2nd General Meeting
University of Naples "L'Orientale"
Naples, Italy, 8-9 February 2024
https://unidive.lisn.upsaclay.fr/

WG2 (WG3)

Multilingual semi-automated identification and annotation of multiword expressions

Ilan Kernerman
Lexicala by K Dictionaries

AIM

Develop tools and datasets for the automatic identification of MWEs and their annotation and integration in multilingual language settings.

METHODOLOGY

Combine information from lexicographic resources and LLMs along with human curation, in creating annotated multilingual lexicons for machine learning-based identification of MWEs.

The variation and quality of MWEs in dictionaries, transformed with cross-lingual embeddings and suggestions from massively multilingual language models, will help machine learning models learn complex MWE patterns (semantic and syntactic) and achieve good generalization for new data.

KEYWORDS

MWE identification/ annotation
ML
models
lexicography
multilingual LLMs
zero/few-shot transfer

STEPS

(1) Extract MWEs (including examples of usage) from lexicographic resources and LLMs
(2) Apply cross-lingual embeddings to match with corpora, and list the top 80 most frequent MWEs
(3) Auto-extract candidate sentences for the top 80 MWEs from corpora and manually annotate them to obtain 25 representative sentences per MWE = 2,000 MWE samples per language
(4) Apply deep learning techniques, particularly LLMs, finetuned on the training data, to discover/detect and tag unseen MWEs
(5) Finetune different LLMs on the training set in few-shot transfer for the languages covered, as well as zero-shot transfer mode for uncovered languages, to reduce the need for large-scale annotation, and produce generally useful MWE detectors for many languages
(6) Export the prediction models that are developed to more languages

OUTCOMES

- ✓ 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).
- ✓ 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, Slovene, Turkish, and possibly others (related to the language competence of the project members and to the ELEXIS annotated multilingual lexicon).
- ✓ UD-based annotated datasets for MWE identification for these languages.
- ✓ New tools, datasets, and trained models available on the ELG platform.

Thank you to
Prof. Marko Robnik Šikonja
University of Ljubljana - FRI

lexicala by K DICTIONARIES
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
UniDive
Funded by the European Union