Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)

Constructing a Derivational Morphology Resource with Transformer Morpheme Segmentation

Łukasz Knigawka
Warsaw University of Technology
Faculty of Electrical Engineering ul. Koszykowa 75
00-662 Warszawa, Poland
lukasz.knigawka.stud@pw.edu.pl

Abstract

This paper describes a framework for the creation of new derivational morphology databases for a selected set of productive affixes in English. The sample resource obtained comprises almost 120k English words with morpheme segmentations generated by Transformer. The model and the database have been compared against other existing solutions. Moreover, this study offers an overview of potentially problematic cases encountered during the process of automatic word segmentation.

1 Introduction

Derivational morphology studies the formation of new words (lexemes) ”rather than forms of a single word (cf. inflection)” (Bauer, 2004). The most common way of deriving new English words is affixation, which involves combining potential bases with affixes so that a new, morphologically complex word can be built. In the present study, two kinds of affixation are considered: suffixation (suffixes are the affixes placed after a base) and prefixation (prefixes precede a base). Affixes, as well as the bases, can be subsumed under morphemes, which are the smallest meaningful morphological units of a language (Hockett, 1958). Morphological segmentation divides words into morphemes, hence automatic morpheme segmentation employs computational methods of morpheme boundary identification. The main focus of this paper is canonical segmentation, first introduced in Cotterell et al. (2016b). It analyses a word as a sequence of canonical morphemes representing the underlying forms of morphemes, which may differ from their orthographic representations. For example, the canonical segmentation of the word *funniest* is *fun-y-est*. In principle, canonical morphological segmentation constitutes a useful, though insufficient, tool for the analysis of morphologically complex words. In this work, methods of automatic morpheme segmentation are reviewed with the aim to create new morphological resources. Initially, a machine learning model is trained to perform canonical morphological segmentation. Subsequently, English words consisting of more than one morpheme are selected for further analysis. All the model input words are potentially affixed, i.e. they contain one of the affixes (prefixes and suffixes) under review. This study also investigates how the trained segmentation model would deal with problematic morphological cases.

2 Related Work

Several recent studies have focused on automatic morphological segmentation. The log-linear model proposed in Cotterell et al. (2016b) is to learn to segment and restore orthographic changes jointly. In Kann et al. (2016), a character-level model consisting of five encoder-decoders is introduced and has become the new state-of-the-art. Convolutional neural networks have been applied in the process of morphological segmentation of Russian words in Sorokin and Kravtsova (2018). A discriminative joint model for canonical segmentation, with a context-free grammar backbone, has been introduced in Cotterell et al. (2016a). After applying it to a subset of the English portion of the CELEX data (Baayen et al., 1996), an annotated treebank consisting of over 7k English words was released. Importantly, Mager et al. (2020) propose two new approaches to obtaining canonical segmentations of words whilst working with limited training data: an LSTM pointer-generator and a neural transducer trained with imitation learning. The two recommended methods outperformed baselines in the low-resource setting while achieving scores close to the best models in the high-resource cases. Another attempt at generating canonical segmentations of lexical items from low-resource languages is described in Moeng et al. (2022), where Transformer obtained not only the highest performance score but also the supervised models outperformed the unsupervised ones. On the other hand, a novel, semi-automatic method of the construction of word-formation networks, focusing mainly on derivation, is proposed in Lango et al. (2021), where sequential pattern mining is used in an unsupervised manner to construct morphological features.

The application of neural networks in different computational morphology tasks, such as morphological segmentation, is delineated in Liu (2021). A model capable of building better word representations for morphologically complex words is proposed in Luong et al. (2013), where RNNs are combined with neural language models to learn morphologically-aware word representations. Other studies, such as Jurdzinski (2017) and El-Kishky et al. (2019), show that performing morpheme segmentation may facilitate the capturing of word properties more efficiently when creating word embeddings. Song et al. (2020) demonstrate that adopting Transformer (Vaswani et al., 2017) to process morpheme information on the input layer may improve performance in the semantic textual similarity task. Hofmann et al. (2021) examine how the input segmentation of BERT (Devlin et al., 2018) affects its interpretations of derivationally complex words and suggests afterwards that the generalisation capabilities of pretrained language models could be improved if a morphologically-informed vocabulary of input tokens has been applied. Hofmann et al. (2020) focus on productive derivational morphology and indicate that pretrained language models, BERT specifically, could generate correct derivatives in a sentence cloze task.

Although many modifications to the standard Transformer architecture have been proposed since the original paper was published, many of them failed to do well across different applications, as demonstrated in Narang et al. (2021). Some Transformer implementations aim explicitly at improving model efficiency. For instance, Primer (So et al., 2021) achieved a smaller training cost thanks to squaring ReLU activations and adding depthwise convolution layers in self-attention. As per Wu et al. (2021), the batch size was crucial in the performance of Transformers on character-level tasks, and with a large enough batch size, recurrent networks are outperformed.

This paragraph presents several recent studies that have attempted to create morphological resources. For instance, Universal Derivations constitutes a collection of harmonised (converted into a common file format and partially converted to a shared schema) word-formation resources (Kyjánek et al., 2020), while DErivBase is a rule-based framework for inducing derivational families for German (Zeller et al., 2013). That approach is further developed for Russian in DerivBase.Ru (Vodolazsky, 2020), whereas almost 70k English words were gathered in the derivational database named MorphoLexEN and presented in Sánchez Gutiérrez et al. (2017). Similar procedures for word segmentation as those used in MorphoLexEN are utilised in MorphoLexFR (Mailhot et al., 2019) which includes almost 39k French words. A derivational and inflectional morphology database (extracted from Wiktionary and consisting of about 519k derivatives in 15 languages) called Morphynet is proposed in Batsuren et al. (2021).

3 Experiments

A transformer model[1] consisting of encoding and decoding blocks was used to obtain word morpheme segmentations. The encoder block comprised positional embedding, multi-head attention, feed-forward and dropout, while the decoder blocks were constructed with the same layers, but the positional embedding layer was masked. The Transformer implementation used for experiments differed slightly from the one proposed in Vaswani et al. (2017). Learned positional encoding was applied instead of a static one, the optimiser's learning rate was static instead of one with warm-up and cool-down, and no label smoothing was utilised. The implementation of the model was inspired by that explored in Moeng et al. (2022). The hidden dimension was set to 256, and the learning rate worked best at 0.0005. A relatively small dropout of 0.1 was applied. Various optimizers available in PyTorch were tested, e.g., Adam (Kingma and Ba, 2014), RAdam (Liu et al., 2019), NAdam (Dozat, 2016), AdamW (Loshchilov and Hutter, 2017), Adadelta (Zeiler, 2012) and Adagrad (Duchi et al., 2011). Adam was chosen in Vaswani et al. (2017) and Moeng et al. (2022), but AdamW led to slightly better results in BERT. NAdam performed best in this research. Different activation functions were tested to replace ReLU (which was used in Moeng et al., 2022), and even though the differences

[1] The code is accessible at https://anonymous.4open.science/r/CanonicalSegmentationTransformers-81ED/

Type	List
Prefix	after, anti, back, circum, contra, counter, de, dis, ex, extra, fore, hyper, im, in, inter, intra, macro, mal, mega, mis, non, out, over, post, pre, pro, pseudo, re, retro, sub, super, supra, trans, ultra, un, under
Suffix	able, age, al, an, ance, ancy, ant, ary, ate, dom, ee, eer, en, er, ess, esque, ette, ful, hood, ian, ic, free, ify, ion, ise, ize, ite, ish, ism, ist, ity, ive, less, let, like, ment, ness, or, ous, ship, some, ster, th, wise, y

Table 1: Lists of considered productive affixes.

in the model scores obtained were not significant, consistently, the best results were obtained with GeLU (Hendrycks and Gimpel, 2016). Squaring ReLU activations, as proposed in Primer slightly decreased performance which decreased even more after trying out Swish units (Ramachandran et al., 2017).

The new derivational morphology resource was built with Transformer word morpheme segmentation. The model was trained on the data from MorphoLexEN. The words used to develop this resource were obtained from the English Lexicon Project (Balota et al., 2007) and were already segmented into morphemes. Inflectional suffixes such as *-s*, *-ing* or *-ed* and contractions such as *'ll* or *'s* were removed manually. Out of 68,624 words in the database, 80% formed the training set, and 10% were assigned to validation and test sets.

A relatively extensive list of English words was compiled out of lexical items from various sources: NLTK corpus (Bird and Loper, 2004), Brown corpus (Francis and Kucera, 1979) and built-in English word lists of macOS and Ubuntu. Each word was case-insensitive. Many words overlapped, so all the duplicates had to be removed. Then, all the individual lists were merged into one list containing 315,404 words. Finally, each word from the list was automatically segmented and entered in the morphological resource, provided that the relevant number of automatically segmented morphemes was greater than one and the lexical item under study started or ended with one of the selected affixes. A set of recognisable productive affixes considered in this study is presented in Table 1.

Model	Accuracy	F1
Semi-CRF	0.54 (.018)	0.75 (.014)
Joint	0.77 (.013)	0.87 (.007)
Joint+Vec	0.82 (.020)	0.90 (.008)
Transformer	0.77 (.015)	0.79 (.015)

Table 2: Results of the canonical segmentation task on a subset of the English part of the CELEX database. Standard deviation is given in parentheses.

4 Results

In this section, model performance is compared to other solutions, the new derivational morphological resource is evaluated, and puzzling morphological cases are analysed.

4.1 Model performance

The model used to create the morphological resource was trained on the subset (Cotterell, 2016) of the English portion of the CELEX lexical database with the view to compare model performance with other modern solutions. The reported results were obtained with 10-fold cross-validation. The training, validation and test sets consisted of 8k, 1k and 1k samples, respectively. Encoder and decoder dropouts were increased to 0.3 to account for limited data issue. Adam optimization and ReLU activations seemed to work best in this low-resource setting. Two metrics were used for comparison: accuracy and morpheme F1 (Van den Bosch and Daelemans, 1999). Segmentation accuracy measured whether every canonical morpheme was identified correctly. This implies that this metric is very harsh, and very close answers are penalized equally as the wrong ones. Morpheme F1 would give credit only if some canonical morphemes were identified correctly. Results are exhibited in Table 2, where the developed model was compared with Semi-CRF (Sarawagi and Cohen, 2004), Joint (Cotterell et al., 2016b) and Joint+Vec (Cotterell and Schütze, 2018).

The Transformer accuracy and F1 measure are close to the scores of other models. More data would probably significantly increase the performance of the tested model. The model used to create the new resource was trained on a several times larger dataset (a subset of MorphoLexEN) and achieved over 94% morpheme F1 and almost 93% segmentation accuracy on the test dataset.

	Morphynet	MorphoLexEN	Morfem (non-strict matching)	Morfem (strict matching)	Combined
Size	67,412	68,624	163,036	118,900	235,579
Precision	0.628	0.592	0.594	0.700	0.561
Recall	0.814	0.848	0.879	0.754	0.929
F1	0.709	0.697	0.709	0.723	0.700

Table 3: Word count, precision, recall and F1 comparison of two chosen linguistic resources, two variants of the proposed one and a combination of Morphynet, MorphoLexEN and Morfem without strict matching.

4.2 The new resource

The obtained morphological resource, named Morfem, consists of 118,900 words supplied with their segmentations[2]. In what follows, the evaluation of the database is discussed.

One thousand random words from the database were manually checked to determine whether their morphological status was correctly recognised. It turned out that over 90% of the randomly selected words constituted complex words derived with one of the selected affixes. The words which were manually marked as simplex yet segmented by the model could be subsumed under different categories. The general list included some proper names, e.g., *Demontez* was segmented as *De-montez*, along with lexical items that were not listed in English dictionaries, e.g., *unie*, or, misspelled words, e.g., *tecnology*. Some morphological cases appeared to be problematic. Certain primarily lexicalized words with potentially divisible internal structures may pose some obstacles, e.g., the words *delay* and *discard* may be treated either as delay and *discard* or *de-lay* and *dis-card*. In MorphoLexEN, *delay* was treated as a single morpheme, while *discard* was divided. The model managed to learn that, and thus only *discard* was included in the resulting database (was divided into *dis* and *card*).

To automatically validate the resource, 901 derivatives containing one of the affixes under study were retrieved manually from Joseph Conrad's Heart of Darkness (Conrad, 1899/2006). Precision and recall measures were calculated for the new database, MorphyNet and MorphoLexEN, to compare the coverage of the created resource with other morphological databases. Words that were present in both, a database and in the manually selected set of derivatives from the book, were marked as true positives. Words that were present in the book and a database, but not in the manually selected set were counted as false positives. Finally, the words that were manually selected, but not found in a resource were designated as false negatives. The test results are presented in Table 3. Two versions were compared with the other databases. One with strict matching, where a word was noted in the resource only if one of the identified morphemes overlapped with an affix from the list. The other, without strict-matching, included all the words which contained more than one morpheme, and started or ended with at least one of the selected affixes. Morfem with strict matching achieved the highest precision while lacking in recall. Morfem with non-strict matching achieved the highest coverage of the derivatives, which is indicated by the highest recall score among the compared databases. Combining the non-strict Morfem with other resources (excluding the strict-matching Morfem) to form a unified vocabulary resulted in even higher recall alongside a significant precision decrease. Deciding which metric is the most relevant depends on the specific application.

5 Conclusion

The proposed framework allows for creating morphological resources larger than those currently available. The automatic morpheme segmentation task results are promising, but there is still some room for improvement. Therefore, a more reliable linguistic resource could be compiled when built upon a more reliable segmentation algorithm. Current state-of-the-art methods of canonical morphological segmentation do not consider the word's context. Knowing that words can be divided differently depending on their context (e.g., *recover* or *re-cover*), methods consulting the context should be developed.

[2] The resource is available at https://anonymous.4open.science/r/CanonicalSegmentationTransformers-81ED/src/CanonicalSegmentationTransformers/experiments/db.txt

References

R Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1996. The CELEX lexical database (cd-rom).

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The English Lexicon Project. *Behavior Research Methods*, 39(3):445–459.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. Morphynet: a Large Multilingual Database of Derivational and Inflectional Morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48.

Laurie Bauer. 2004. *A Glossary of Morphology*. Washington, D.C.: Georgetown University Press.

Steven Bird and Edward Loper. 2004. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Joseph Conrad. 1899/2006. *Heart of Darkness*. Project Gutenberg, https://www.gutenberg.org/ebooks/219.

Ryan Cotterell. 2016. Canonical segmentation data. `https://github.com/ryancotterell/canonical-segmentation/tree/master/english`. [Online; accessed 14-November-2021].

Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016a. Morphological Segmentation Inside-Out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330.

Ryan Cotterell and Hinrich Schütze. 2018. Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association for Computational Linguistics*, 6:33–48.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016b. A Joint Model of Orthography and Morphological Segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Timothy Dozat. 2016. Incorporating Nesterov Momentum into Adam. *ICLR Workshop*.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Ahmed El-Kishky, Frank F. Xu, Aston Zhang, and Jiawei Han. 2019. Parsimonious Morpheme Segmentation with an Application to Enriching Word Embeddings. *2019 IEEE International Conference on Big Data (Big Data)*, pages 64–73.

W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor*, 5(2):7.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian Error Linear Units (GELUs). *arXiv preprint arXiv:1606.08415*.

Charles F. Hockett. 1958. *A course in Modern Linguistics*. New York: The Macmillan Company.

Valentin Hofmann, Janet Pierrehumbert, and Hinrich Schütze. 2021. Superbizarre Is Not Superb: Derivational Morphology Improves BERT's Interpretation of Complex Words. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3594–3608.

Valentin Hofmann, Janet B Pierrehumbert, and Hinrich Schütze. 2020. DagoBERT: Generating Derivational Morphology with a Pretrained Language Model. *arXiv preprint arXiv:2005.00672*.

Grzegorz Jurdzinski. 2017. Word Embeddings for Morphologically Complex Languages. *Schedae Informaticae*, 25.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural Morphological Analysis: Encoding-Decoding Canonical Segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lukáš Kyjánek, Zdenek Zabokrtsky, Magda Sevcikova, and Jonáš Vidra. 2020. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, 115:5–30.

Mateusz Lango, Zdenek Zabokrtsky, and Magda Sevcikova. 2021. Semi-automatic construction of word-formation networks. *Language Resources and Evaluation*, 55.

Ling Liu. 2021. Computational Morphology with Neural Network Approaches. *arXiv e-prints*, pages arXiv–2105.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv preprint arXiv:1908.03265*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the Low-resource Challenge for Canonical Segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250.

Hugo Mailhot, Maximiliano Wilson, Joël Macoir, Hélène Deacon, and Claudia Sánchez Gutiérrez. 2019. Morpholex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, 52.

Tumi Moeng, Sheldon Reay, Aaron Daniels, and Jan Buys. 2022. Canonical and Surface Morphological Segmentation for Nguni Languages. In *Artificial Intelligence Research*, pages 125–139, Cham. Springer International Publishing.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. 2021. Do Transformer Modifications Transfer Across Implementations and Applications? *arXiv preprint arXiv:2102.11972*.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for Activation Functions. *arXiv preprint arXiv:1710.05941*.

Sunita Sarawagi and William W Cohen. 2004. Semi-Markov Conditional Random Fields for Information Extraction. *Advances in Neural Information Processing Systems*, 17.

David So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. 2021. Searching for Efficient Transformers for Language Modeling. *Advances in Neural Information Processing Systems*, 34:6010–6022.

Yuncheng Song, Shuaifei Song, Juncheng Ge, Menghan Zhang, and Wei Yang. 2020. Incorporating Morphological Compostions with Transformer to Improve BERT. *Journal of Physics: Conference Series*, 1486:072071.

Alexey Sorokin and Anastasia Kravtsova. 2018. *Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language: 7th International Conference, AINL 2018, St. Petersburg, Russia, October 17–19, 2018, Proceedings*, pages 3–10.

Claudia Sánchez Gutiérrez, Hugo Mailhot, Hélène Deacon, and Maximiliano Wilson. 2017. MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, http://link.springer.com/article/10.3758/s13428-017-0981-8:1–13.

Antal Van den Bosch and Walter Daelemans. 1999. Memory-Based Morphological Analysis. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 285–292.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Daniil Vodolazsky. 2020. DerivBase.Ru: a Derivational Morphology Resource for Russian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3937–3943, Marseille, France. European Language Resources Association.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the Transformer to Character-level Transduction. In *EACL*.

Matthew D Zeiler. 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv preprint arXiv:1212.5701*.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1201–1211.