Multi-word lexical units recognition in WordNet*

Marek Maziarz (Wrocław University of Science and Technology, Poland)

Ewa Rudnicka (Wrocław University of Science and Technology, Poland)

Łukasz Grabowski (University of Opole, Poland)

** This research was funded by the Polish National Science Centre (NCN) under agreement no. UMO-2019/33/B/HS2/02814*.

MWE Workshop

Goal

- devise a method for recognising **multi-word lexical units** (**MWLUs**) from **multi-word expressions (MWEs)** found in:
- Princeton WordNet (Fellbaum, 1998)
- and in enWordNet (Rudnicka et al., 2015), a small extension of WordNet developed by the plWordNet team on the basis of mapping

where

- **MWEs** - (PWN and enWN lemmas consisting of) at least two graphic words separated by space(s) (cf. Sag et al., 2002)
- **MWLUs** - lexicalised MWEs (recorded by dictionaries) (Maziarz et al., in print)

MWEs vs MWLUs

- MWEs: ‘idiosyncratic interpretations that cross word boundaries (or spaces)’’ (Sag et al. 2002)
- varying degree of syntactic and/or semantic idiosyncrasy leads to a varying degree of lexicality of different types of MWEs such as:
    - idioms,
    - proper names,
    - fixed phrases,
    - compound nouns,
    - collocations.
- Which MWEs can be treated as lexicalised *multi-word lexical units (MWLU)s*?

What do we need the MWE/MWLU distinction for?

- To know which strings of words function as words themselves and cannot be annotated through their component parts only (McCrae et al. 2020)

- NLP tasks that need MWE/MWLU identification:
  - morpho-syntactic tagging
  - parsing
  - sense annotation
  - word-sense disambiguation
  - text understanding

Direct motivation

- the list of WordNet MWEs treated as gold standard for NLP applications (Pearce, 2001; Farahmand et al., 2014; Schneider et al., 2014; Riedl & Biemann, 2016)
- many of these MWEs of questionable lexicality, such as:
    - elements of wordnet taxonomy, e.g. *biological group*, *animal group*
    - quantifier phrases, e.g. *piece/article of furniture*
    - collocations: *rich people*, *psychology department*
    - …
- Which MWEs do we want in a wordnet and how shall we tag them?

Towards a method for MWLU recognition

- Building an MWE dataset
- Annotating samples
- Applying statistical models

Building an MWE dataset

- *Step 1*: extract MWEs from WordNet and enWordNet, understood as *wordnet senses with lemmas built of at least two graphic words separated by space(s)*
- *Step 2*: filter out proper names:
- MWEs from synsets holding *Instance* and/or *I-instance* relations
- *Step 3*: filter out specialist terminology of biology and chemistry:
- MWEs from synsets with hyponymy relation to {biological group 1}, {chemical element 1}, {chemical 1}
- Results:
- nouns 33.7k, verbs 4.4k, adjectives 0.5k, adverbs 0.8k

Annotating a 200 MWE sample

- a random 200 MWE sample drawn from the 39.4k MWE dataset
- MWEs annotated by a pair of lexicographers for their presence in general use English dictionaries:
    - Oxford Lexico,
    - Merriam Webster,
    - Collins,
    - Longman
- Crucially, both MWE lemmas and their PWN and enWN senses checked
- MWEs with lemmas and senses present in any of the dictionaries considered *lexicalised*.

Rule-based approach (1)

- a 200 MWE sample checked for:
  - I-synonymy,
  - the presence of an MWE lemma in a conglomerate Polish-English ‘cascade’ dictionary (Kędzia et al., 2013)

- These features were annotated automatically.

Rule-based approach (2): Making use of the I-synonymy relation

- **I(**nterlingual) **synonymy** relation links unique pairs of synsets from plWordNet and WordNet and enWordNet (Rudnicka et al. 2012)
  - understood as *large correspondence between meanings and relation structures of the synsets from the two wordnets*
- *Hypothesis:*
  - Senses from synsets holding I-synonymy relation likely to be lexicalised in the two languages
- *Reservation:*
  - the degree of correspondence between specific pairs of English-Polish senses may not the same within a given pair of Polish-English synsets.

Rule-based approach (3)

- Results (200 MWE sample):
  - Precision for the MWLU class = **76%**, Recall = 26% (too low), [“surefire”]
  - Precision for the non-MWLU class = 42%, Recall = **87%** [“trash”]

- Results (whole PWN):
  - 6,390 potential MWLUs / 39,406 English MWEs.
  - Additional evaluation (18 MWEs randomly sampled from potential MWLUs):
    - Precision for the MWLU class = 76%.

Statistical approach (1)

- a 200 MWE sample checked for
  - 6 lexicality features,
  - ridge logistic regression.

Statistical approach (2): Lexicality features

- l-synonymy;
- the presence of an MWE lemma in a conglomerate Polish-English ‘cascade’ dictionary (Kędzia et al., 2013);
- the length of an MWE in terms of the number of its characters (excluding spaces);
- the length of an MWE in terms of the number of spaces between component words;
- the cosine similarity between (MP sentence transformer vectors, calculated separately for an MWE lemma itself and its WordNet gloss);
- the ordinal number of an MWE sense in PWN.

Statistical approach (3)

- Results (200 MWE sample):
  - Precision for the MWLU class = **83%**, Recall = 45%, [“surefire”]
  - Precision for the non-MWLU class = **49%**, Recall = **83%**. [“trash”]

- Results (the whole PWN):
  - 18,971 potential MWLUs / 39,406 MWEs.
  - Additional evaluation (50 MWEs randomly sampled from potential MWLUs):
    - Precision for the MWLU class = 81%.

Conclusions

- both models perform well with respect to singling out non-MWLUs
- the models achieved good precision with respect to MWLU recognition
- still about a half of MWLUs were not found
- better models needed:
  - better features e.g. I-synonymy replaced with a more detailed sense-level relation of *strong and regular equivalence* (Rudnicka et al., 2019)
  - collocation strength measures could be added
- we obtained a gold standard-like list of MWLUs from PWN
- open question: is our dictionary-based definition of lexicality useful?

Datasets

- We publish the datasets used in this research under the CC BY-SA 4.0 licence on GitHub:
  - https://github.com/MarekMaziarz/Multi-word-lexical-units
  - https://clarin-pl.eu/dspace/handle/11321/853

References (1)

- Farahmand, M., & Martins, R. T. (2014). A supervised model for extraction of multiword expressions, based on statistical context features. In Proceedings of the 10th workshop on multiword expressions (MWE) pp. 10-16.
- Fellbaum, Ch. (Ed.) (1998). WordNet: An electronic lexical database, Cambridge, MA: MIT Press, 1998.
- Kędzia, P., Piasecki, M., Rudnicka, E. and K. Przybycień. (2013). Automatic prompt system in the process of mapping plWordNet on Princeton WordNet." Cognitive Studies, 13: 123-141.
- Maziarz, M., Grabowski, Ł, Piotrowski, T., Rudnicka, E. & Piasecki, M. (in print). "Lexicalisation of Polish and English word combinations: an empirical study". Poznań Studies in Contemporary Linguistics.
- McCrae, J., Rademaker, A., Rudnicka, E., and F. Bond. (2020). English WordNet 2020: improving and extending a wordnet for English using an open-source methodology. Proceedings of LREC 2020.
- Pearce, D. (2001). Synonymy in collocation extraction. In Proceedings of the workshop on WordNet and other lexical resources, Second meeting of the North American chapter of the Association for Computational Linguistics, pp. 41-46.

References (2)

- Riedl, M., & Biemann, Ch. (2016). Impact of MWE resources on multiword recognition. In Proceedings of the 12th Workshop on Multiword Expressions, pages 107–111, Berlin, Germany. Association for Computational Linguistics.
- Rudnicka, E., Maziarz, M., Piasecki, M. and S. Szpakowicz. (2012). A strategy of mapping Polish WordNet onto Princeton WordNet. In Kay, M. and Boitet, Ch. (eds.), Proceedings of COLING 2012: Posters. Mumbai, India, pp. 1039-1048. www.aclweb.org/anthology/C12-2101
- Rudnicka, E. Witkowski, W, and M. Kaliński. (2015). Towards the methodology for extending Princeton WordNet. Cognitive Studies 15.
- Rudnicka et al. (2019). Sense equivalence in plWordNet-Princeton WordNet mapping. International Journal of Lexicography, Oxford.
- Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In International conference on intelligent text processing and computational linguistics, pp. 1-15. Springer, Berlin, Heidelberg.
- Schneider, N., Danchik, E., Dyer, Ch., & Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the MWE gamut. Transactions of the Association for Computational Linguistics, 2:193–206.
- Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. (2020). MPNet: Masked and permuted pre-training for language understanding. Advances in Neural Information Processing Systems, 33: 16857-16867.

Acknowledgements

- This research has been funded by the Polish National Science Centre under the grant agreement No UMO-2019/33/B/HS2/02814.

Q/A

- Thank you very much for your attention

- Contact: marek.maziarz@pwr.edu.pl, ewa.rudnicka@pwr.edu.pl, lukasz@uni.opole.pl

Lexicality of MWEs in dictionaries

- We used Collins, Longman, Lexico and Merriam-Webster (the fantastic four):
  - large and renowned
- Collins COBUILD (ca ⅔ of the dictionary MWLUs):
  - verbalised MWE policy = semantic <u>and</u> syntactic idiosyncrasy.
- Lexico & Merriam-Webster (Maziarz et al., in print):
  - the same conclusions: semantic non-compositionality <u>and</u> strong collocations were added.
- Problem:
  - the size of a dictionary affects the number of MWEs treated as lexicalised,
  - solution: take many different large dictionaries.