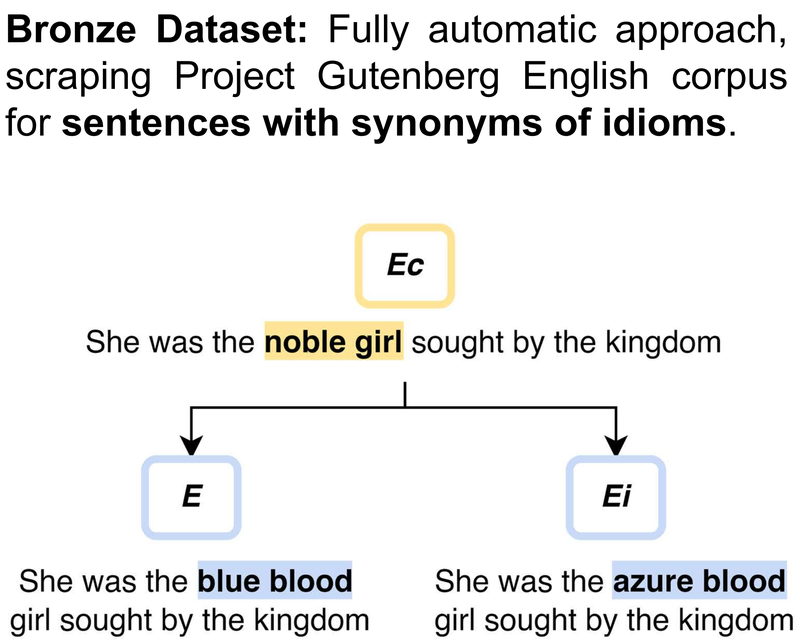
Is Less More?
Quality, Quantity and Context in Idiom Processing.

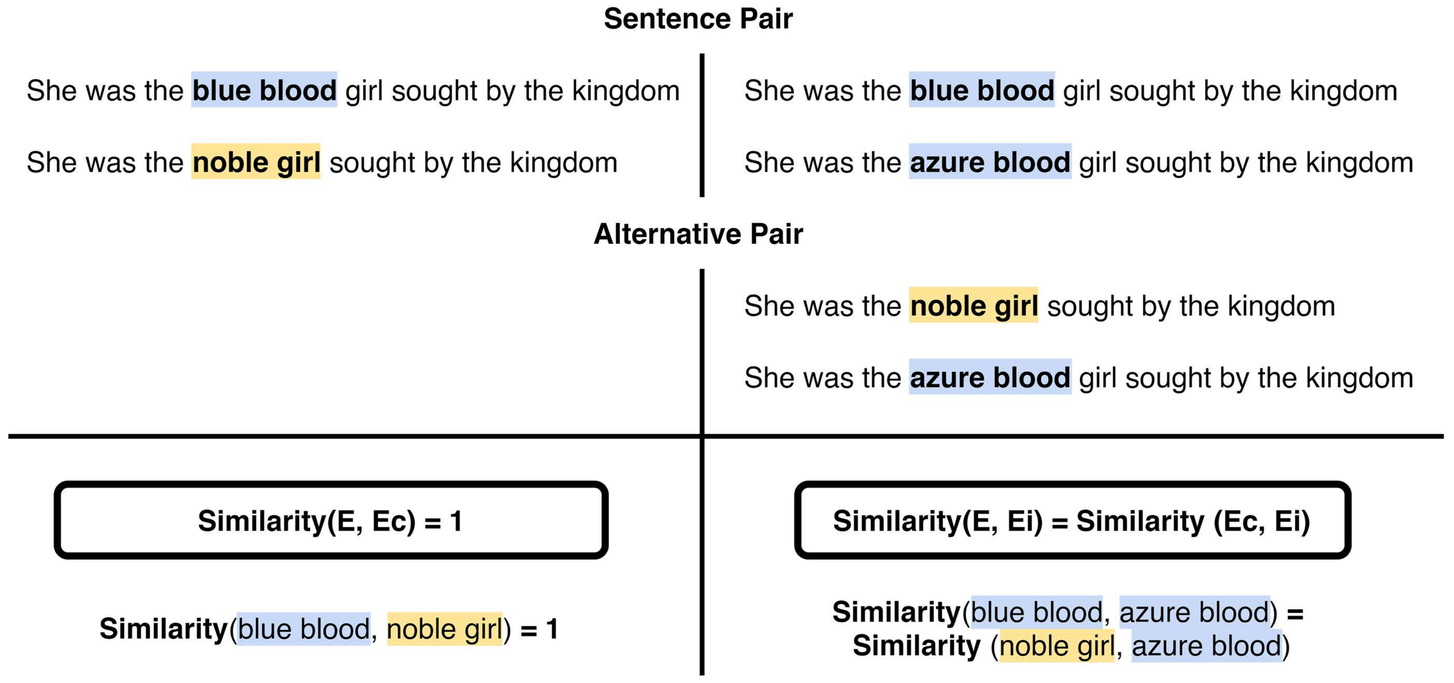
Agne Knietaite, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas Pickard, Aline Villavicencio

Presented by: Dylan Phelps

Datasets: Noun Compound Synonym Substitution in Books – NCSSB

Bronze Dataset: Fully automatic approach, scraping Project Gutenberg English corpus for **sentences with synonyms of idioms**.





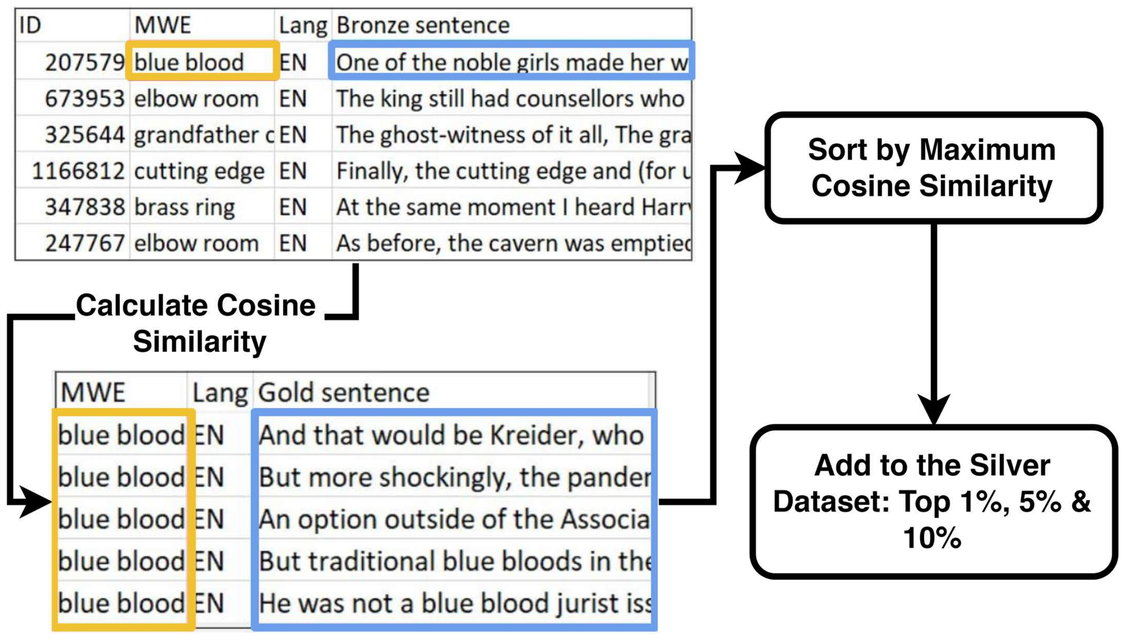
Silver Dataset: Top 1%, 5% and 10% of the Bronze dataset when **ranked according to cosine similarity** with frequency count vectors of sentences for a given MWE in the SemEval dataset.

Gold Dataset: Manual approach, **hand-labelled** by 2 to 3 annotators, where Silver dataset is used as a base.

Gold
Size: ~1,500
Quality: High

Silver
Size: ~10,000, ~50,000, ~100,000
Quality: Moderate

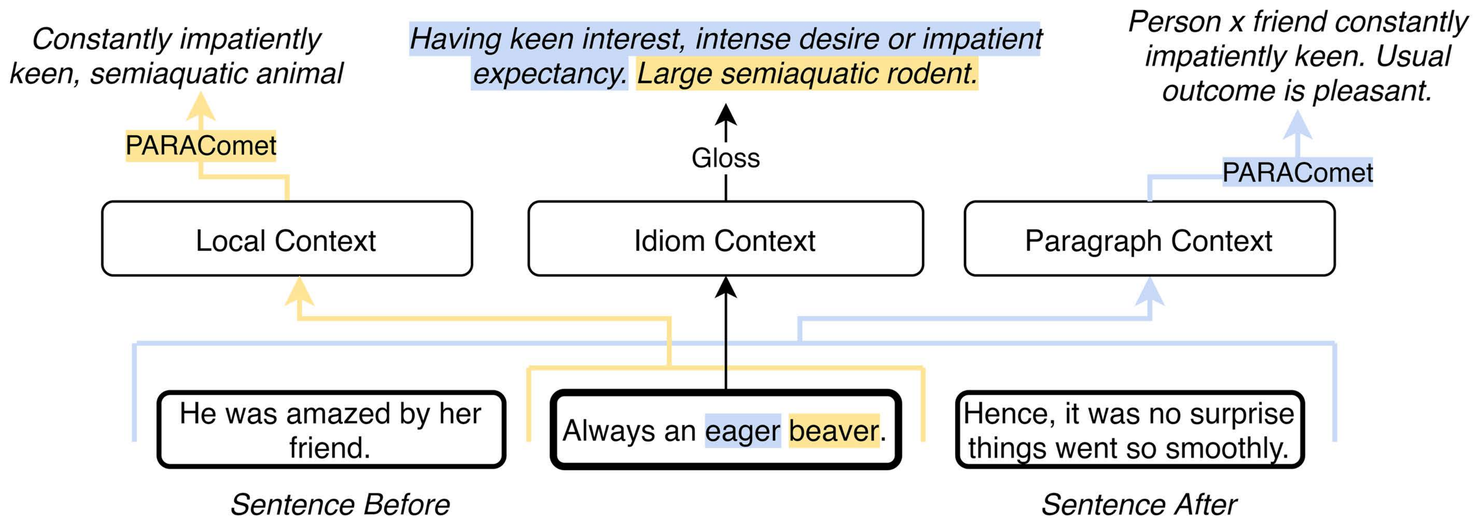
Bronze
Size: ~1,400,000
Quality: Low



Models: Context & Knowledge

Pretrained **mBERT model** is enhanced with **3 types of context**:

- **Idiom constituent word** knowledge
- **Sentence-wide context** knowledge
- **Paragraph-wide context knowledge**, including sentences coming before and after



Results

Dataset Quality vs Quantity?

For non-enhanced models, quantity is important

For enhanced models, quality is important

Local Knowledge vs External Knowledge?

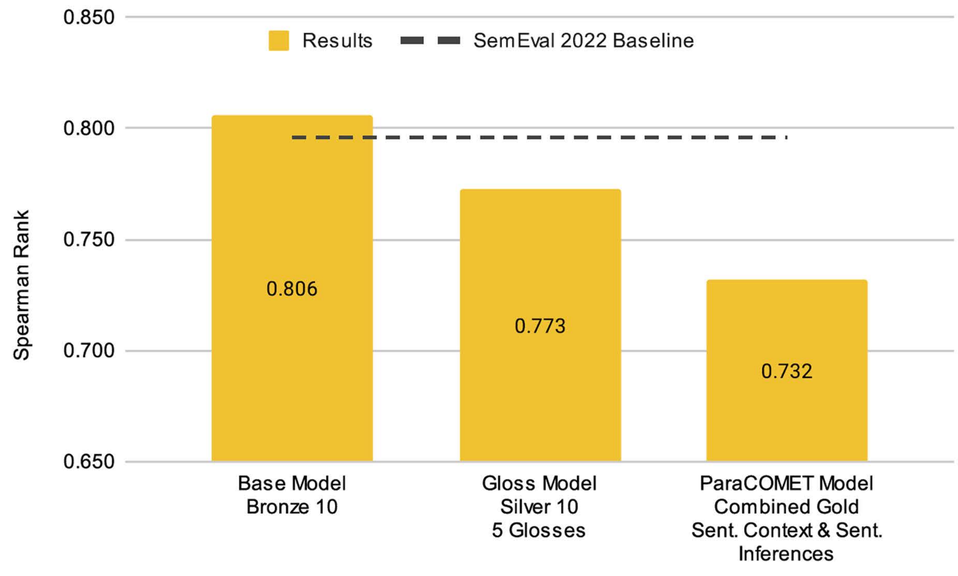
Paragraph & surrounding sentences context is generally not useful

Idiom constituent word & target sentence knowledge is the most useful

Quality Data + Well-targeted Context = Best Models?

If model enhancement quality decreases, dataset size needs to increase

Enhanced models still need a quality dataset of considerable size to outperform basic approaches



Best-performing model configurations