MinWikiSplit: A Sentence Splitting Corpus with Minimal Propositions

Christina Niklaus, André Freitas and Siegfried Handschuh

University of St.Gallen

SENTENCE SPLITTING

Complex source

Starring Meryl Streep, Bruce Willis, Goldie Hawn and Isabella Rossellini, the film focuses on a childish pair of rivals who drink a magic potion that promises eternal youth.

The film is a partly fictionalized presentation of the tragedy that occurred in Kasaragod District of Kerala in India, as a result of endosulfan, a pesticide used on cashew plantations owned by the government.

Simplified target

The film is starring Meryl Streep.
The film is starring Bruce Willis.
The film is starring Goldie Hawn.
The film is starring Isabella Rossellini.
The film focuses on a childish pair of rivals.
These rivals drink a magic potion.
This magic potion promises eternal youth.

The film is a partly fictionalized presentation of the tragedy.
This tragedy occurred in Kasaragod District of Kerala in India.
This was as a result of endosulfan.
Endosulfan is a pesticide.
This pesticide is used on cashew plantations.
These cashew plantations are owned by the government.

- hard to comprehend by **human readers**
- difficult to analyze by **semantic applications**

simpler and more regular structure that is **easier to process**

Problem

- Many Text Simplification approaches make use of a set of **hand-crafted transformation rules** to decompose complex sentences into a sequence of structurally simplified utterances, requiring a **complex rule engineering process**.
- To overcome this expensive manual effort, Narayan et al. [1] compiled the **WebSplit** corpus, the first TS dataset that **explicitly addresses the task of sentence splitting**, while abstracting away from other TS operations.
- Based on this corpus, they presented a first attempt at modelling a **data-driven sentence splitting approach** where simplification rewrites are learned automatically from examples of aligned complex source and simplified target sentences.

Limitations

corpus	size	domain	pros	cons	construction
WebSplit v1 [1]	> 1 million pairs	verbalisations of triples from DBPedia	first TS dataset that **explicitly addresses the task of sentence splitting**, while abstracting away from other TS operations	99% of the simple sentences contained in the validation and test sets **also appear in the training set**	derived from the WebNLG corpus [6] in a **semi-automatic** way
WebSplit v2 [3]	> 1 million pairs	*see above*	no simple sentence that is contained in the development or test set occurs verbatim in the training set	**unnatural linguistic expressions** over only a small vocabulary and a uniform sentence structure	new train-development-test data split
WikiSplit [4]	> 1 million pairs	Wikipedia	**rich and varied vocabulary** over naturally expressed sentences that show a **diverse linguistic structure**	only a **single split** per complex source sentence	**mined from Wikipedia edit histories** by comparing adjacent snapshots and using a high-precision BLEU-based heuristic to filter out misaligned pairs
HSplit [5]	359 pairs	Wikipedia	**high-quality** manually compiled gold standard corpus	**small-scale**	**human generated** (each complex source sentence was modified by 4 annotators according to two sets of sentence splitting guidelines)

Corpus Construction

- We ran **DisSim** [2], a syntactic TS framework, over the complex input sentences from **WikiSplit**.
- DisSim applies a small set of **35 hand-written transformation rules** to recursively decompose clausal and phrasal elements and transform them into stand-alone sentences.

In that way, **each complex source sentence is broken down into a set of minimal propositions**, i.e. a sequence of sound, self-contained utterances with each of them presenting a minimal semantic unit that cannot be further decomposed into meaningful propositions.

Problem Definition

SPLIT INTO MINIMAL PROPOSITIONS

- *Input*: A **complex sentence** *C*.
- *Goal*: Produce a **sequence of simple sentences** $T_1, \ldots, T_n$, $n >= 2$, such that

1. **syntactic correctness**: each simple sentence T_i is grammatically sound.
2. **semantic correctness**: the output sentences $T_1, \ldots, T_n$ convey all and only the information in *C*.
3. **minimality**: each simple sentence T_i presents a minimal semantic unit, i.e. it cannot be further decomposed into meaningful propositions.

Quality Control

dependency parse and POS-based heuristics to **filter out malformed simplifications**, i.e.

- sequences that contain grammatically incorrect sentences
- sentences that mix multiple semantic units

Conclusion

- MinWikiSplit is a **large-scale sentence splitting corpus** that consists of **203K pairs** of complex source sentences and their simplified counterparts.
- We augmented the Split-and-Rephrase task [1] by the notion of **minimality**: each complex source sentence is broken down into a set of minimal propositions.

Application

- MinWikiSplit can be used to **train NLG applications** that perform a syntactic TS, simplifying sentences with a complex linguistic structure into a **fine-grained representation of short sentences that present a simple and more regular structure**.
- This output may serve as an intermediate representation that is **easier to process** for downstream semantic applications and may support a **faster generalization in ML tasks**.
- In that way, the performance of a wide range of AI tasks may be facilitated and improved.

Results

manual analysis

grammaticality	meaning preservation	structural simplicity
Is the output fluent and grammatical?	Does the output preserve the meaning of the input?	Is the output simpler than the input, ignoring the complexity of the words?
4.36	4.10	4.43

automatic metrics

	basic statistics				syntactic complexity	
	#T/S	#S/C	%SAME	LD_{SC}	SAMSA	$SAMSA_{abl}$
Complex	30.75	1.18	100	0.00	0.36	0.94
MinWikiSplit	12.12	3.84	0.00	17.73	0.40	0.48

WebSplit:

1. A Loyal Character Dancer was published by Soho Press, in the United States, where some Native Americans live.
2. Dead Man's Plack is in England and one of the ethnic groups found in England is the British Arabs.

WikiSplit:

1. The film is a partly fictionalized presentation of the tragedy that occurred in the Kasaragod District of Kerala in India. *** This tragedy was a result of endosulfan, a pesticide used on cashew plantations owned by the government.
2. Starring Meryl Streep, Bruce Willis, Goldie Hawn and Isabella Rossellini. *** The film focuses on a childish pair of rivals who drink a magic potion that promises eternal youth.

Christina Niklaus
christina.niklaus@unisg.ch

Institute of Computer Science
University of St. Gallen

References

[1] Narayan et al., 2017. Split and Rephrase. EMNLP.
[2] Niklaus et al., 2019. Transforming Complex Sentences into a Semantic Hierarchy. ACL.
[3] Aharoni and Goldberg, 2018. Split and Rephrase: Better Evaluation and Stronger Baselines. ACL.
[4] Botha et al., 2018. Learning To Split and Rephrase From Wikipedia Edit History. EMNLP.
[5] Sulem et al., 2018. BLEU is Not Suitable for the Evaluation of Text Simplification. EMNLP.
[6] Gardent et al., 2017. The WebNLG challenge: Generating text from RDF data. INLG.