Word-level Morpheme segmentation using Transformer neural network

Tsolmon Zundui and **Chinbat Avaajargal**
National University of Mongolia
tsolmonz@num.edu.mn, chinbat.carl@gmail.com

Abstract

This paper presents the submission of team NUM DI to the SIGMORPHON 2022 Task on Morpheme Segmentation Part 1, word-level morpheme segmentation. We explore the transformer neural network approach to the shared task. We develop monolingual models for world-level morpheme segmentation and focus on improving the model by using various training strategies to improve accuracy and generalization across languages.

1 Introduction

Morphological analysis is the heart of nearly all natural language processing tasks, such as sentiment analysis, machine translation, information retrieval, etc. Such natural language processing tasks become infeasible without any morphological analysis. One reason is the sparsity resulting from a high number of word forms that introduce out-of-vocabulary (OOV). Morphological segmentation is a way to deal with language sparsity by introducing the standard segments within the words rather than dealing with word forms (having multiple morphemes).

Morpheme segmentation is a type of morphological analysis in which words are divided into surface forms of morphemes, for example, successfulness = success @@ful @@ness. Automated morpheme segmentation was studied in the early years of natural language development (NLP). However, significant progress has been made in recent years in using various machine learning techniques.

Since morphemes are the smallest meaningful language units, information about the morphemic structure of words is already used in various NLP applications and additional tasks, including machine translation and recognition of semantically related words (cognates).

In this paper, we proposes a supervised method for word-level morphological segmentation using a transformer neural network. The task of machine translation has seen significant progress in recent times with the advent of Transformer-based models (Vaswani et al., 2017) for this year's SIGMORPHON 2022 shared task on morpheme segemntation (Batsuren et al., 2022a) which at the word level, participants will be asked to segment a given word into a sequence of morphemes. Input words contains all types of word forms: root words, derived words, inflected words, and compound words. However, to the best of our knowledge, there has not been work that applies such morpheme segmentation transformer-based models.

The paper is organized as follows: Section 2 addresses the related work on supervised morpheme segmentation, Section 3 describes the data used in training, Section 5 describes the model architecture, and section 6 presents the experiment results.

2 Related work

Z. Harris in (Harris, 1970) proposed the earliest method of morpheme segmentation. It detects morpheme boundaries by letter variety statistics (LVS) (Çöltekin, 2010). Morfessor system (Creutz and Lagus, 2007), (Smit et al., 2014) exploits unsupervised machine learning methods to be trained on a large unlabelled text. Another kind of semi-supervised machine learning for morpheme segmentation (Ruokolainen et al., 2014) was based on conditional random fields; the task was considered as sequential classifying and labeling letters of a given the word. A pure supervised method with significantly better quality for the twofold task of morpheme segmentation with classification was proposed in (Sorokin and Kravtsova, 2018); it was effective due to applying a convolutional neural network and training on the representative labeled data. The model outperforms all previous morpheme segmentation models, giving F-measure up

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 139 - 143

to 98% on morpheme boundaries. Recent works developed two more supervised machine learning models for morpheme segmentation with classification for Russian words (Bolshakova and Sapin, 2019a), (Bolshakova and Sapin, 2019b). The first is based on decision trees with gradient boosting, while the second applies Bi-LSTM neural network. However, they were developed for morpheme segmentation applied CNN, Bi-LSTM, not applied transformer neural network. Therefore, to study possible ways to build a more broad supervised model with a transformer neural network.

3 Data

A dataset for this task, the organizer integrated all basic types of morphological databases (including UniMorph (Kirov et al., 2018; McCarthy et al., 2020; Batsuren et al., 2022b) – inflectional morphology; MorphyNet (Batsuren et al., 2021) – derivational morphology; Universal Dependencies (Nivre et al., 2017) and ten editions of Wiktionary – compound morphology and root words) cover 9 languages. 8 of these languages were available initially, while 1 surprise language, Mongolia, was released one week before the submission deadline. Each language had split a train and a development sample. The amount of data for the different languages vary in size, from 18966 (Mongolian) to 926098 (Hungarian). Each sample occupies a single line and consists of input word, the corresponding morpheme sequence, and the corresponding morphological category. Except for Spanish, eight languages have morphological word categories shown in table 1. All the data is available on the Github[1] page.

(1) **Example Training Set**

```
pentazole penta @@azo @@ole 010
nyala nyala 000
biots biot @@s 100
```

(2) **Example Development Set**

```
newspaper new @@s @@paper 011
players play @@er @@s 110
congruity congruent @@ity 010
```

(3) **Example Test Set**

```
hyperonym
distance
```

[1]https://github.com/sigmorphon/2022SegmentationST

To preprocess the dataset, we used the fairseq command-line tool to binarize the training data, making it easy for developers and researchers to directly run operations from the terminal.

4 Model architecture

We use the character level Transformer implementation of *fairseq* (Ott et al., 2019). Our model is composed of one encoder input word, and one decoder output segmentation of the word. We train a monolingual word segmentation model for each given language with identical parameters, 50 epochs, 1 encoder layer, 1 decoder layer, 0.0001 learning rate, using the Adam optimizer (Kingma and Ba, 2014) and the cross-entropy loss. Various hyperparameters of our Transformer model were experimentally tested in several experiments. The resulted model has the encoder and decoder layer with 128 hidden units, and the batch size is 32. Encoder and decoder more layers slightly improve the quality (less than 0.5%), but the model became too heavy both for training and evaluation. We also use created checkpoints to save the checkpoint the latest and the best ones. It is also a safe guard in case the training gets disrupted due to some unforeseen issue.

4.1 Evaluation

For the word-level segmentation shared task, the following evaluation metrics are provided.

- Precision: fraction of correctly predicted morphemes on all predicted morphemes
- Recall: ratio of correctly predicted morphemes on all gold morphemes
- F-measure: the harmonic mean of the precision and recall
- Edit distance: average Levenshtein distance between the predicted output and the gold instance.

We compare our results with the baseline model, in which the multilingual Bert tokenizer is shown in table 2.

5 Results

Results of the evaluation are shown in Table 2, where the leftmost column stands for the ISO-639 language code, the next one for the number of train data, the next one for the number of test data, rest

Word class	Description	Example
000	Root words	Vivian - Vivian
001	Compound only	snowfight - snow @@fight
010	Derivation only	unafraid - un @@afraid
011	Derivation and Compound	peacekeeper - peace @@keep @@er
100	Inflection only	descendents - descendent @@s
101	Inflection and Compound	setbacks - set @@back @@s
110	Inflection and Derivation	brandishing - brand @@ish @@ing
111	Inflection, Derivation, Compound	faultfinders - fault @@find @@er @@s

Table 1: Word categories.

Lang.	Train size	Test size	Models	Precision	Recall	F-measure	Distance
eng	458692	57755	Transformer	84.02	83.12	83.56	0.48
			Baseline	20.99	28.79	24.28	2.69
ces	30694	4000	Transformer	88.49	87.52	88.00	0.35
			Baseline	22.10	19.72	20.84	2.94
fra	252671	31588	Transformer	87.48	84.14	85.78	0.72
			baseline	11.08	14.00	12.37	4.32
hun	742239	95278	Transformer	96.33	95.50	95.91	0.21
			baseline	20.88	27.81	23.85	3.54
ita	369208	46153	Transformer	90.38	88.74	89.55	0.58
			baseline	8.12	10.54	9.18	5.35
lat	705862	88234	Transformer	97.03	95.68	96.35	0.08
			baseline	6.76	13.17	8.94	4.14
mon	15171	1900	Transformer	87.99	83.32	85.59	0.58
			baseline	5.89	10.59	7.57	4.51
rus	627367	78425	Transformer	95.6	93.42	94.5	0.46
			baseline	13.23	14.13	13.67	7.62
spa	688673	86088	Transformer	96.33	94.33	95.32	0.29
			baseline	15.76	17.91	16.76	5.20

Table 2: Comparison of our model and baseline for morpheme segmentation

of the columns stand for the evaluation metrics provided by shared task. It is clearly seen that our model performs much better in all evaluation metrics than the baseline model. We expected rich morphological language models to get lower scores than others. However, the results show that the English word segmentation model has a lower recall, precision, and f-measure scores than other language models; even Mongolian has fewest training data. In all metrics, the Latin word segmentation model had the highest score. All models trained on more than 60,000 training data have more than 90 points in the recall, precision, and f-measure score. In table 3, we compare the f-measure score of our model with team DeepSPIN-3 (Peters and Martins, 2022). Although our model performed poorly in all languages, it performed competitively.

6 Conclusion

We have presented the monolingual models for morpheme segmentation in 9 languages. Our model run outperforms the baseline. Even though our models as implemented prior to submission failed to attain reasonable evaluations scores on the word-level morpheme segmentation task, our results indicate that our model has the potential to have a better performance after fine-tuning and the good performance of our model under varying morphological complexity languages.

In future work, we plan on exploring multilingual word-level morpheme segmentation a model.

Language	Teams	F-measure
eng	NUM DI	83.56
	DeepSPIN-3	**93.63**
ces	NUM DI	88.0
	DeepSPIN-3	**93.84**
fra	NUM DI	85.78
	DeepSPIN-3	**95.73**
hun	NUM DI	95.91
	DeepSPIN-3	**98.72**
ita	NUM DI	89.55
	DeepSPIN-3	**97.43**
lat	NUM DI	96.35
	DeepSPIN-3	**99.38**
mon	NUM DI	85.59
	DeepSPIN-3	**98.51**
rus	NUM DI	94.5
	DeepSPIN-3	**99.35**
spa	NUM DI	95.32
	DeepSPIN-3	**99.04**

Table 3: Comparison of our model and model of the best team for word-level morpheme segmentation

Author contribution

All authors equally contributed to this work.

Acknowledgement

Thanks to Chinbat Avaajargal for dedicated work on this task. Thanks to several anonymous reviewers for their constructive feedback.

References

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.

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinović, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022a. The sigmorphon 2022 shared task on morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina J. Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022b. Unimorph 4.0: Universal morphology.

Elena Bolshakova and Alexander Sapin. 2019a. Bi-lstm model for morpheme segmentation of russian words. In *Conference on Artificial Intelligence and Natural Language*, pages 151–160. Springer.

Elena Bolshakova and Alexander Sapin. 2019b. Comparing models of morpheme analysis for russian words based on machine learning. In *Proc. of the International Conference Dialogue*, volume 2019, pages 104–113.

Çağrı Çöltekin. 2010. Improving successor variety for morphological segmentation. *LOT Occasional Series*, 16:13–28.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

Zellig S Harris. 1970. Morpheme boundaries within words: Report on a computer test. In *Papers in Structural and Transformational Linguistics*, pages 68–77. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy,

Sandra Kübler, et al. 2018. Unimorph 2.0: universal morphology. *arXiv preprint arXiv:1810.11101*.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. Unimorph 3.0: Universal morphology.

Joakim Nivre, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, et al. 2017. Universal dependencies 2.1.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Ben Peters and André F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 84–89.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of russian language. In *Conference on Artificial Intelligence and Natural Language*, pages 3–10. Springer.

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.