Towards a Shared Task on MWE detection

Veronika Vincze, Antoine Doucet and Agata Savary

University of Szeged Université de La Rochelle Université François Rabelais Tours

23 September 2015

Introduction

- Wide range of researchers interested in MWEs and parsing
- Several corpora and tools are available
- It is problematic to directly compare results obtained on different datasets and/or different methods
- Shared task with standardized annotation principles, corpora and evaluation metrics

Towards a shared task: aims and goals

- An earlier proposal: application for organizing a CoNLL shared task on verbal MWE detection by University of Szeged in 2014 – rejected:
  - not multilingual enough
  - no international organizing committee
- As multilingual as possible
- Focused scope: verbal MWEs
  - light verb constructions (**have a shower**)
  - verb-particle constructions (**set up**)
  - idioms (**kick the bucket**)
  - other verbal MWEs (**drink and drive**)
- Task: identify their occurrences in running text
- Evaluation: standardized evaluation metrics

What we need

- Annotated datasets
- Standardized annotation guidelines
- Annotation tool
- Evaluation methodology
- Evaluation metrics

Annotated datasets & guidelines

- Data with MWE annotation on the basis of standardized guidelines
- (Re)annotation efforts are required
- Possibly texts from newspapers
- Corpus size: 3500-4000 MWE occurrences per language (approx. 18-20K sentences, based on earlier annotation for English)
- Annotation guidelines written in English
- Harmonizing theoretical and computational linguistic considerations
- Basic principles:
  - Each verbal MWE occurrence is annotated
  - Subcategories are annotated
  - Lexically fixed elements that can form a separate token are only annotated (e.g. prepositions are but case suffixes are not: **take a photo of** sg but **döntést hoz** vmiről)
  - Non-contiguous elements are also annotated
- To be adapted to the given language by the annotator team
- A part of the data should be double-annotated to check IAA

Annotation tool & technical issues

- Able to mark MWEs in running texts
- Different tools available: proposals are welcome
- Annotation format (XML?) and validation scripts
- Annotation platform

Call for interest

- A call for interest was sent out in July
  - contributors to corpus annotation
  - future participants
- 17 teams have expressed their interest so far
- About 17 languages: Croatian, French, Romanian, English, Italian, Hebrew, Yiddish?, Hungarian, Spanish, Polish, Turkish, Greek, Bulgarian, Slovenian, Farsi, German, Swedish?...

Management

- Similar to Universal Dependencies
- One language – one team
- There are general guidelines
- There can be language specific extensions if necessary
- Each team is responsible for documenting language specific issues

Shared task organization

- Participants will be provided:
  - Training datasets annotated for verbal MWEs (∼3K MWE occurrences)
  - Data can be morphologically and syntactically parsed (if automatic tools are available)
  - Standardized annotation guidelines
  - Language-specific extensions
  - Annotation tool
  - Blind test datasets (∼1K MWE occurrences)
- Additional data and resources might be useful:
  - A large pool of unannotated data for each language
  - Lists and lexicons of MWEs for each language
  - Other useful resources
  - Provided by the language teams?
  - Contributed by participants and then used collectively?
- Participants should provide:
  - Output of their system (i.e. output of their system run on the evaluation dataset)

Evaluation methodology

- Training and test data are annotated for subcategories of verbal MWEs
- At the evaluation, only one category will be used (MWE) – data sparsity problems and unbalanced distribution of subcategories across languages can be avoided
- Precision, recall and F-score as metrics
- Nearmisses?
- Relaxed metrics?
- Evaluation scripts are needed
- Official results: separate ranking for each language
- (Unofficial) cross-linguistic comparisons can also be made

Possible venues

- ~~ACL 2016~~
- EACL 2017 (Valencia)
- CoNLL-2017

A tentative timeline

- October 2015: confirmation of teams responsible for each language
- December 2015: annotation tool set up and annotation guidelines finalized
- January 2016: annotation starts
- June 2016: version 1.0 of training and test datasets and evaluation tools
- September 2016: datasets finalized
- 2017: shared task workshop

Questions to discuss in Iași

- Formation of teams
- Annotation guidelines
- Technical issues – WG3 meeting

Contributions: WG3

- Discovery subgroup: planning & participation
- Evaluation metrics & methodology
- Technical issues:
- annotation tool (also non-adjacent MWEs + for linguists!)
- annotation platform

Evaluation methodology

- **Tracks**: open, closed or semi-closed?
- How to treat **additional resources** (if any)?
- Separate **ranking** for different methods? (supervised-unsupervised etc.)
- Separate ranking for different tracks?
- Separate ranking for subcategories of verbal MWEs?

Evaluation metrics

- Phrase-based or token-based?
- Exact match or more relaxed metrics?
- **take this issue into account**
- FP and FN at the same time
- Precision/Recall/F-score ?