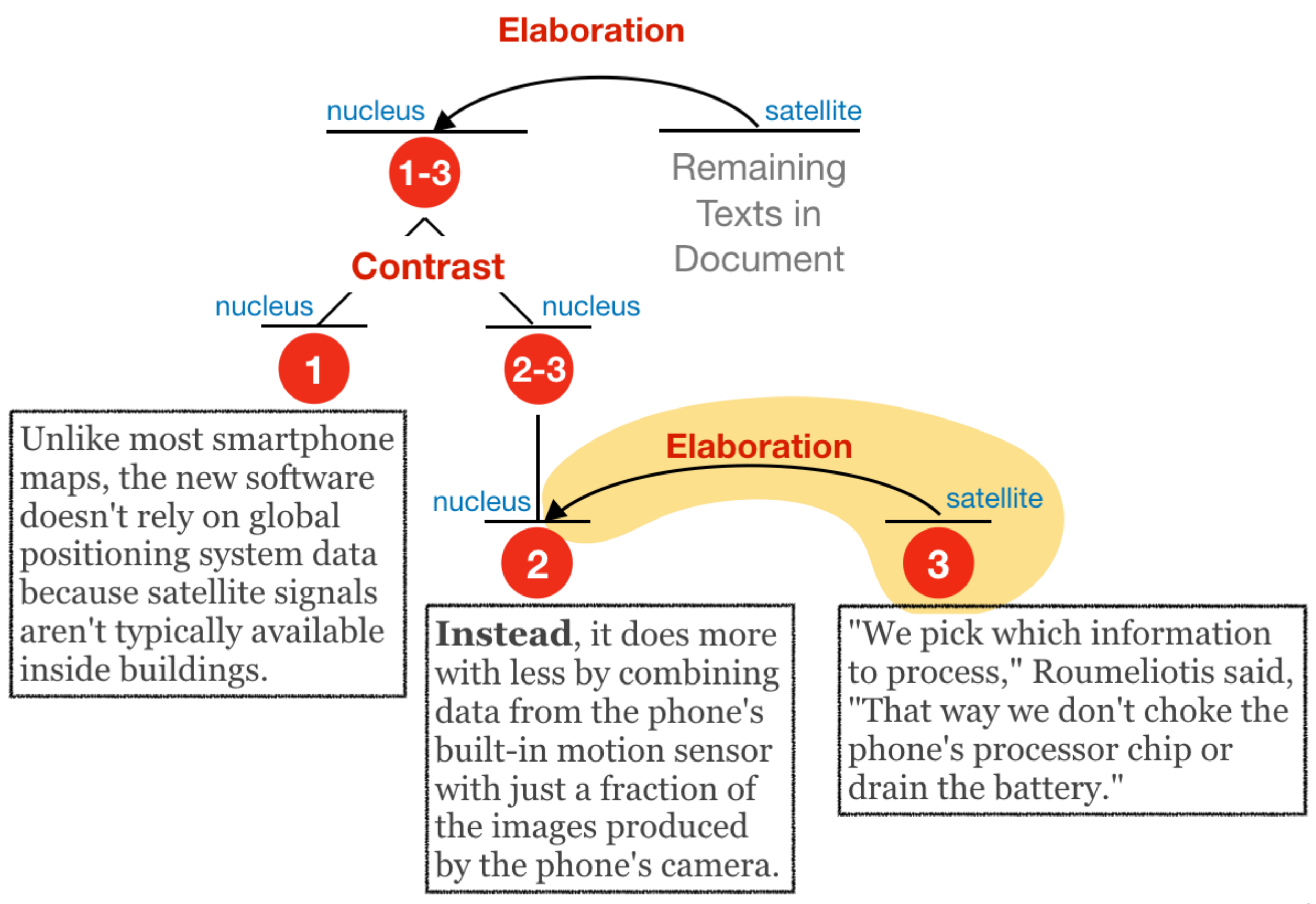
Equal Meaning | Partial Overlap | Mismatch

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

Link for Newsela Corpus
https://newsela.com/data/

Discourse factors Analysis

Rhetorical Structure Theory (RST) Tree



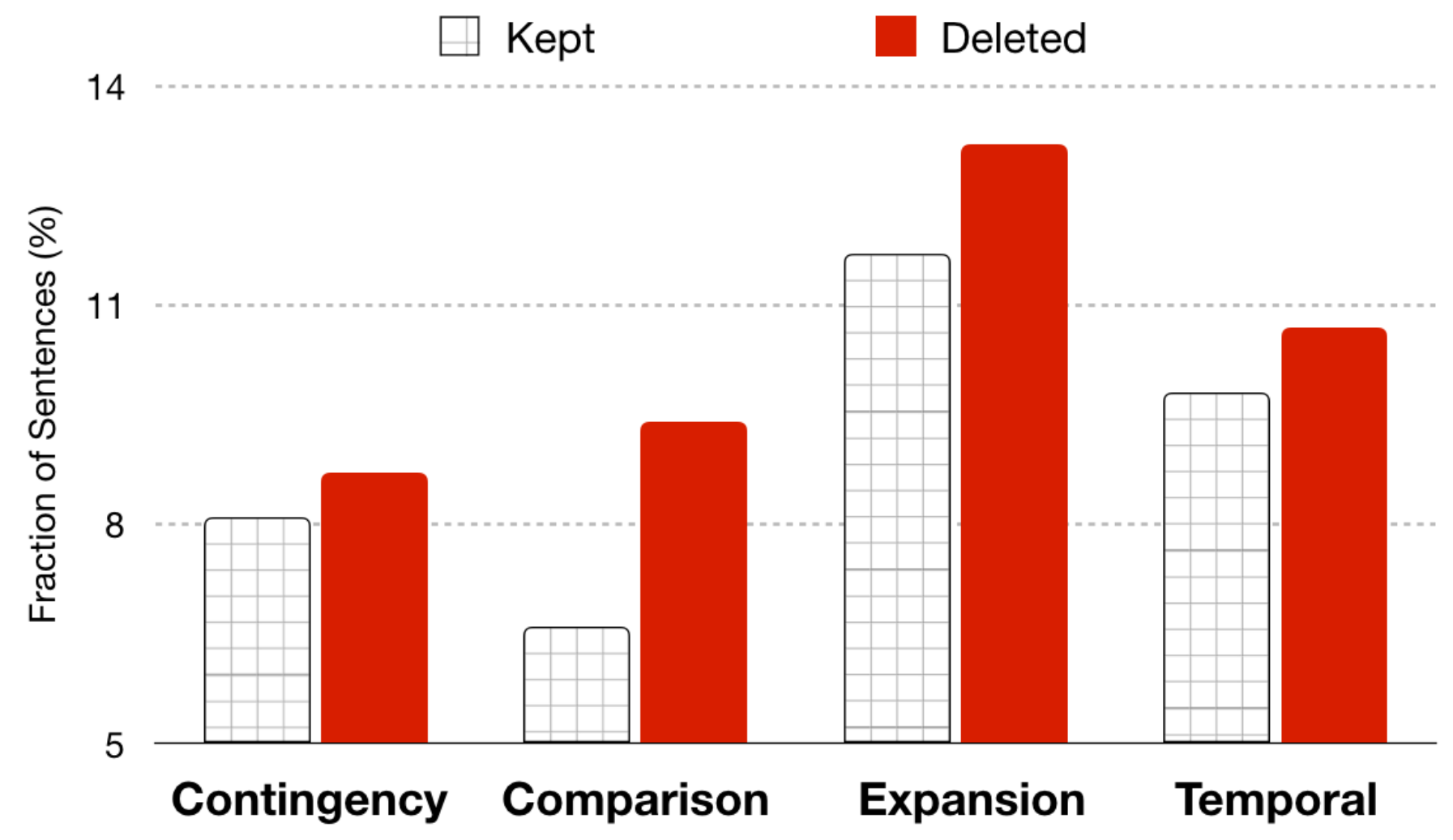
Unlike most smartphone maps, the new software doesn't rely on global positioning system data because satellite signals aren't typically available inside buildings.

Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by the phone's camera.

"We pick which information to process," Roumeliotis said, "That way we don't choke the phone's processor chip or drain the battery."

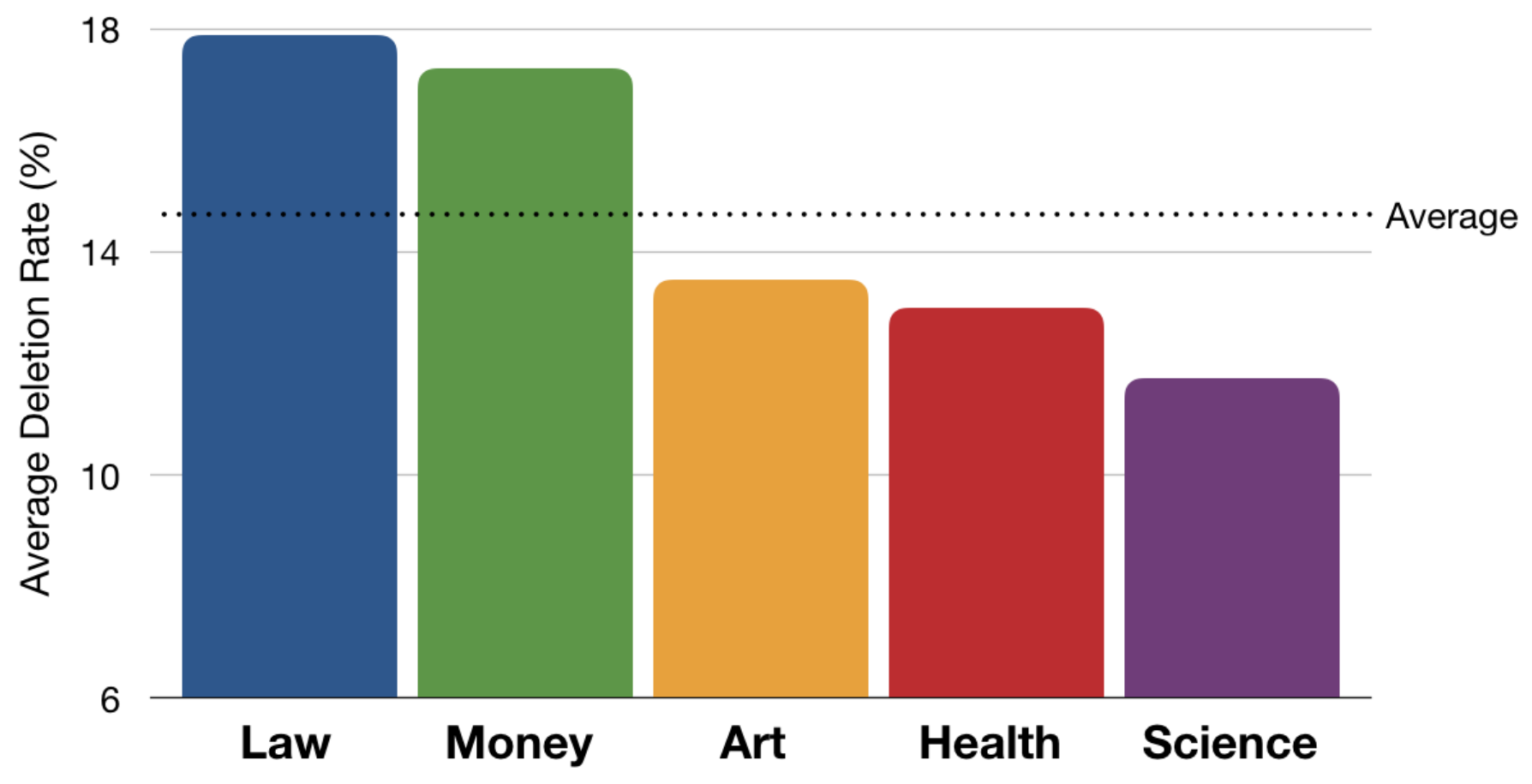
- Sentences governed by elaboration relation are more likely to be deleted.
- Explanations are more likely to keep.
- Sentences near root are mostly kept.

Discourse Connectives



- Sentences with discourse connectives are more likely to be deleted.

Topics



- Topics affect sentence's deletion ratio.

References

- Xu, W.; Napoles, C.; Pavlick, E.; Chen, Q.; and Callison-Burch, C. 2016. Optimizing statistical machine translation for text simplification. TACL
- Petersen, S. E., and Ostendorf, M. 2007. Text simplification for language learners: a corpus analysis. In SLaTE
- Štajner, S., Drndarević, B. and Saggion, H., 2013. Corpus-based sentence deletion and split decisions for Spanish text simplification.
- Maddela, M. and Xu, W., 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In EMNLP

Features and Modeling

Document characteristics

- Number of tokens
- Topics

Discourse features

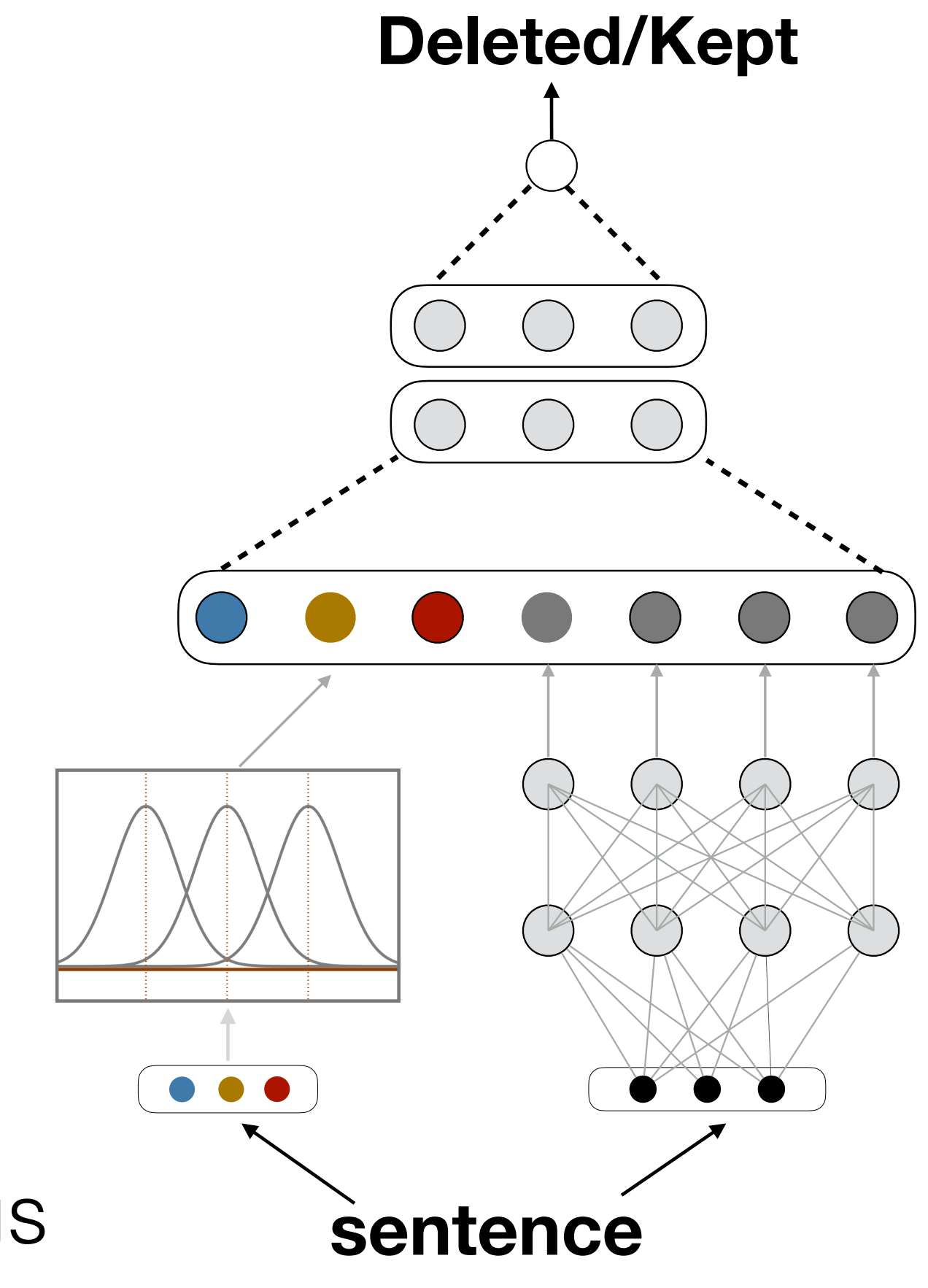
- Indictor of connectives
- Governing relation

Position features

- Sentence position
- Paragraph position

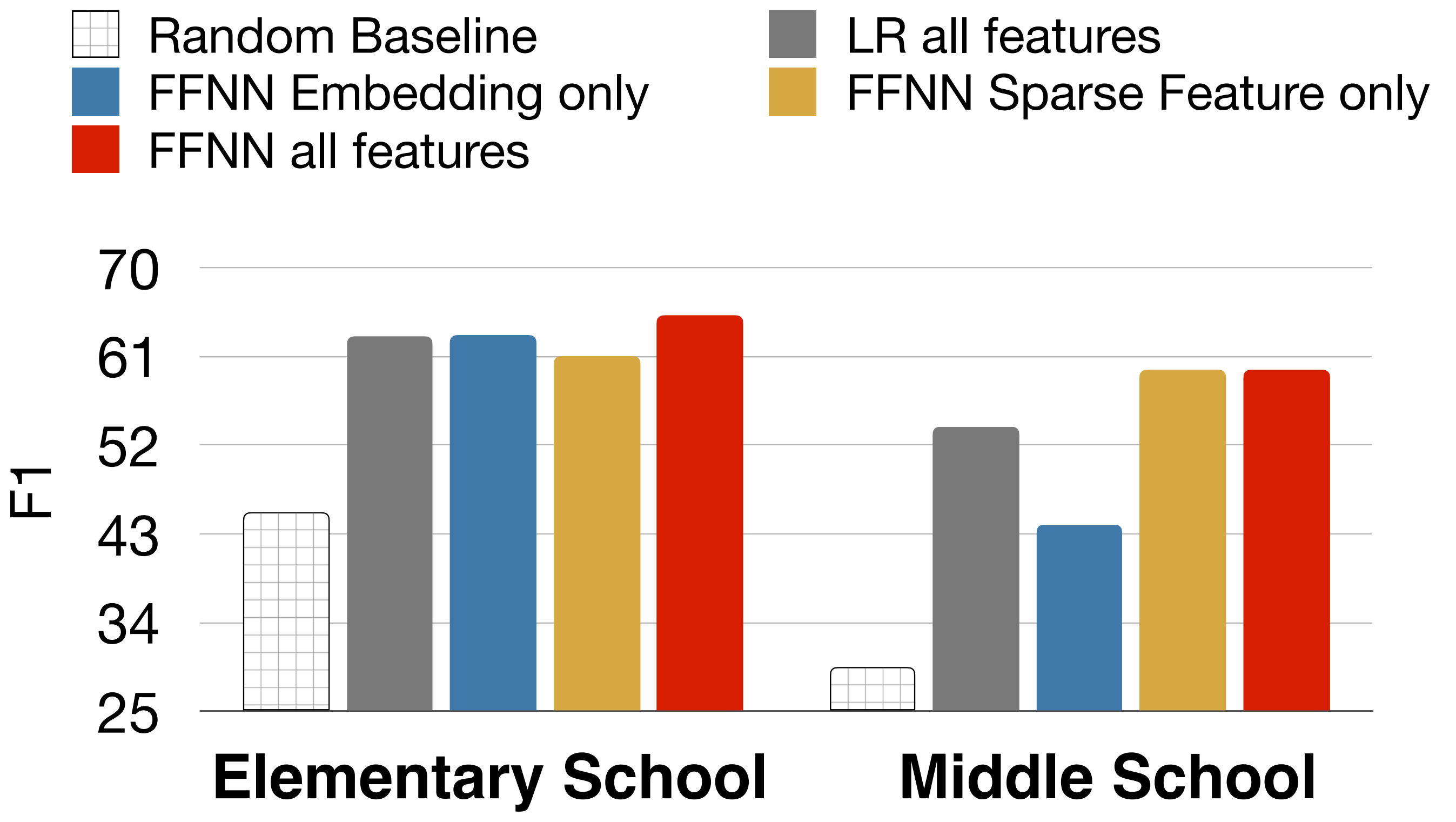
Semantic features

- 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.

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.