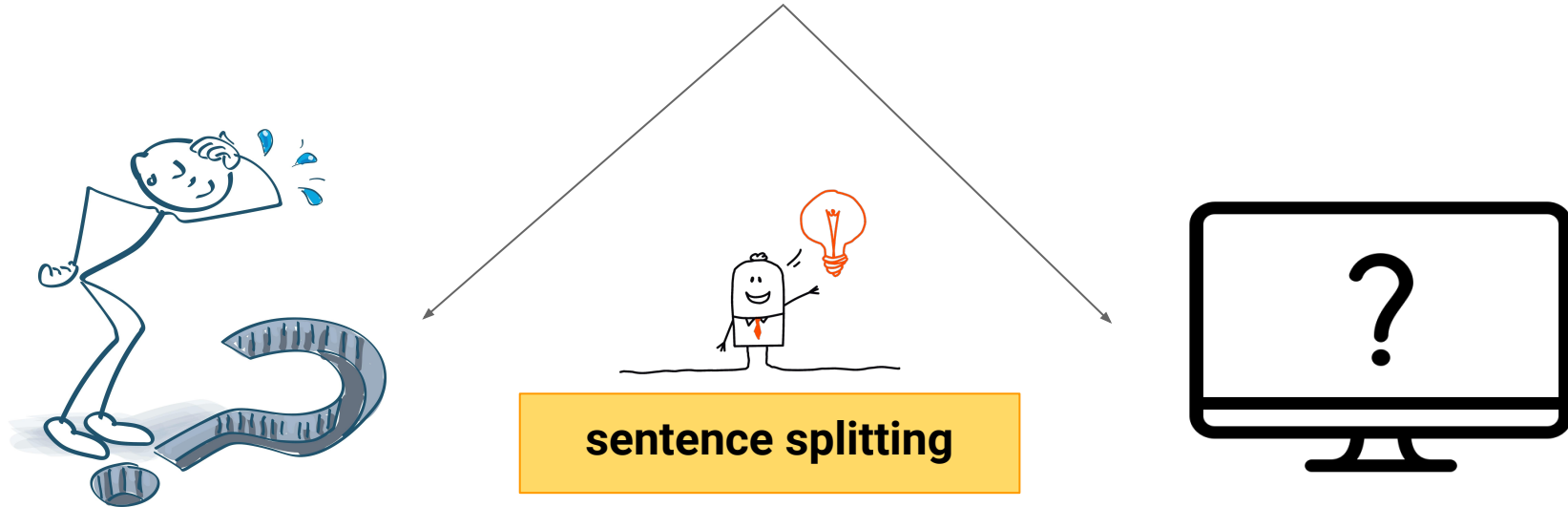


Transforming Complex Sentences into a Semantic Hierarchy

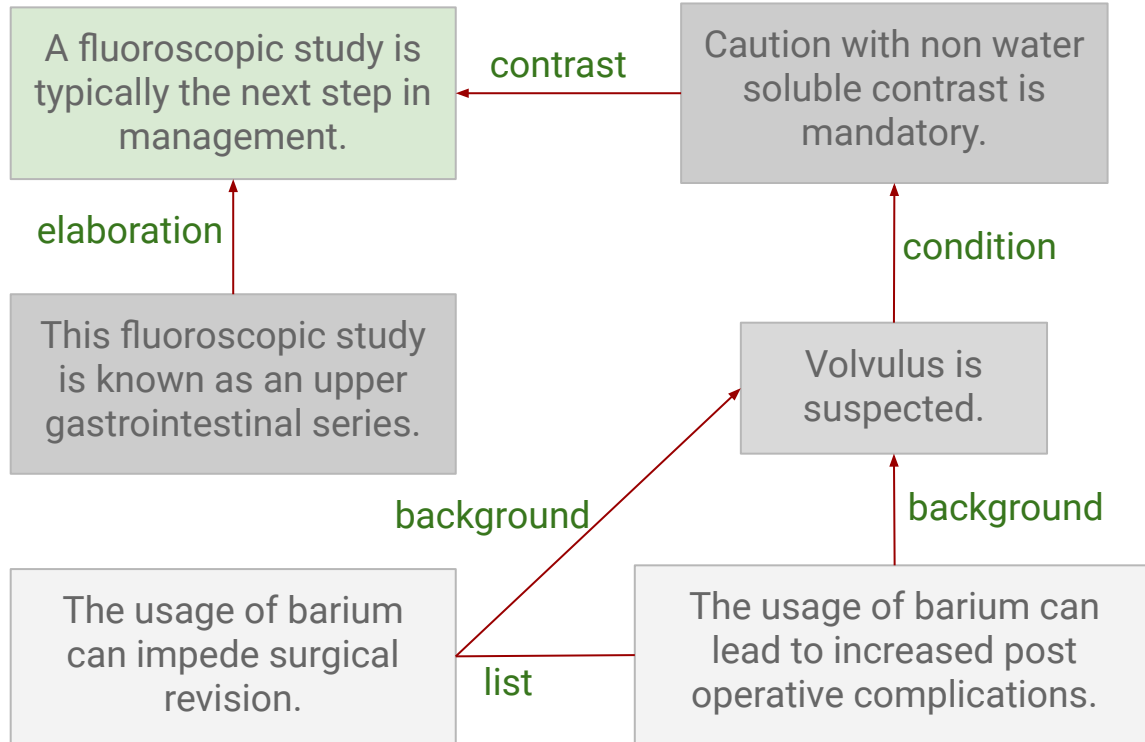
Christina Niklaus, Matthias Cetto, André
Freitas and Siegfried Handschuh

Text Simplification

“A fluoroscopic study known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications.”



Semantic Hierarchy



- (1) sentence splitting
- (2a) contextual hierarchy
- (2b) rhetorical relations

**minimal
propositions**

**preservation of
coherence structure**

Discourse-aware Sentence Simplification

- Recursive transformation stage
- **35 hand-crafted grammar rules**
- Encode syntactic and lexical features

(1) How to **split up and rephrase** the input?
 (2) How to set up a **semantic hierarchy**?

RULE	TREGEX PATTERN	EXTRACTED SENTENCE
SubordinationPostExtractor	ROOT <<: (S < (NP \$.. (VP < +(VP) (SBAR <, (IN \$+ (S < (NP \$.. VP))))))))	S < (NP \$.. VP .

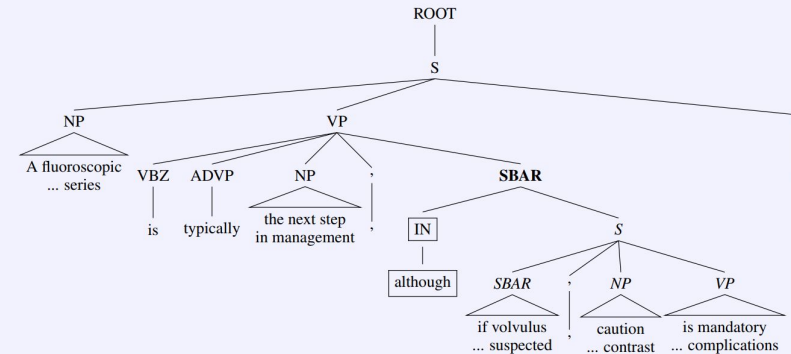
(1) Sentence Splitting

- Rules encode both the **splitting points** and **rephrasing procedure**

	CLAUSAL/PHRASAL TYPE	# RULES
Clausal disembedding		
1	Coordinate clauses	1
2	Adverbial clauses	6
3a	Relative clauses (non-defining)	8
3b	Relative clauses (defining)	5
4	Reported speech	4
Phrasal disembedding		
5	Coordinate verb phrases (VPs)	1
6	Coordinate noun phrases (NPs)	2
7a	Appositions (non-restrictive)	1
7b	Appositions (restrictive)	1
8	Prepositional phrases (PPs)	3
9	Adjectival and adverbial phrases	2
10	Lead NPs	1
	Total	35

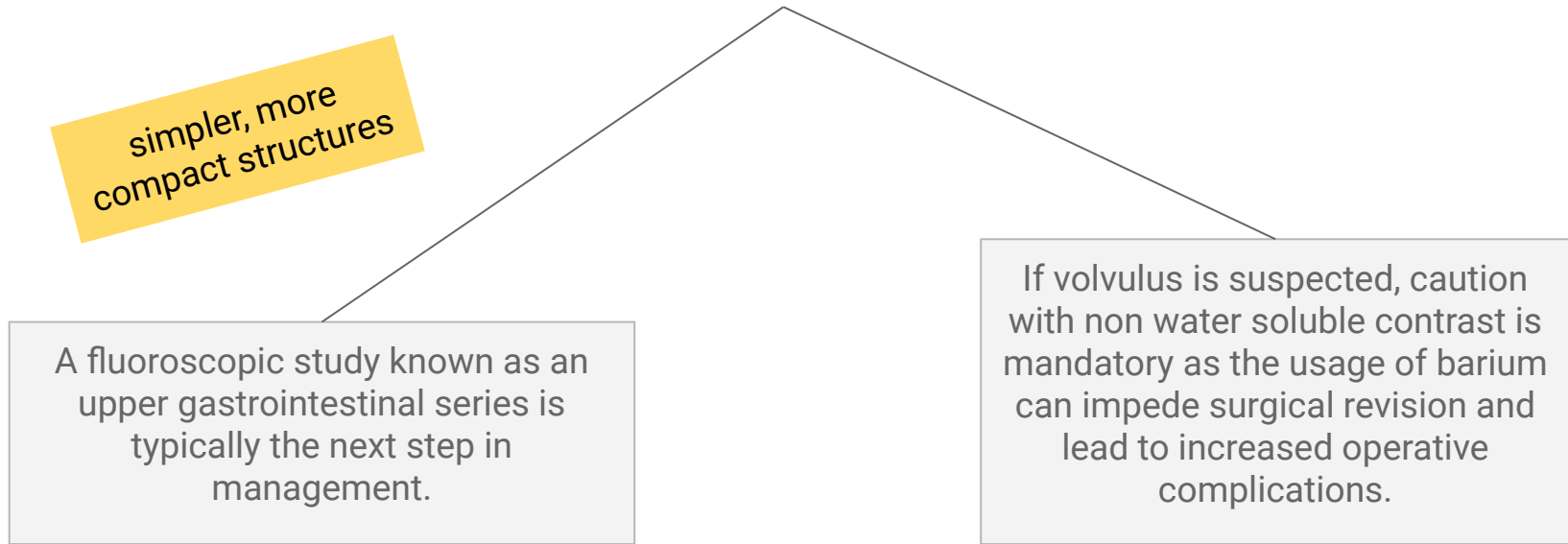
Example: SUBORDINATIONPOSTEXTRACTOR

Matched Pattern:



(1) Sentence Splitting: Example

“A fluoroscopic study known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications.”



(2a) Constituency Type Classification

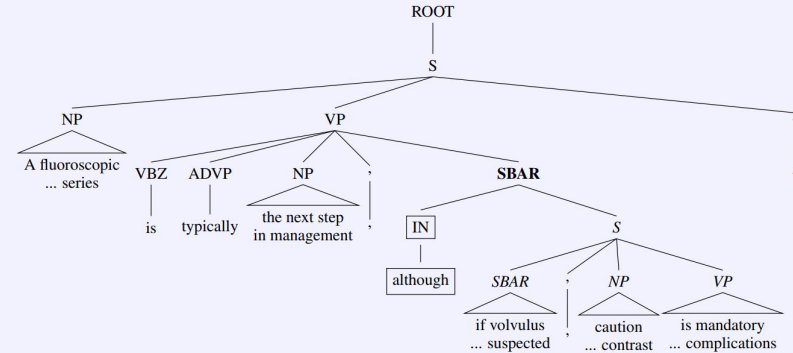
- Establishes a **contextual hierarchy** between the split sentences (based on syntax)
- Adopts the concept of nuclearity from **RST**

coordinations:
CORE

subordinations:
CONTEXT

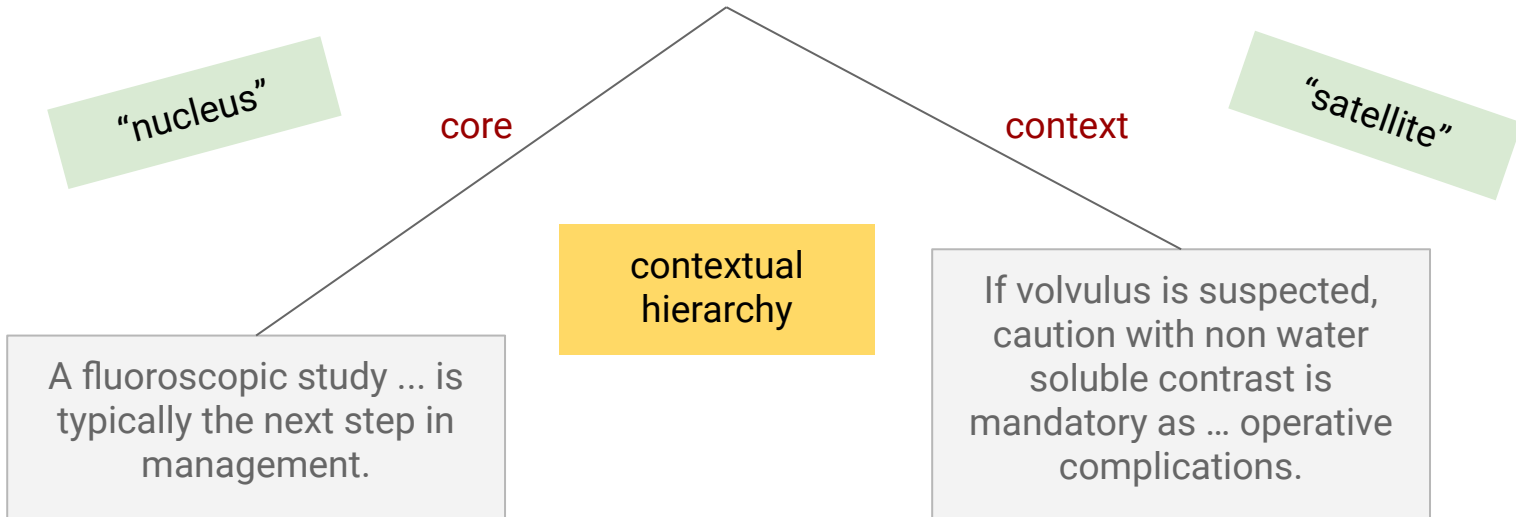
Example: SUBORDINATIONPOSTEXTRACTOR

Matched Pattern:



(2a) Constituency Type Classification: Example

“A fluoroscopic study known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications.”



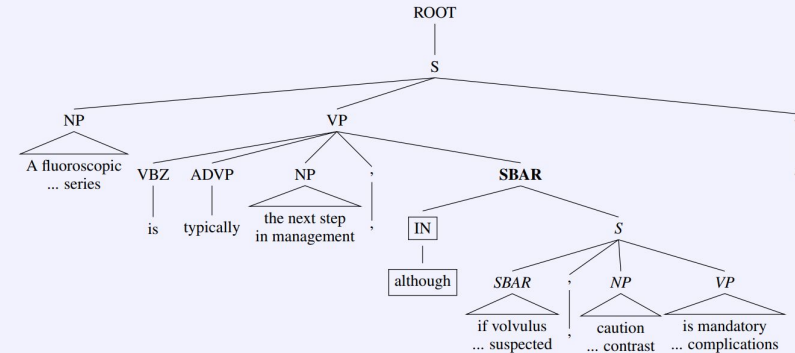
(2b) Rhetorical Relation Identification

- Identifies and classifies **rhetorical relations** that hold between a pair of split sentences
- Based on syntactic and lexical features

mapping of
rhetorical cue
words

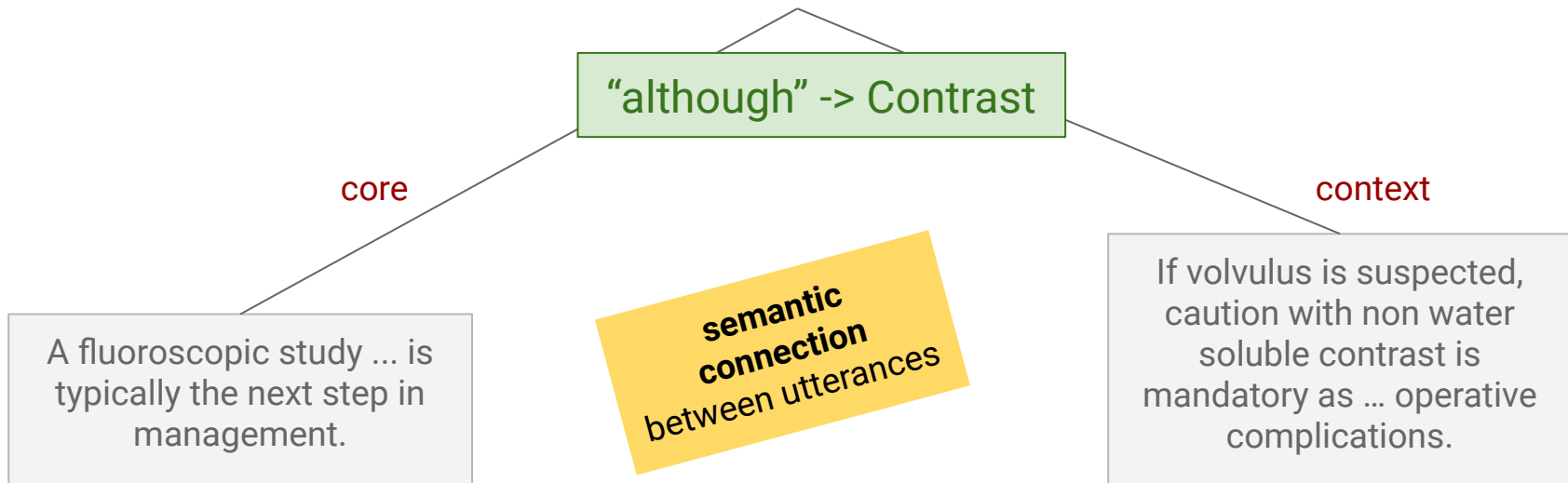
Example: SUBORDINATIONPOSTEXTRACTOR

Matched Pattern:



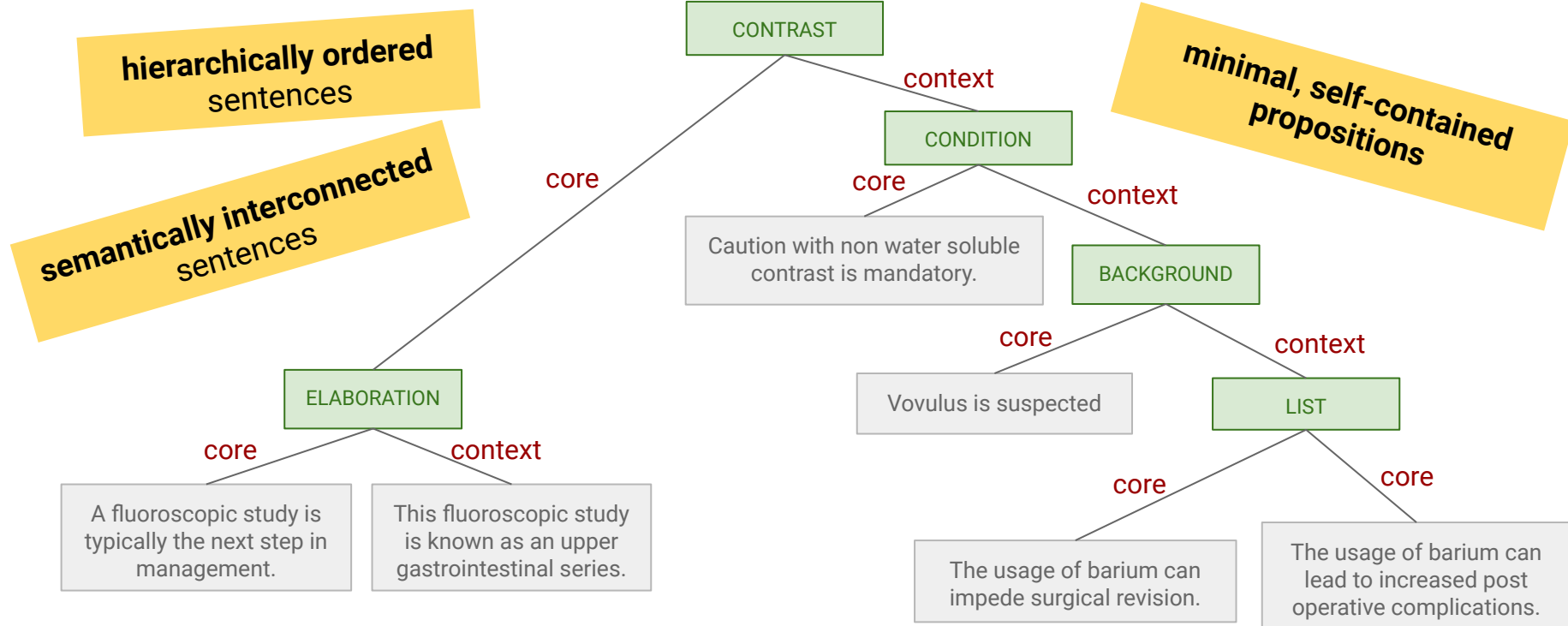
(2b) Rhetorical Relation Identification: Example

“A fluoroscopic study known as an upper gastrointestinal series is typically the next step in management, although if volvulus is suspected, caution with non water soluble contrast is mandatory as the usage of barium can impede surgical revision and lead to increased post operative complications.”



Discourse Tree

- **Recursive** simplification of leaf nodes in a **top-down** fashion until no more rule matches



Evaluation: Baselines

- Focus on **splitting subtask**
- Comparison with state-of-the-art syntactic TS systems that **explicitly model splitting operations**

syntax-driven **rule-based** approaches:

1. RegenT (Siddharthan and Mandya, 2014)
2. YATS (Ferrés et al., 2016)

manual definition of a set of grammar rules based on syntactic information

decomposition of a sentence into its **main semantic constituents**

approaches based on **semantic parsing**:

3. Hybrid (Narayan and Gardent, 2014)
4. DSS (Sulem et al., 2018)

data-driven approaches:

5. Seq2Seq (Botha et al., 2018)

model **learns simplification rewrites** from examples of aligned complex source and simplified target sentences

Evaluation: Corpora

Wikilarge
(359 test sentences)
(Xu et al., 2016)

WikiSplit
(5000 test sentences)
(Botha et al., 2018)

Newsela
(1077 test sentences)
(Xu et al., 2015)



Evaluation: Basic Statistics

- Average **sentence length** of the simplified sentences
- Average **number of simplified sentences** per complex input
- Percentage of sentences that are **copied** from the source
- Average **Levenshtein distance** from the input

How **conservative** are the systems?

Wikilarge	#T/S	#S/C	%SAME	LD _{sc}
Complex	22.06	1.03	100	0.00
Simple reference	20.19	1.14	0.00	7.14
DisSim	11.01	2.82	0.00	11.90
DSS	12.91	1.87	0.00	8.14
Hybrid	13.44	1.03	0.00	13.04
YATS	18.83	1.40	18.66	4.44
RegenT	18.20	1.45	41.50	3.77

Evaluation: Syntactic Complexity

- SAMSA: high correlation with **simplicity** and **grammaticality** (Sulem et al., 2018)
- SAMSA_{abl}: high correlation with **meaning preservation** (Sulem et al., 2018)

SAMSA is **maximized**
when each split
sentence represents
exactly 1 semantic unit
in the input.

Wikilarge	SAMSA	SAMSA _{abl}
Complex	0.59	0.96
Simple reference	0.48	0.78
DisSim	0.67	0.84
DSS	0.64	0.75
Hybrid	0.47	0.76
YATS	0.56	0.80
RegenT	0.61	0.85

Evaluation: Human Annotation

- **Grammaticality:** Is the output fluent and grammatical? (1 - 5)
- **Meaning preservation:** Does the output preserve the meaning of the input? (1 - 5)
- **Structural simplicity:** Is the output simpler than the input, ignoring the complexity of the words? (-2 - 2)

- 50 randomly sampled sentences
- 2 annotators

Wikilarge	G	M	S	avg.
Simple reference	4.70	4.56	-0.2	3.02
DisSim	4.36	4.50	1.30	3.39
DSS	3.44	3.68	0.06	2.39
Hybrid	3.16	2.60	0.86	2.21
YATS	4.40	4.60	0.22	3.07
RegenT	4.64	4.56	0.28	3.16

Extrinsic Evaluation: Open IE

- **Open IE**: turn unstructured information into a structured representation in the form of **relational tuples**

Supervised-OIE (alone) : (Stanovsky et al., 2018)

- (1) (A fluoroscopic study; known; as an upper gastrointestinal series)
- (2) (caution with non water soluble contrast; is; mandatory as the use of barium is contraindicated in patients with a known or suspected perforation)
- (3) (as the usage; of barium can impede; surgical revision)
- (4) (; to increased; post operative complications)

loose arrangement of tuples that lack the expressiveness needed for a proper interpretation of complex assertions

- **Integration of DisSim as a preprocessing step** into state-of-the-art Open IE approaches
 - enrich extractions with **contextual information**
 - allows to restore the semantic relationship between a set of propositions

Extrinsic Evaluation: Open IE

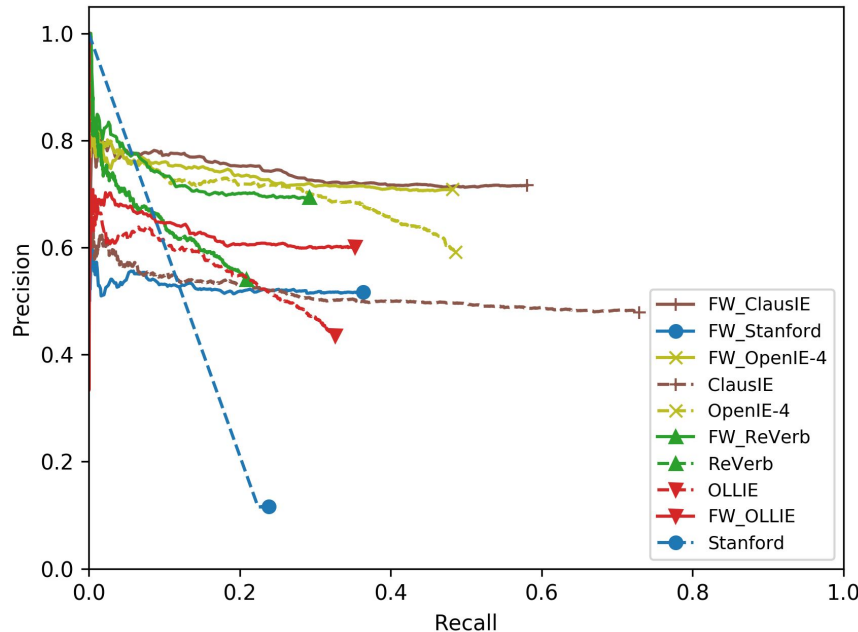
Supervised-OIE (using framework):

```
(5) #1 0 (A fluoroscopic study; is; typically, the next step in management)
(5a)      L:ELABORATION      #2
(5b)      L:CONTRAST         #3
(6) #2 1 (This; fluoroscopic study is known; as an upper gastrointestinal series)
(7) #3 0 (Caution with non water soluble; is; mandatory)
(7a)      L:CONTRAST         #1
(7b)      L:CONDITION        #7
(7c)      L:BACKGROUND       #4
(7d)      L:BACKGROUND       #5
(7e)      L:BACKGROUND       #6
(8) #4 1 (The usage of barium; can impede; surgical revision)
(8a)      L:LIST             #5
(8b)      L:LIST             #6
(9) #5 1 (The usage of barium; can lead; to increased post operative complications)
(9a)      L:LIST             #4
(9b)      L:LIST             #6
(10) #6 1 (The usage of barium; to increased; post operative complications)
(10a)     L:LIST             #4
(10b)     L:LIST             #5
(11) #7 1 (Volvulus; is suspected; )
```

semantic hierarchy:
preserve the
interpretability in
downstream tasks

Extrinsic Evaluation: Open IE

- **Minimality:** improve performance of state-of-the-art Open IE approaches in terms of **precision and recall**

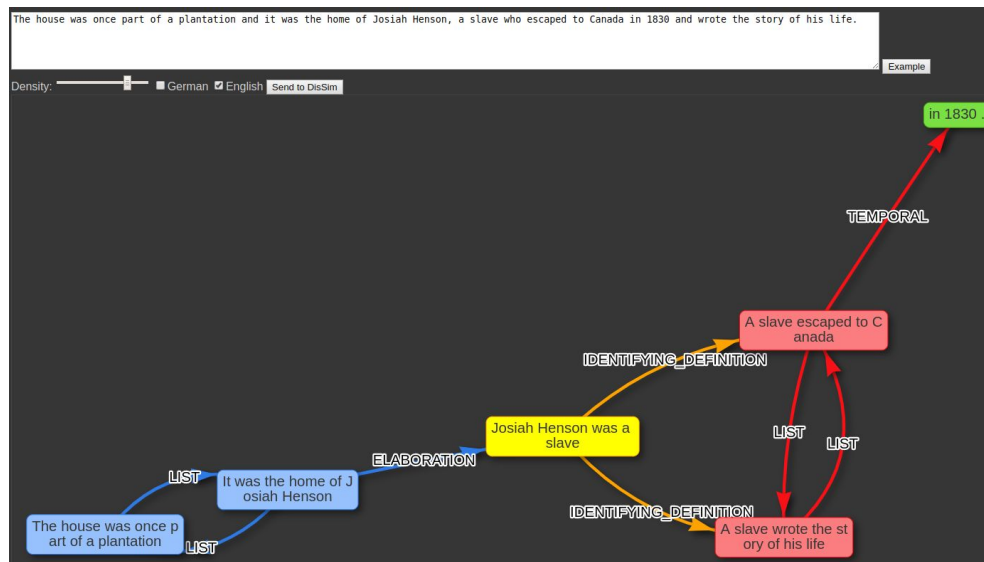


System	Precision	Recall	AUC
Stanford Open IE (Angeli et al., 2015)	+346%	+52%	+597%
ReVerb (Fader et al., 2011)	+28%	+40%	+57%
OLLIE (Mausam et al., 2012)	+38%	+8%	+20%
ClausIE (Del Corro and Gemulla, 2013)	+50%	-20%	+15%
OpenIE-4 (Mausam, 2016)	+20%	-1%	+3%

Cetto et al., 2018

Conclusion

- Transformation of complex sentences into a set of hierarchically ordered and semantically interconnected sentences that present a simplified syntax
 - minimal semantic units
 - semantic hierarchy
- DisSim **outperforms the state of the art** in syntactic TS
 - fine-grained output with high level of grammaticality and meaning preservation
 - improvement of 5%, 4% and 6% in SAMSA against the second best-performing approach
 - domain independence
- Application as a **preprocessing step**:
 - improves the performance of downstream applications in precision and recall
 - enriches their output with important contextual information



Code:

<https://github.com/Lambda-3/DiscourseSimplification>