



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.

STEPS

- (1) Extract MWEs from lexicographic resources and LLMs
- (2) Apply cross-lingual embeddings and list the top 80 most frequent MWEs
- (3) Auto-extract candidate and manually annotate them
- (4) Apply deep learning techniques to discover/detect and tag unseen MWEs
- (5) Finetune different LLMs on the training set in few-shot transfer for the languages covered and produce generally useful MWE detectors

KEYWORDS

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

OUTCOMES

- ✓ A framework for MWE discovery and annotation
- ✓ Trained models for MWE identification
- ✓ UD-based annotated datasets for MWE identification