Verbal Multiword Expression Identification: Do We Need a Sledgehammer to Crack a Nut?

Caroline Pasquer¹  Agata Savary¹  Carlos Ramisch²  Jean-Yves Antoine¹

¹University of Tours, LIFAT, France; ²Aix Marseille University, LIS, France

Multitask expressions (MWEs)

- Definition: Word combinations with idiosyncratic properties (lexical, syntactic, semantic, etc.)
  - cut corners ‘do an incomplete job’
- MWE identification – challenges:
  - Variability
  - Companies were cutting corners to save costs.
  - The field would look uneven if corners were cut.
  - Literal and coincidental occurrences
  - Start with cutting one corner of the disinfectant bag.
  - If you cut along the lines, you get two acute corners.
- Our idea: simple method to identify verbal MWEs (VMWEs) previously seen in training data

PARSEME Shared Task

- Corpora:
  - Unified typology for VMWEs in 19 languages:
    - verbal idioms (VID: cut corners),
    - light-verb constructions (LVC: full: take a walk),
    - inherently reflexive verbs (IRV: help oneself)
  - VMWE + morpho-syntactic annotation
  - TRAIN/DEV/test split
  - Seen VMWEs: 74–94% in DEV; 50–92% in TEST
- Shared task on VMWE identification (edition 1.1):
  - Systems are given VMWE-annotated TRAIN/DEV,
  - Given test with morphosyntax, predict VMWEs
  - Best cross-language macro-average F1 < 0.6

MWEs’ Nature as a Guiding Principle

- MWEs are regular at the level of tokens, but idiosyncratic at the level of types
- Strong lexicalization: the combination of lexemes (not their senses) makes a MWE
  - cut corners vs. trim corners, cut edges
- Critical difficulty of unseen data
  - Room for improvement on seen MWEs
  - Literal readings are extremely rare in practice

Hypotheses

- H1 Search co-occurrences of precise lexemes (+POS) annotated as VMWEs in train
- H2 Allow only for those morphosyntactic variants which were previously seen in train
- H3 Not rely heavily on automatic POS tagging and syntactic parsing, which may be noisy
- H4 Consider syntactic coherence to eliminate coincidental occurrences

The Seen2020 system

1. Extract from dev (training) or test (testing) previously seen lemma multisets (H1)
2. Filter by applying a combination of:
   - [f1] Components should be disambiguated by their POS (H1, H3)
   - [f2] Components should appear in specific orders (H2)
   - [f3] Components and inserted POS should appear in specific orders (H2)
   - [f4] Components should not be too far (H4, H3)
   - [f5] Closer components are preferred over distant ones (H4, H3)
   - [f6] Components should be syntactically connected (H4)
   - [f7] Nominal components should appear with a seen infection (H2)
   - [f8] Nested VMWEs should be annotated as in training
3. Select the optimal filter combination (among 2⁸ possibilities) using dev

Results

- System
  - Macro-ave. F1
  - SHOMA
  - TRAVERSAL
  - TRAPACC₁₀
  - VarIDE
  - Seen2020

- Ex. seen data
  - La lumière sera faite sur ce drame
    - ‘The light will be done on this drama.’
  - La porte a résolument été fermée aux initiatives
    - ‘The door was firmly closed on initiatives.’
  - Il ferme la porte à une loi
    - ‘He closes the door on a law.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’

Conclusions

- Seen2020’s focus on (variants of) seen data boosts global performances
  - Seen2020 outperforms best (neural) systems on seen and seen+unseen data
  - Seen2020 is very simple: 8 morphosyntactic on/off filters (vs. millions of real numbers)
  - Seen2020’s results are interpretable:
    - Identify filters responsible for errors → incremental development and customization
    - Filters reflect known, pervasive and generic linguistic properties of VMWEs
  - Drawbacks: sensitivity to scarcity and quality of VMWE annotations
  - Perspectives: new languages, out-of-domain generalizability

Focus on filters: example corpus

- Ex. Candidate
  - SHOMA
  - TRAVERSAL
  - TRAPACC₁₀
  - VarIDE
  - Seen2020

- Ex. seen data
  - La lumière sera faite sur ce drame
    - ‘The light will be done on this drama.’
  - La porte a résolument été fermée aux initiatives
    - ‘The door was firmly closed on initiatives.’
  - Il ferme la porte à une loi
    - ‘He closes the door on a law.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’

- Ex. seen data
  - La lumière sera faite sur ce drame
    - ‘The light will be done on this drama.’
  - La porte a résolument été fermée aux initiatives
    - ‘The door was firmly closed on initiatives.’
  - Il ferme la porte à une loi
    - ‘He closes the door on a law.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’

- Ex. seen data
  - La lumière sera faite sur ce drame
    - ‘The light will be done on this drama.’
  - La porte a résolument été fermée aux initiatives
    - ‘The door was firmly closed on initiatives.’
  - Il ferme la porte à une loi
    - ‘He closes the door on a law.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’
  - La lumière a enfin été faite sur ce drame.
    - ‘The light was finally shed on this drama.’