

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

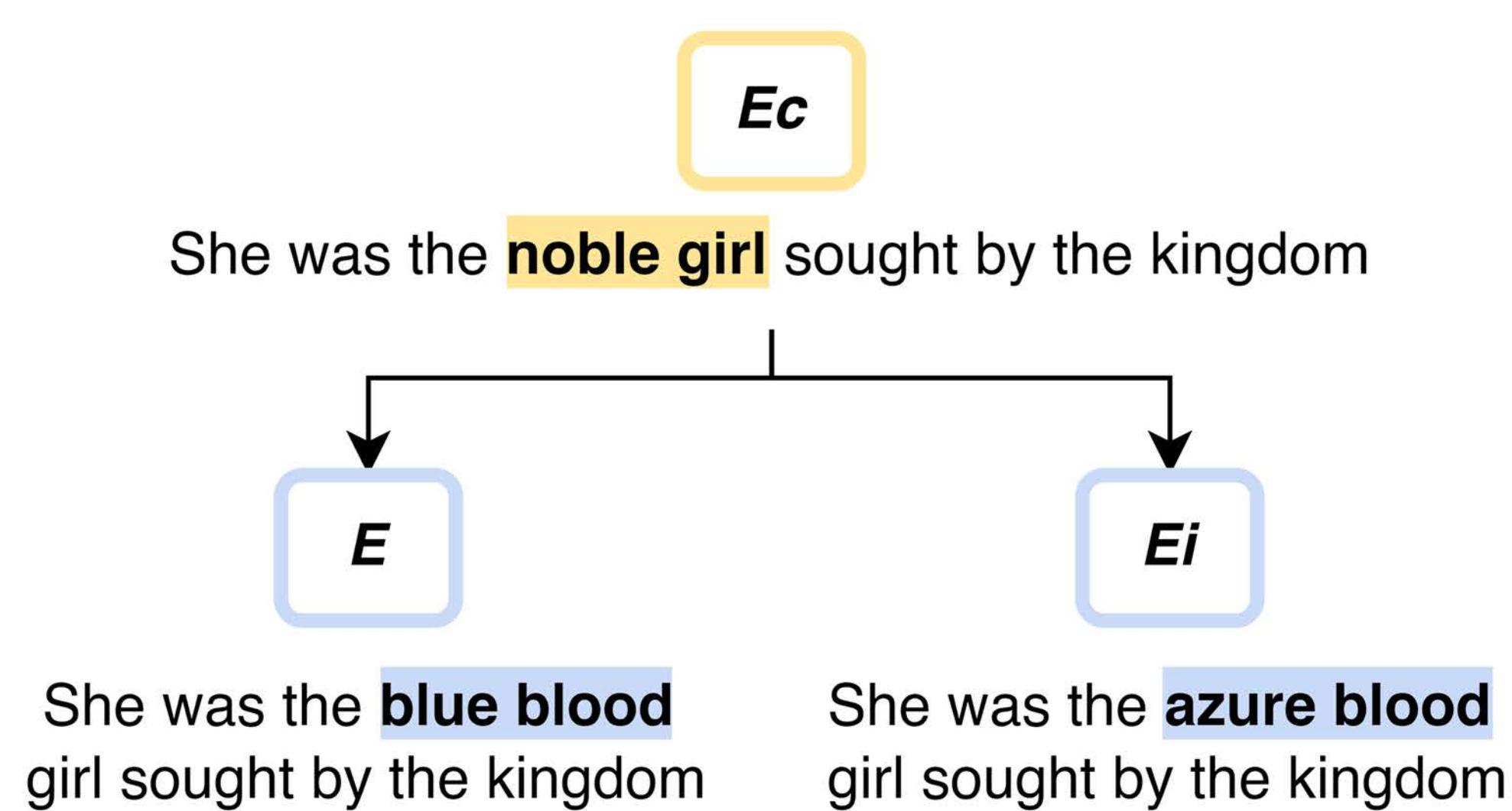


Agne Knietaite, Adam Allsebrook, Anton Minkov, Adam Tomaszewski, Norbert Slinko, Richard Johnson, Thomas Pickard, Aline Villavicencio

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**.



Sentence Pair
She was the **blue blood** girl sought by the kingdom
She was the **noble girl** sought by the kingdom
She was the **azure blood** girl sought by the kingdom

Alternative Pair

She was the **noble girl** sought by the kingdom
She was the **azure blood** girl sought by the kingdom

Similarity(E, Ei) = Similarity (Ec, Ei)

Similarity(blue blood, noble girl) = Similarity (blue blood, azure blood) = Similarity (noble girl, azure blood)

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

ID	MWE	Lang	Bronze sentence
207579	blue blood	EN	One of the noble girls made her w
673953	elbow room	EN	The king still had counsellors who
325644	grandfather	cEN	The ghost-witness of it all, The gra
11166812	cutting edge	EN	Finally, the cutting edge and (for u
347838	brass ring	EN	At the same moment I heard Harr
247767	elbow room	EN	As before, the cavern was emptied

Calculate Cosine Similarity		
MWE	Lang	Gold sentence
blue blood	EN	And that would be Kreider, who
blue blood	EN	But more shockingly, the pander
blue blood	EN	An option outside of the Associa
blue blood	EN	But traditional blue bloods in the
blue blood	EN	He was not a blue blood jurist iss

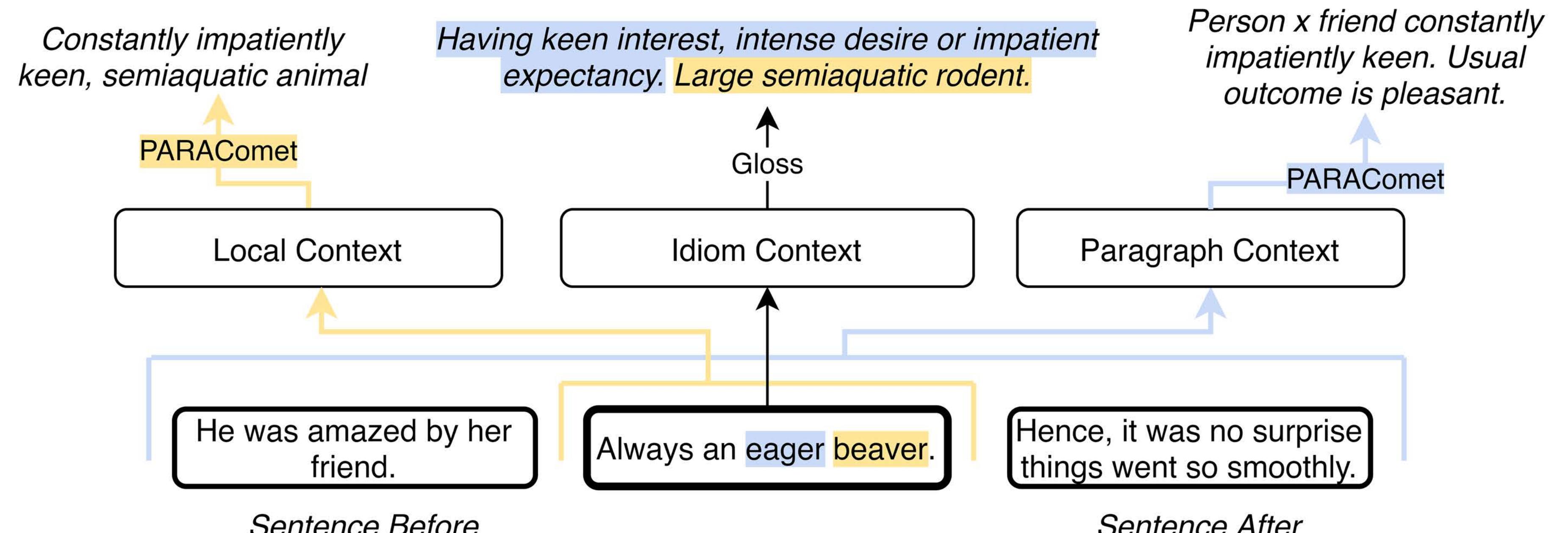
Sort by Maximum Cosine Similarity

Add to the Silver Dataset: Top 1%, 5% & 10%

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

