UniDive 2nd general meeting

Naples, 8-9 February 2024

WG2, WG3

Multilingual semi-automated identification and annotation of multiword expressions

Ilan Kernerman

Abstract

This paper presents a new project aiming to develop tools and datasets for the automatic identification of multiword expressions (MWEs) and their annotation and integration in multilingual language settings. We will combine information from quality lexicographic resources and large language model (LLM) outputs along with human curation, to create annotated multilingual lexicons for machine learning-based identification of MWEs. The new resources will serve to identify MWEs and include them in corpus-driven frequency lists for lexicons and lexicographic content that enhance the performance of multiple NLP applications (e.g. (neural) machine translation, word sense disambiguation, parsing, etc), to build a framework that will serve to identify and process MWEs in numerous languages. The project is planned primarily in the context of UniDive WG2 and concerns aspects of multilinguality in WG3 as well.

The main outcomes include:

(1) A framework for MWE discovery and detection (of word form variants), with easy-to-use command line interface (CLI) and a software library offering the key features of model training (based on annotated datasets) and MWE identification (based on the trained models).

(2) Trained models for MWE identification, available for download or via web service, for diverse (types of) languages, such as Dutch, English, Estonian, French, Hebrew, Italian, Polish, Portuguese, Russian, Slovenian, Turkish, and possibly others (related to the language expertise of the project members and to the annotated multilingual lexicon developed as part of the H2020 ELEXIS project (cf. https://elex.is)).

(3) UD-based annotated datasets for MWE identification for the languages above.

The prediction models that are developed will be exported to other languages and will benefit them too.

NEEDS**.** While the automatic retrieval of single-word lexicons from corpora is relatively straightforward, extracting MWEs is non-trivial and far more complex, due to the (i) lack of agreement on terminology, (ii) large variety of syntactic structures and semantic ambiguity (e.g. *break the ice* can be idiomatic but also compositional), (iii) lack of relevant (annotated) data and annotation standards (cf. Rosén et al. 2015), and (iv) lack of effective evaluation methods (mostly for MWE identification). Therefore, MWE lexicons are usually created fully manually, or by using tools that consider only co-occurrence features.

According to Jackendoff (1997), the number of MWEs in a speaker's lexicon is possibly "of the same order of magnitude as the number of single words of the vocabulary", and according to Sag et al. (2002) "it seems likely that this is an underestimate." This emphasizes the importance of substantial coverage and precision of MWE frequency lists and lexicons, for both lexical resources and NLP applications.

WORK. We will first extract MWEs (including examples of usage) from (i) the Global series of K Dictionaries (https://lexicala.com/dictionaries/) and (ii) LLMs (e.g. ChatGPT, LLaMa, Falcon, mT5-XL), then apply cross-lingual embeddings to match them against corpora (e.g. from Sketch Engine, https://sketchengine.eu/), to list the top 80 most frequent ones per language. Sentences containing candidates for these top 80 MWEs will be automatically extracted from corpora and manually annotated by experts to obtain 25 representative sentences per MWE; more MWEs will be annotated if present in example sentences, so the total number of MWE samples per language will reach at least 2,000. Alternatively, we might initially pilot fewer languages, depending on those spoken by the participants. The Estonian data could stem from resources of the Estonian Language Institute (https://eki.ee/) and the Slovenian from those of Jožef Stefan Institute (https://ijs.si/).

We will then apply deep learning techniques, particularly LLMs, fine-tuned on the training data to discover/detect and tag unseen MWEs. The variation and quality of MWEs that appear in dictionaries, transformed with cross-lingual embeddings and suggestions from massively multilingual language models, will help machine learning models to learn complex MWE patterns (semantic and syntactic) and achieve good generalization for new data.

Recent massively multilingual language models implicitly encode text representations in a latent joint space of many languages. Cross-lingual transfer between different languages is possible by training the models in resource-rich languages, and then the acquired knowledge is transferred to target languages via zero-shot or few-shot transfer. This approach supersedes previous techniques based on explicit static and contextual embeddings that generate explicit numeric vectors for words. We will fine-tune different LLMs on our training set and use them in few-shot transfer for the covered languages as well as in zero-shot transfer mode for uncovered languages. This will reduce the need for large-scale annotation and produce generally useful MWE detectors for many languages.

INTEGRATION. The new tools, datasets, and trained models will become available on the ELG platform.

Keywords

MWE identification, MWE annotation, machine learning, models, lexicography, multilingual large language models, zero-shot transfer, few-shot transfer

References

Rosén, V., De Smedt, K., Losnegaard, G. S., Bejček, E., Savary, A., and Osenova, S. 2016.. MWEs in treebanks: From survey to guidelines. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. https://aclanthology.org/L16-1368

Jackendoff, R. 1997. *The architecture of the language faculty* (No. 28). Boston, MA: MIT Press.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In *International conference on intelligent text processing and computational linguistics*. Berlin, Heidelberg: Springer, 1-15.