Lexicalised and Non-lexicalized Multi-word Expressions in WordNet: a Cross-encoder Approach

Marek Maziarz♠, Łukasz Grabowski♢♠, Tadeusz Piotrowski♣,
Ewa Rudnicka♠ and Maciej Piasecki♠

♠ Wroclaw University of Science and Technology, Poland, ♢ University of Opole, Poland,
♣ University of Wrocław, Poland
{marek.maziarz, ewa.rudnicka, maciej.piasecki}@pwr.edu.pl,
lukasz@uni.opole.pl, tadeusz.piotrowski@uwr.edu.pl

Wrocław
University
of Science
and Technology

Outline

- Introduction
- Sample annotation
- ML approach
- Results
- Conclusions

Introduction

- Goal: recognise lexicalised MWEs in WordNet (which we will call MWLUs - multi-word lexical units),
- Procedure: combine rule-based and statistical approaches,
- Improvement: a cross-encoder approach.

Introduction

- MWEs in PWN and in enWN (:= at least one space),
- no proper names, no biological taxonomy and chemistry terms,
- 39,406 MWEs in total.

Sample annotation: MWEs from WN

- The vast majority of MWEs in the dataset were nouns:

nouns	verbs	adjectives	adverbs
33713	4389	540	764
86%	11%	2%	1%

Table: POS statistics for the MWE dataset.

- Nearly 1% of the total data set was randomly sampled,
- 387 MWEs were chosen
  - 250 from our previous experiment Maziarz et al. (2022)
  - 137 new MWEs (to balance the sample to get MWLU/non-MWLU ratio as in the original data set)

Sample annotation: Lexicality status

- Multi-word lexical unit (MWLU) := MWE that was given the headword status in any of our reference dictionaries.
  - they are treated as multi-word lexical units by lexicographers (native speakers of English, whose lexical competence surpasses that of any native speaker of English).
  - 243/387 MWLUs in our sample.
- 6 dictionaries were inquired
  - New Oxford Dictionary of English (NODE, British),
  - Merriam-Webster Collegiate (M-W, USA),
  - Collins Dictionary (CED, British),
  - New World Dictionary (N-W, USA),
  - Collins COBUILD (COBUILD),
  - Longman Dictionary of Contemporary English (Longman).
- online versions (updated quite regularly in contrast to printed versions),
- dictionaries for both American or British English speakers.

Sample annotation: Lexicality status

- Four of those dictionaries (NODE, M-W, N-W, CED) are so-called **medium**, or **desktop**, dictionaries
  - intended to be used primarily by educated native speakers of English,
  - include most of the vocabulary that educated native speakers can find in texts
  - and which they may not know (that is why they reach for a dictionary), though they do not use them on their own.
- Two are so-called **pedagogical** dictionaries (COBUILD, Longman)
  - intended to be used primarily by advanced learners of English or non-native speakers of English (Jackson, 2022; Cowie, 2009),
  - include vocabulary of high frequency (the active vocabulary of English native speakers),
  - mainly for a non-English user,
  - a balanced selection of British and American items.

ML approach: Cross-encoder

The task of distinguishing MWLUs from non-MWLU in WordNet:

- In (Maziarz et al., 2022) we applied logistic regression to the task.
- We use now a cross-encoder (Reimers and Gurevych, 2019).
- The setu4993/smaller-LaBSE model (Feng et al., 2020)
  - smaller model — for relatively small manually annotated sample,
  - multilinguality — for future applications (especially for plWordNet).
- We used default settings for model learning.
- Four epochs were arbitrarily chosen.

ML approach: Cross-encoder

Tokenizer and model inputs were truncated to 48 tokens.

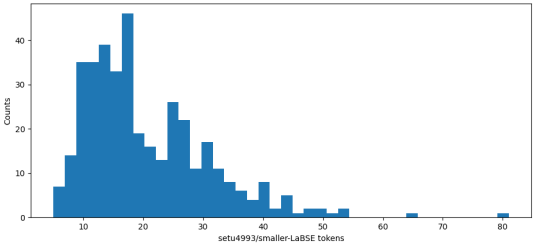

Figure: Histogram of lengths of sample definitions (enriched with hypernyms) in terms of LaBSE tokens. The 95th percentile for the empirical distribution equals 41, while the maximal length is 81 tokens.

ML approach: Cross-encoder

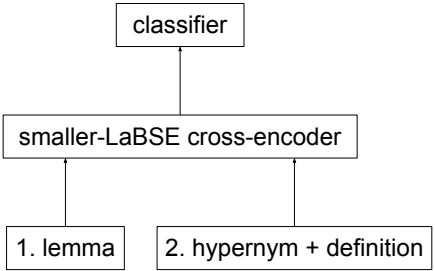

Figure: Cross-encoder for MWLU recognition in WordNet.

ML approach: Cross-encoder

lemma	hypernym, definition	label
jest at	mock, subject to laughter or ridicule	0
take back	disown, take back what one has said	1

Table: Two examples from the sample passed to the cross-encoder. Zero means 'non-lexicalised multi-word expression', while one stands for 'multi-word lexical unit'.

Hypernymic lemmas are added to teach the model to discover semantic compositionality of a MWE, cf. Bauer (2019).

ML approach: Bootstrap cross-validation

Efron's .632 bootstrap estimator (Efron, 1983; Jiang and Simon, 2007).

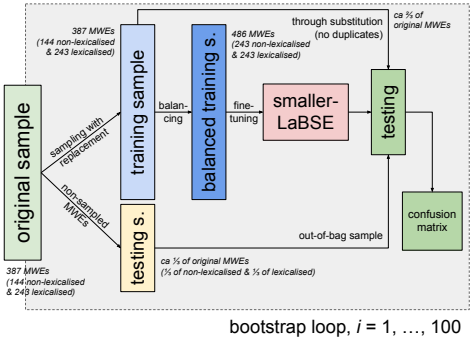

Figure: Efron's bootstrap cross-validation scheme ($B = 100$ iterations).

ML approach: Bootstrap cross-validation

The confusion matrices were obtained from Efron's .632 bootstrap rule:

$$N_i(j) = n \times Pr_i(j) = n \times [0.632 \times Pr_i^{test}(j) + 0.368 \times Pr_i^{subst}(j)] \quad (1)$$

- $B = 100$, the number of bootstrap iterations,
- $n = 387$, i.e. the whole sample size,
- i $(= 1, ..., B)$ — i-th confusion matrix,
- j $(= 1, 2, 3, 4)$ — j-th cell of the i-th confusion matrix,
- $P_i(j)$ — the proportion of each cell counts,
- the superscript ^{test} — a testing data (out-of-bag sample) confusion matrix,
- the subscript ^{subst} — a training sample confusion matrix (through substitution).

Results

		real		efficiency		
LaBSE model		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	89.3	52.3	$\mathbf{.63}^{-}_{*}$	$\mathbf{.62}^{**}_{**}$	$\mathbf{.62}^{-}_{**}$
	MWLU	54.5	190.9	$\mathbf{.78}^{**}_{**}$	$.78_{**}$	$\mathbf{.78}_{**}$
majority baseline		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	0	0	—	0	—
	MWLU	144.0	243.0	.63	$\mathbf{1}^{**}$	**.77**
random baseline		non-MWLU	MWLU	P	R	F
prediction	non-MWLU	69.7	71.9	.49	.36	.41
	MWLU	124.0	121.4	.50	.63	.55

Results

Caption to the previous table. Confusion matrix and cross-encoder (setu4993/smaller-LaBSE) classification results for the discrimination of multi-word lexical units ("MWLUs") and non-lexicalised MWEs ("non-MWLU") in bootstrap cross-validation. Differences between the model and a random/majority baseline are statistically significant at *) $< .025$ or **) $< .01$ significance level. Comparisons with the random baseline are presented in subscript, while differences from the majority baseline are given in superscript. The presented values are averaged out over all bootstrap iteration rounds. Please note that the significance level less than 0.01 was obtained when none of the bootstrap trials (out of $B = 100$ samples) found a result supporting the null hypothesis.

Results

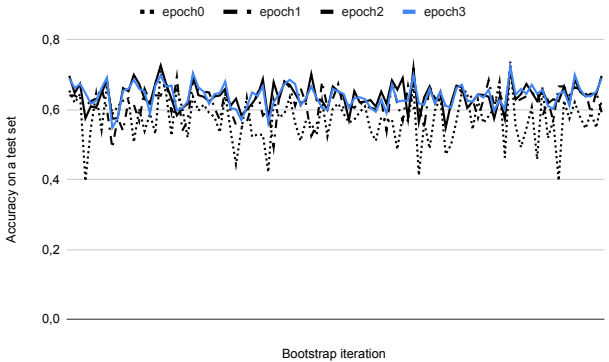

Figure: Accuracy gain/loss on testing sets throughout four epochs and one hundred bootstrap iterations.

Conclusions

The smaller-LaBSE language model:

- is better than the uniform distribution random baseline,
- has beaten the majority class baseline
  - with regard to the precision of the 'MWLU' class
  - and the recall of the 'non-MWLU' class,
  - the F1 measures were indistinguishable.
- has beaten our previous model (Maziarz et al., 2022):
  - F1 for the 'MWLU' class is much better (78% vs. 58%, $p < 0.01$),
  - the measure for the 'non-MWLU' class is not worse (62% vs. 61%, $p = .31$)

Conclusions

- Interestingly, the cross-encoder model was given no more than bare lemmas and their synset definitions enriched only with hypernyms.
- No corpus frequency (a feature important in MWE recognition) was provided.
- We assume that the smaller-LaBSE cross-encoder (the black box *par excellance*) relied on semantic discrepancies between a word combination and its semantic description in the definition, that is, on semantic opacity/compositionality.
- This assumption should be further verified in consecutive experiments in the future.
- The rationale for our experiment is pivoted on lexicographic descriptions taken manually from dictionaries. A few words must be said to address possible shortcomings of this approach.

Conclusions

Operationalization of the definition of MWLUs has its own limitations:

- Native-speaker dictionaries can include items because these items were included in some dictionaries that had been published earlier and which were quite influential.
- And these items are not lexical units, even though they are quite frequent in texts but the users might expect them in a dictionary.
- M-W and Oxford dictionaries are such influential dictionaries.
- Unfortunately, this also works in the other direction: a MWLU that is not very rare in texts may not be recorded in dictionaries because no previous dictionary recorded it.
- In contrast, editors of pedagogical dictionaries are not constrained by tradition and one may believe that the items they include are genuine lexical items.
- Clearly there is room for improvement both for wordnets and for "traditional" dictionaries.

Acknowledgements

This research was funded by the Polish National Science Centre (NCN) under agreement no. UMO-2019/33/B/HS2/02814. Also, we would like to thank Mirosława Podhajecka for her invaluable help with data annotation.

Bibliografia I

Aho, A. V. and Ullman, J. D. (1972). *The Theory of Parsing, Translation and Compiling*, volume 1. Prentice-Hall, Englewood Cliffs, NJ.

American Psychological Association (1983). *Publications Manual*. American Psychological Association, Washington, DC.

Ando, R. K. and Zhang, T. (2005). A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Andrew, G. and Gao, J. (2007). Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Bauer, L. (2019). *Complex lexical units. Compounds and multi-word expressions*, chapter Compounds and multi-word expressions in English. de Gruyter.

Chandra, A. K., Kozen, D. C., and Stockmeyer, L. J. (1981). Alternation. *Journal of the Association for Computing Machinery*, 28(1):114–133.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Cowie, A., editor (2009). *The Oxford History of English Lexicography. Clarendon Press*. Clarendon Press, London.

Bibliografia II

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: the 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Farahmand, M. and 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)*, pages 10–16.

Fellbaum, C., editor (1998). *WordNet – An Electronic Lexical Database*. The MIT Press.

Feng, F., Yang, Y., Cer, D., Arivazhagan, N., and Wang, W. (2020). Language-agnostic bert sentence embedding.

Gantar, P., Colman, L., Parra Escartín, C., and Martínez Alonso, H. (2018). Multiword expressions: Between lexicography and NLP. 32(2):138–162. _eprint: https://academic.oup.com/ijl/article-pdf/32/2/138/29012810/ecy012.pdf.

Gusfield, D. (1997). *Algorithms on Strings, Trees and Sequences*. Cambridge University Press, Cambridge, UK.

Bibliografia III

Jackson, H. (2022). *The Bloomsbury Handbook of Lexicography. Second ed.* Bloomsbury Academic, London.

Jezek, E. (2016). *The Lexicon: An Introduction (Oxford Textbooks in Linguistics*. Oxford University Press, Oxford.

Jiang, W. and Simon, R. (2007). A comparison of bootstrap methods and an adjusted bootstrap approach for estimating the prediction error in microarray classification. *Statistics in medicine*, 26(29):5320–5334.

Lipka, L. (1990). *An Outline of English Lexicology; Lexical Structure, Word Semantics, and Word-formation*. Max Niemeyer, Tuebingen.

Maziarz, M., Grabowski, Ł., Piotrowski, T., Rudnicka, E., and Piasecki, M. (2023). Lexicalisation of polish and english word combinations: an empirical study. *Poznan Studies in Contemporary Linguistics*. in print.

Maziarz, M., Rudnicka, E., and Grabowski, Ł. (2022). Multi-word lexical units recognition in wordnet. In *Proceedings of the 18th Workshop on Multiword Expressions@ LREC2022*, pages 49–54.

Rasooli, M. S. and Tetreault, J. R. (2015). Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. version 2.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Bibliografia IV

Riedl, M. and Biemann, C. (2016). Impact of mwe resources on multiword recognition. In *Proceedings of the 12th Workshop on Multiword Expressions*, pages 107–111.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002a). Multiword expressions: A pain in the neck for NLP. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer Berlin Heidelberg.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002b). Multiword expressions: A pain in the neck for nlp. In *International conference on intelligent text processing and computational linguistics*, pages 1–15. Springer.

Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: running the mwe gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Van Rompaey, T., Davidse, K., and Petré, P. (2015). Lexicalization and grammaticalization: The case of the verbo-nominal expressions be on the/one's way/road. *Functions of Language*, 22(2):232–263.

Woźniak, M. (2017). *Jak znaleźć igle w stogu siana? Automatyczna ekstrakcja wielosegmentowych jednostek leksykalnych z tekstu polskiego*. IJP PAN, Kraków.