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

Nishan Chatterjee[1,2,3] **Antoine Doucet**[3] **Senja Pollak**[1]
[1]Jožef Stefan Institute, Slovenia
[2]Jožef Stefan International Postgraduate School, Slovenia
[3]University of La Rochelle, France
nchatter@etudiant.univ-lr.fr

Relevant UniDive working groups: WG1, WG3

1 Introduction

Multi-word expressions (MWEs) play a crucial role in natural language understanding, yet their discovery, especially in evolving domains like social media can present several challenges (Ramisch et al., 2023). Additionally, social media platforms generate massive amounts of informal text with novel and evolving expressions that existing rule-based systems and traditional approaches often miss (Zampieri et al., 2022).

This project aims to bridge this gap by developing an automatic framework for discovering MWE trends from large social media and assigning confidence scores to potential MWEs. This can then be used with a human-in-the-loop system to train models for detecting emerging MWEs.

This can provide valuable insights into emerging language trends, enabling further research into evolving colloquialisms, idiomatic expressions, and other forms of multi-word structures. The confidence prediction will be part of the project's subtask, allowing for a more granular assessment of expression validity and reliability.

2 Research Objectives

The core objectives of the project are (1) MWE Discovery: Identify potential MWEs from large-scale social media corpora using statistical measures and machine learning techniques; (2) Confidence Scoring: Assign confidence scores to the discovered MWEs based on statistical strength, context, and annotator agreement, providing insights into trends and language use evolution; (3) Shared Task Release: Develop a shared task with the annotated dataset, allowing the NLP community to validate and extend MWE discovery techniques.

3 Related Works

The MERGE project by Gries and Wahl (2017) focuses on discovering MWEs by iteratively merging bigrams with high association strength using log-likelihood scores, forming longer sequences with each iteration. Schneider et al. (2014) introduced a comprehensive annotation approach, manually grouping tokens into MWEs in a social web corpus to handle the diversity and ambiguity of MWEs. Zampieri et al. (2021) proposed using MWE features in a deep neural network for hate speech detection, demonstrating improvements in performance by integrating MWE embeddings like word2vec and BERT. Zampieri et al. (2022) further extended this work, using a DNN-based approach to identify MWEs in tweets and improve hate speech detection, confirming the effectiveness of MWE features. Samadarshi et al. (2024) evaluated LLMs' abstract reasoning using the New York Times Connections game, showing that even advanced models like GPT-4o struggle with clustering and categorizing words compared to human players.

4 Proposed Methodology

We draw inspiration from the types of MWEs identified in Villavicencio and Idiart (2019) and any other expression that shows repetition trends.

4.1 Data Collection

The project will start with a large collection of social media corpora that include TweetsKB (Fafalios et al., 2018) for English Tweets, xLiMe (Rei et al., 2016) for German, Italian, and Spanish Tweets, and the Edinburgh Twitter Corpus (Petrović et al., 2010) for multilingual Tweets. Additionally, we also want to investigate distributions from Reddit (Henderson et al., 2019) [1]. The data will be preprocessed by removing noise (e.g., URLs, user mentions, and irrelevant content) while retaining relevant linguistic information.

4.2 N-Gram Extraction

We will extract n-grams of varying lengths (n=1 to n=10 for starters) from the cleaned corpus as candidate MWEs. To ensure that only meaningful

[1]A collection of large datasets for conversational response selection including Reddit, OpenSubtitles, and AmazonQA

n-grams are considered, frequency thresholds will be applied using the following statistical association measures as highlighted by Villavicencio and Idiart (2019): (1) Pointwise Mutual Information (PMI) that measures the co-occurrence of word pairs beyond random chance; (2) Specific Total Correlation (STC) that captures the total interaction among words in multi-word phrases; (3) Specific Information Interaction (SII), that quantifies the shared information between words in a phrase; (4) Student's T-Test (t-statistic) that tests the significance of word co-occurrences; (5) Dice Coefficient, that measures the similarity between word pairs; (6) Chi-Square Test (χ^2) that assesses whether co-occurrence is statistically significant or not; (7) LogDice Rychlý (2008), which adjusts the Dice coefficient to better capture low-frequency MWEs in large corpora; and (8) Longest-Commonest Match (LCM) Kilgarriff et al. (2015), which identifies recurring sequences of varying lengths in the corpus to capture longer MWEs.

4.3 Filtering Existing MWEs

We will filter out already identified MWEs using the PARSEME corpus (Savary et al., 2023), the MWE-CWI dataset (Kochmar et al., 2020), and the Streusle (Zampieri et al., 2022) dataset to focus on novel and potentially emerging expressions.

We will also apply a Named Entity Recognition (NER) model (Wang et al., 2020) [2] to automatically filter out proper names and other named entities. Additionally, we also plan to filter out Idiomatic Expressions using the Saxena and Paul (2020) dataset.

This will help capture any missed commonly occurring NERs, Idioms, and other forms of MWEs that show repetition trends.

4.4 Leveraging Pre-Trained Embeddings and Tokenization

To handle the unique challenges of social media language, we will compare Byte Pair Encoding (BPE) vs Morphological Segmentation[3] as tokenization methods to capture subword units, enabling us to discover MWEs that might not align with traditional word boundaries. Pre-trained embeddings from models like BERT or GPT will be used to (1) Cluster similar expressions that differ in surface form but share meaning, and (2) Disambiguate context-dependent MWEs using contextual embeddings to capture semantic nuances.

4.5 LLM-Assisted Annotation with Human-in-the-Loop

We plan to use a collection of large language models (LLMs) as initial annotators mimicking humans to identify MWEs from the n-gram candidates. A human-in-the-loop feedback system will allow human annotators to review and refine these suggestions, providing a hybrid annotation system that combines machine speed with human judgment. This will include assigning confidence scores to each identified MWE based on agreement between the models and human annotators, as well as statistical association strength. This should help reduce prototyping time and test the system's validity with lower costs.

For evaluation, we can use the standard metrics of Precision, Recall, Accuracy, F1 score, and Cohen's/Fleiss Kappa for the inter-annotator agreements (human-human, human-LLM, LLM-LLM).

4.6 Challenges and Mitigation Strategies

We expect the following challenges to the said task:

- **Ambiguity in MWEs:** Handling context-sensitive MWEs is difficult. Pre-trained embeddings and models can help address this (Kim et al., 2023).

- **Scalability:** Annotating large amounts of data is time-consuming. LLMs assisting with initial annotations and employing human annotators for validation can improve scaling.

- **Existing MWEs:** The filter system assumes existing MWE resources remain valid, but this may oversimplify language evolution and semantic shifts.

- **Feasibility of LLMs:** As Samadarshi et al. (2024) show, LLMs struggle to automatically address this task, however, a human-in-the-loop system should improve this system.

5 Conclusion

This project aims to create a robust pipeline for discovering and annotating novel MWEs from social media by assigning confidence scores to po-

[2] This is the current SOTA model for NER on the CoNLL 2003 (English) dataset with an F1 score of 94.6%.

[3] BPE seems to be the more adopted choice, however for multilingual settings and low-resource languages, Morphological Segmentation has been shown to perform better Mager et al. (2022)

tential MWEs and identifying emerging trends in informal online language. By combining statistical association measures, LLMs, and a human-in-the-loop feedback system, this approach will contribute to linguistic research by providing insights into evolving language use. Additionally, the annotated dataset will be released as part of a shared task, encouraging further research and development in the field of MWE discovery.

References

Pavlos Fafalios, Vasileios Iosifidis, Eirini Ntoutsi, and Stefan Dietze. 2018. Tweetskb: A public and large-scale RDF corpus of annotated tweets. *CoRR*, abs/1810.10308.

Stefan Th. Gries and Alexander Wahl. 2017. Merge : A new recursive approach towards multiword expression extraction and four small validation case studies.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulic, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the Workshop on NLP for Conversational AI*. Data available at github.com/PolyAI-LDN/conversational-datasets.

Adam Kilgarriff, Vít Baisa, Miloš Jakubíček, and Pavel Rychlý. 2015. Longest-commonest match. *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom*, page 397–404.

Jong Myoung Kim, Young-jun Lee, Sangkeun Jung, and Ho-jin Choi. 2023. Semantic ambiguity detection in sentence classification using task-specific embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 425–437, Toronto, Canada. Association for Computational Linguistics.

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. Detecting multiword expression type helps lexical complexity assessment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2022. Bpe vs. morphological segmentation: A case study on machine translation of four polysynthetic languages.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, Los Angeles, California, USA. Association for Computational Linguistics.

Carlos Ramisch, Abigail Walsh, Thomas Blanchard, and Shiva Taslimipoor. 2023. A survey of MWE identification experiments: The devil is in the details. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 106–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Luis Rei, Simon Krek, and Dunja Mladenić. 2016. xLiMe twitter corpus XTC 1.0.1. Slovenian language resource repository CLARIN.SI.

Pavel Rychlý. 2008. A lexicographer-friendly association score. *Proc. 2nd Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN*, 2:6–9.

Prisha Samadarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game.

Agata Savary, Cherifa Ben Khelil, Carlos Ramisch, et al. 2023. PARSEME corpus release 1.3. In *Proceedings of the 19th Workshop on Multiword Expressions (MWE 2023)*, pages 24–35, Dubrovnik, Croatia. Association for Computational Linguistics.

Prateek Saxena and Soma Paul. 2020. Epie dataset: A corpus for possible idiomatic expressions.

Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T. Mordowanec, Henrietta Conrad, and Noah A. Smith. 2014. Comprehensive annotation of multiword expressions in a social web corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 455–461, Reykjavik, Iceland. European Language Resources Association (ELRA).

Aline Villavicencio and Marco Idiart. 2019. Discovering multiword expressions. *Natural Language Engineering*, 25(06):715–733.

Xinyu Wang, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2020. Automated concatenation of embeddings for structured prediction. *CoRR*, abs/2010.05006.

Nicolas Zampieri, Irina Illina, and Dominique Fohr. 2021. Improving automatic hate speech detection with multiword expression features.

Nicolas Zampieri, Carlos Ramisch, Irina Illina, and Dominique Fohr. 2022. Identification of multiword expressions in tweets for hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 202–210, Marseille, France. European Language Resources Association.