Developing a Potentially Idiomatic Expressions (PIE) Corpus for Turkish

Murathan Kurfalı

Objective

The goal of this project is to develop the first comprehensive corpus of Turkish Potentially Idiomatic Expressions (PIE), annotated for varying degrees of compositionality. This resource will significantly enhance natural language processing applications by addressing the unique challenges presented by idiomatic expressions in Turkish.

Background

Potentially Idiomatic Expressions (PIEs) are linguistic constructs where the meaning can vary drastically between literal and idiomatic interpretations based on the context in which they are used. These expressions present unique challenges in natural language processing due to their dual interpretive nature, which can affect the outcome of tasks like sentiment analysis, machine translation, and information retrieval.

For example, consider the Turkish expression "Hapı yutmak":

- Literal Usage: "*To swallow the pill*"
  - "Doktorun verdiği *hapı yuttum*."
  - "I swallowed the pill the doctor gave me."
- Idiomatic Usage: "To be in a serious trouble"
  - "Sınıfta kavga çıkaran öğrenciler *hapı yuttu*."
  - "The students who started a fight in the classroom got in serious trouble."

The ability to discern these meanings is crucial for effective language understanding systems. As such, Google translate, as of May 3, 2024, cannot correctly translate the idiomatic example above, highlighting the importance of recognition of PIEs in downstream tasks.

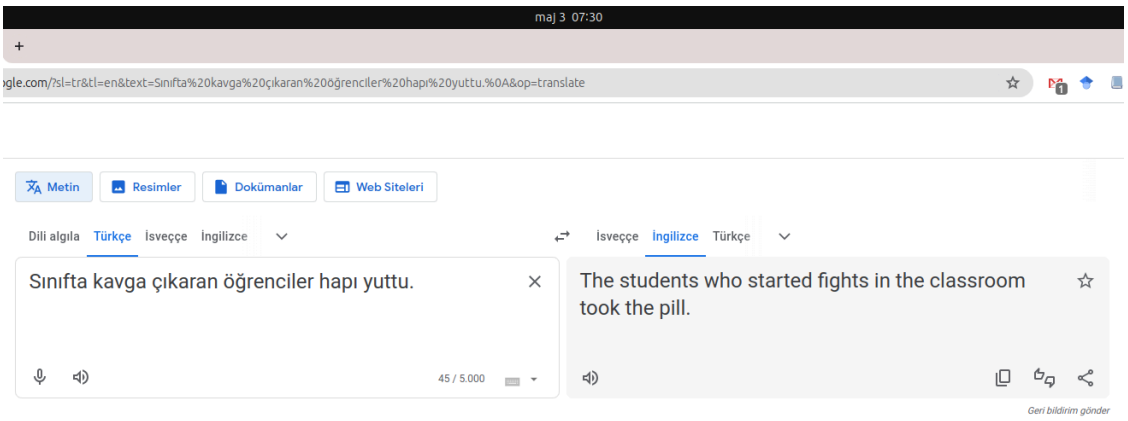

Figure 1: Google translation does not recognize the idiomatic usage of Turkish PIE

Project Methodology

Developing a linguistic resource from scratch for a new language has certain difficulties. In this case, the first challenge is to compile the list of PIEs in Turkish as currently, there is no such list available. Therefore, the project will start with compilation of a comprehensive list. To this end, I will initially go through the Turkish Parseme lexicon (Erden et al., 2018). However, to ensure completeness, a more thorough n-gram extraction, potentially up to 3-grams, will be performed on a large raw corpus, with the most frequent extracted n-grams to be manually reviewed to identify potential PIEs. The recognition and annotation of PIEs is a relatively straightforward task that can be undertaken by a single researcher, as there is little ambiguity involved that would necessitate multiple annotators.

The second step involves creating a corpus of PIEs annotated for idiomatic or literal usage. This will be achieved through a semi-automatic method using contextual language models like BERT. As demonstrated by Kurfali (2020), BERT (Devlin et al., 2019) can effectively distinguish between different usages of PIE expressions in German and English, even in an unsupervised manner through clustering. This approach can be applied to Turkish to obtain initial annotations, which will then be reviewed and corrected by the researcher. This process will ensure a large annotated corpus of identified PIEs in an efficient manner. Modern sentence embedding models, such as mixedbread ai embeddings[1], will be employed in this stage to inform future studies.

Once sufficiently large corpus is created, a supervised computational model to distinguish idiomatic usages of Turkish PIEs will be built. Moreover, the constructed corpus will also be used to assess understanding of the modern LLMs of PIEs in a relatively low-resource language in a zero-shot setting.

Utility of the Training School

The UniDive training will provide crucial insights into corpus annotation techniques, which are fundamental for the success of this project. The courses on multiword expressions and annotation infrastructure will specifically guide the development of semi-automatic annotation processes. Additionally, one of the potential outcomes of this project is to enhance the Turkish PARSEME lexicon with a new annotation layer where the training at UniDive will be instrumental in achieving.

Open Questions

- What are the most effective methods for systematically identifying potentially idiomatic expressions (PIEs) for a new language?
- To what extent can we develop and validate automated tools to accurately annotate the usages of Turkish PIEs using contextual language models?
- How do large language models perform in identifying and interpreting potentially idiomatic expressions in Turkish?

Project Phase: This project is currently in the planning stage, set to begin with the collection of a comprehensive list of Turkish PIEs.

References

Berk, G., Erden, B., & Güngör, T. (2018, May). Turkish verbal multiword expressions corpus. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[1] https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Kurfali, M., & Östling, R. (2020). Disambiguation of potentially idiomatic expressions with contextual embeddings. In Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona, Spain (Online), December 13, 2020 (pp. 85-94).