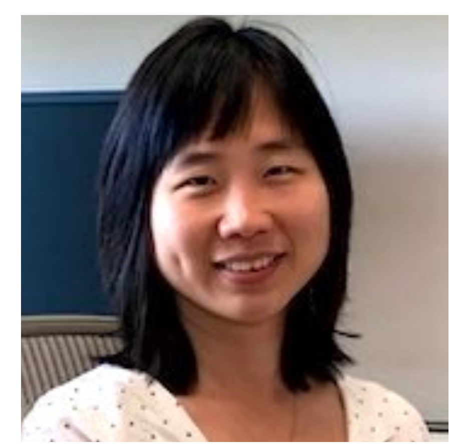
Discourse Level Factors for Sentence Deletion in **Text Simplification**

Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li



The Ohio State University
Department of Computer Science and Engineering

The University of Texas at Austin
Department of Linguistics

More than 65% of 8th graders in American public schools were not proficient in reading and writing.

— National Assessment of Educational Progress
released by the U.S. Department of Education

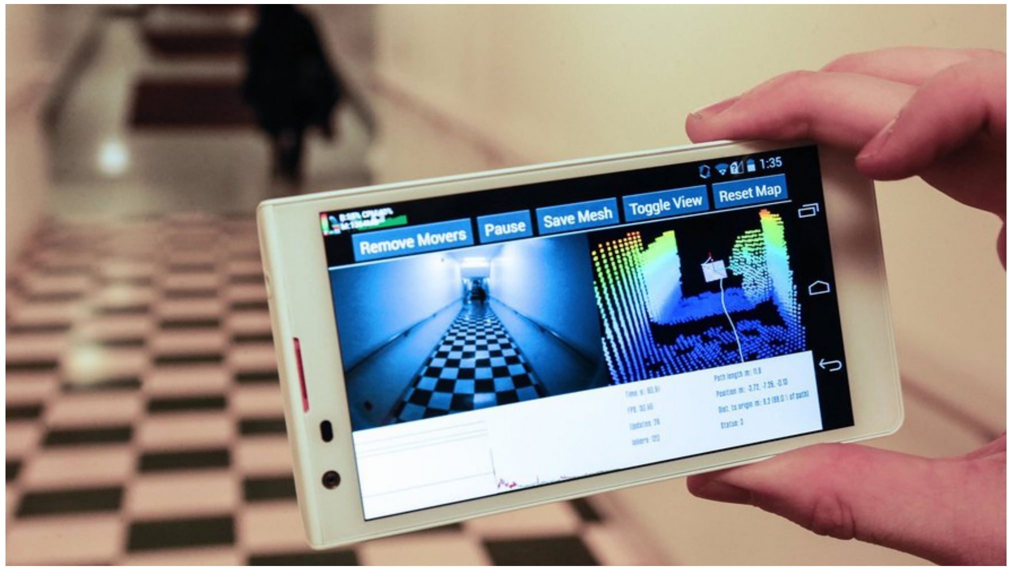
Text Simplification

Goal: rewrite text to be easier to read,
while remaining truthful in content

Science

Building an indoor 3-D map on the spot, via smartphone

By Steve Alexander, Minneapolis Star Tribune
Published: 03/31/2014 Word Count: 777



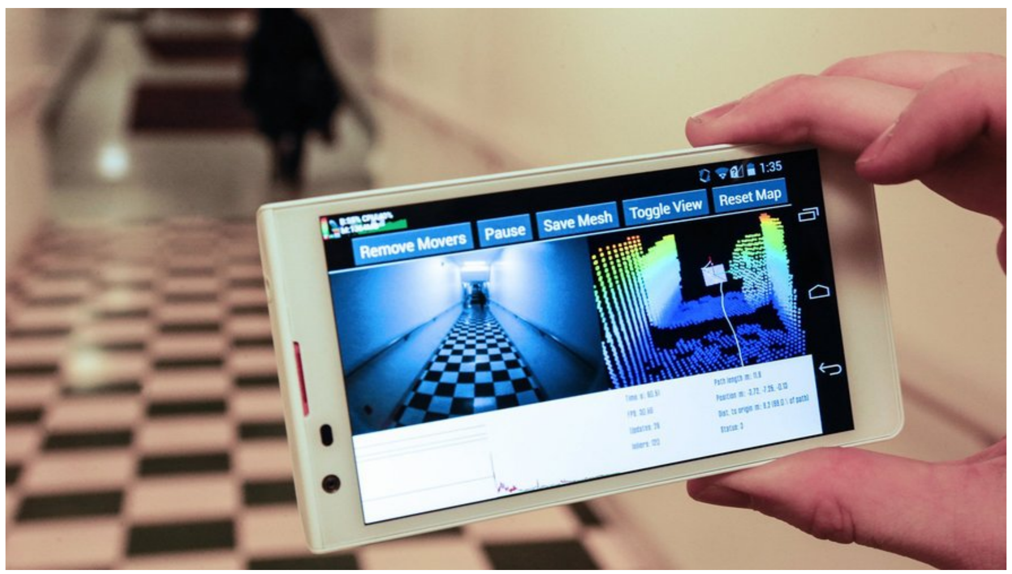
Unlike most smartphone maps, the new software doesn't rely on global positioning system data because satellite signals aren't typically available inside buildings. Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by 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."

Science

Building an indoor 3-D map on the spot, via smartphone

By Steve Alexander, Minneapolis Star Tribune
Published: 03/31/2014 Word Count: 777



Describing a technique a mobile software uses to build indoor map

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."

Text Simplification

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."

Text Simplification

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."

Text Simplification

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."

Text Simplification

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."

Text Simplification

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."

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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."

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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."

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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.

Instead, the software uses the phone's camera and built-in motion sensor.

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

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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.

Instead, the software uses the phone's camera and built-in motion sensor.

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

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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.

Instead, the software uses the phone's camera and built-in motion sensor.

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

Text Simplification

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

The new software doesn't use global positioning satellite (GPS) signals.

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.

Instead, the software uses the phone's camera and built-in motion sensor.

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

Sentence Deletion

Our Work

- Manually annotated corpus with sentence alignments.

Our Work

- Manually annotated corpus with sentence alignments.
- Analysis of discourse level factors affecting the deletion of sentences.
  - Governing relation of sentence in RST tree.
  - Discourse connectives in sentence.

Our Work

- Manually annotated corpus with sentence alignments.
- Analysis of discourse level factors affecting the deletion of sentences.
    - Governing relation of sentence in RST tree.
    - Discourse connectives in sentence.
- Automatic prediction of sentence's deletion.

Our Work

- **Manually annotated corpus with sentence alignments.**
- Analysis of discourse level factors affecting the deletion of sentences.
  - Governing relation of sentence in RST tree.
  - Discourse connectives in sentence.
- Automatic prediction of sentence's deletion.

Newsela Corpus (Xu et al. 2015)

- Newsela is a U.S. Education company based in New York City.
- **1,932 news articles** rewritten by professional editors for schools children.
- Each document (~47 sentences) is simplified to 4 different reading levels.
- But, only document aligned

newsela

https://newsela.com/data/

Newsela Corpus (Xu et al. 2015)

- Newsela is a U.S. Education company based in New York City.
- **1,932 news articles** rewritten by professional editors for schools children.
- Each document (~47 sentences) is simplified to 4 different reading levels.

- But, only document aligned

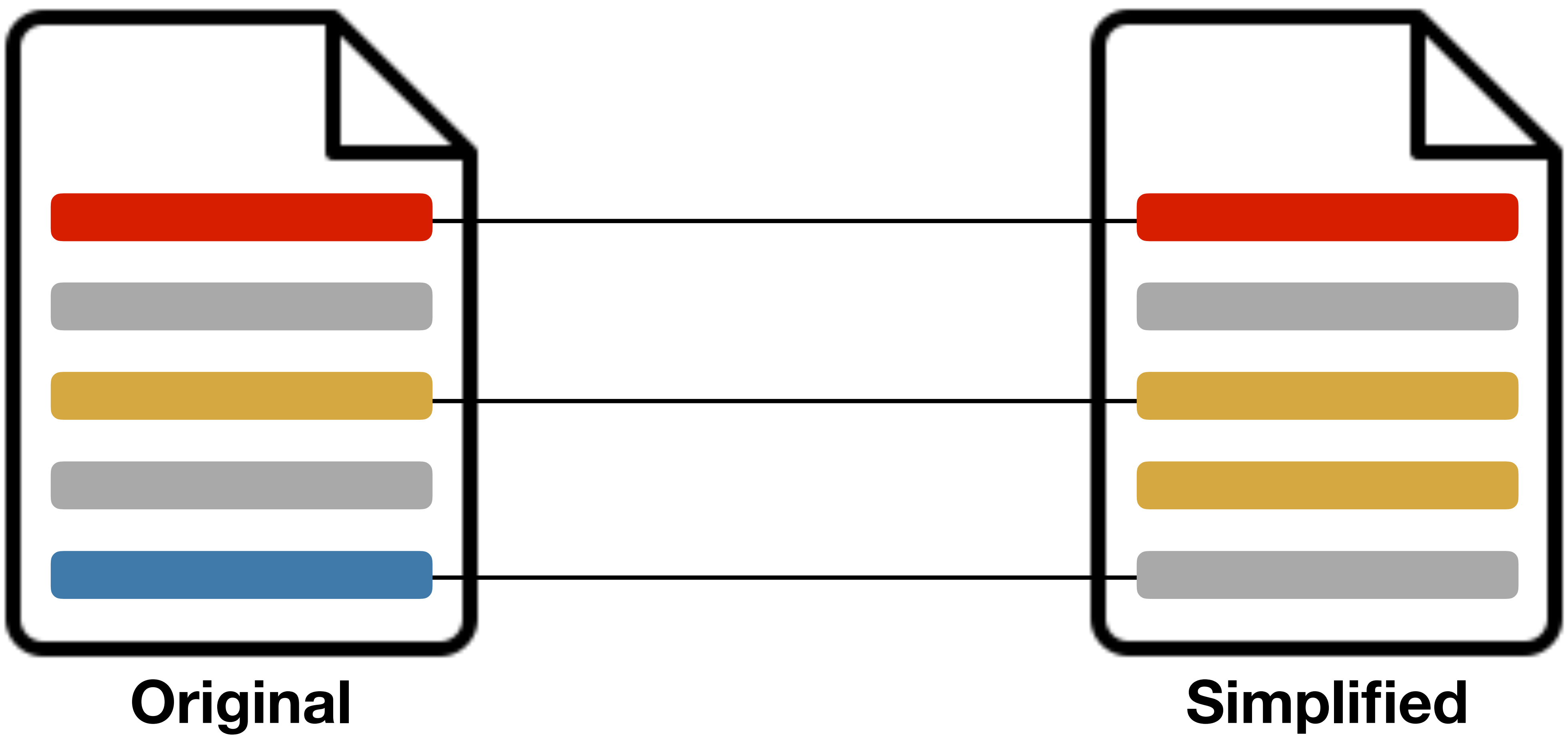
We manually annotated 50 sets of articles across three reading levels to analyze what sentences get deleted.

newsela

https://newsela.com/data/

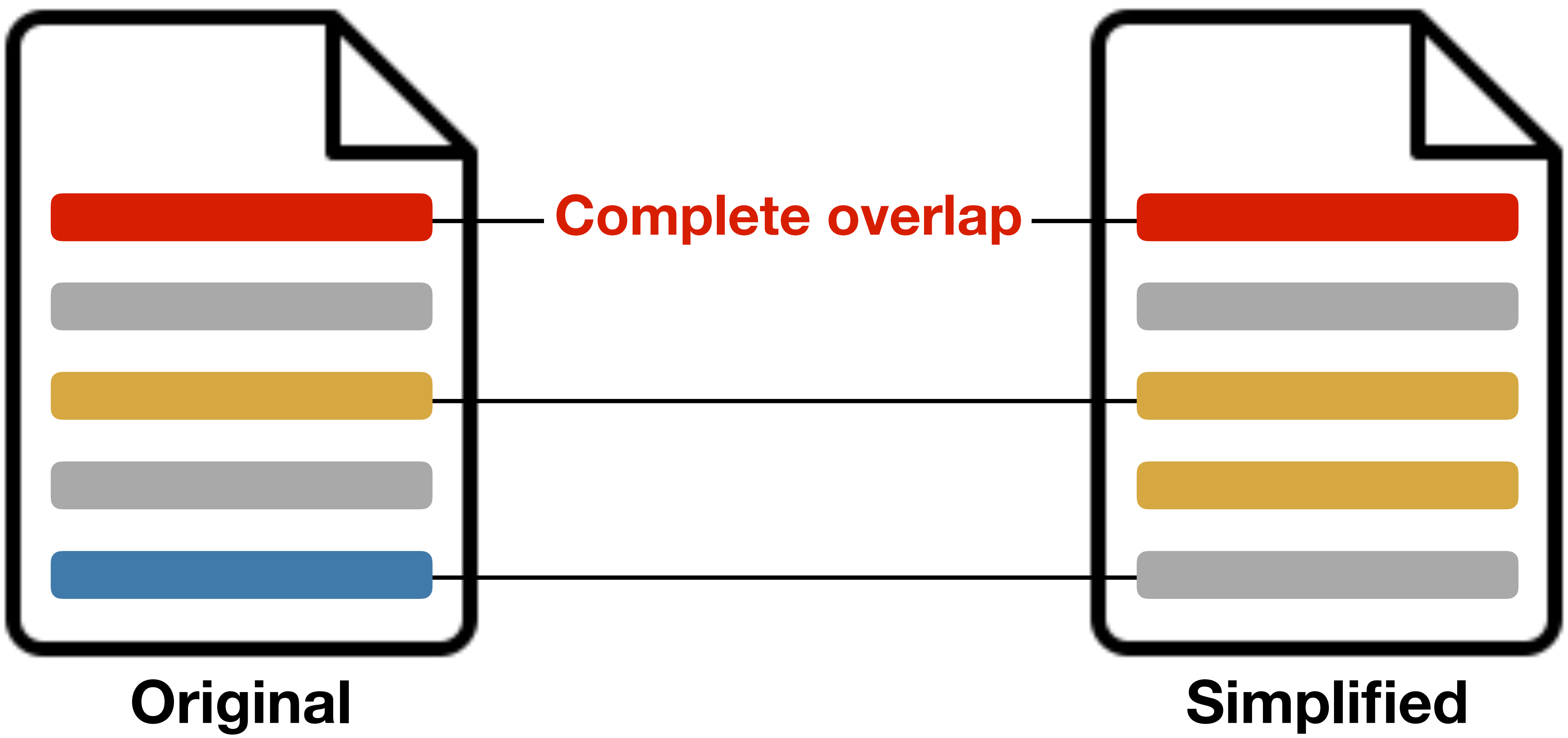
Manual Annotation

- Classification on sentence pairs.
- inter-annotator agreement at **0.807** by **Cohen's kappa**.
- Annotations aggregated by majority vote from 5 workers.
- Lastly verified by in-house annotators (not the authors).



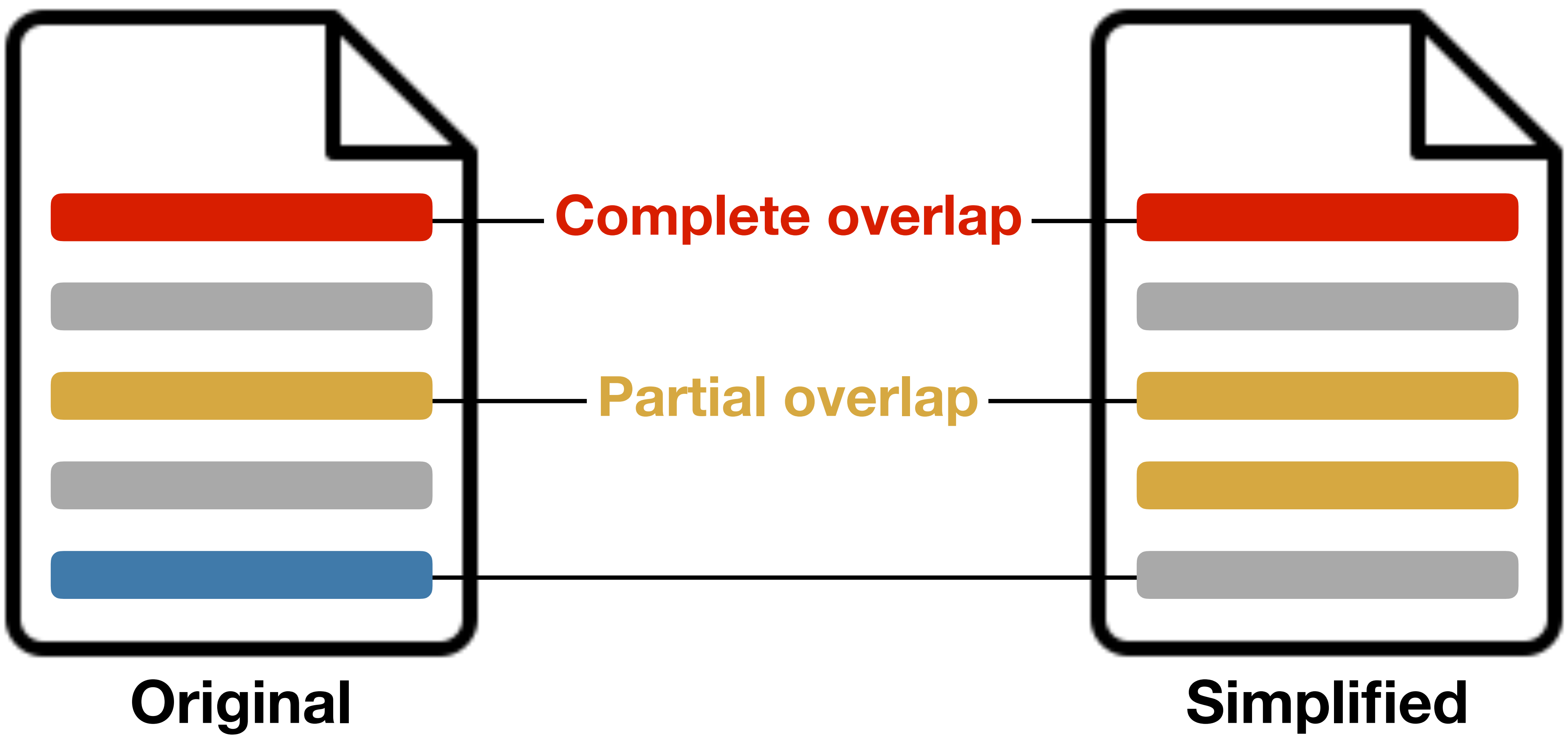
Manual Annotation

- Classification on sentence pairs.
- inter-annotator agreement at **0.807** by **Cohen's kappa**.
- Annotations aggregated by majority vote from 5 workers.
- Lastly verified by in-house annotators (not the authors).



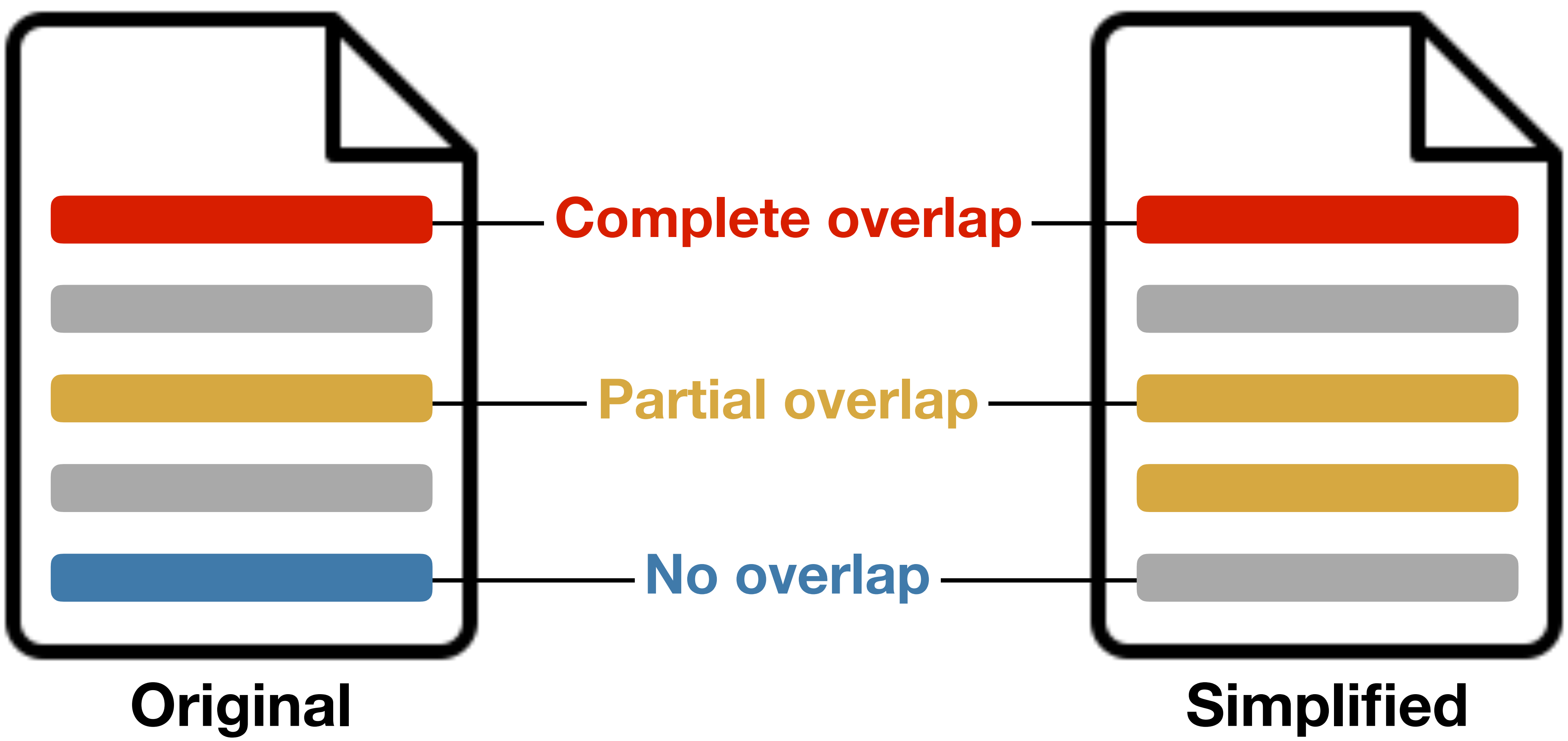
Manual Annotation

- Classification on sentence pairs.
- inter-annotator agreement at **0.807** by **Cohen's kappa**.
- Annotations aggregated by majority vote from 5 workers.
- Lastly verified by in-house annotators (not the authors).



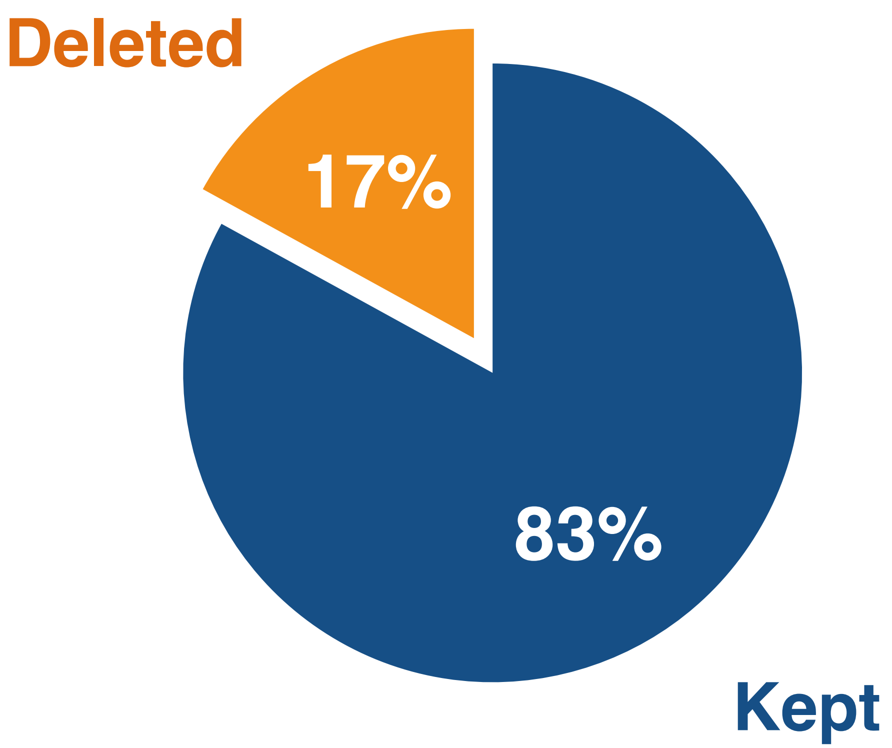
Manual Annotation

- Classification on sentence pairs.
- inter-annotator agreement at **0.807** by **Cohen's kappa**.
- Annotations aggregated by majority vote from 5 workers.
- Lastly verified by in-house annotators (not the authors).



Sentence Deletion

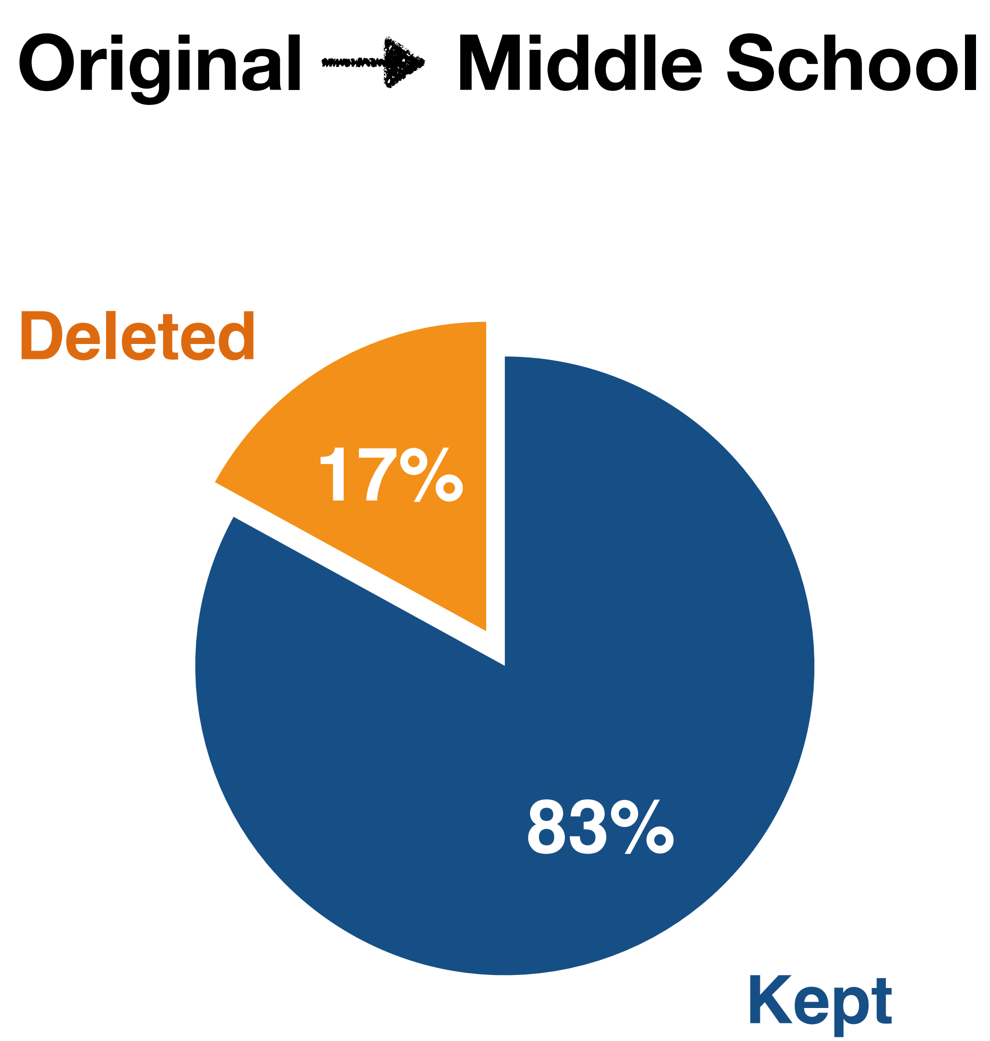
Original → Middle School

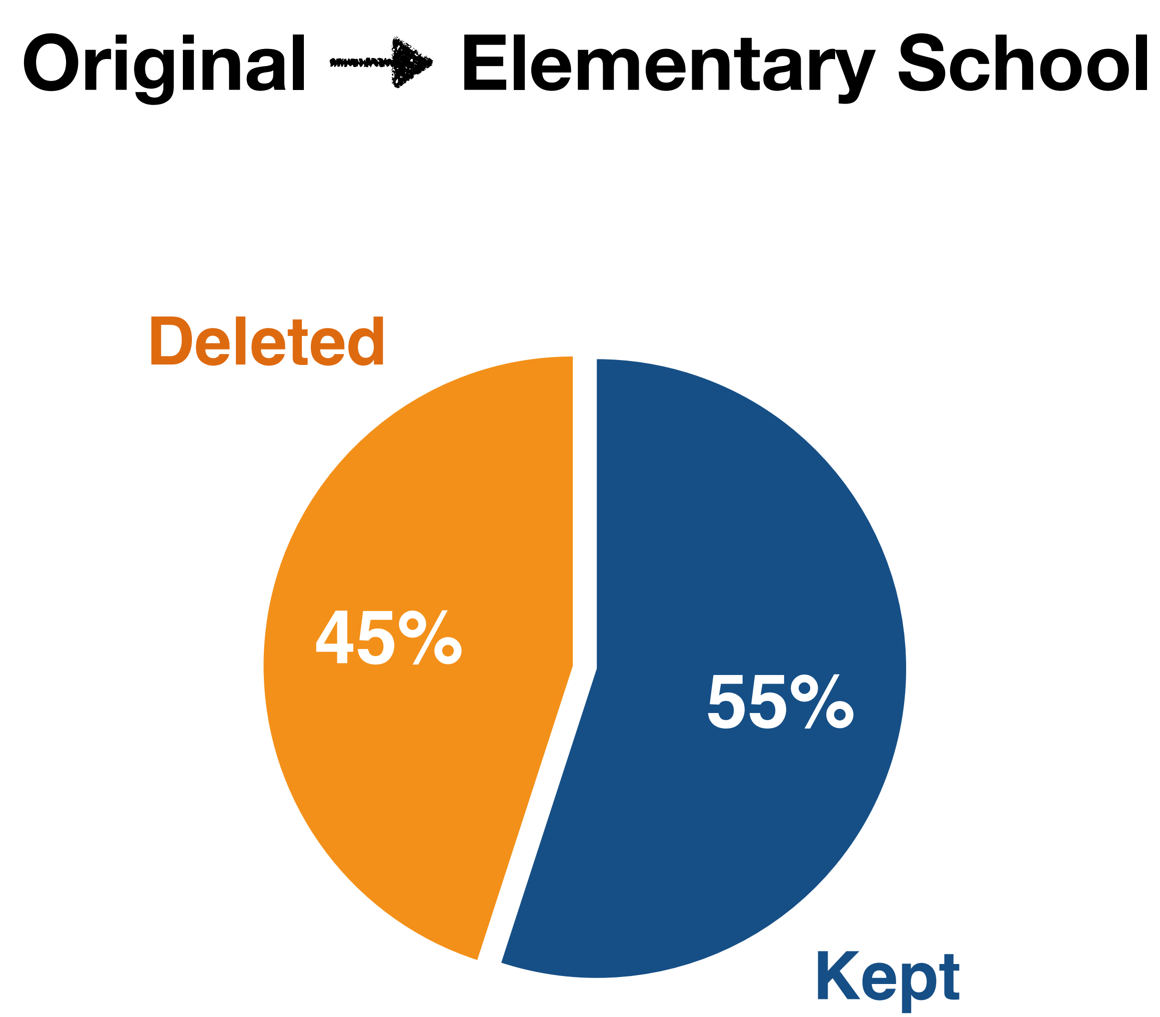


Professional editors remove entire sentences when simplifying news articles.

* based on 50 articles we manually sentence-aligned in the Newsela corpus

Sentence Deletion





Professional editors remove entire sentences when simplifying news articles.

* based on 50 articles we manually sentence-aligned in the Newsela corpus

Our Work

- Manually annotated corpus with sentence alignments.
- **Analysis of discourse level factors affecting the deletion of sentences.**
  - Governing relation of sentence in RST tree.
  - Discourse connectives in sentence.
- Automatic prediction of sentence deletion.

Discourse Analysis

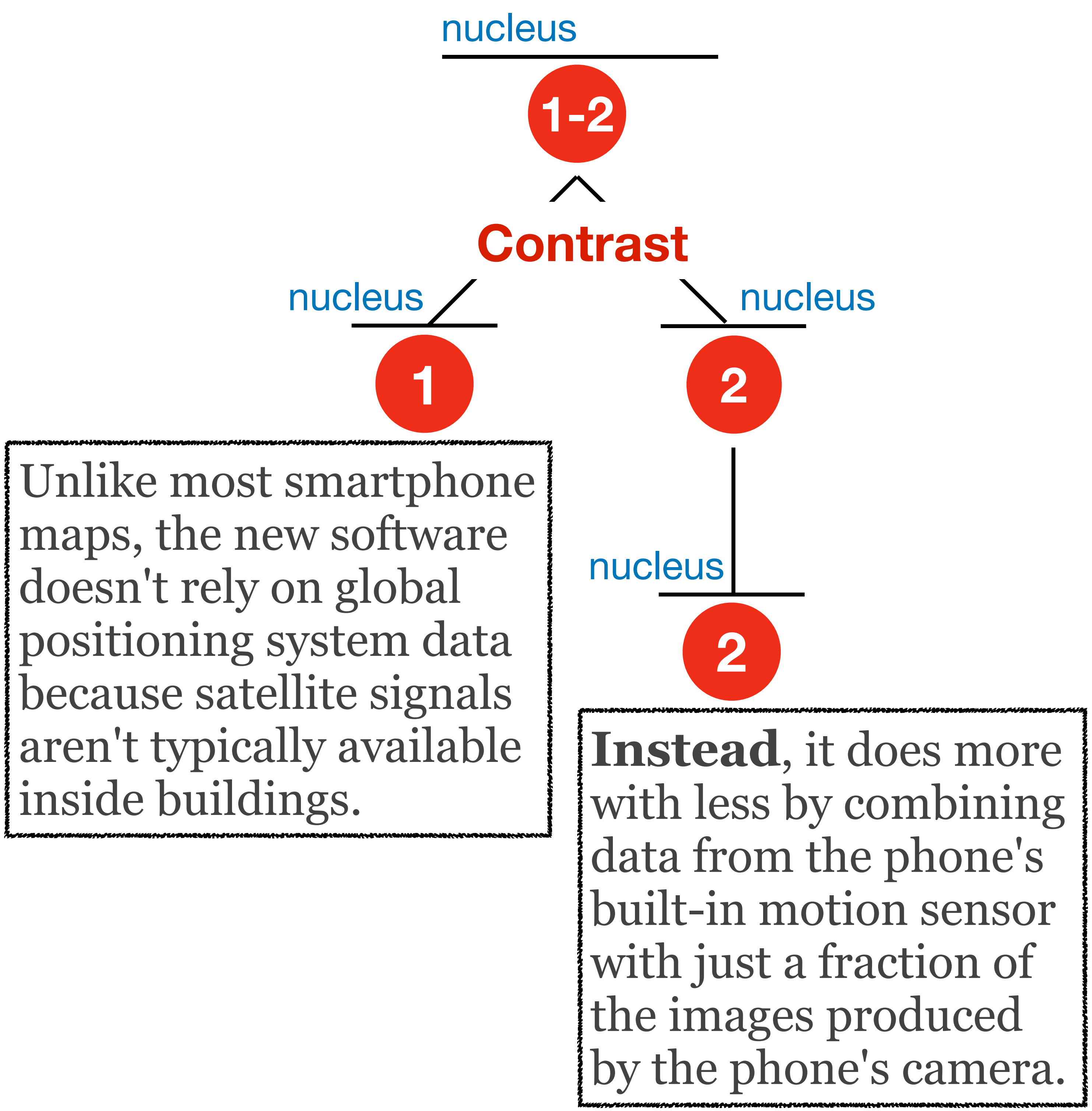
Unlike most smartphone maps, the new software doesn't rely on global positioning system data because satellite signals aren't typically available inside buildings.

Discourse Analysis

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

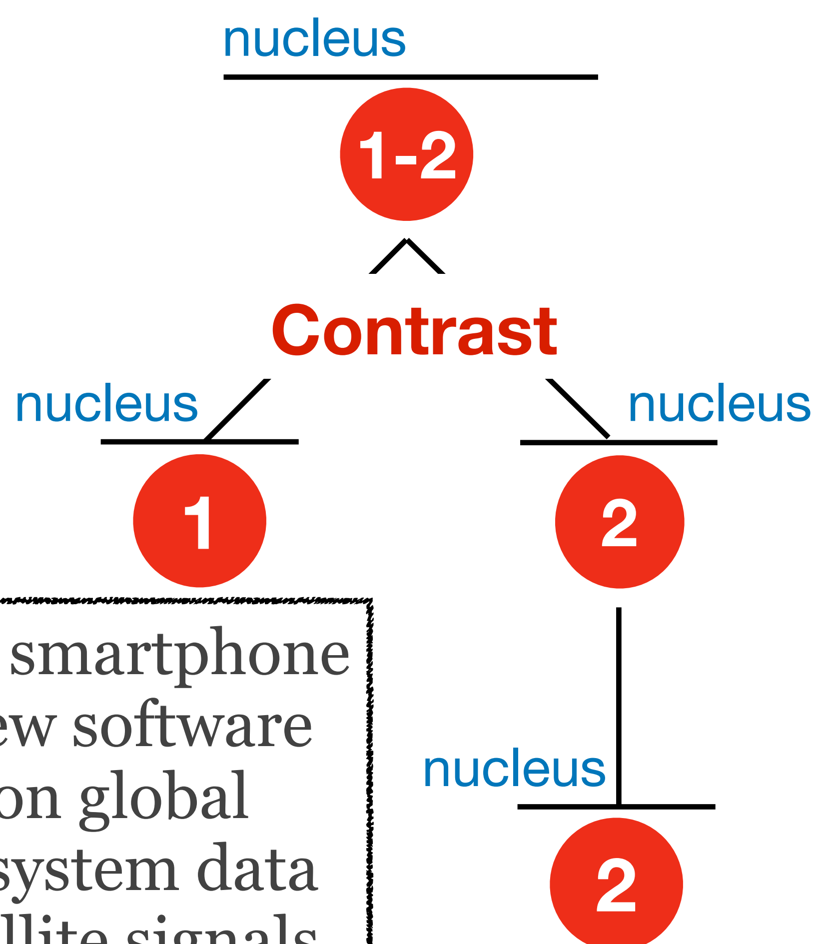
Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by the phone's camera.

Rhetorical Structure Theory (RST) Tree



* using the discourse parser from (Surdeanu, Mihai, et al, 2015)

Inter-sentential RST

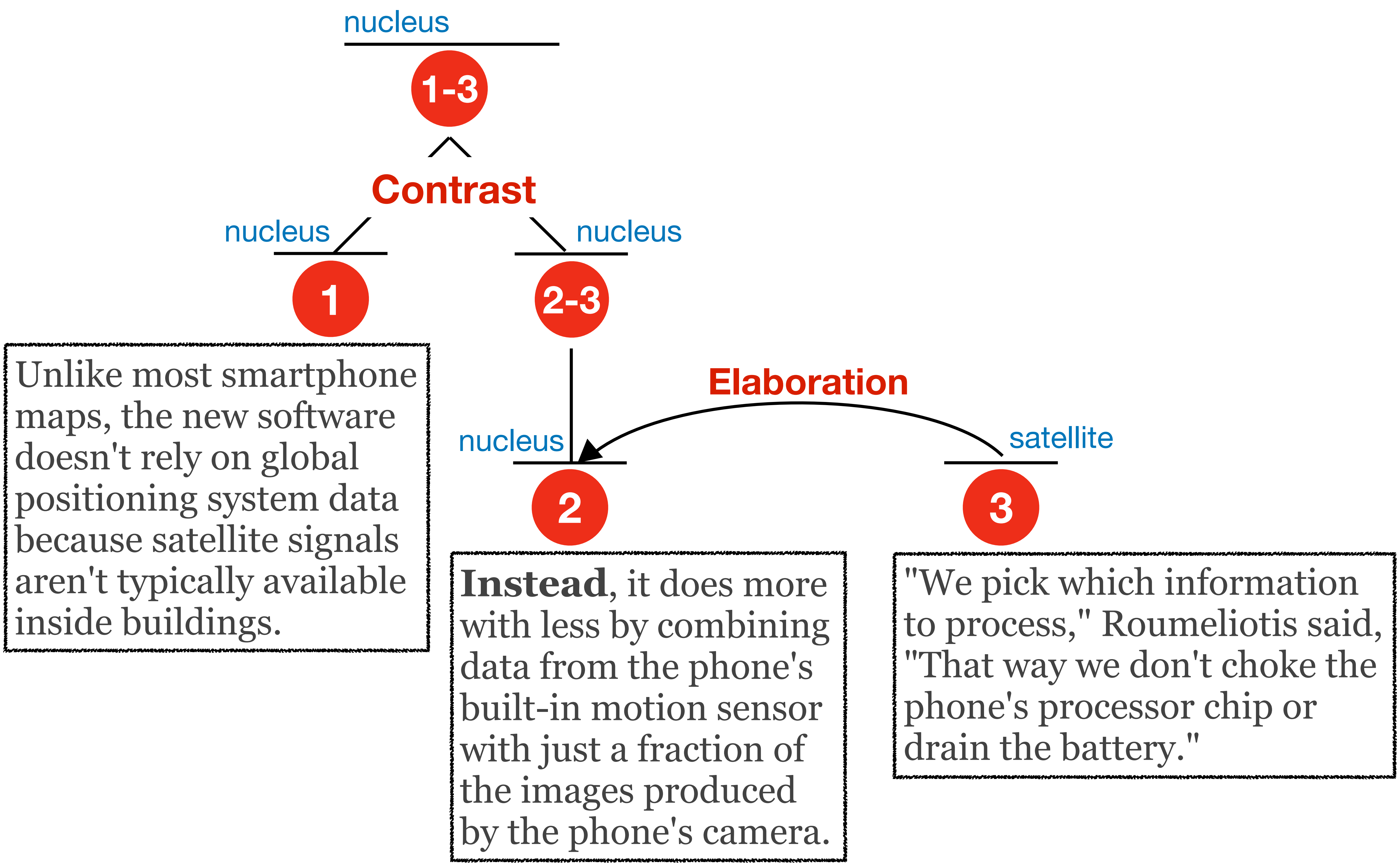


Unlike most smartphone maps, the new software doesn't rely on global positioning system data because satellite signals aren't typically available inside buildings.

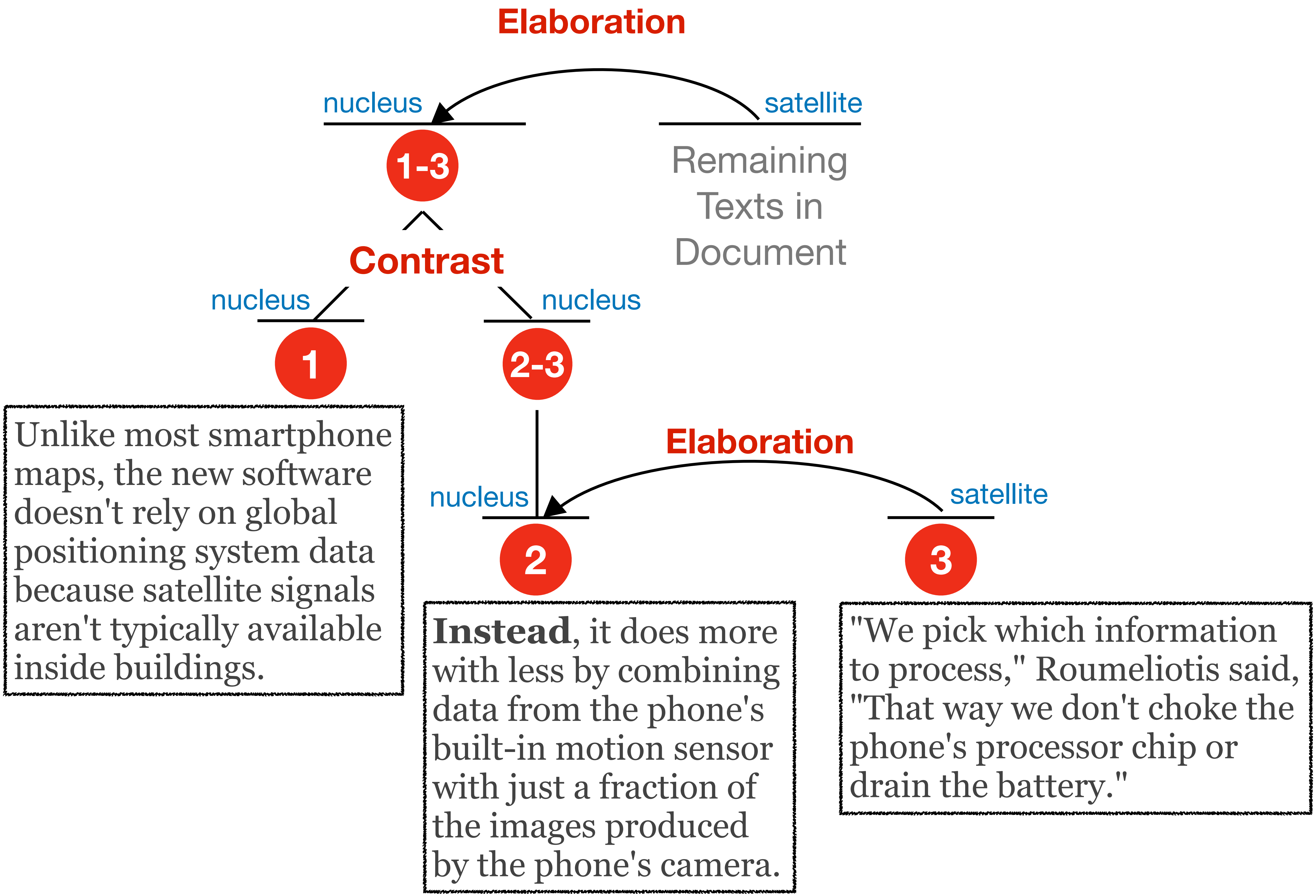
Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by the phone's camera.

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

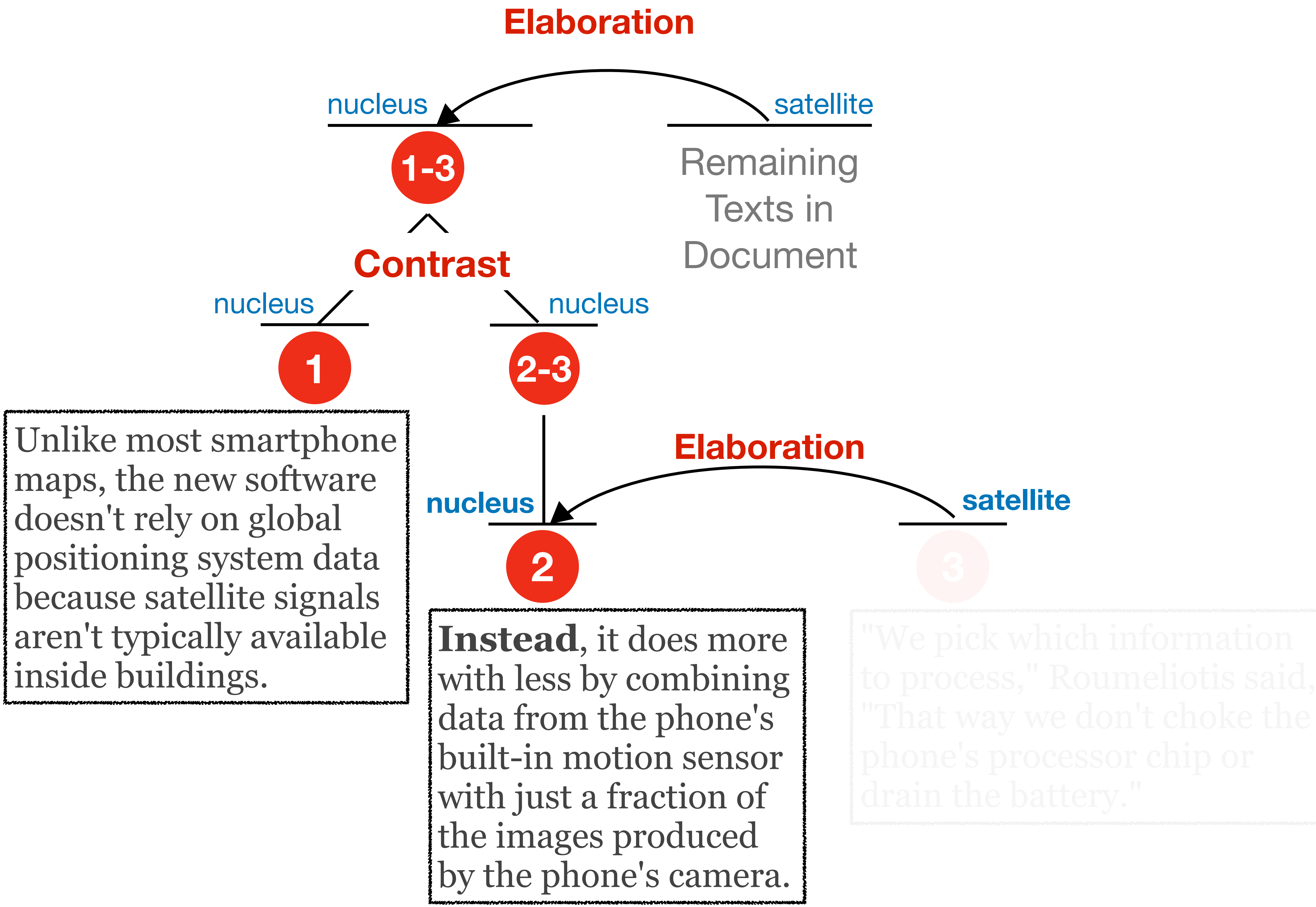
Inter-sentential RST



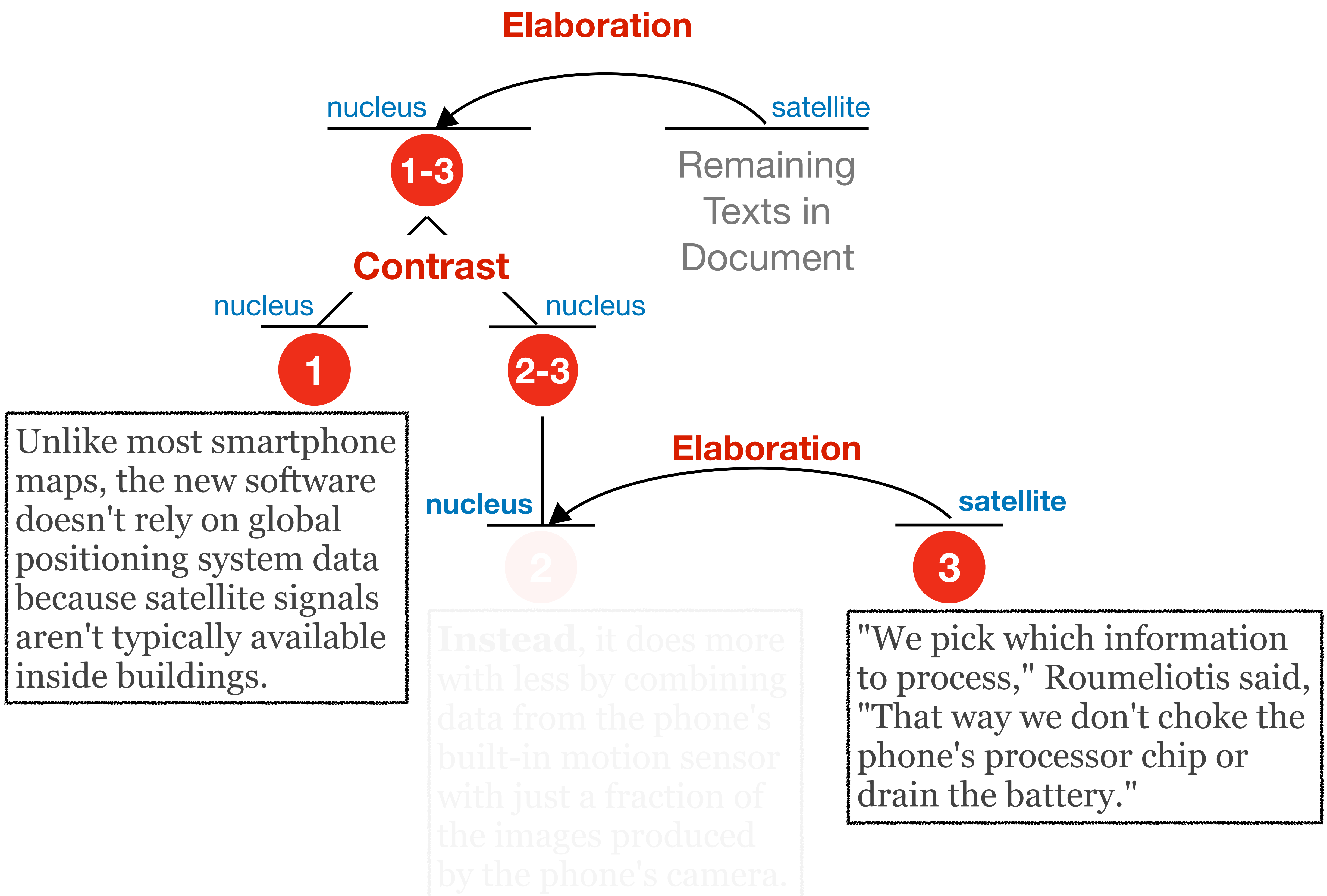
Inter-sentential RST



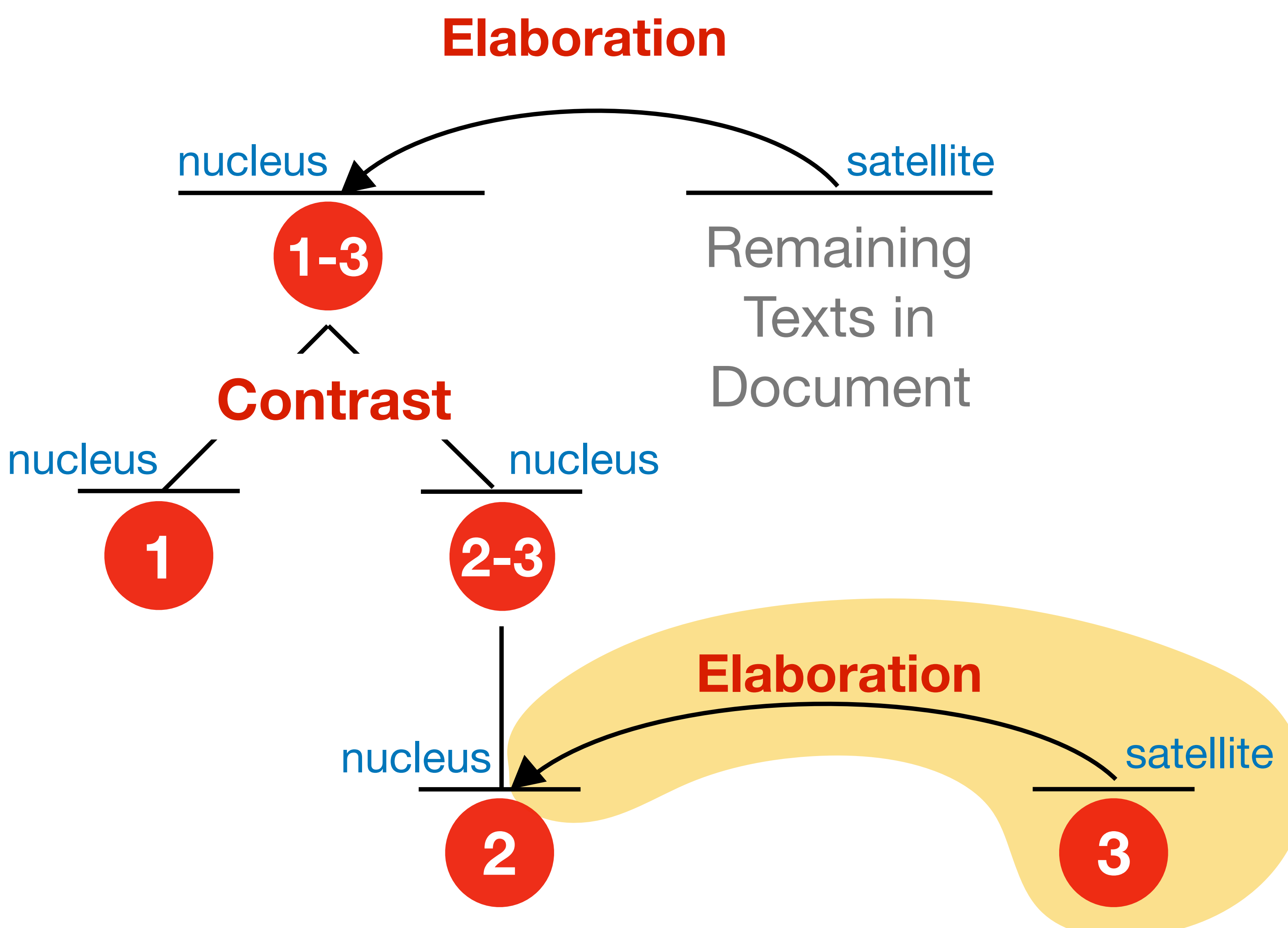
Inter-sentential RST



Inter-sentential RST

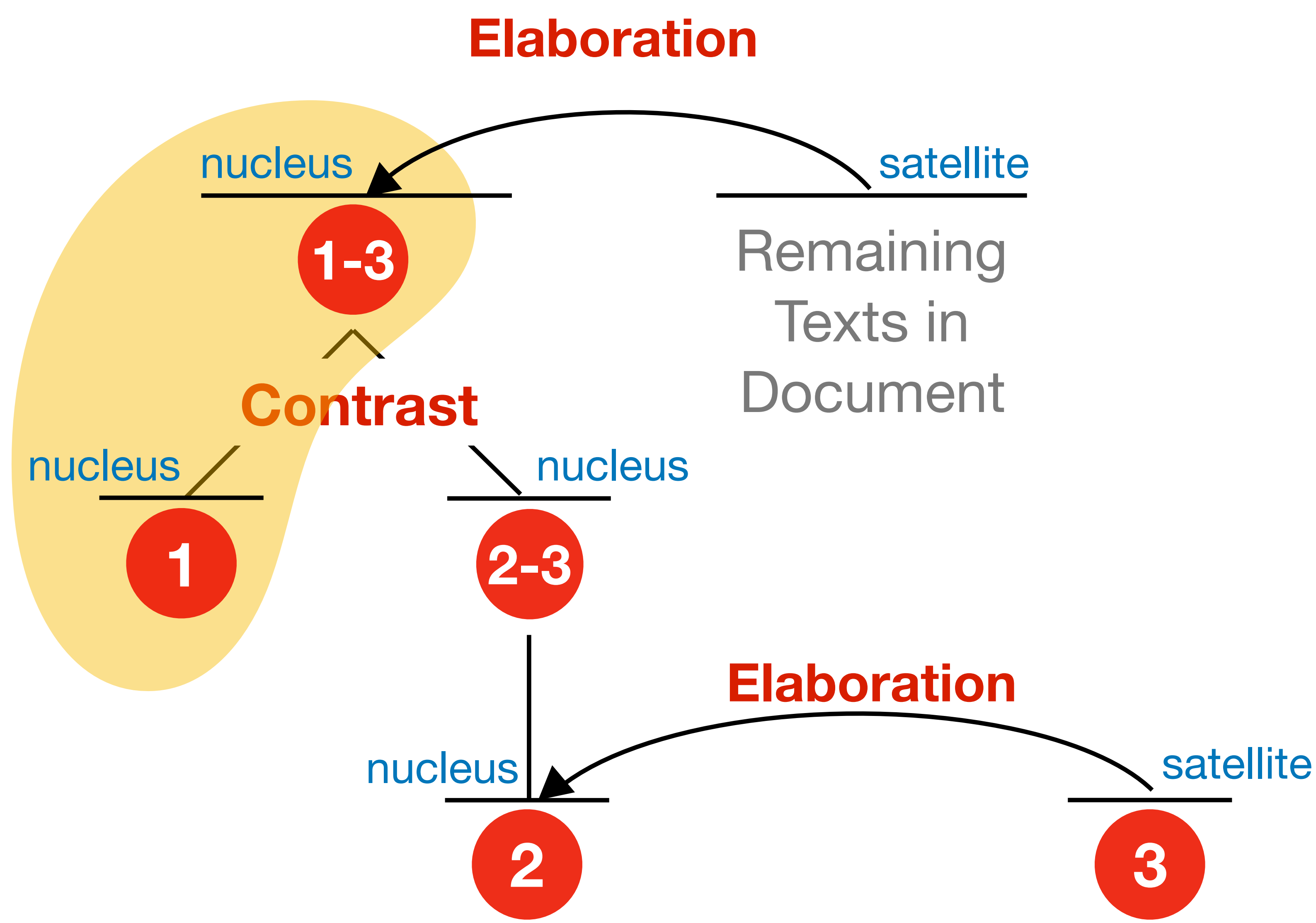


Inter-sentential RST



The **lowest governing relation** of sentence 3 is **elaboration**.

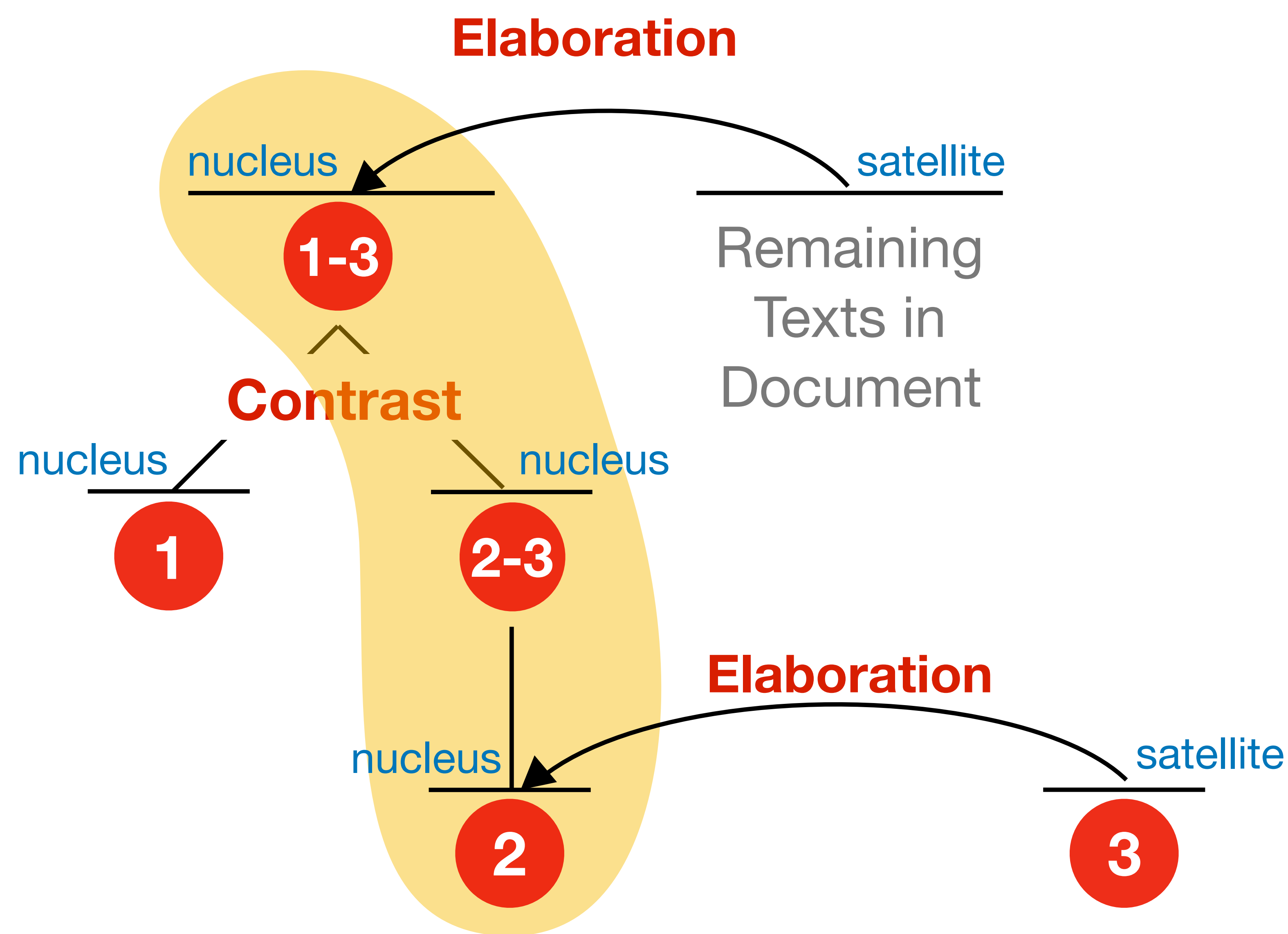
Inter-sentential RST



Sentence 1 and 2 have **no governing relation** in the discourse tree.

(i.e., they are nucleuses that are close to the root)

Inter-sentential RST



Sentence 1 and 2 have **no governing relation** in the discourse tree.

(i.e., they are nucleuses that are close to the root)

Governing Relations in the Discourse

#sentences	Middle School	
	Kept	Deleted
No Relation	8.4%	5.7% ↓

Sentences that are nuclei and close to the root are less likely to be deleted.

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

Governing Relations in the Discourse

#sentences	Middle School	
	Kept	Deleted
No Relation	8.4%	5.7% ↓

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

Governing Relations in the Discourse

#sentences	Middle School	
	Kept	Deleted
No Relation	8.4%	5.7% ↓
Elaboration	79.3%	81.6%

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

Governing Relations in the Discourse

#sentences	Middle School	
	Kept	Deleted
No Relation	8.4%	5.7% ↓
Elaboration	79.3%	81.6%
Explanation	1.9%	1.1% ↓
Background	1.9%	1.2%

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

Governing Relations in the Discourse

#sentences	Middle School	
	Kept	Deleted
No Relation	8.4%	5.7% ↓
Elaboration	79.3%	81.6%
Explanation	1.9%	1.1% ↓
Background	1.9%	1.2%

Sentences used for explanations are less likely to be deleted.

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

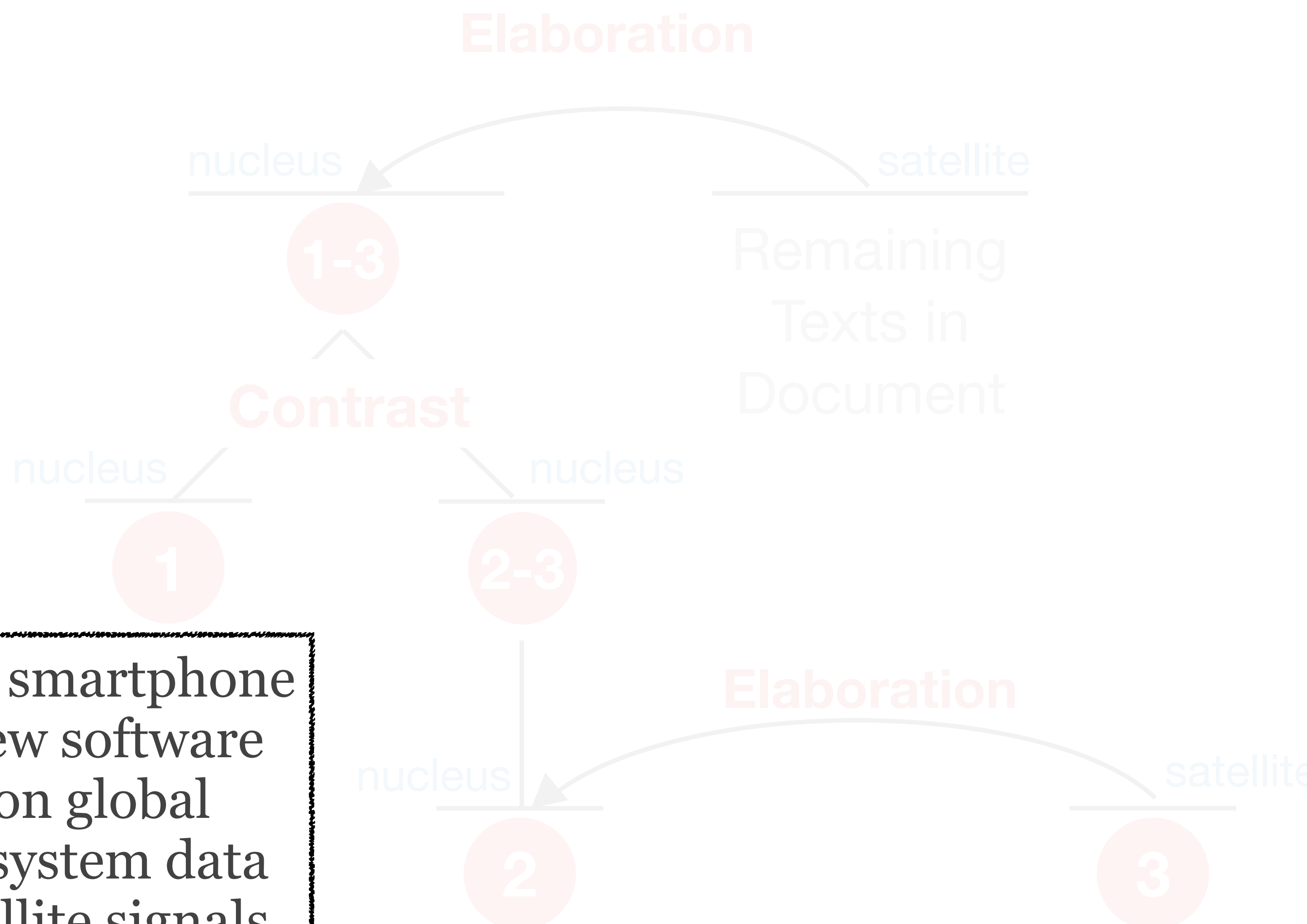
Governing Relations in the Discourse

#sentences	Middle School		Elementary School	
	Kept	**Deleted**	**Kept**	**Deleted**
No Relation	8.4%	5.7% ↓	11.5%	3.8% ↓
Elaboration	79.3%	81.6%	75.2%	84.0% ↑
Explanation	1.9%	1.1% ↓	2.0%	1.6%
Background	1.9%	1.2%	2.2%	2.1%

Sentences used for explanations are less likely to be deleted.

↓ : significantly lower presence among deleted sentences than the kept ones ↑ : higher.

Discourse Connectives



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

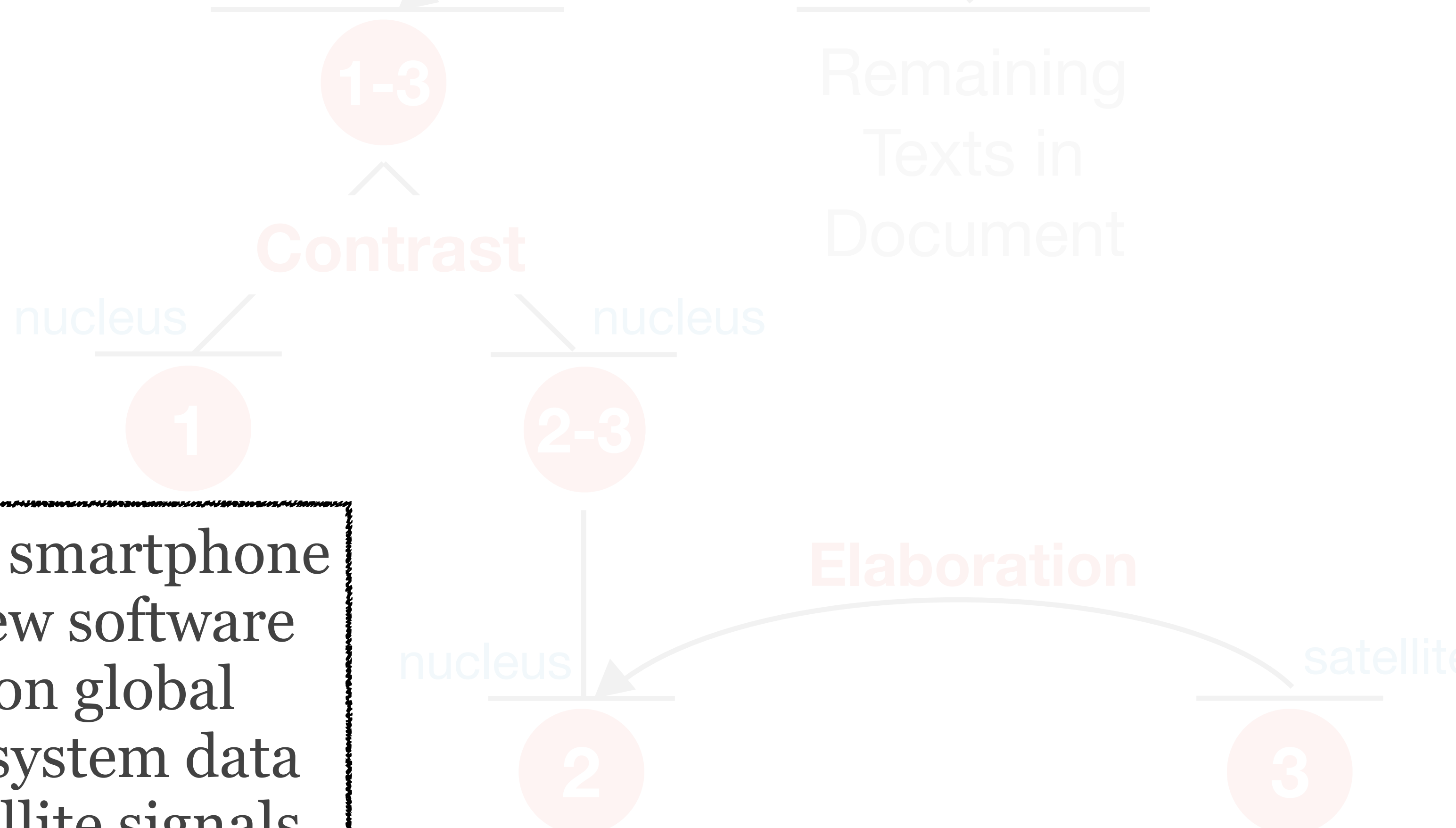
Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by the phone's camera.

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

Following the style of Penn Discourse Treebank (Miltsakaki, Eleni, et al. 2004.)

Discourse Connectives

Words or **phrases** that connect or relate two coherent sentences or phrases and indicate the presence of **discourse** relations.



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

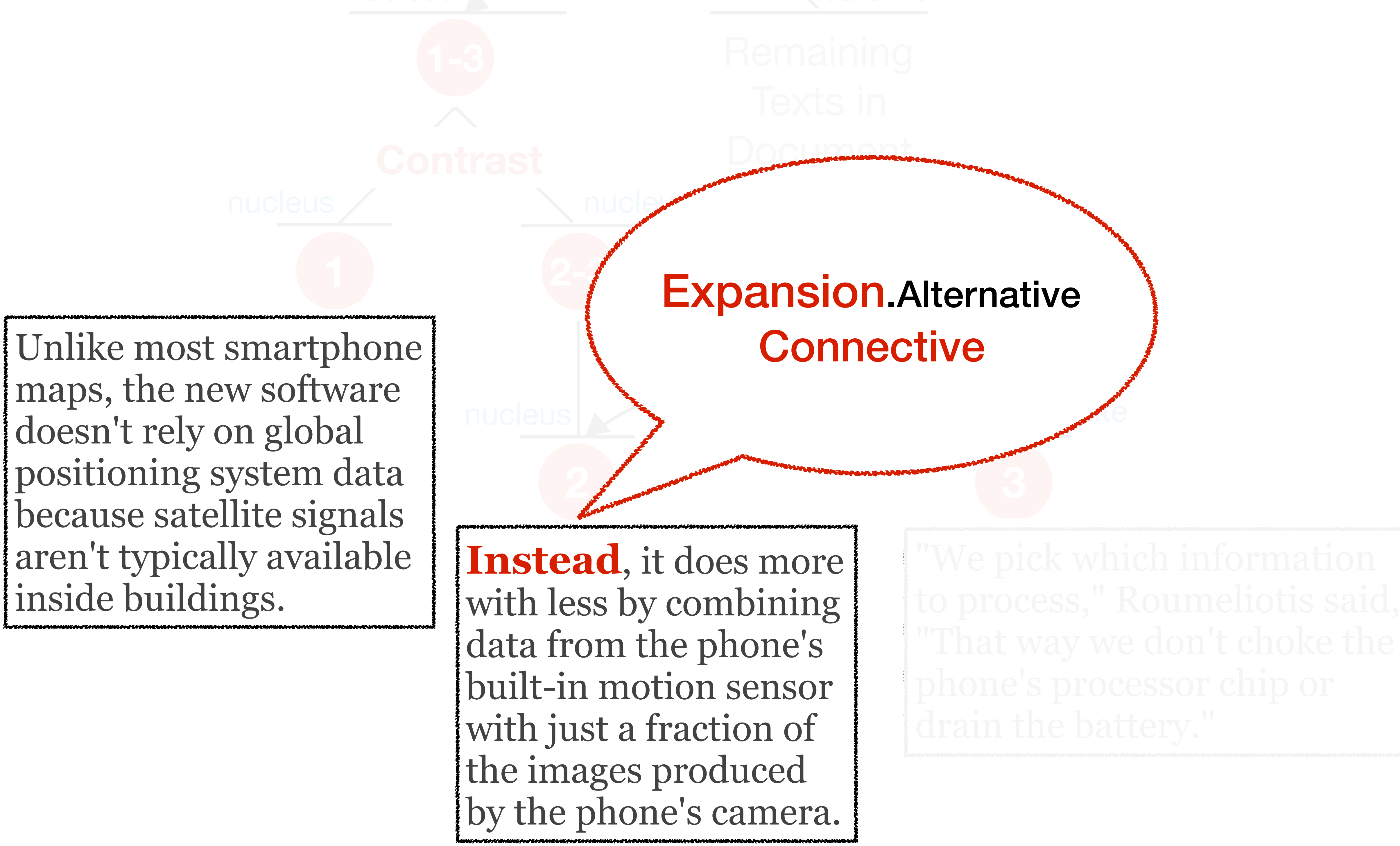
Instead, it does more with less by combining data from the phone's built-in motion sensor with just a fraction of the images produced by the phone's camera.

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

Following the style of Penn Discourse Treebank (Miltsakaki, Eleni, et al. 2004.)

Discourse Connectives

Words or **phrases** that connect or relate two coherent sentences or phrases and indicate the presence of **discourse** relations.



Following the style of Penn Discourse Treebank (Miltsakaki, Eleni, et al. 2004.)

Discourse Connectives

Expansion

indeed, or, as if, instead, rather, further, besides, and, for example, otherwise, for instance, overall, in fact, if then, also, in addition, similarly, moreover, nor

Comparison

meantime, however, while, on the contrary, although, as if, but, still, nevertheless, by contrast, yet, though

Contingency

because, thus, so that, if, when, so, since, as long as, as a result, therefore

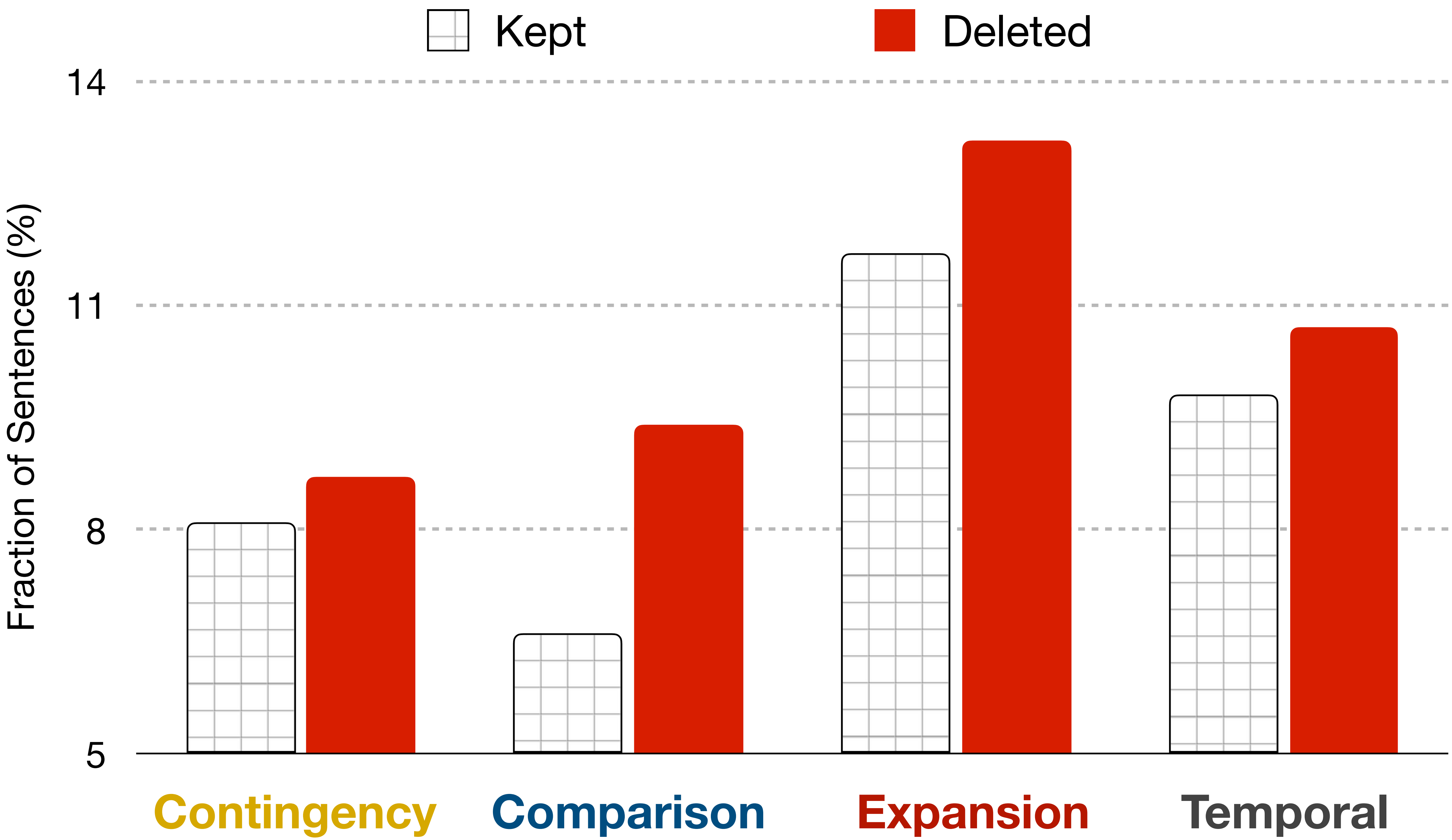
Temporal

later, in turn, when, before, once, while, then, meanwhile, previously, thereafter, since, after, as, ultimately, afterward, until

Following the style of Penn Discourse Treebank (Miltsakaki, Eleni, et al. 2004.)

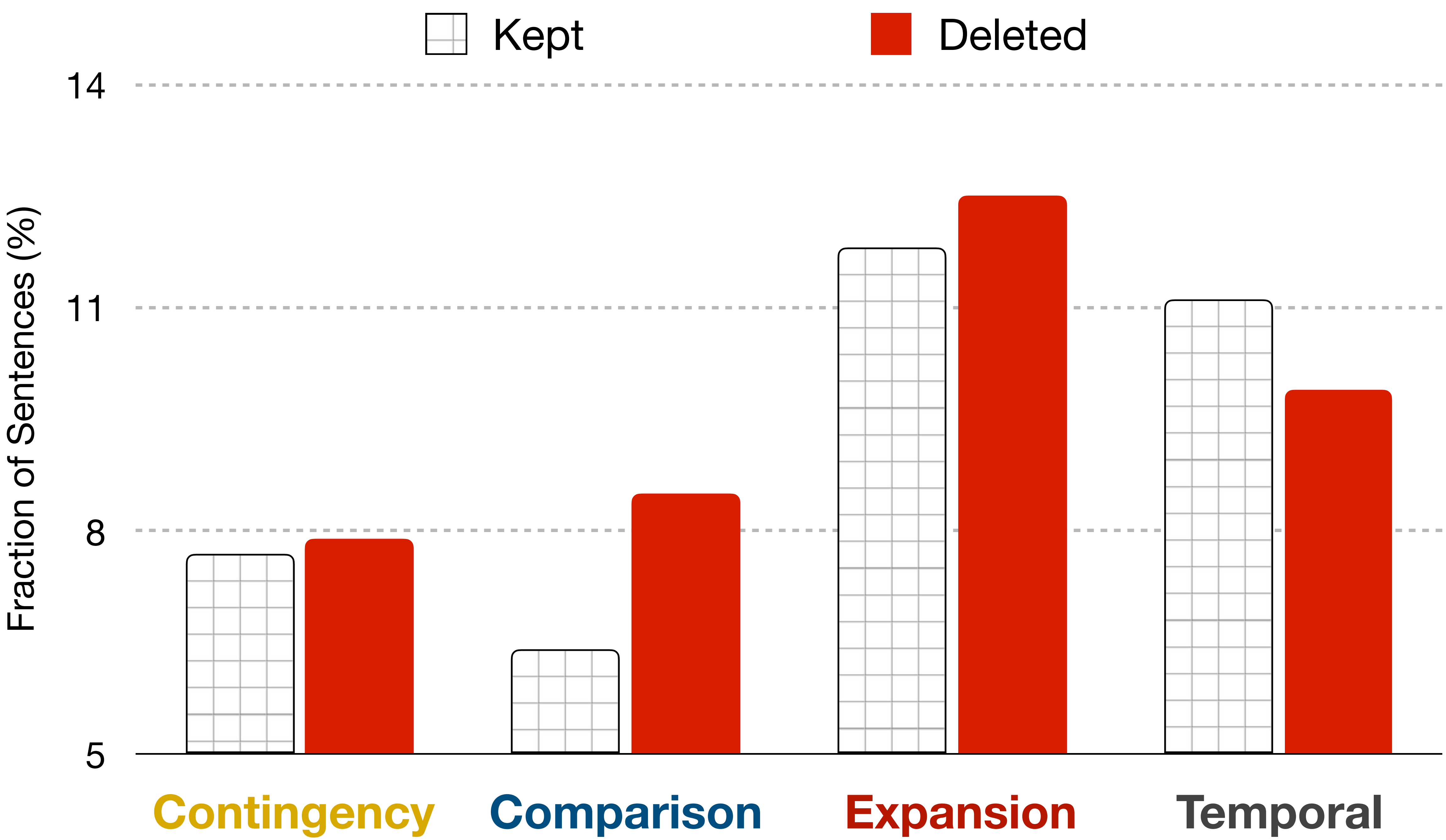
Discourse Connectives in Elementary School

Discourse Connectives in Elementary School



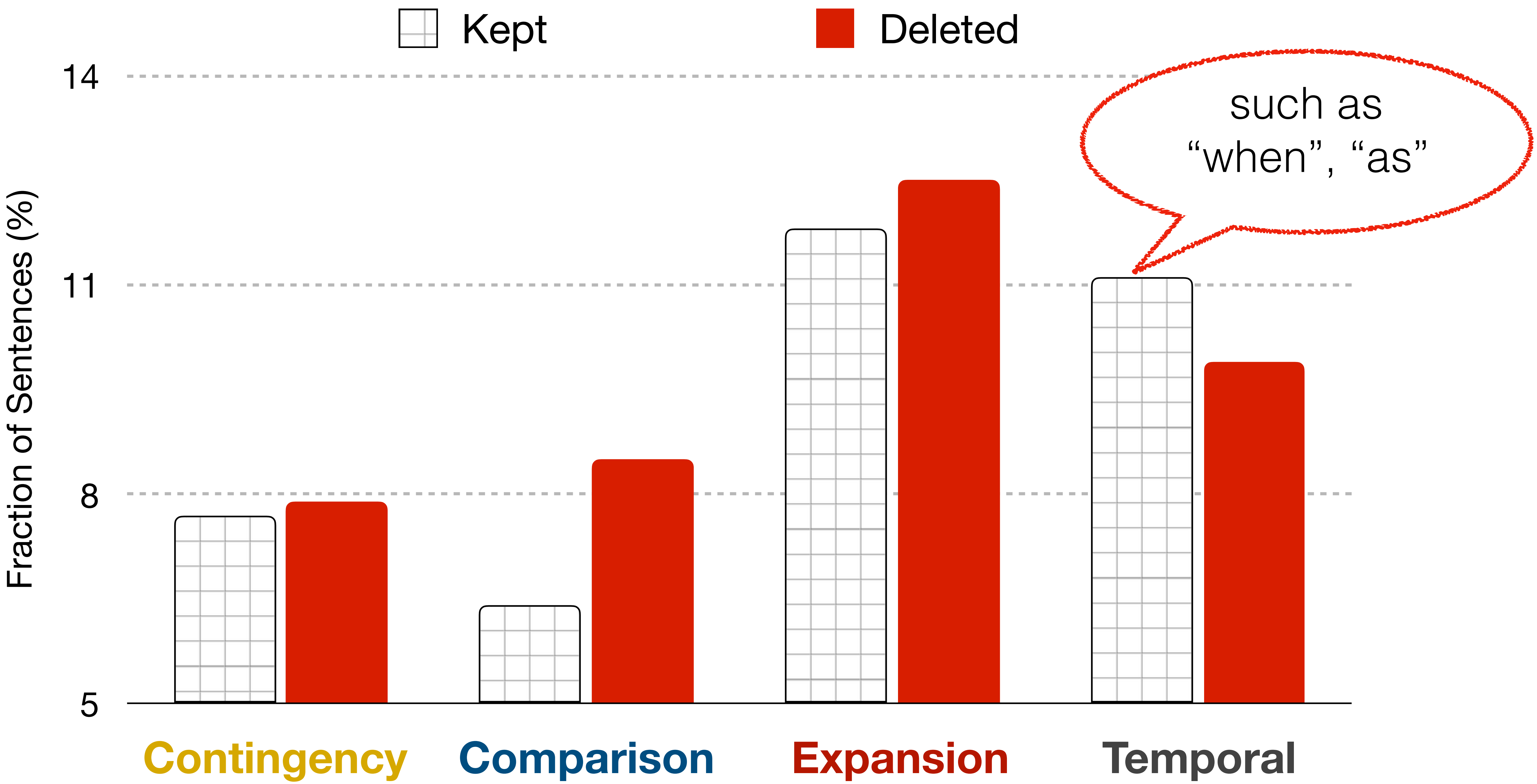
Sentences with discourse connectives are more likely to be deleted.

Discourse Connectives in Middle School



Sentences with discourse connectives are more likely to be deleted.
But, less so for middle school than elementary school.

Discourse Connectives in Middle School

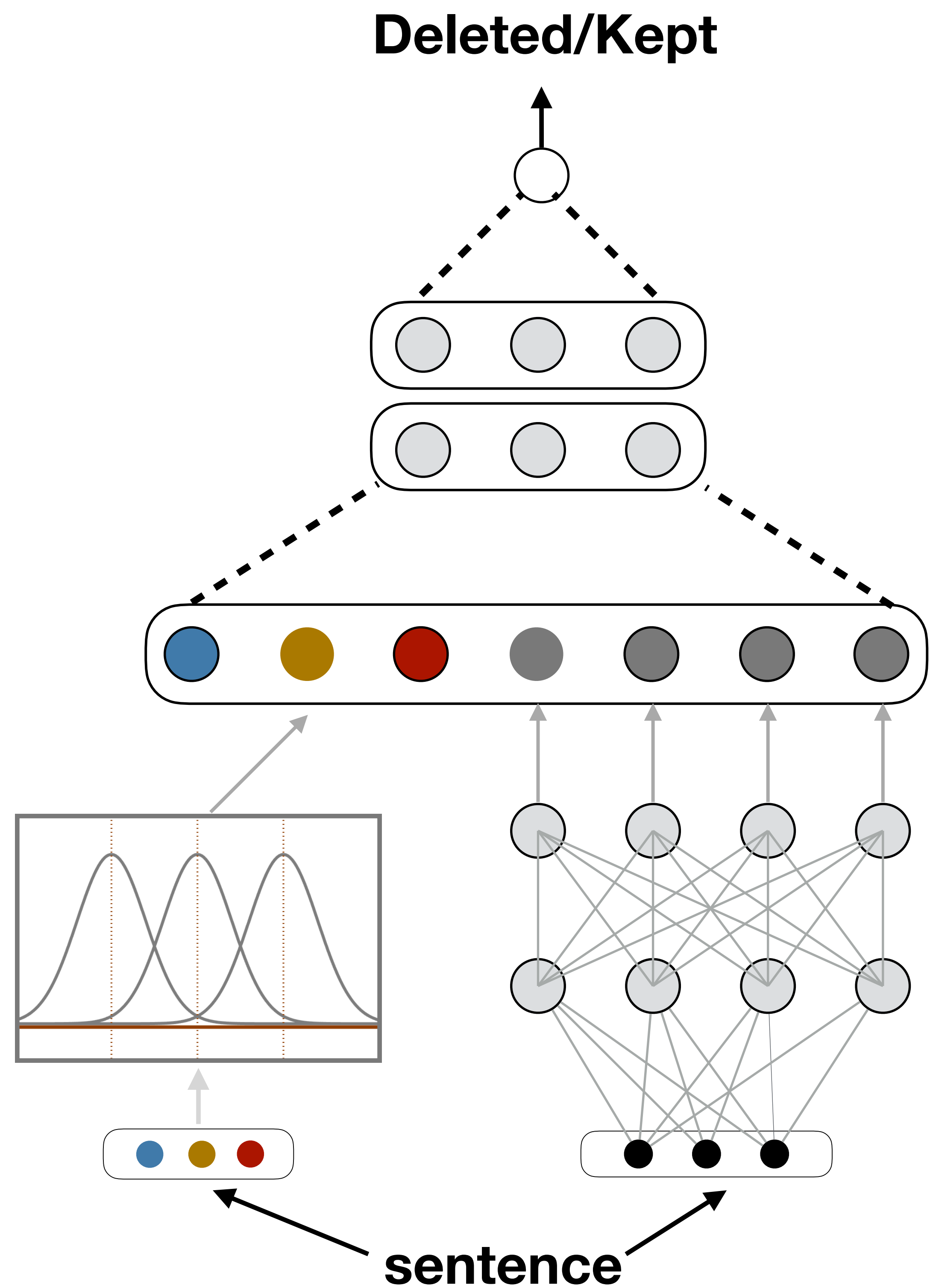


Sentences with discourse connectives are more likely to be deleted.
But, less so for middle school than elementary school.

Our Work

- Manually annotated corpus with sentence alignments.
- Analysis of discourse level factors affecting the deletion of sentences.
  - Governing relation of sentence in RST tree.
  - Discourse connectives in sentence.
- **Automatic prediction of sentence's deletion.**

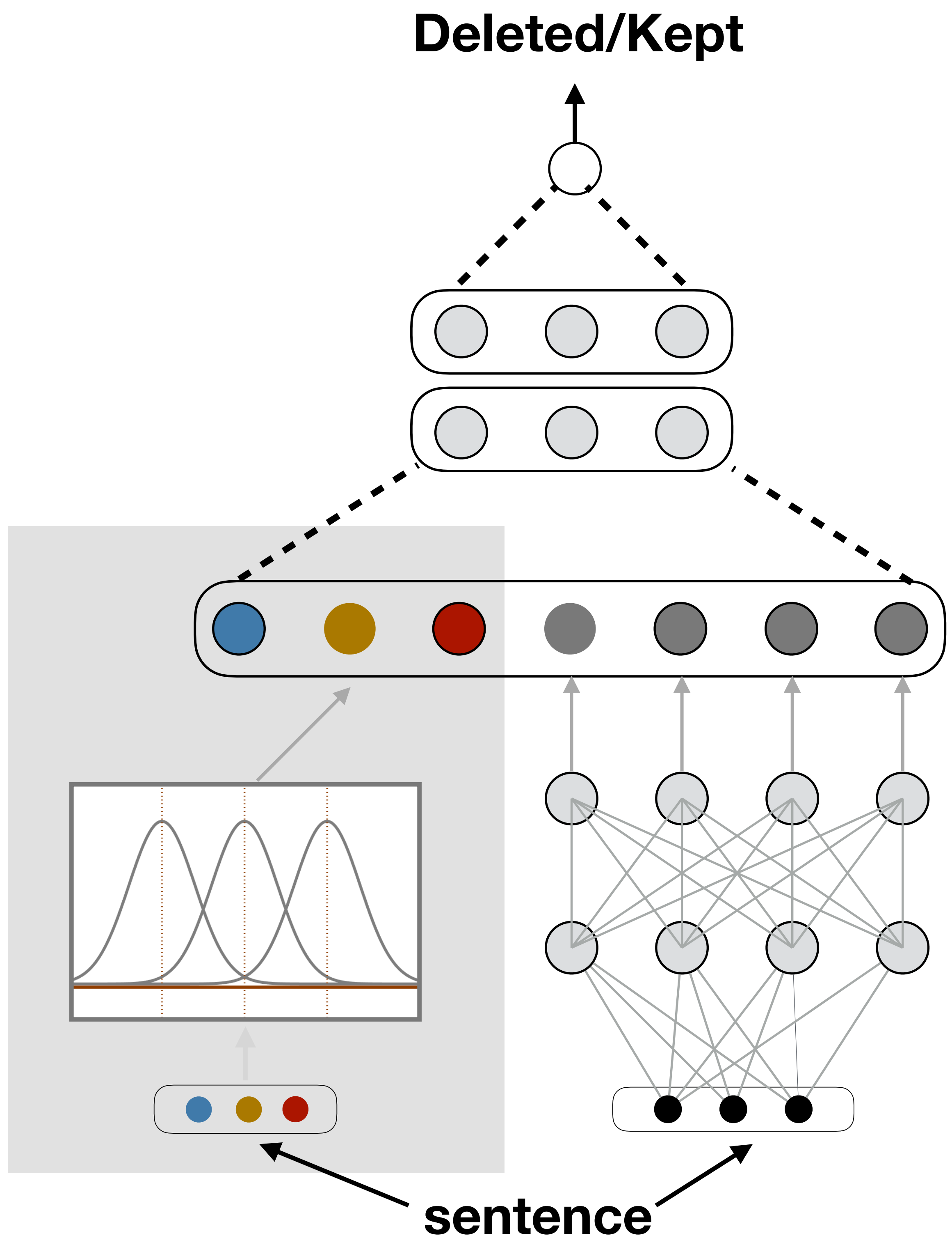
Predicting Sentence Deletion



Features

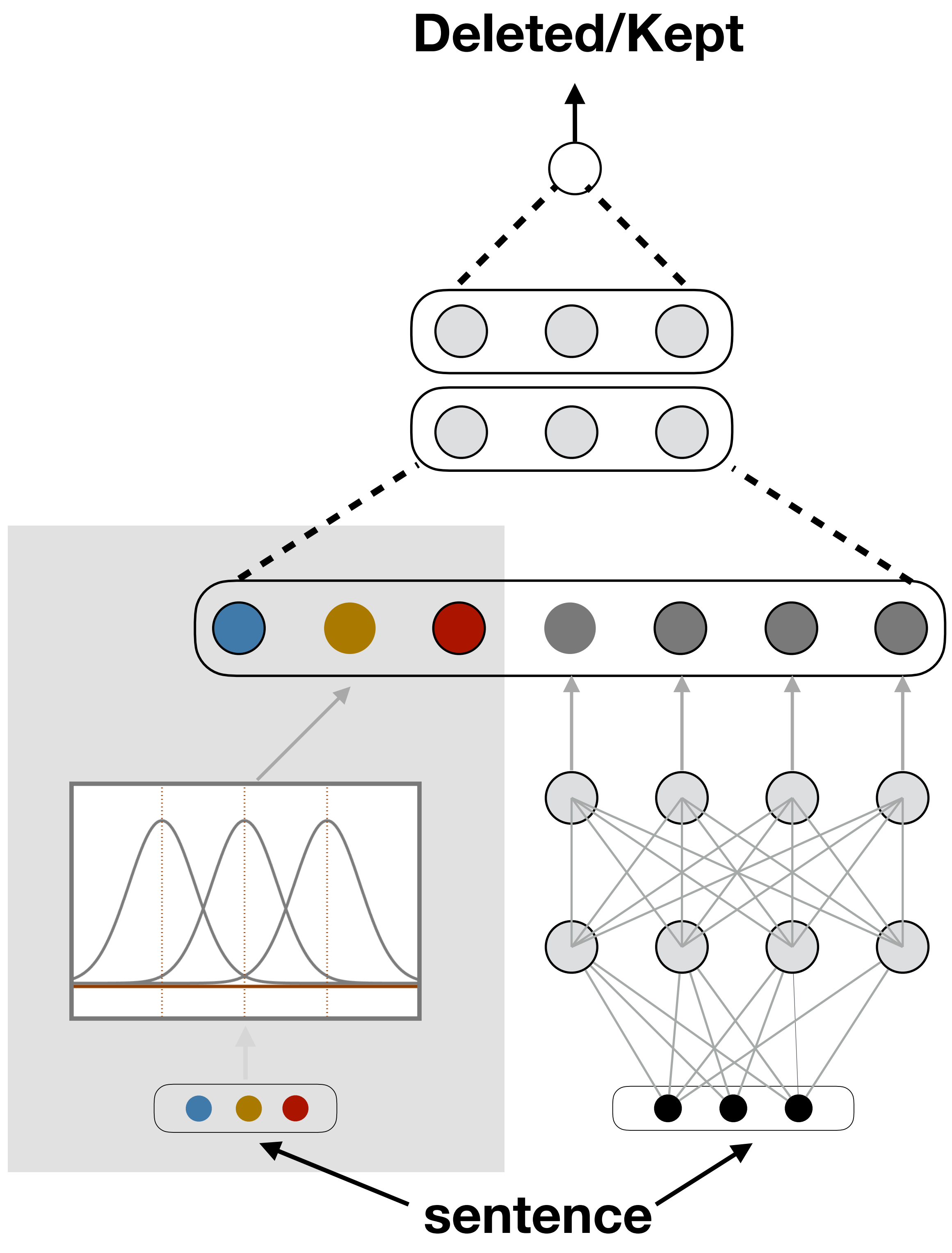
Document characteristics

- Number of sentences
- Number of tokens
- Topic



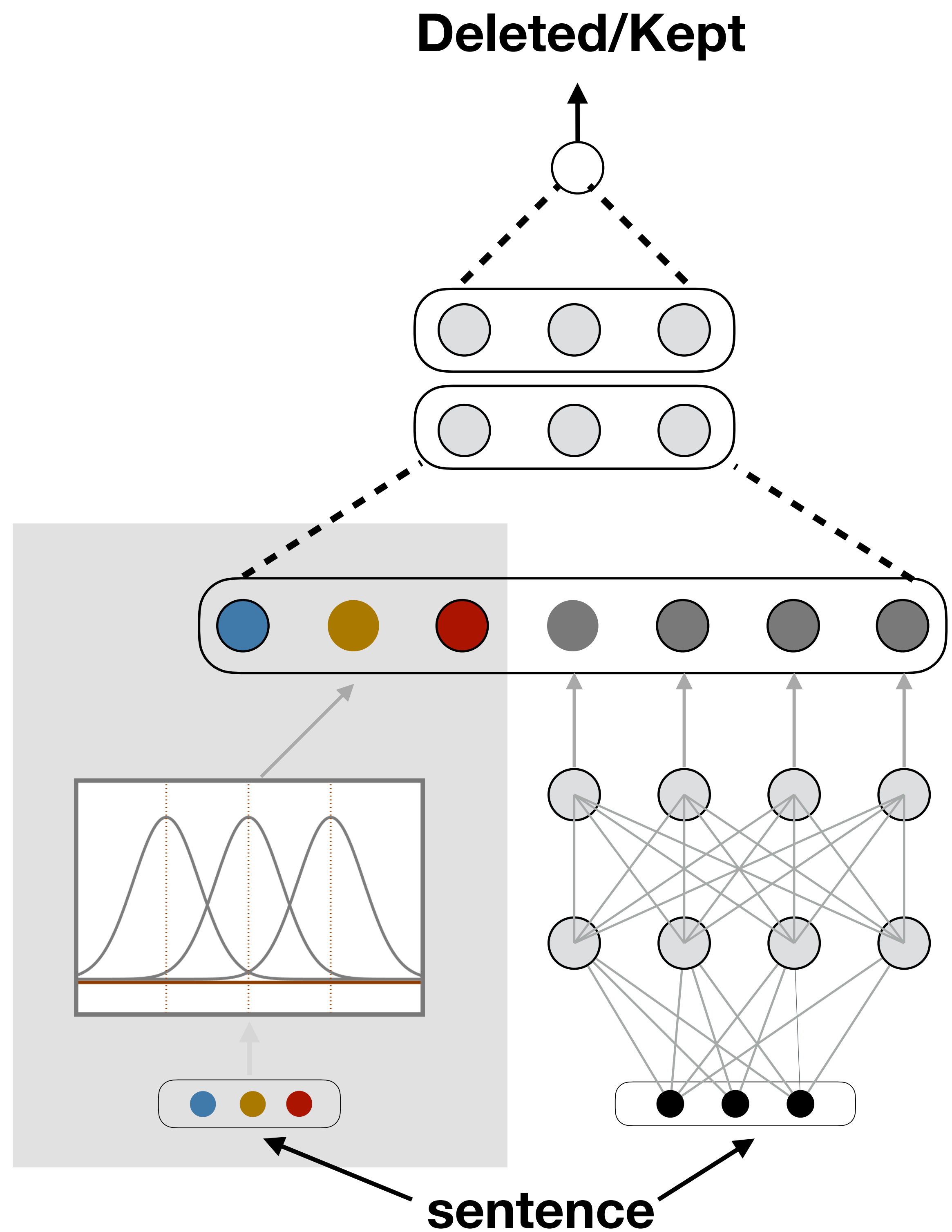
Features

- Document characteristics
- **Discourse features**
  - Depth of sentence in RST tree
  - Indicator of nuclearity
  - Governing relation
  - Indicator of explicit connectives
  - Position of discourse connectives



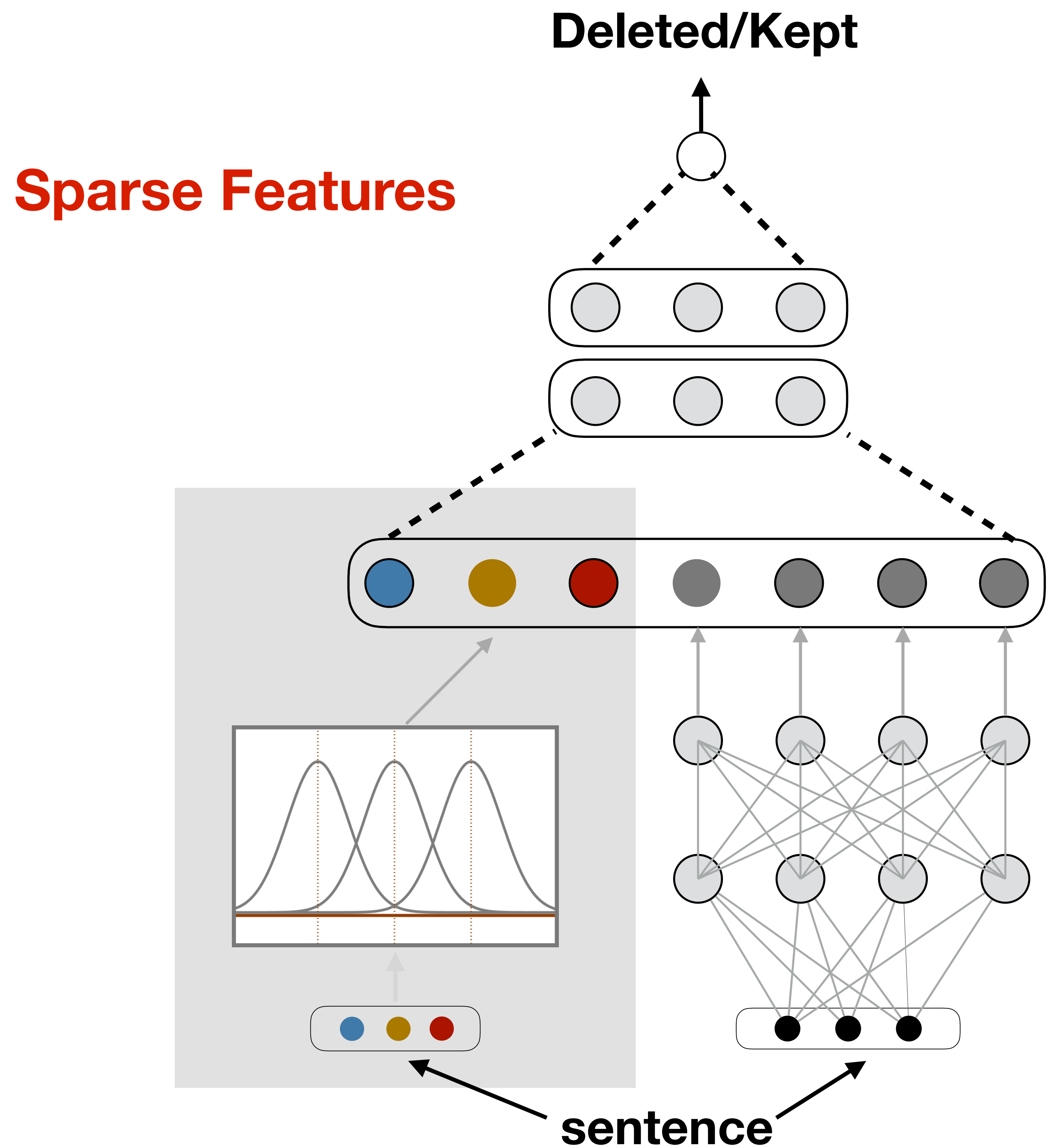
Features

- Document characteristics
- Discourse features
- **Position features**
  - Sentence's position in document
  - Paragraph's relative position
  - Sentence's position inside paragraph



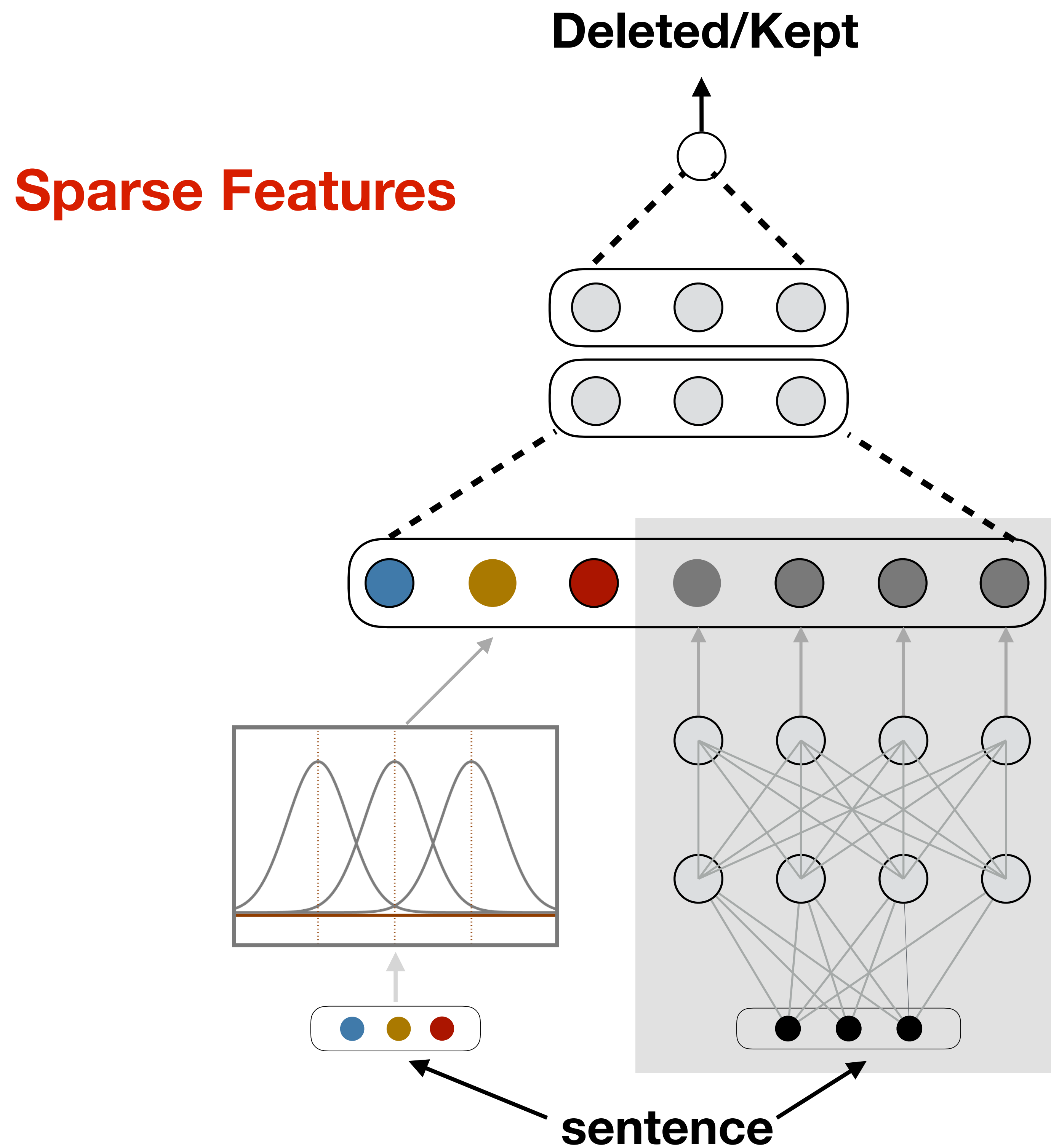
Features

- Document characteristics
- Discourse features
- Position features



Features

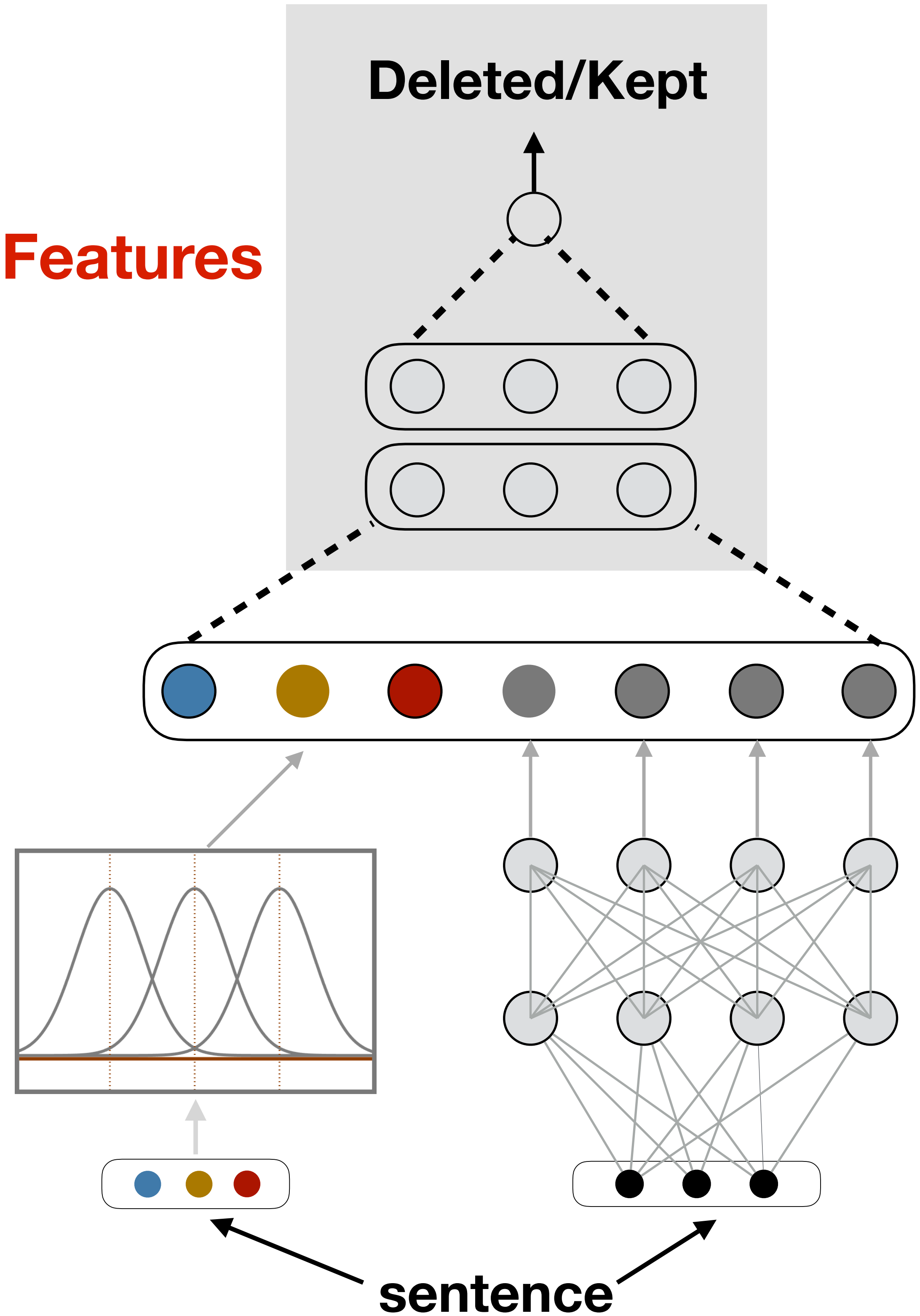
- Document characteristics
- Discourse features
- Position features
- **Semantic features**
  - 300D GloVe Embeddings



Features

- Document characteristics
- Discourse features
- Position features
- Semantic features

Sparse Features

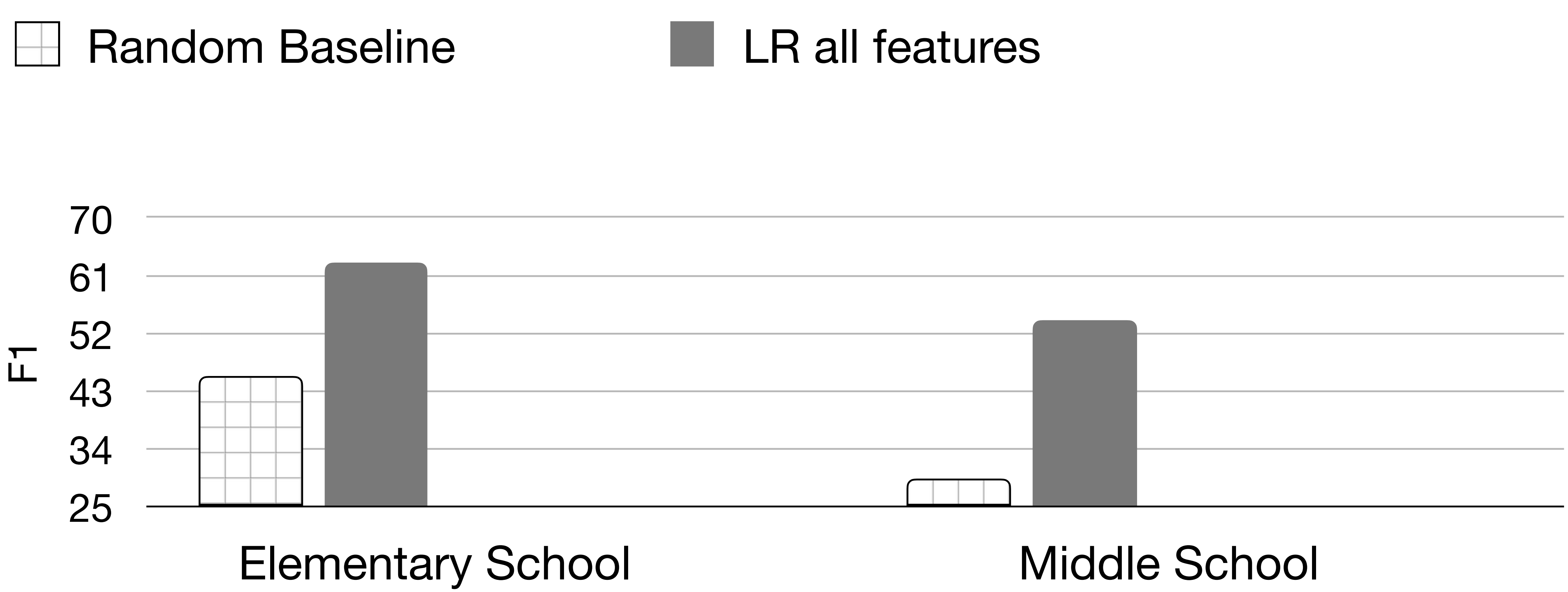


Dataset & Evaluation

- **Training set:** 42,264 sentences in 886 articles **automatically** aligned using Sent2Vec from the Newsela dataset (Pagliardini, Gupta, and Jaggi 2018).
- **Dev/Test sets:** 450/1838 sentences in the 50 articles **manually** aligned.

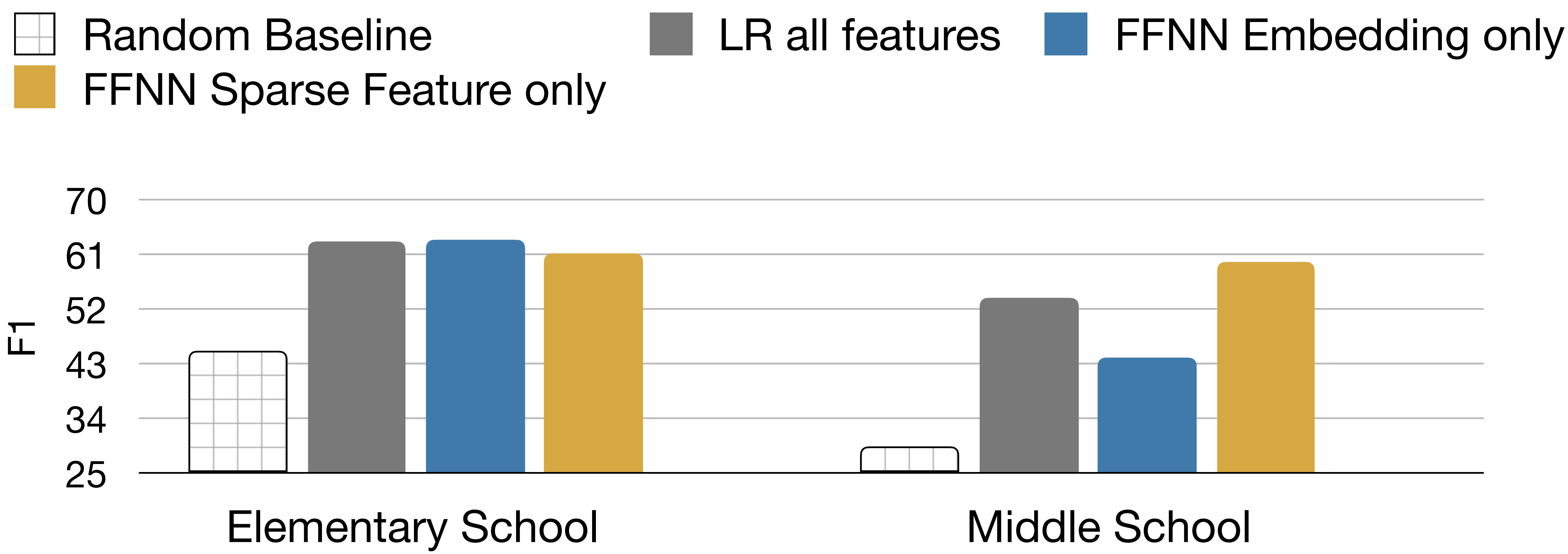
Results (predicting which sentence will be deleted)

- Middle school is harder to predict than elementary school.
- Both sparse features and word embeddings can help.
- FFNN+Gaussian Layer works better than Logistic Regression Model.



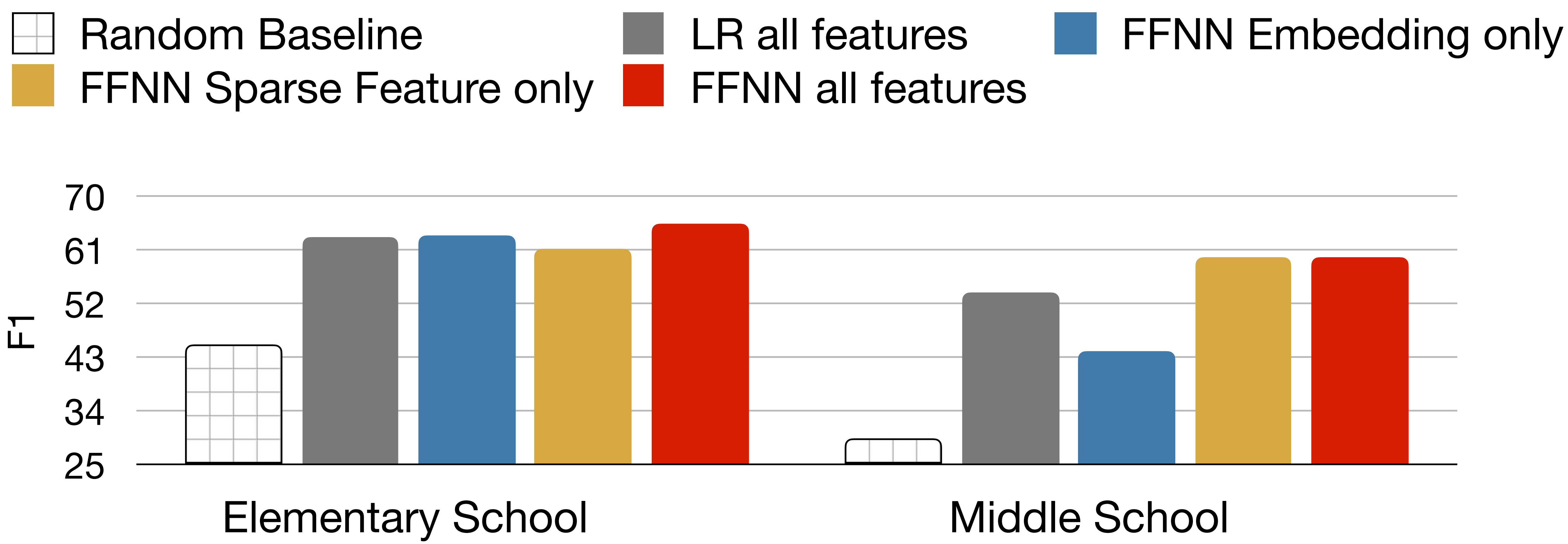
Results (predicting which sentence will be deleted)

- Middle school is harder to predict than elementary school.
- Both sparse features and word embeddings can help.
- FFNN+Gaussian Layer works better than Logistic Regression Model.



Results (predicting which sentence will be deleted)

- Middle school is harder to predict than elementary school.
- Both sparse features and word embeddings can help.
- FFNN+Gaussian Layer works better than Logistic Regression Model.



Takeaways

- Manually aligned corpus can help text simplification task.
- Discourse level factors are associated with sentence deletion.
- Discourse level factors contribute to the challenging task of predicting sentence deletion for simplification

Discourse Level Factors for Sentence Deletion in Text Simplification

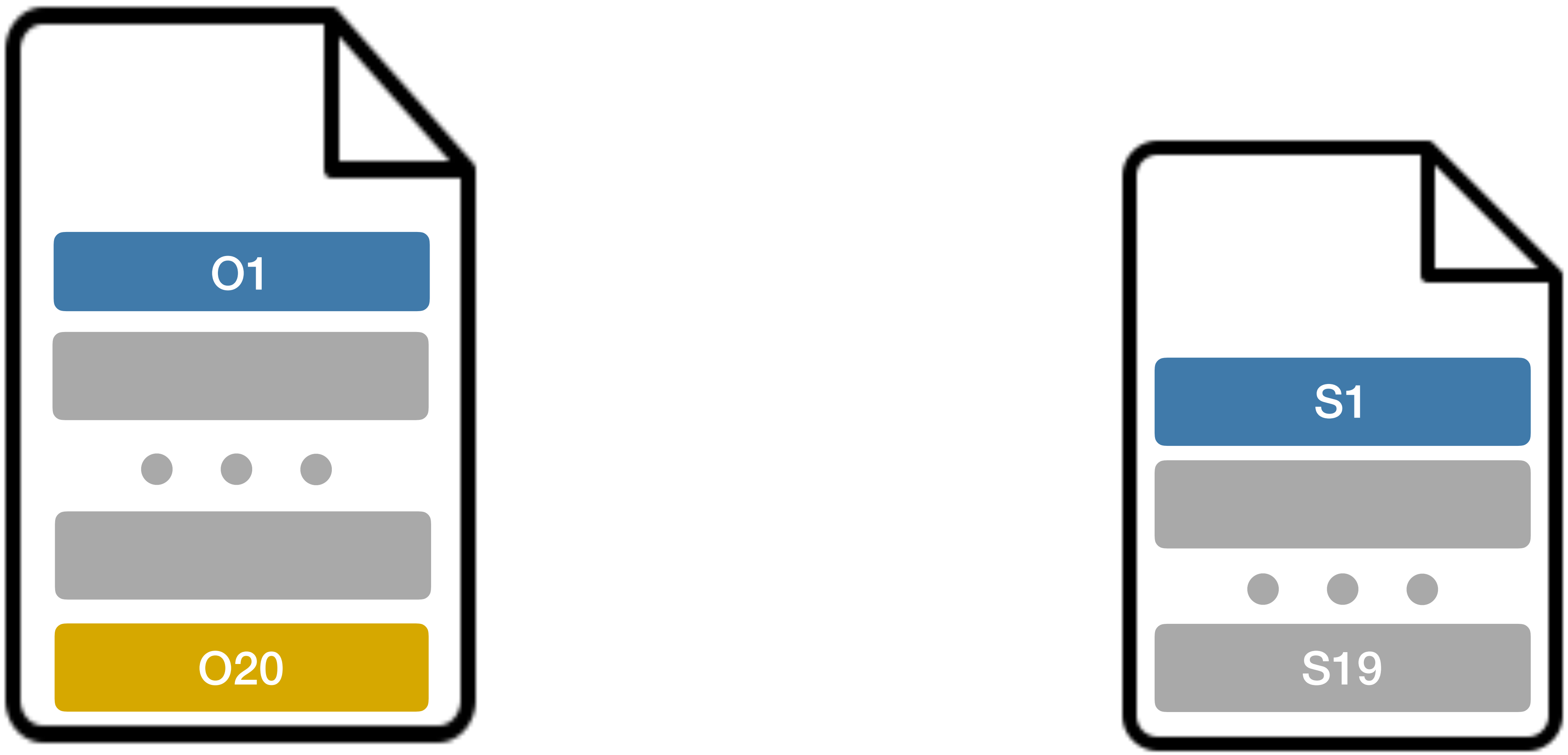
Yang Zhong, Chao Jiang, Wei Xu, and Junyi Jessy Li

Thank you!

Q & A

Backup Pages

Automatic Alignment

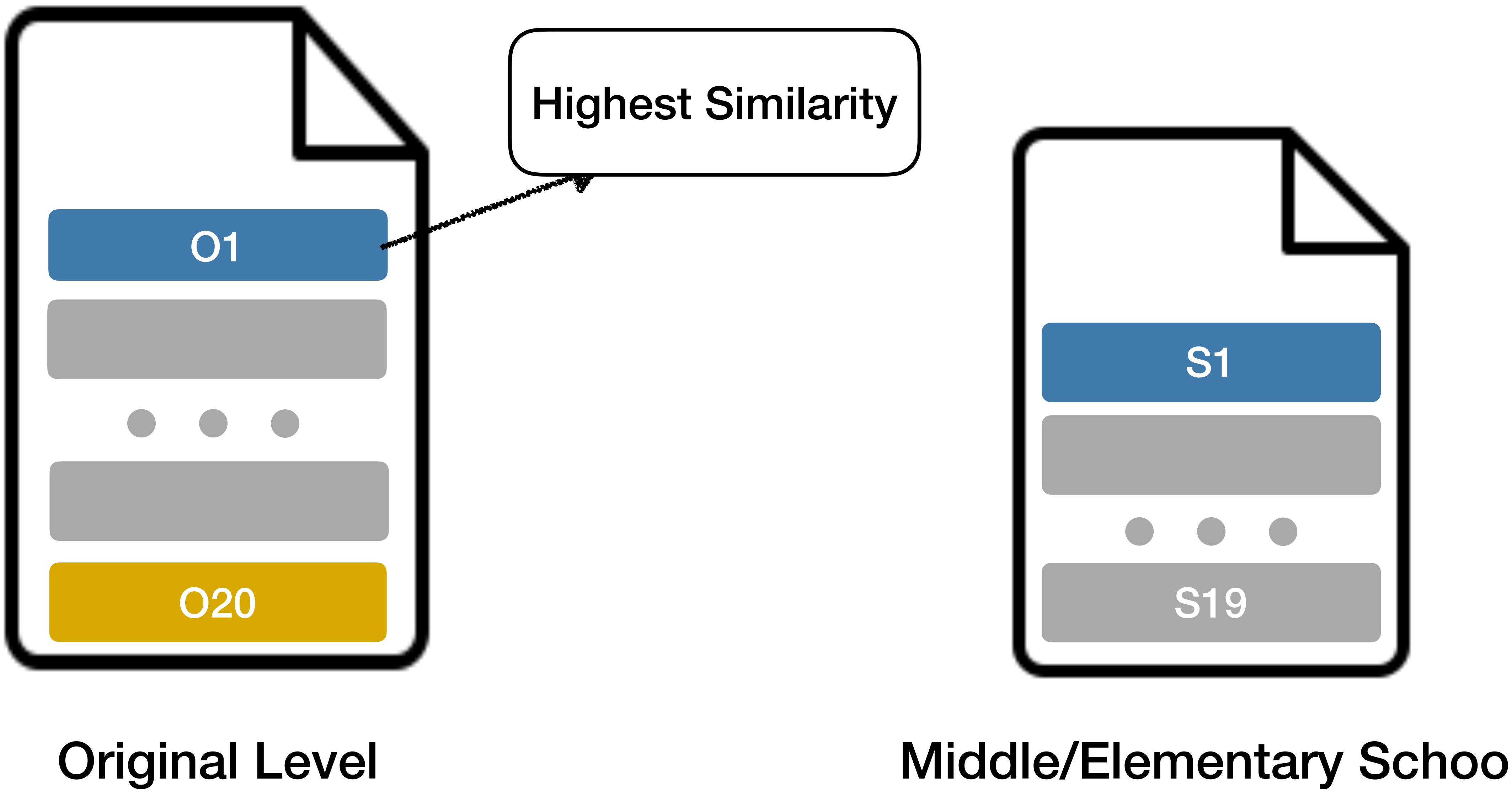


Original Level

Middle/Elementary School

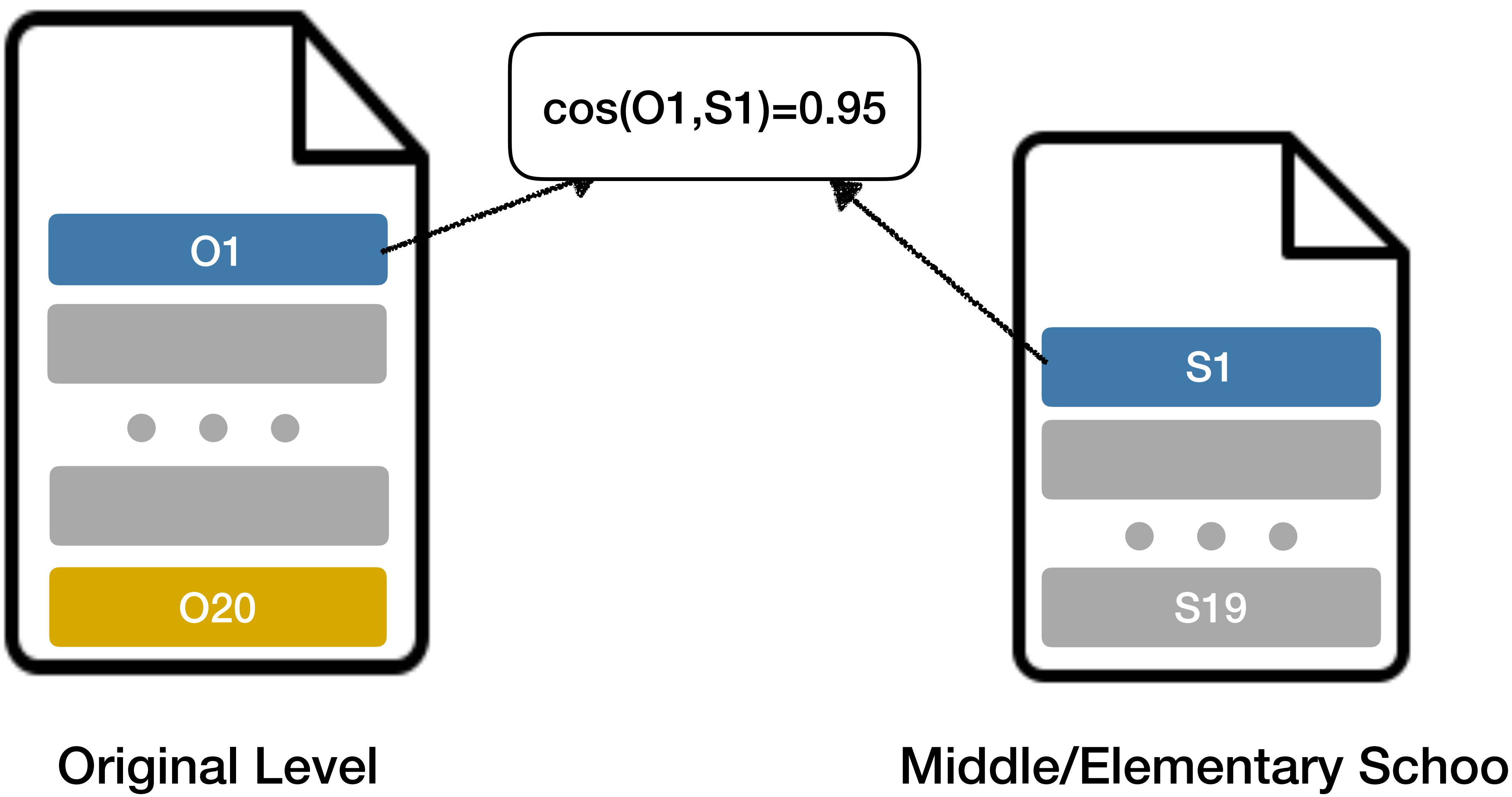
cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment



cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment



cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

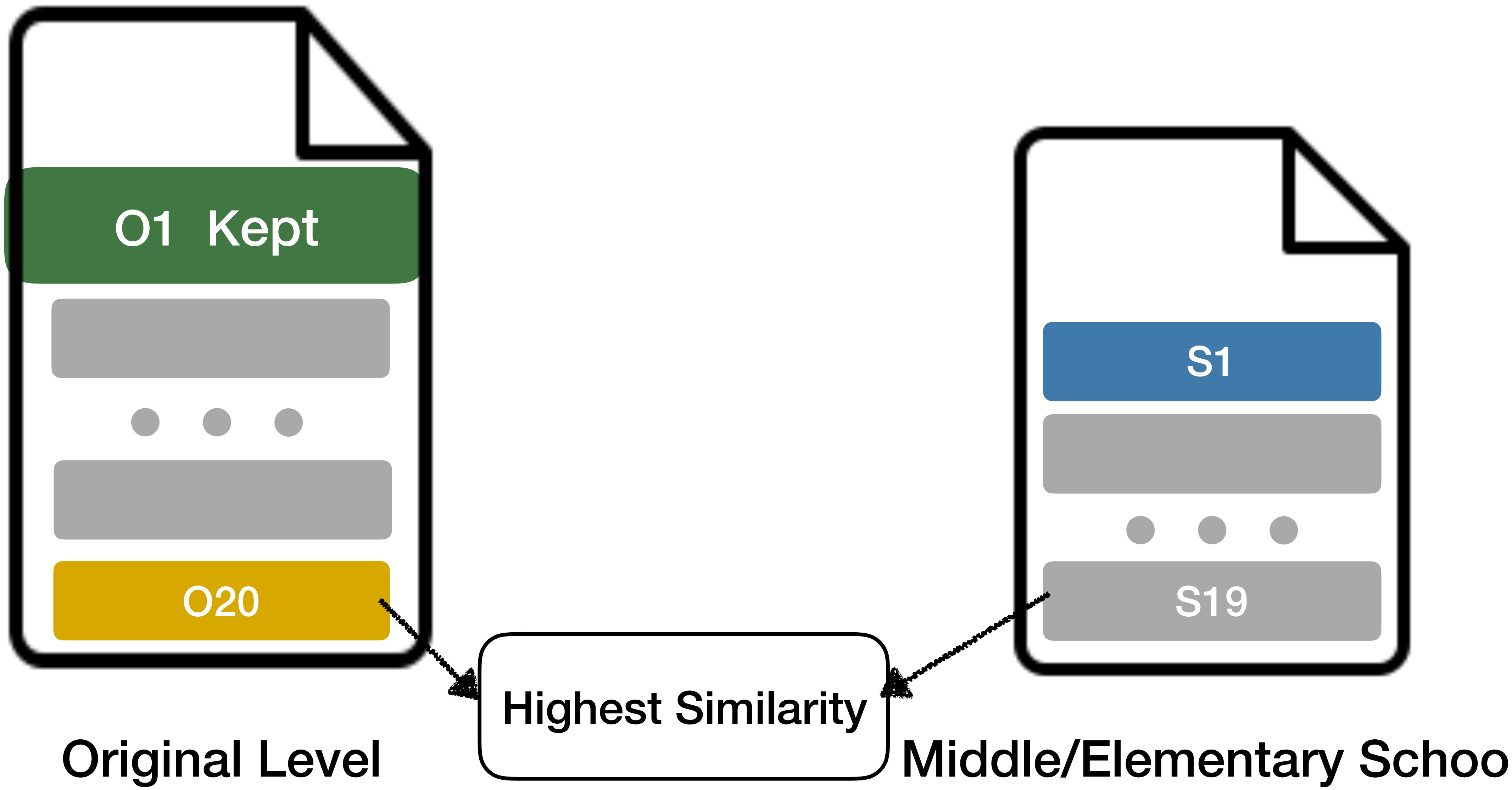
Automatic Alignment

Original Level

Middle/Elementary School

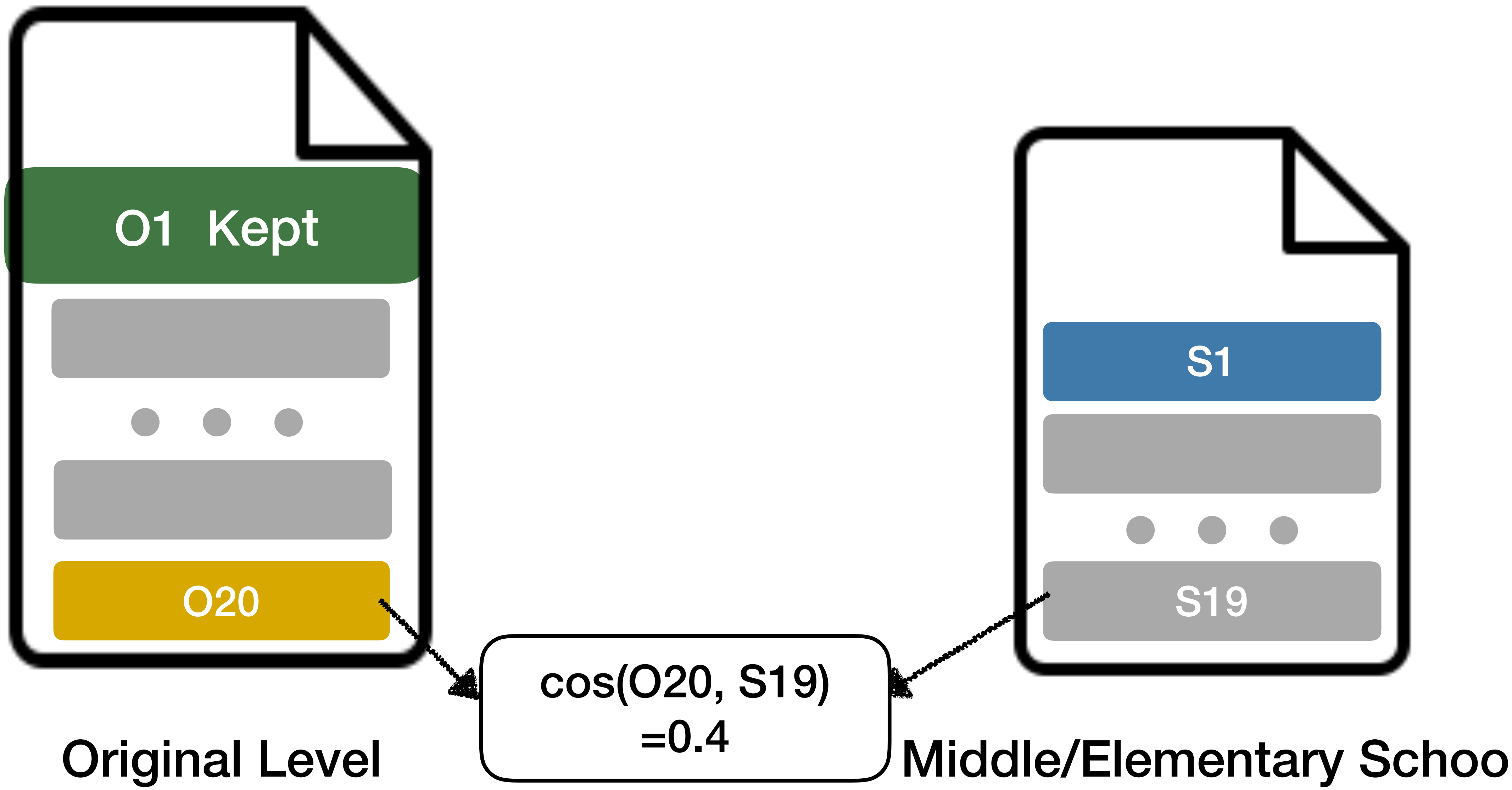
cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment



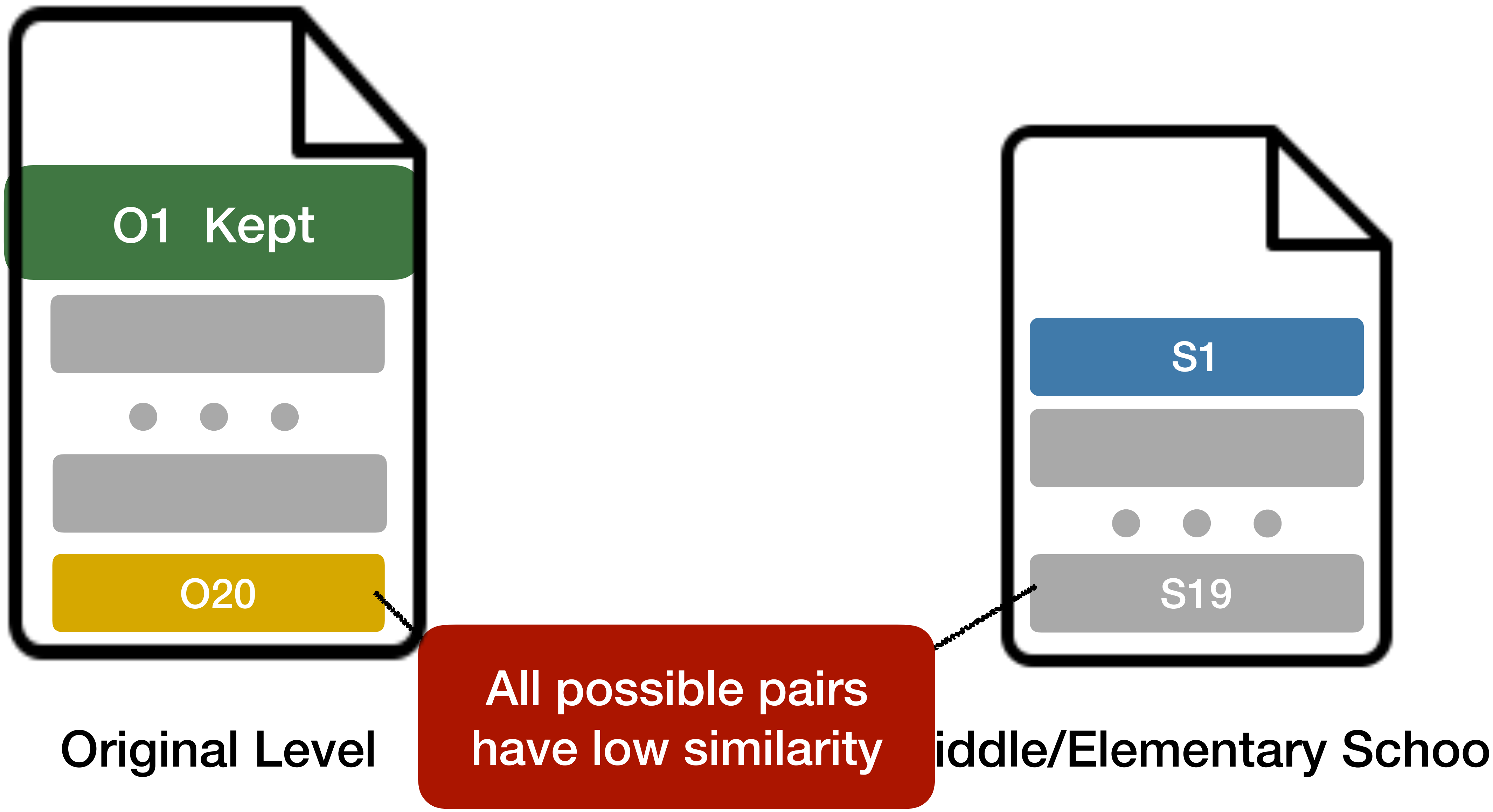
cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment



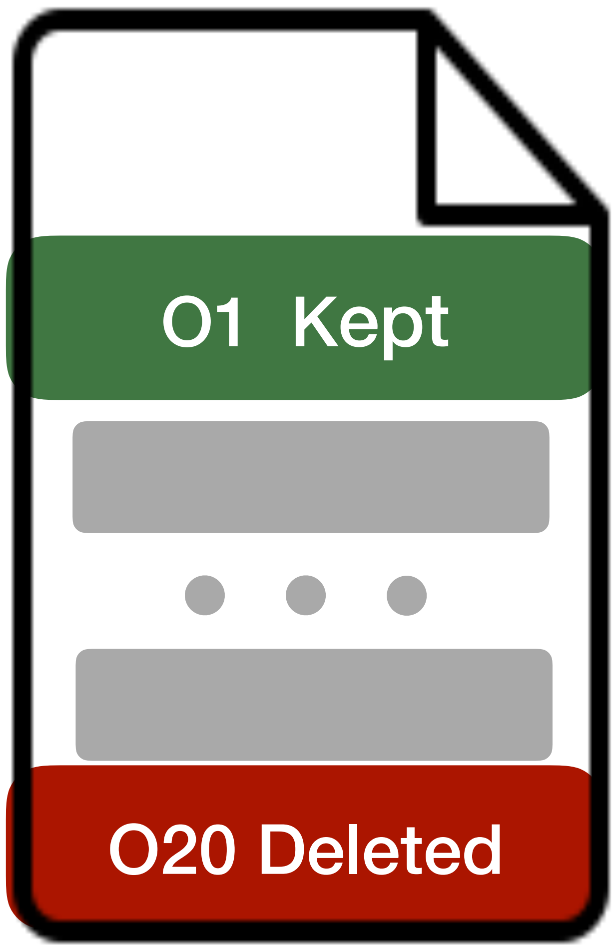
cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment

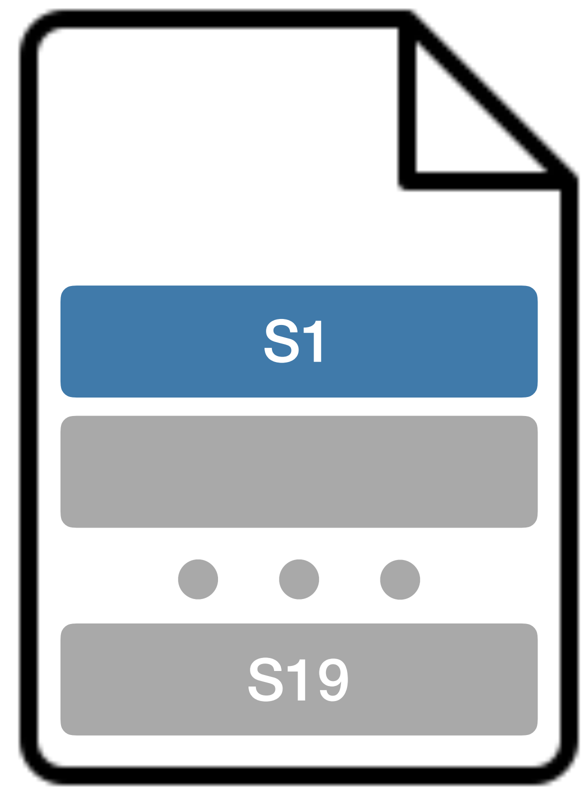


cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Automatic Alignment



Original Level



Middle/Elementary School

cosine similarity based on 700D sentence embedding Sent2Vec (Pagliardini, Gupta, and Jaggi 2018)

Case Study on Temporal connectives

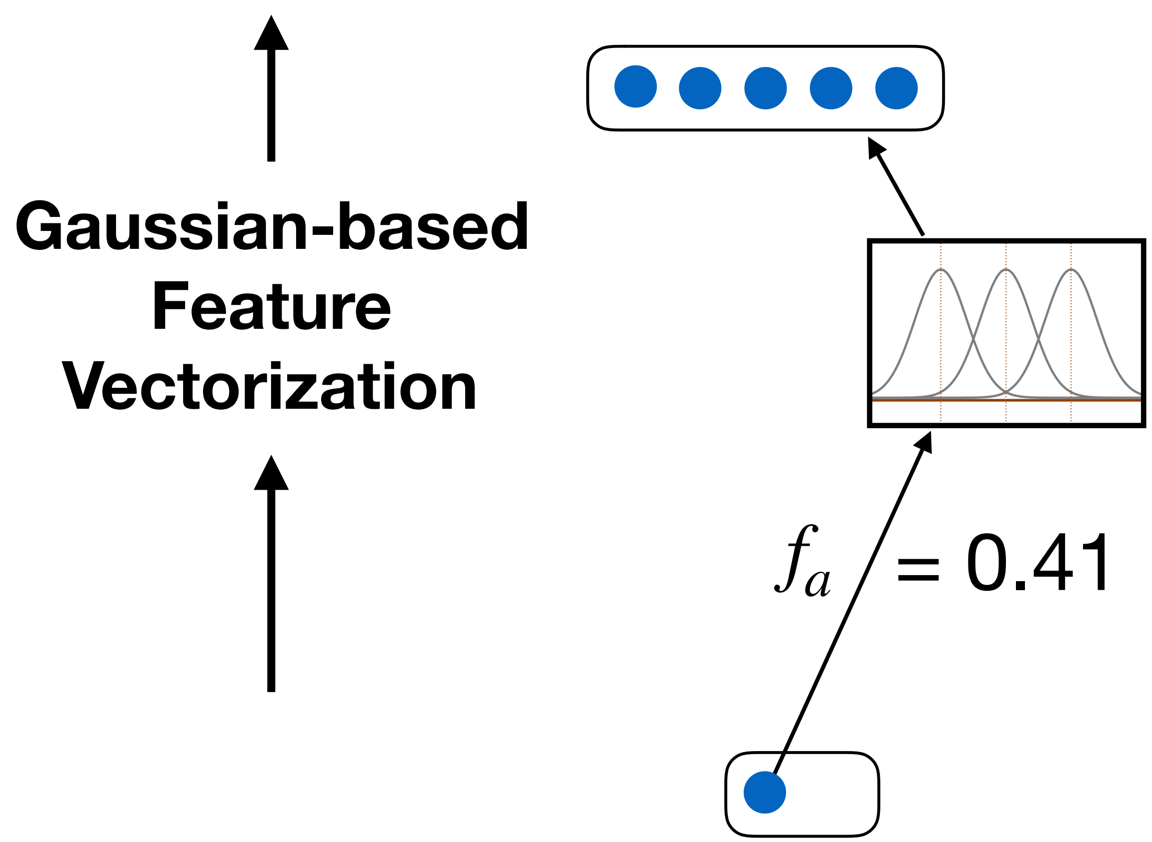
Still, the attention to the issue is a shift from decades ago, **when** Los Angeles and other major cities battled crippling smog and treated it as a local matter.

Now that climate change has put the spotlight on the global rise of carbon dioxide, other pollutants are increasingly being viewed in the same way, as international concerns.

Temporal connectives will presuppose the event involved in the whole context.

Gaussian binning Vectorization

$\overrightarrow{f_a}$ = [~0.0, **0.44**, **0.54**, ~0.02, ~0.0]



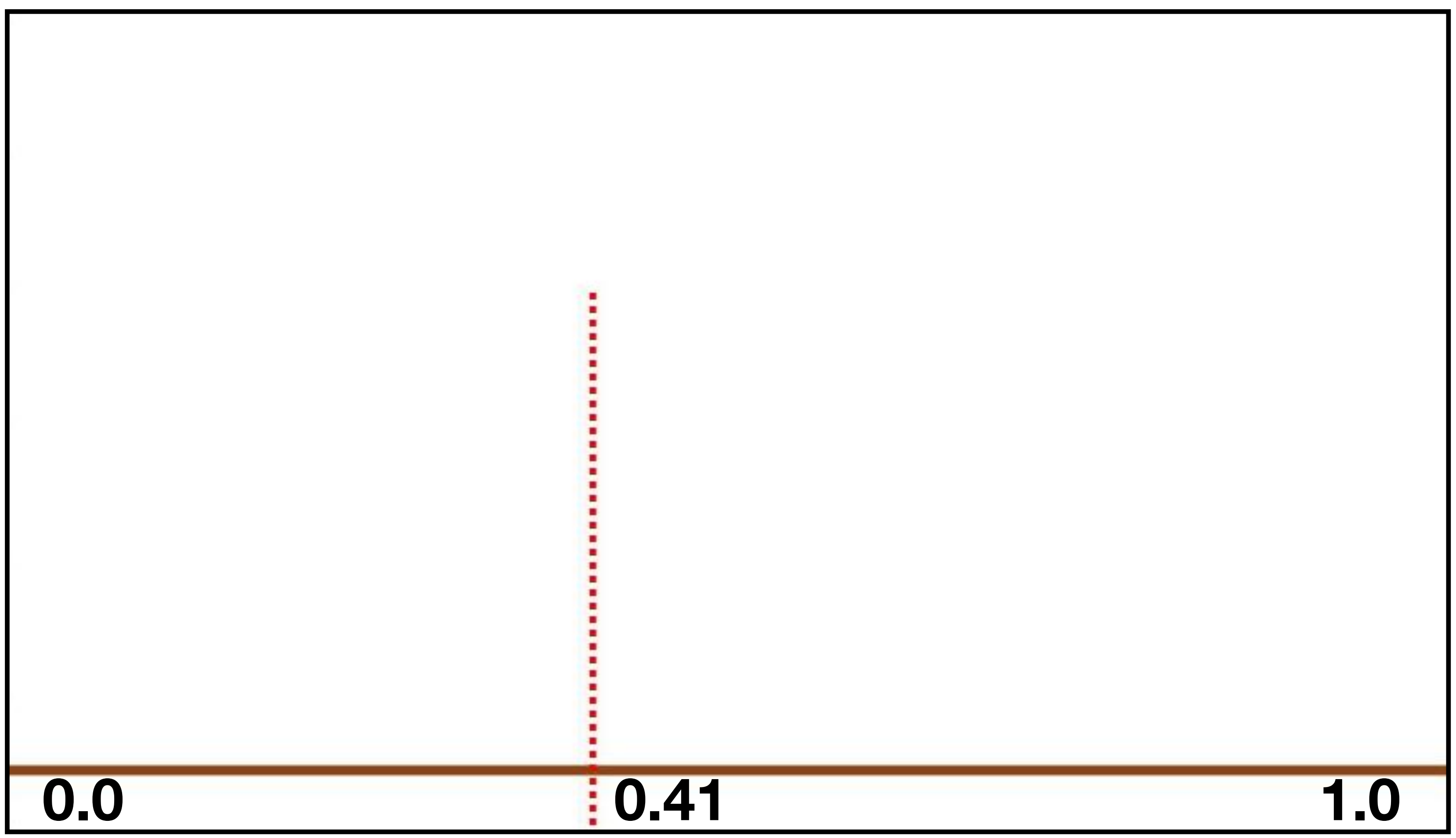
$$d_j(f) = e^{-\frac{(f-\mu_j)^2}{2\sigma^2}}$$

*smooth binning approach (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w)$ = 0.41, $f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [~0.0, 0.44, 0.54, ~0.02, ~0.0]

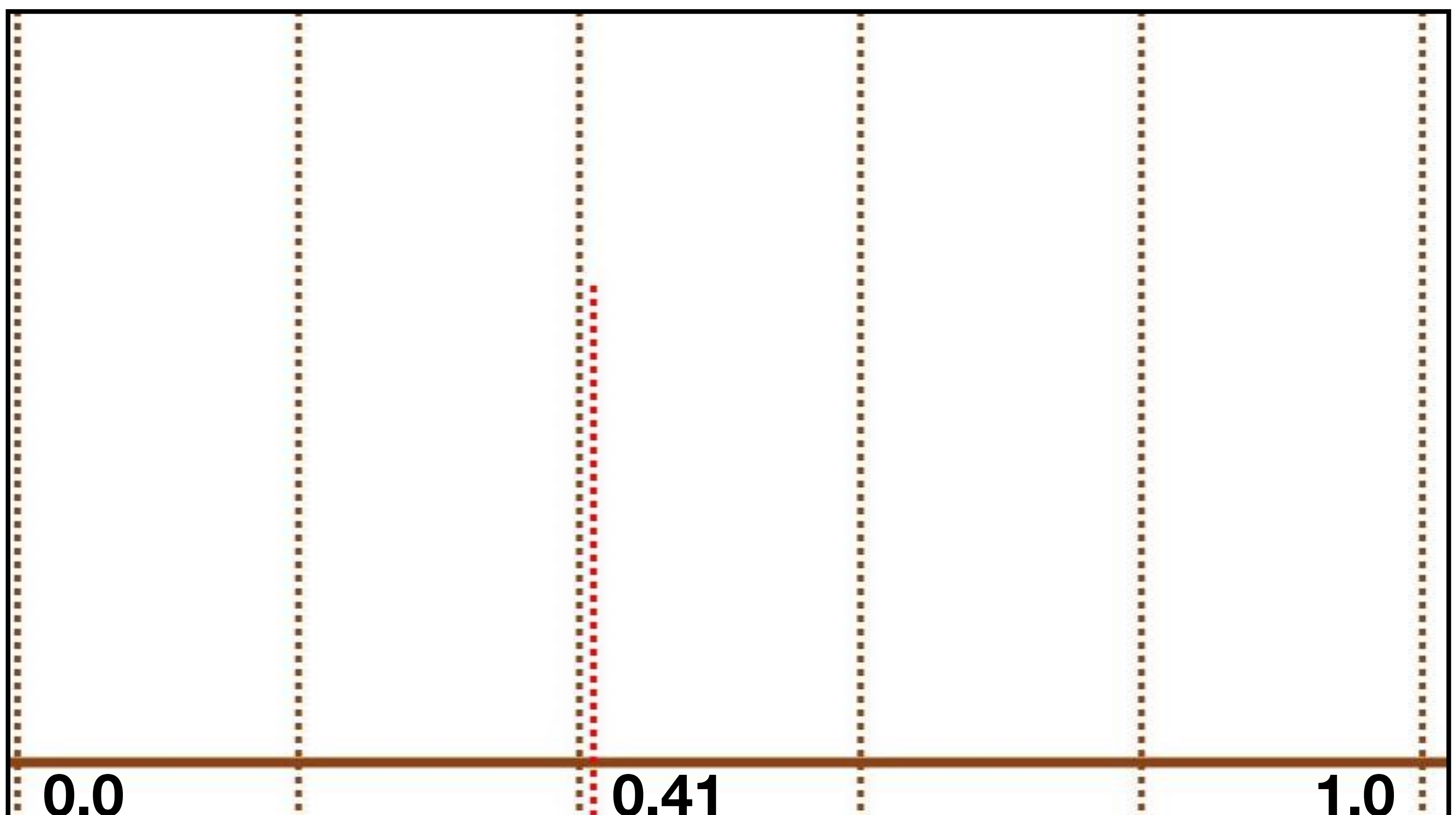


*slides credit to (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w)$ = 0.41, $\quad f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [~0.0, 0.44, 0.54, ~0.02, ~0.0]

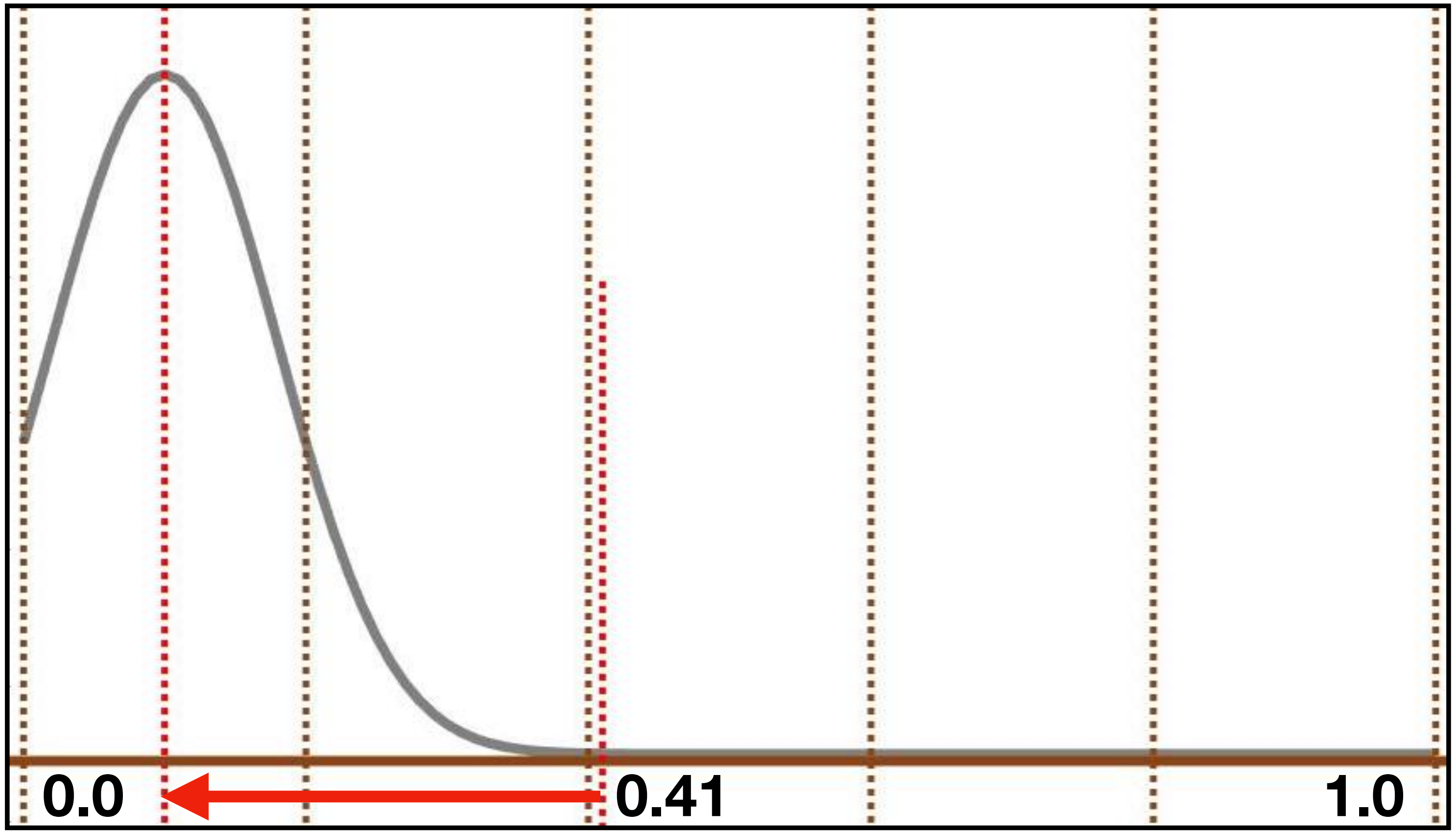


*slides credit to (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w)$ = 0.41, $f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [**~0.0**,]

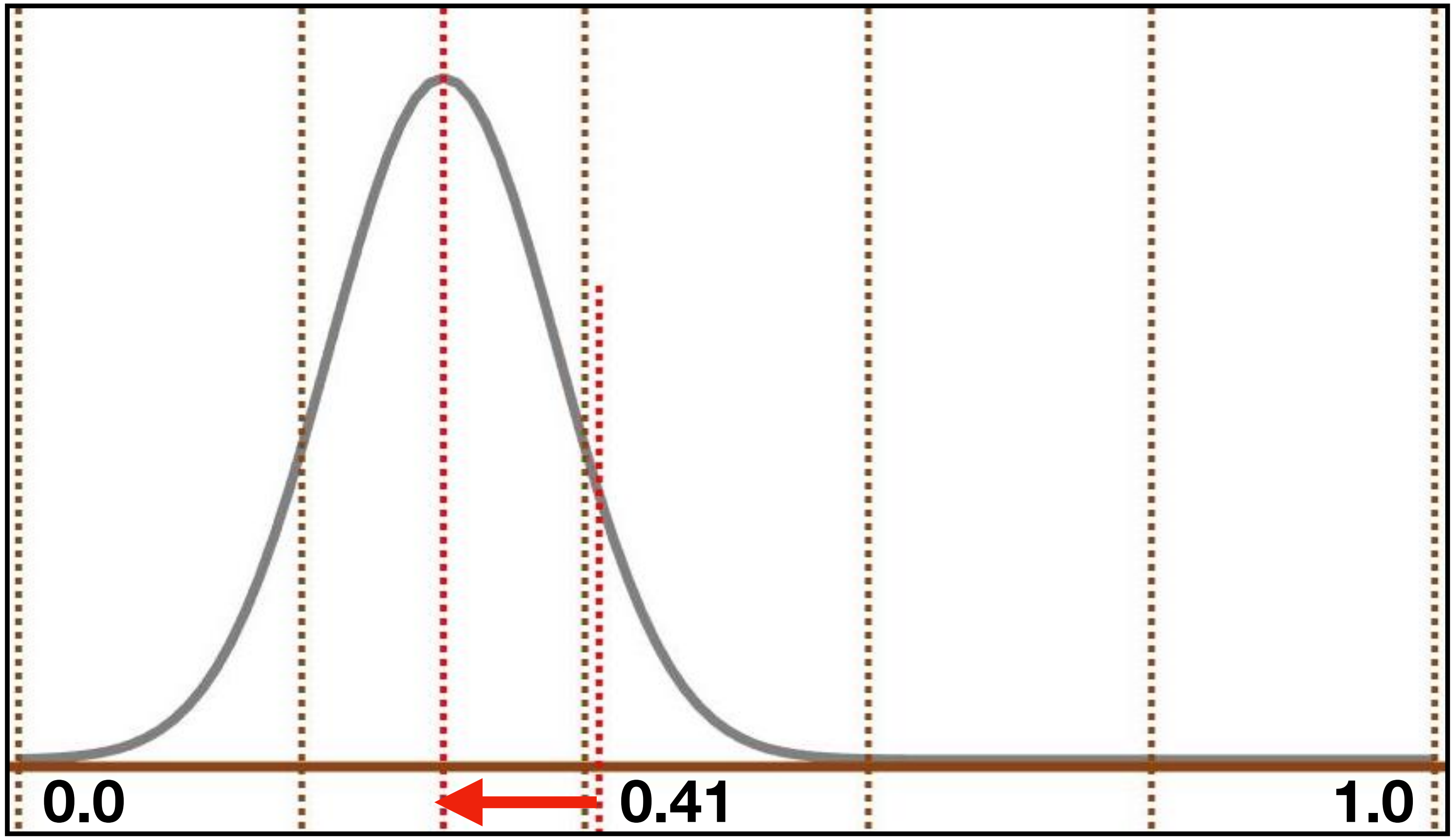


*slides credit to (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41, \quad f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [~0.0, **0.44**,]

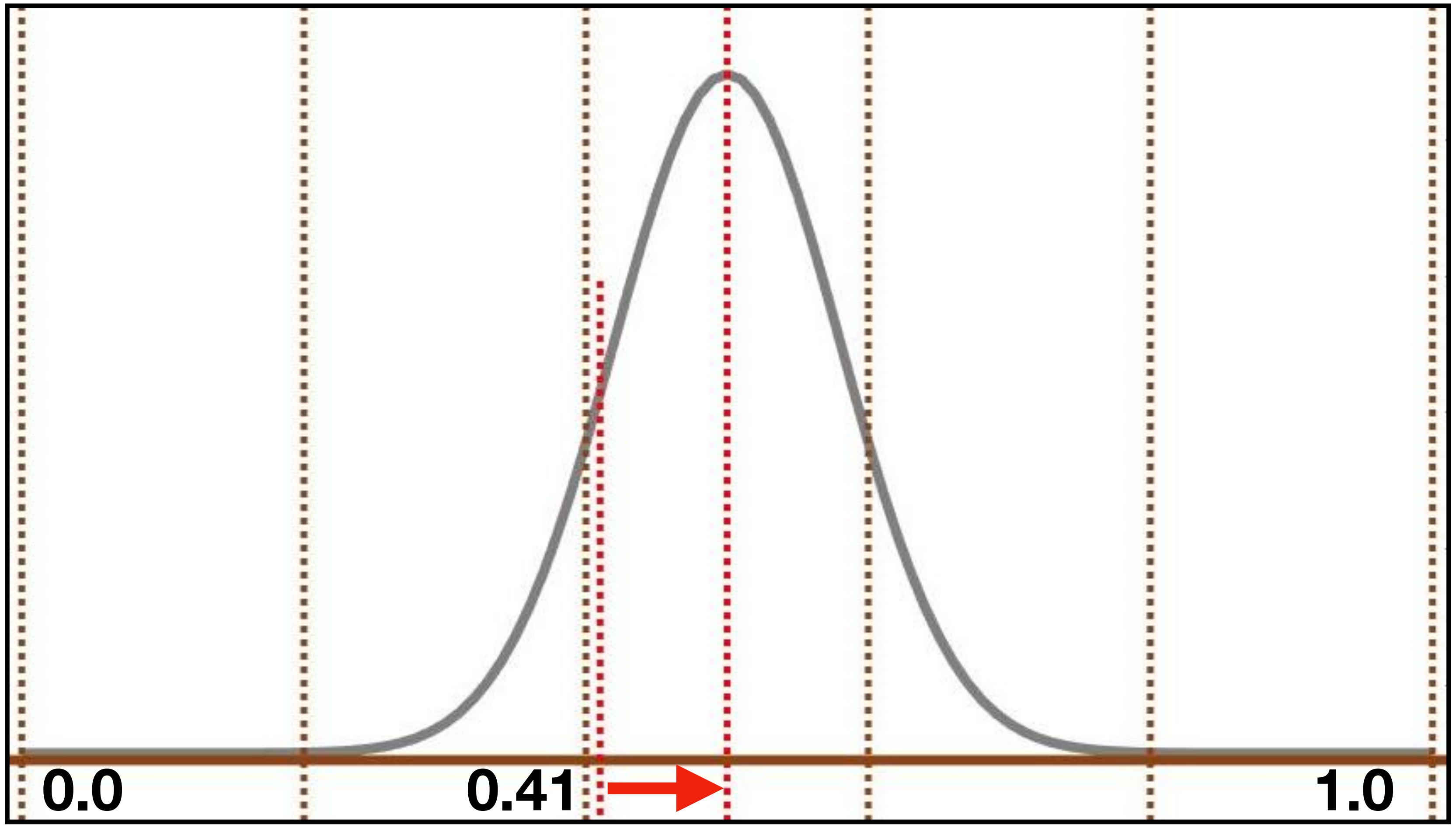


*slides credit to (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41, \quad f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [~0.0, 0.44, **0.54**,]

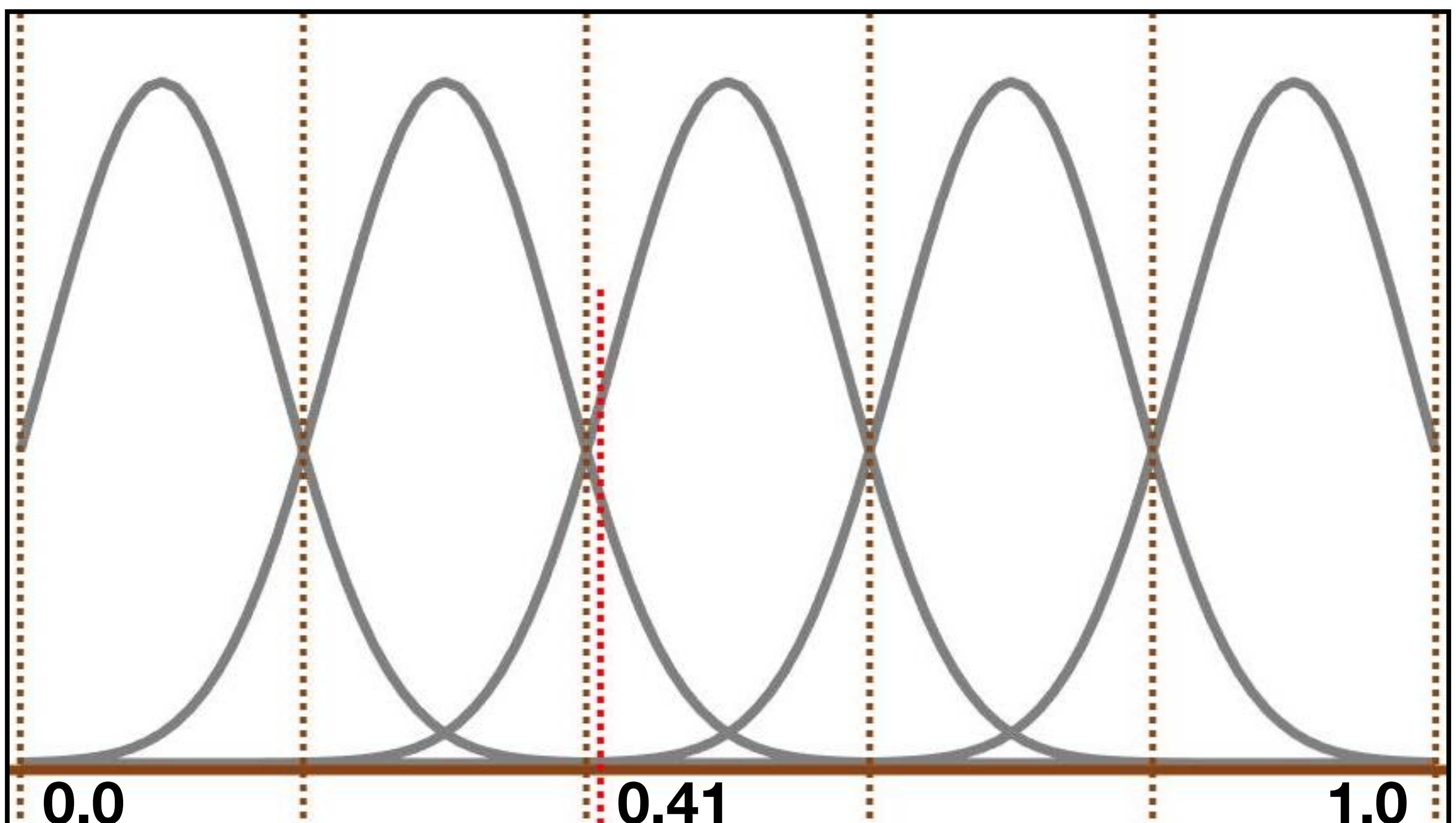


*slides credit to (Maddela and Xu 2018)

Gaussian Feature Vectorization

Single feature value : $f(w) = 0.41$, $f(w) \in [0,1]$

Vectorized feature : $f(w)$ = [~0.0, 0.44, 0.54, ~0.02, ~0.0]



*slides credit to (Maddela and Xu 2018)