WikiMorph: Learning to Decompose Words into Morphological Structures

Jeffrey T. Yarbro, Andrew M. Olney

[1] University of Memphis, Memphis, TN 38152, USA
{jyarbro2, aolney}@memphis.edu

Abstract. This paper presents WikiMorph, a tool that automatically breaks down words into morphemes, etymological compounds (morphemes from root languages), and generates contextual definitions for each component. It comes in two flavors: a dataset and a deep-learning-based model. The dataset was extracted from Wiktionary and contains over 450k entries. We then used this dataset to train a GPT-2 model to generalize and decompose any word into morphemes and their definitions. We find that the model accurately generates complex breakdowns when given a high-quality initial definition.

Keywords: morphemes, GPT-2, Wiktionary, etymological compounds

1 Introduction

The ability to recognize the morphological structure of words and the meaning of the morphemes within that structure positively correlates with vocabulary development and reading comprehension [4, 7, 8, 27]. Unfortunately, there are not many tools designed to increase morphological awareness. While morpheme segmentation tools and datasets are available [2, 6, 18, 20, 22], these lack critical elements for learning, such as definitions for each morpheme or a sense of its etymology. We attempt to fill this void by introducing WikiMorph: a dataset and deep-learning model. The dataset was collected by extracting user-inputted morphological data from a December 2020 version of the English Wiktionary XML dump file. This dataset contains morphemes (from English and root languages), PoS tags, and contextual definitions for each morpheme. Since Wiktionary lacks morpheme entries for some words, we also train a GPT-2 model on this dataset to generalize and break down any English word.

The model receives two inputs: a word and, optionally, its definition. If a definition is not received, the model will attempt to generate a definition for the input word. From there, it autoregressively generates a word breakdown which includes morphemes and contextualized definitions. See Section 3 for results.

2 WikiMorph: Dataset & Model

Wiktionary is an online, multilingual dictionary sponsored by the Wikimedia Foundation that contains a wide variety of information useful for NLP tasks. For this paper, we are primarily interested in the definition and etymology sections of Wiktionary. The

etymology section is of particular importance since it often contains annotated morphological segmentations for words. These segmentations can either be in English or from root languages such as Latin or Ancient Greek. We will refer to morphemes from root languages as etymological compounds throughout this paper. These compounds are useful since they give additional insights into English words and often allow further morpheme segmentation within the root language.

Extracting data from Wiktionary comes with many challenges. Most notably, standardization. Wiktionary was primarily designed to allow for flexible formatting to make it easy for authors across the web to contribute. This flexibility makes it essential first to regularize the formatting of Wiktionary. We do this by looping through the XML file and applying many regular expressions. These regular expressions aim to remove markup codes and allow our morpheme extraction algorithm to grab all relevant data.

Wiktionary does not require authors to input morpheme segmentations when a word falls under a common rule. Meaning that some affixes are regularly void of morphological entries, and therefore, unacceptable for this work. Most of these missing affixes are suffixes that change the grammatical context of the word. (e.g., making dog plural by adding -s). To combat this, we created a list of common suffixes that did not have regular entries in Wiktionary and used a series of heuristics to find the root morpheme. We then check a word corpus to see whether the root morpheme is an actual word and use DistillBERT word embeddings to see whether it is similar to the base word.

To extract morphemes, we deploy a recursive methodology. This methodology first attempts to find English morphemes within the Etymology section of Wiktionary. If found, we proceed to search Witkionary's entry for each of these found morphemes to see whether they too contain annotated morpheme segmentations. We repeat this process until the word cannot be broken down further in English. We then perform a similar lookup in root languages we deemed as "good" for each English morpheme. With "good" in this context meaning that we found examples where the language gave additional insights not seen in the English breakdown alone. If multiple etymological breakdowns were found, we chose only one with two criteria in mind. (1) Does the compound have a complete Wiktionary entry? (2) How insightful is the root language for English words? While rankings varied based on criteria 1, the system typically prefers Latin and Ancient Greek compounds since they are well-represented in Wiktionary and many morphologically complex words are derived from them.

Words often have different meanings depending upon the context. The same is true for morphemes within a word. We account for this by choosing the best definition entry for each morpheme using word embeddings from two models: DistillBERT and Spacy's Core model [10, 21]. We perform two operations for each morpheme definition. (1) Definition Similarity: Cosine similarity between the base word and morpheme's definition. (2) Addition Similarity: Adds word vectors from other morphemes within the base word to the current morpheme's definition vector, then takes the cosine similarity between the new vector and the base word's definition vector. We then perform a weighted average operation over the values and choose the definition with the highest average.

Since Wiktionary does not guarantee that word entries have a complete morpheme breakdown, it is necessary to filter out any of our extractions containing incomplete breakdowns. We do this by looping through the extracted morphemes and using a series

of heuristics on each root morpheme to ensure completeness. These heuristics consist of the following checks: (1) Checks the number of syllables within the word [1]. (2) Checks the word frequency [19]. (3) Checks the number of etymological compounds. (4) Checks to see whether there are any common affixes.

We then train the WikiMorph model by extending the large variant of GPT-2 made available by Hugging Face [24]. GPT-2 is an autoregressive model that uses the decoding blocks of the transformer architecture [11, 23]. It contains 36 decoding blocks with 774M parameters. We use 16-bit precision for lesser memory requirements and greater training speed. We then fine-tune the pretrained GPT-2 model for three epochs.

To assess the model, we removed 1500 samples from the dataset prior to training. We then perform an ablation test on these samples to see how the model performs when it receives an input definition vs. when it does not. For both conditions, the aim is to test how well the model segments morphemes and its ability to generate contextualized definitions. To test its segmentation ability, we use accuracy and character-level ROUGE1 as a sanity check to ensure that the model did not produce wildly different morphemes [26]. To evaluate how well the model generates contextualized definitions, we use word-level ROUGE1 (with stemming) and cosine similarity between the generated definition and ground-truth definition using RoBERTa word embeddings [15]. For definition evaluation, the metrics are only performed when both the generated and ground-truth sample have an instance of the same morpheme to ensure alignment.

3 Results & Discussion

Table 1. Results showing model performance and differences between when the model receives an input definition (+) vs. when it does not (-).

		English Morphemes		**Etymology Morphemes**	
	Metrics	+ Def	- Def	+ Def	- Def
Morpheme Segmentation	Accuracy	0.925	0.887	0.890	0.854
	Character ROUGE1	0.992	0.985	0.572	0.559
Definition generation	ROUGE1 Score	0.808	0.528	0.945	0.931
	RoBERTa Sim.	0.754	0.421	0.933	0.913

The results in Table 1 show that the model performed well at segmenting morphemes with over 85% accuracy for both English and etymology. For English morphemes, the model also did well at matching the characters within the ground-truth's segmentation, demonstrating that it is unlikely to give wildly different results even when the segmentations are different. The only notable differences in morpheme characters came in examples such as the ones shown in Figure 1. Here the model adds characters to properly form the root morpheme "perceive".

It is important to note that some differences between the sample and generated breakdowns are not errors. As stated in section 2, Wiktionary does not guarantee a full morphological breakdown for word entries, leading to some entries within our dataset not having a complete segmentation. However, since morphemes are repeatable in words (e.g., a- appears in arise and amoral), there is a probability that the morpheme is

correctly segmented in other entries of the dataset—allowing the model to pick up on this probability and act as a denoiser, as shown in Figure 1B.

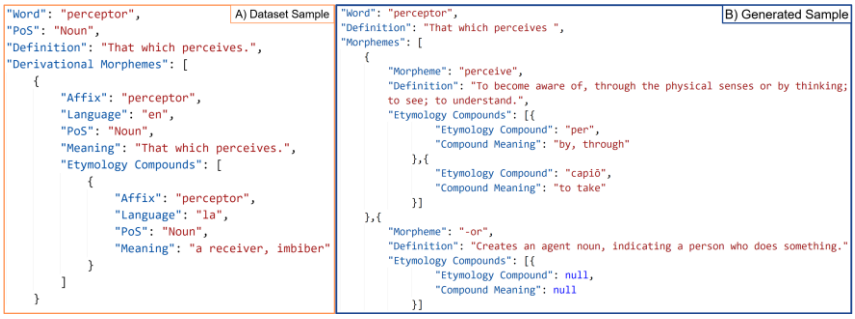

Fig. 1. A) Real sample from WikiMorph dataset. B) Generated output when given a definition.

The model showed considerable improvements while generating English definitions when given an input definition. This result is not surprising. When the model receives a definition, it has a good initialization—allowing it to pay attention and reference that good definition during the generation of each morpheme's definition. If the model does not receive an input definition, it has no context about the word aside from what it might have learned during training. Without this context, it can hallucinate while generating the initial definition, cascading additional errors across subsequent morphemes.

Interestingly, while the English contextualized definitions were significantly worse when the model did not receive a definition, the generated definitions for etymological compounds only saw a slight degradation. We speculate the reasons for this are due to three reasons. (1) The definitions are often much shorter for root languages than in English, thereby decreasing the probability that the model makes an error on an early token leading to subsequent errors. (2) There are fewer definition entries for each etymological compound. (3) These affixes frequently appear throughout many different words in our dataset, giving the model many opportunities to memorize the result.

4 Conclusion

This paper presents WikiMorph, a novel dataset and GPT-2-based model designed to help students learn morphology. The dataset extracted is one of the largest morpheme datasets to date and the only large-scale dataset containing contextualized definitions and etymological compounds. The trained WikiMorph model displayed an impressive ability to generate word breakdowns; however, further evaluation is required to determine its effectiveness in learning environments.

Acknowledgments

This material is based upon work supported by the National Science Foundation (1918751, 1934745) the Institute of Education Sciences (R305A190448)

References

1. Ash, Steve. 2018. "Jg2p." https://github.com/steveash/jg2p.
2. Balota, David A., Melvtn J. Yap, Michael J. Cortese, Keith A. Hutchison, 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): 445–59. https://doi.org/10.3758/BF03193014.
3. Blais, Caroline, Daniel Fiset, Martin Arguin, Pierre Jolicoeur, Daniel Bub, and Frédéric Gosselin. 2009. "Reading between Eye Saccades." *PLoS ONE* 4 (7). https://doi.org/10.1371/journal.pone.0006448.
4. Bowers, Peter N., John R. Kirby, and S. Hélène Deacon. 2010. "The Effects of Morphological Instruction on Literacy Skills: A Systematic Review of the Literature." *Review of Educational Research* 80 (2): 144–79. https://doi.org/10.3102/0034654309359353.
5. Burani, Cristina, Stefania Marcolini, Maria De Luca, and Pierluigi Zoccolotti. 2008. "Morpheme-Based Reading Aloud: Evidence from Dyslexic and Skilled Italian Readers." *Cognition* 108 (1): 243–62. https://doi.org/10.1016/j.cognition.2007.12.010.
6. Creutz, Mathias and Lagus, K. (2002). "Unsupervised Discovery of Morphemes." *Proceedings of the {ACL}-02 Workshop on Morphological and Phonological Learning*, 21–30. https://doi.org/10.3115/1118647.1118650
7. Duncan, Lynne G. 2018. "Language and Reading: The Role of Morpheme and Phoneme Awareness." *Current Developmental Disorders Reports* 5 (4): 226–34. https://doi.org/10.1007/s40474-018-0153-2.
8. Goodwin, Amanda P., and Soyeon Ahn. 2010. "A Meta-Analysis of Morphological Interventions: Effects on Literacy Achievement of Children with Literacy Difficulties." *Annals of Dyslexia* 60 (2): 183–208. https://doi.org/10.1007/s11881-010-0041-x.
9. Gwilliams, Laura. 2020. "How the Brain Composes Morphemes into Meaning." *Philosophical Transactions of the Royal Society B: Biological Sciences* 375 (1791). https://doi.org/10.1098/rstb.2019.0311.
10. Honnibal, Matthew and Montani, Ines and Van Landeghem, Sofie and Boyd, Adriane. 2021. "SpaCy: Industrial-Strength Natural Language Processing in Python." Zenodo. https://doi.org/10.5281/zenodo.1212303.
11. Hoppe, Sabrina, and Marc Toussaint. 2020. "Qgraph-Bounded q-Learning: Stabilizing Model-Free off-Policy Deep Reinforcement Learning." *ArXiv*.
12. Kirov, Christo, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. "Very-Large Scale Parsing and Normalization of Wiktionary Morphological Paradigms." *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, 3121–26.
13. Krizhanovsky, A. A. 2010. "Transformation of Wiktionary Entry Structure into Tables and Relations in a Relational Database Schema," 10. http://arxiv.org/abs/1011.1368.
14. Bensoussan, M., & Laufer, B. 1984. "Lexical guessing in context in EFL reading comprehension." Journal of Research in Reading, *7*(1), 15–31. https://doi.org/10.1111/j.1467-9817.1984.tb00252.x
15. Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *ArXiv*, no. 1.
16. Luong, Minh-thang, and Christopher D Manning. 2003. "Better Word Representations with Recursive Neural Networks for Morphology Minh-Thang." *CoNLL-2013*, 104–13.

17. Metheniti, Eleni, Guenter Neumann, and Josef van Genabith. 2020. "Linguistically Inspired Morphological Inflection with a Sequence-to-Sequence Model." http://arxiv.org/abs/2009.02073.
18. Metheniti, Eleni, and Günter Neumann. n.d. "Wikinflection : Massive Semi-Supervised Generation of Multilingual Inflectional Corpus from Wiktionary," no. Tlt 2018: 147–61.
19. Robyn Speer, Joshua Chin And, Andrew Lin And, Sara Jewett And, and Lance Nathan. 2018. "LuminosoInsight/Wordfreq." https://doi.org/10.5281/zenodo.1443582.
20. Sánchez-Gutiérrez, Claudia H., Hugo Mailhot, S. Hélène Deacon, and Maximiliano A. Wilson. 2018. "MorphoLex: A Derivational Morphological Database for 70,000 English Words." *Behavior Research Methods* 50 (4): 1568–80. https://doi.org/10.3758/s13428-017-0981-8.
21. Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *ArXiv*, 2–6.
22. Smit, Peter, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2015. "Morfessor 2.0: Toolkit for Statistical Morphological Segmentation," 21–24. https://doi.org/10.3115/v1/e14-2006.
23. Vaswani, Ashish, 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* 2017-Decem (Nips): 5999–6009.
24. Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, et al. 2019. "Transformers: State-of-the-Art Natural Language Processing." *ArXiv*. https://doi.org/10.18653/v1/2020.emnlp-demos.6.
25. Zhu, Yi, Ivan Vuli, and Anna Korhonen. 2013. "For Learning Word Representations."
26. Lin, Chin-Yew. 2004. ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.
27. Hayashi, Yuko, and Victoria Murphy. 2011. "An Investigation of Morphological Awareness in Japanese Learners of English." *Language Learning Journal* 39 (1): 105–20. https://doi.org/10.1080/09571731003663614.