Discovery of MWEs

WG3 Report on MWE Processing

Tomas Krilavičius, Justina Mandravickaitė, Michael Oakes,
Carlos Ramisch, Federico Sangati, Veronika Vincze

Abstract

This poster summarizes the WG3 survey on tools and techniques for automatic MWE discovery. It serves as a ground for the state-of-the-art report of WG3 on hybrid & multilingual MWE processing. Keywords and pointers provided here are further detailed in our shared document: `http://goo.gl/IE0hLC`

Association measures

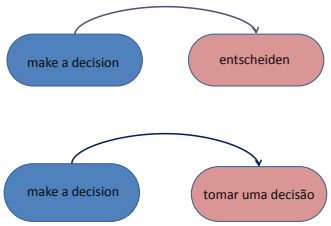
Lexical measures that estimate the association strength between words are one of the main tools employed in unsupervised discovery of MWEs in corpora. There are different ways of measuring this strength of word association:

- Measures based on raw frequency
- Measures based on information theory, e.g. pointwise mutual information
- Measures based on the contingency tables, e.g. chi-square
- Statistical significance
- Measures of association between 3 or more words
- Measures which use linguistic information in addition to word frequencies, e.g. affinity of a word to a syntactic pattern

$\Longrightarrow$ No consensus about best type of measure to use in each case

Parallel Corpora

One-to-many correspondences are exploited in MWE detection and cross-lingual MWE detection is also enabled:



Techniques based on word alignment, dependency parsing, alignment mismatches and/or decision trees have been used for MWE detection. Statistical MT systems also exploit MWE-annotated paralell corpora.

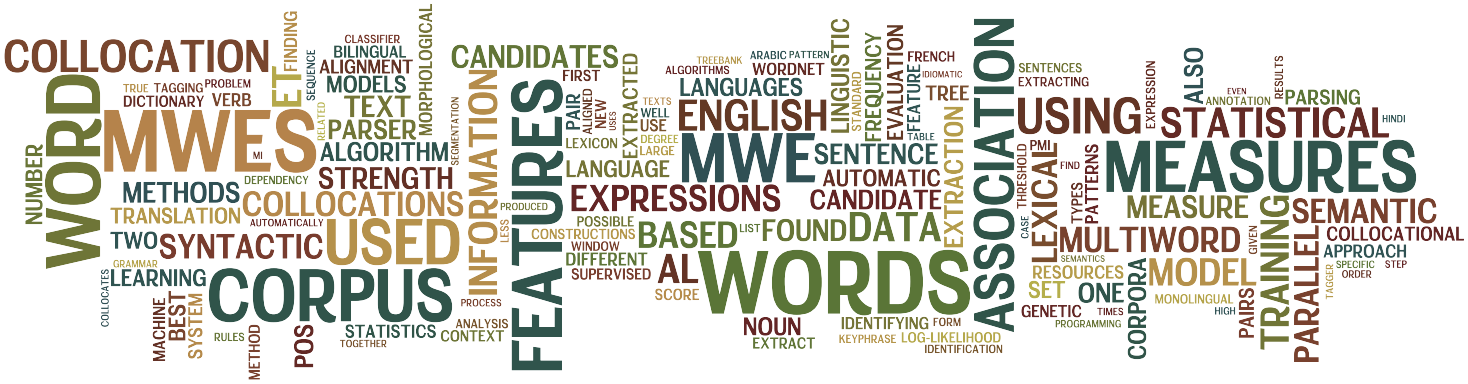
Supervised Machine Learning

Most machine learning methods use lexical resources, i.e., corpora, treebanks, dictionaries, lexicons, etc. $\rightarrow$ Dependency on certain resources makes supervised machine learning approaches partly language-dependent Primary lexical resources can be complemented with web dictionaries and WordNet.

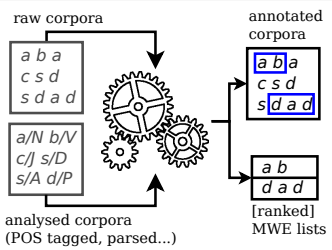
Features typically employed:

- n-gram frequencies
- lemmas
- orthographic variations
- association measures
- morphosyntactic patterns

Common techniques for complementing supervised machine learning: filtering, pregrouping, re-ranking, thresholds, combination, POS tags, chunks, chunk sequences, manual annotation and evaluation.



Tools



Corpus searches and concordancers
Sketch engine, AntConc, WordSmith

Association measures and patterns
UCS, Text::NSP, mwetoolkit, LocalMaxs, ACCURAT toolkit, Xtract (Dragon), bgMWE

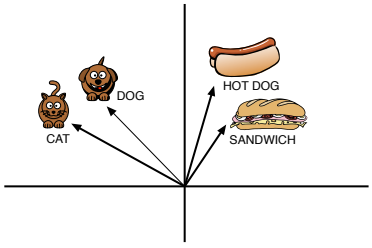
Token-based annotation/tagging
jMWE, AMALGr, FIPS-Co, StringNet

Recurring tree fragments
FragmentSeeker, DiscoDOP, Varro

Semantic Properties

Based on the *non-decomposability* property: the meaning of an MWE cannot be derived from the meanings of its component words.

Context distribution methods (hot dog $\neq$ dog)

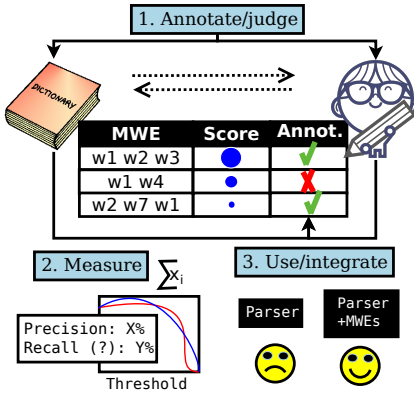


Substitution methods

Expression	Substitution	MWE
Break the vase	Break the cup	NO
Break the ice	Break the snow	YES

Evaluation

How to evaluate a lexicon of automatically discovered MWEs?



Challenging and open question