Simplification by Lexical Deletion

Matthew Shardlow, Piotr Przybyła
Manchester Metropolitan University, Universitat Pompeu Fabra
m.shardlow@mmu.ac.uk, piotr.przybyla@upf.edu

Examples

Naturalization makes them **naturalized** citizens of their new country.

Plants include **familiar** types such as tree, herb, bushes, grass, vine, fern, moss, and green algae.

There were many brooks providing **fresh** water.

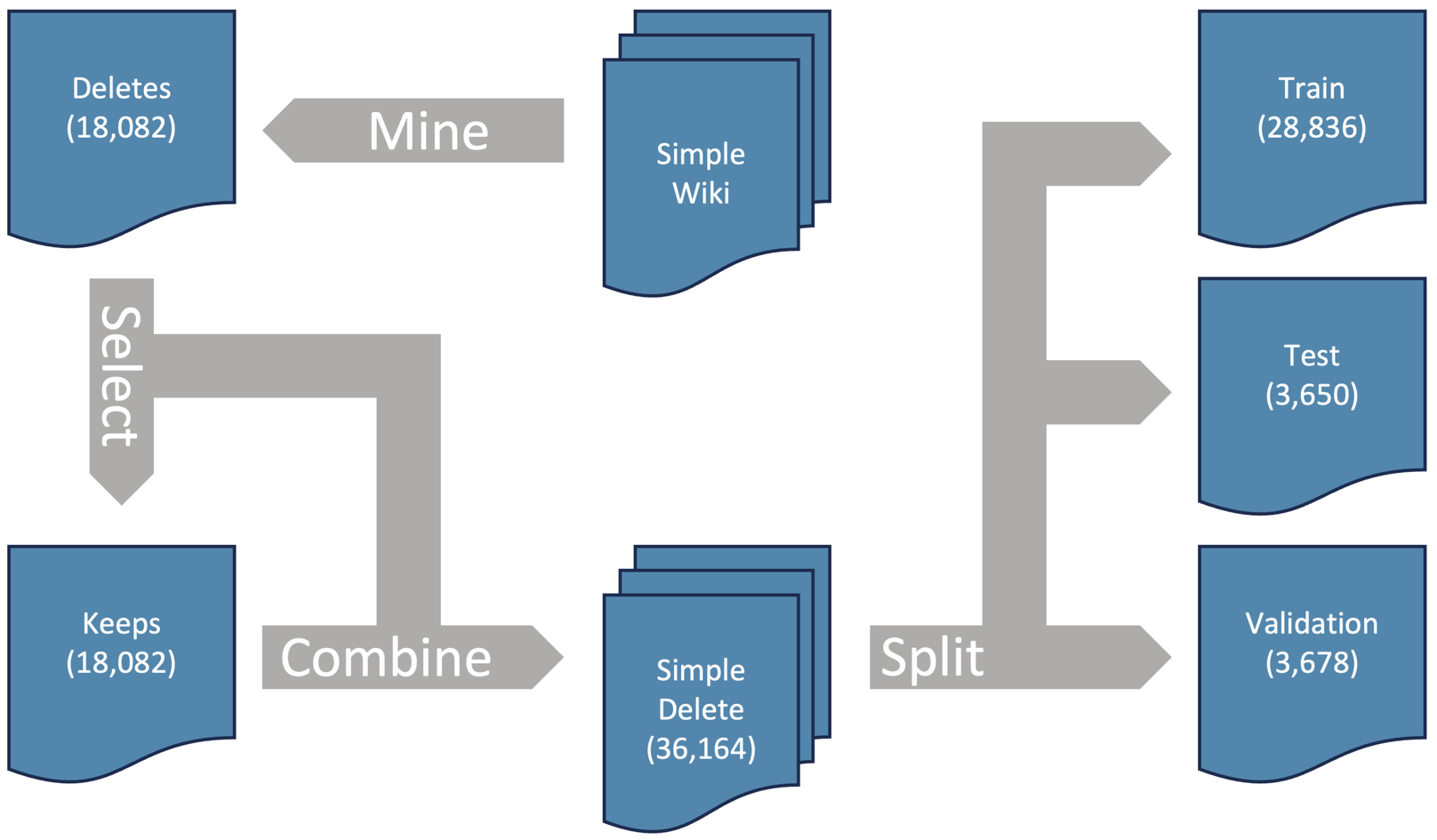
Abstract

- Lexical simplification traditionally focuses on replacing tokens with simpler alternatives.
- However, in some cases we may remove a word rather than replace it.
- We propose supervised and unsupervised solutions for lexical deletion.
- We contribute a new silver-standard corpus of 18,082 lexical deletions: **SimpleDelete**, mined from Simple Wikipedia edit histories.
- Deletion is one part of the wider lexical simplification puzzle, which we isolate and investigate.

Background

Simplification by deletion has been studied as an emergent property of systems which perform simplification through sentence to sentence translation [1, 2]. It is also possible to force systems to provide certain types of operations through the use of control tokens [3]. Our work leverages simple English Wikipedia edit histories, drawing on a long line of prior simplification studies to generate corpora using this resource.

SimpleDelete



Methods

TerseBERT: a custom version of the BERT model, originally developed for multi-word lexical simplification [4]. A special `[NONE]` token reflects the probability that the left and right context of the given mask position occur directly after each other, with no words between them.
SVM: We use a linear kernel SVM with fastText embeddings for the candidate token, whole context, left-context and right-context.
Bert-Large: We fine-tune for 5 epochs on our training partition using the given parameters (Adam optimiser, warmup steps = 500, weight decay = 0.01, learning rate = 0.001).
ACCESS: Capable of lexical or clausal deletion. We ran ACCESS over all contexts in our test set using the default control token parameters. For each context we identify whether a target word was deleted in the simplified output.

Results

Type	System	P	R	F1
U	TerseBert$_{0.03}$	0.677	0.942	0.788
U	TerseBert$_{0.27}$	0.746	0.850	0.795
U	ACCESS	0.719	0.472	0.570
S	SVM	0.766	0.666	0.712
S	BERT-large	0.870	0.830	0.850

Deletion prediction performance of different approaches on our dataset. TerseBert$_X$ refers to the deletion score being thresholded at X to give a binary classification. U and S refer to unsupervised and supervised systems with respect to our corpus.
Fine-tuned Bert-Large outperforms the SVM. Supervised systems outperform unsupervised. Unsupervised TerseBert performs competitively with supervised systems, given threshold selection. ACCESS gets competitive precision but low recall, indicating reluctance to delete.

References

[1] Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[2] Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Online, July 2020. Association for Computational Linguistics.

[3] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. Controllable sentence simplification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4689–4698, Marseille, France, May 2020. European Language Resources Association.

[4] Piotr Przybyła and Matthew Shardlow. Multi-Word Lexical Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING 2020)*, pages 1435–1446, Barcelona, Spain, 2020. International Committee on Computational Linguistics.

Manchester Metropolitan University

upf.