MEMES: Mining Emerging Multi-word Expressions from Social-Media

3rd General Meeting
Hungarian Research Centre for Linguistics,
Budapest: 28-30 January, 2025

Nishan Chatterjee[1, 2], Antoine Doucet[2], Senja Pollak[1]
Institute Jožef Stefan, Slovenia[1],
University of La Rochelle, France[2]

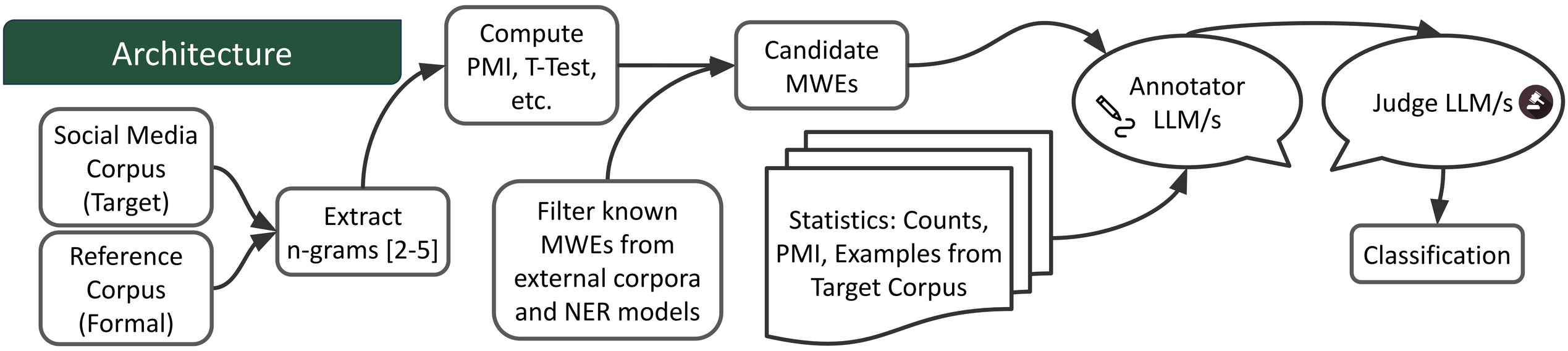
Motivation

- Language is constantly evolving based on social interactions and trends which give rise to phrases or units of expressions.
- Continually annotate and track these trends are difficult using traditional methods: *“no cap”*, *“in your <noun> era”*.
- Automating the process with data-driven expression extraction combined with LLMs as annotators and judges to scale MWE discovery.

Novelty

- Comparative Analysis between a Reference Corpus (Formal Language) vs a Target Corpus (Social Media).
- Using statistical methods (PMI, T-test) + LLMs to reason, detect, validate, and classify MWEs iteratively.

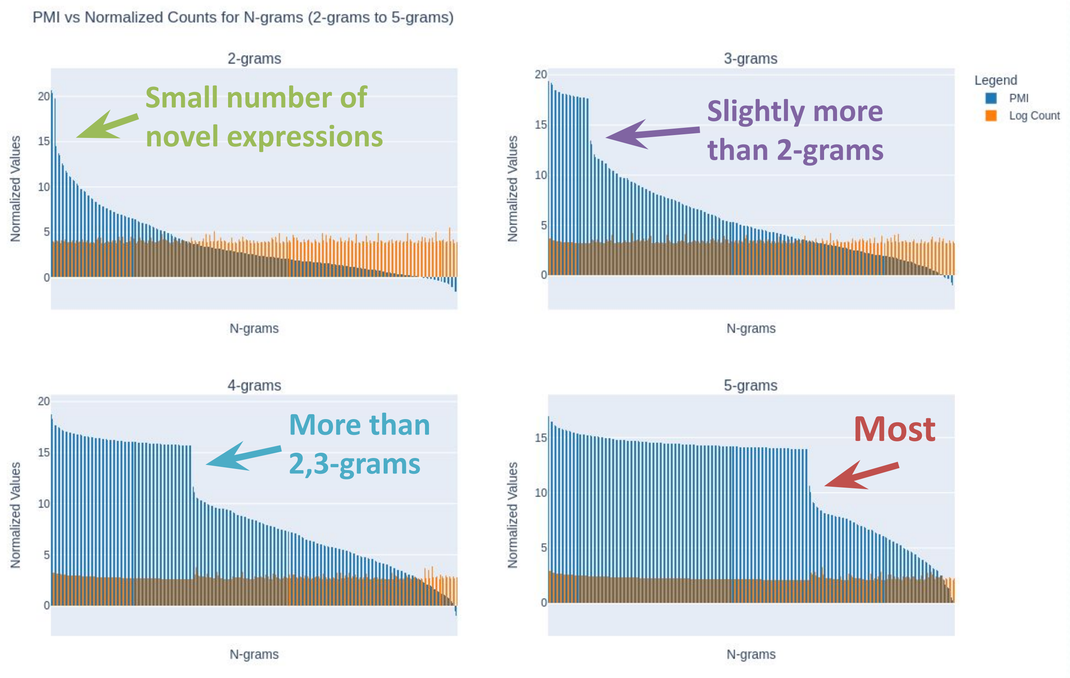
Architecture



Preliminary Findings

Reference Corpus: 2000 English Books between 1502 and 1823 with ≈ 123.4 million words with:

- 27.5 million 2-grams
- 61.5 million 3-grams
- 81.7 million 4-grams
- 86.8 million 5-grams



Target Corpus: 0.38 million Reddit posts and comments from TLDR subreddit between 2006 and 2016 with ≈ 77.6 million words with:

- 9 million 2-grams
- 30 million 3-grams
- 48.6 million 4-grams
- 55.6 million 5-grams

Limitations

Time-Intensive, Limited Syntactic Coverage, Corpus Limitations, Bootstrapping Gaps, prompt fitting.

Check out our GitHub Repo for more details!!

Next Steps

- Refine algorithms by integrating additional metrics & accounting for Syntactic-Ngrams.
- Alternative LLM prompts for better granularity and nuance.
- Expand and improve Reference and Target Corpora for better and broader coverage.

Key Takeaways

- Combining statistical metrics with LLMs ensures broad coverage to detect new social MWE expressions.
- Easy reconfiguration of discovery and annotation guidelines.
- Computationally expensive and relies on finding the right prompts.

Jožef Stefan Institute
Informatique Image Interaction Univ. La Rochelle
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union