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]

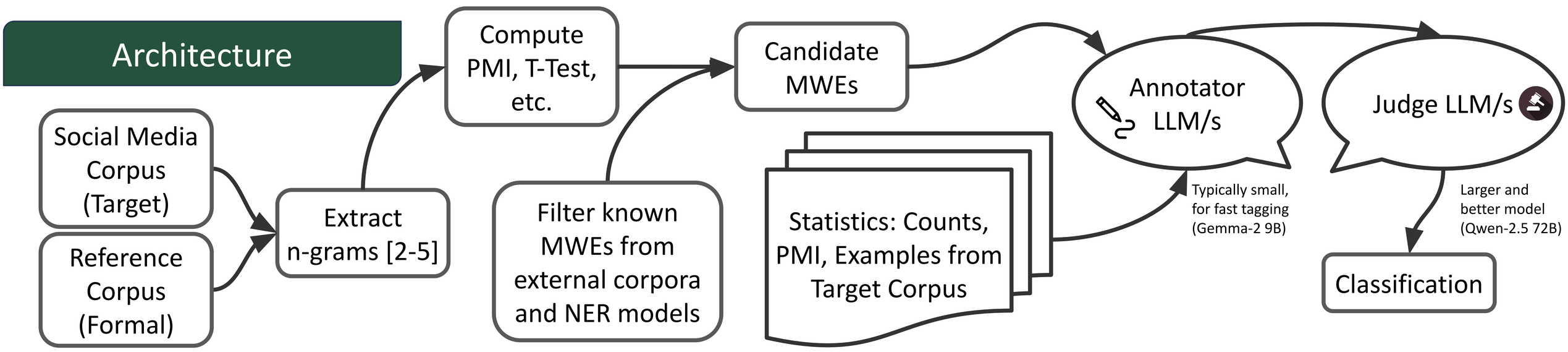
Motivation

- 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

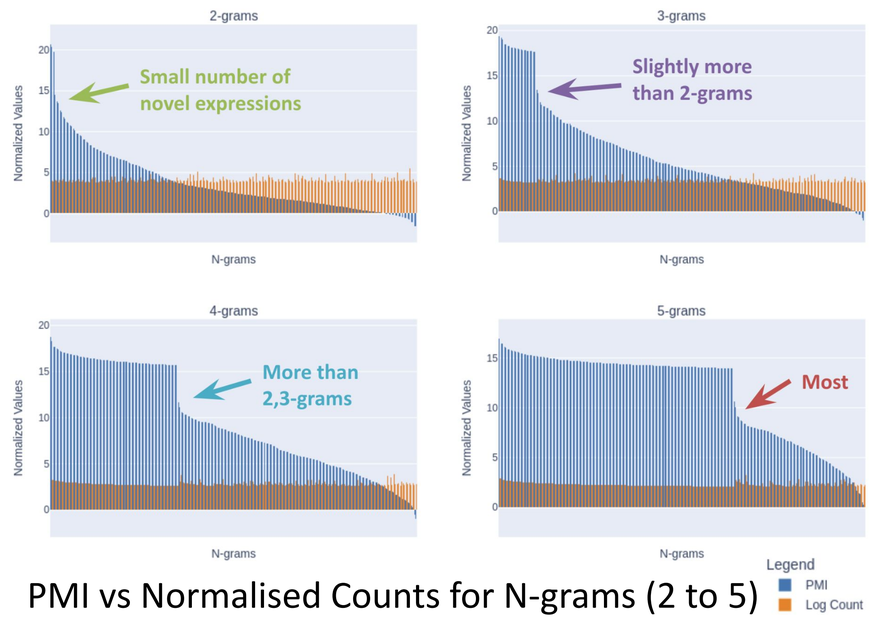
- Comparative Analysis between a Reference Corpus (Formal Language) vs a Target Corpus (Social Media).
- Using statistical methods (PMI, T-test) + LLMs to reason, detect, validate, and classify MWEs iteratively.

Architecture



Preliminary Findings

N-gram	Label	Frequency	PMI	Occurrences
'shit out of'	Novel MWE	2055	18.01	• Beat the shit out of me, on my own. • I'm upvoting the shit out of anyone who recognizes this beer.
'felt like i was'	Novel MWE	729	16.52	• I'm not the type to slack off on the job, but I felt like I was just in crunch time mode from hour one through hour 8, just so I could get as much money as I could.
'as opposed to'	Variation of an existing MWE	2260	18.15	• It's enforced by praising kids for their hard work (as opposed to their intelligence).



PMI vs Normalised Counts for N-grams (2 to 5)

Corpus Statistics

Reference Corpus: 2000 English Books between 1502 and 1823 with ≈ 123.4 million words.
Target Corpus: 380k Reddit posts and comments from TLDR subreddit between 2006 and 2016 with ≈ 77.6 million words.

MWEs Discovered

Out of 5605 shortlisted (high frequency and PMI),
We find 53 Novel MWEs, 277 variations of Existing MWEs, and 3847 Not MWEs.
Example: 'shit out of' (frequency: 2055, PMI: 18.01)

Check out our GitHub Repo for more details!!

Next Steps

- Refine algorithms by integrating additional metrics & accounting for Syntactic-Ngrams.
- Iteratively find good 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