Automatic identification of Potential Multi-word Expressions using Naturalness of scenes

Nishan Chatterjee

May 6, 2024

1 Project Proposal

The goal of the project is to advance the automated identification of multi-word expressions (MWEs) within multilingual corpora, with a particular focus on English (for prototyping) and thereby extending to languages with limited resources. MWEs, being phrases composed of multiple words that act as a single semantic entity, present a notable challenge in the realm of computational linguistics due to their complex and idiomatic characteristics. By leveraging large language models/ large language and vision models, visual scene comprehension, generating artificially possible scenes, and analysing MWEs in diachronic windows, the goal of this project is to construct a system that will contribute to the enhancement of natural language understanding and processing.

Keywords: Multi-Word Expressions (MWEs), Computational Linguistics, Scene Representation, Diachronic Analysis

2 Potential Methodologies

1. Use large language models (LLMs) to process descriptions of scenes from various corpora (literature, annotated datasets, etc.) and identify realistic physical grounding. Extract potential phrases that may form MWEs. 2. Employ large language models or language and vision models to analyse natural descriptions and corroborate the identification of MWEs through visual

context. 3. Transform natural descriptions into a scene graph representation to enumerate and simulate all possible relationships using either graph-based methods [2], or borrowing inspiration from Phi-3 [1], thereby facilitating the generation of a range of potential scenes for MWE identification. 4. Conduct diachronic analysis to examine the evolution of linguistic multi-word expressions over time. Evaluate model performance in capturing contemporary multi-word expressions commonly used in the current period.

3 Training School Relevance

The courses offered by the training school are highly pertinent to the project. The sessions on dependency and universal dependency parsing along with the session on MWE annotation should be beneficial to this project and towards broadening my scope since I haven't worked with MWE annotation schemas before. Additionally, the brainstorming and hackathon-like structure sounds like a fantastic setup to get some initial results for a work like this.

4 Current Project Phase and Open Questions

The project is currently in the initial planning stage. While we have identified our primary goals, we don't have a concrete idea of which methodologies to start prototyping with.

1. What are the easiest questions to immediately start prototyping towards? 2. What datasets might be relevant? 3. In what ways can we model the diachronic windowing of MWEs using current linguistic datasets?

References

[1] et al. Abdin, Marah. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[2] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. 2022.