One Size Fits Few: Finding the Optimal Subword Sizes for FastText Models across Languages

An article for the SIGIR 2021 Doctoral Consortium

Vítek Novotný
witiko@mail.muni.cz

Faculty of Informatics, Masaryk University

March 11, 2021

The article

- FastText *subword embeddings* are useful for many NLP tasks:
  1. word similarity, [1, 2]
  2. word analogy, [1]
  3. language modeling, [1]
  4. word sense disam., [3, 2]
  5. text classification, [4]
  6. semantic text similarity, [5]
  7. information retrieval, [6]
  8. and others … [7]
- *Subword sizes* are a crucial hyperparameter, but often not optimized to save time [1, 8].
- In our work, we:
  1. find the optimal subword sizes on English, German, Czech, and Italian word analogy tasks,
  2. Show a 5% improvement on the Czech word analogy task with optimal subword sizes.
  3. describe the *optimal subword coverage* statistic and estimate its population mean,
  4. show that optimal subword coverage predicts optimal subword sizes on Spanish (+0.60%), French (+0.95%), Hindi (+2.16%), and Turkish (+0.30%) word analogy tasks.
  5. predict optimal subword sizes for languages with word boundaries,
  6. publish a software package for predicting the optimal subword sizes for a language.
- Since our preprint [9], point 4 has been added to our experiment.

The reviews

TSD 2020, Review 1

Main contributions The paper deals with word embeddings. Authors try to find the optimal size of n-gram characters for four languages.

Positive aspects The paper shows a relation with n-gram coverage and an optimal n-gram size. This method can be (probably) used to quickly estimate the optimal n-gram sizes for additional languages.

Negative aspects [. . .] The biggest problem of the paper is that all the results are *evaluated only on a synthetic task of word analogies*. The results must be confirmed on several downstream tasks. The same is valid for the hypothesis that the n-gram coverage may estimate the optimal n-gram size.

Further comments [. . .] I am not sure if such a paper is fit for a scientific conference without any further in-depth analysis of the results.

The reviews

TSD 2020, Review 2

Main contributions Authors try to obtain the optimal size of subword n-grams for the fastText word embeddings algorithm and devise a simple, computationally inexpensive method of estimating the size of the subwords.

Positive aspects […] The presented method of estimating the optimal n-gram sizes is useful especially if one wants to create embeddings in bulk (e.g. from many languages) or play with other parameters (as opposed to e.g. a grid search). The improvement of 5% is already noticeable. The article can be a good starting point for further work on optimization of fastText hyperparameters.

Negative aspects I am *not quite convinced* that n-gram coverage correlates with the optimal n-gram sizes (though I can think of some arguments why it holds across languages, at least roughly). […] *Optimization is performed on synthetic benchmarks*, not real-life tasks.

Further comments none

The reviews

TSD 2020, Review 3

Main contributions The paper studies the optimal size of fastText ngrams for several languages in the analogy task

Positive aspects The idea of the paper is that the n-gram coverage correspond to optimal sizes of ngrams for using fasttext in nlp tasks. This gives the possibility to find optimal size of ngrams without extensive calculations, which is very important for improving the performance of applications in NLP tasks.

Negative aspects The optimal hyperparameters were *studied only for one task: the analogy task*. The results can change in different tasks.

Further comments no

The reviews

*SEM 2020, Reviewer Recommendations

> *The author(s) never compared their performance (trained without optimized) across different word analogy categories.*
>
> *[…] the embeddings could have been tested also on the word similarity task as in Bojanowski et al. 2017, that also includes languages other than the ones taken into consideration by the authors.*
>
> *The experimental setup could also benefit from highlighting more clearly the results of the original embeddings […]*

The venue

- I am planning to submit our article to the SIGIR 2020 Doctoral Consortium:

Doctoral Consortium papers due Tue, Mar 16, 2021
Doctoral Consortium papers notification Mon, Apr 19, 2021

- The paper should be in a long paper format with a one-page appendix:

> *To apply for a place at the Doctoral Consortium, candidates should submit a paper up to 5+1 pages in length, solely authored by the student [...]*

The submitted paper, solely authored by the student, will be the basis for detailed discussions at the Consortium. To get the most out of the discussion, it should include:

- Motivation for the proposed research.
- Background and related work (including key references).
- Description of proposed research, including main research questions.
- Research methodology and proposed experiments (where appropriate).
- Specific research issues for discussion at the Doctoral Consortium.

Thank you for your attention!

Bibliography I

[1] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146. URL: `https://www.aclweb.org/anthology/Q17-1010.pdf` (visited on 03/10/2021).

[2] Eniafe Festus Ayetiran, Petr Sojka, and Vít Novotný. “EDS-MEMBED: Multi-sense embeddings based on enhanced distributional semantic structures via a graph walk over word senses”. In: *Knowledge-Based Systems* (2021), p. 106902. DOI: `10.1016/j.knosys.2021.106902`.

Bibliography II

[3] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. “A unified model for word sense representation and disambiguation”. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1025–1035. URL: `https://www.aclweb.org/anthology/D14-1110.pdf` (visited on 03/10/2021).

[4] Matt Kusner et al. “From word embeddings to document distances”. In: *International conference on machine learning*. PMLR. 2015, pp. 957–966. URL: `http://proceedings.mlr.press/v37/kusnerb15.pdf` (visited on 03/10/2021).

Bibliography III

[5] Delphine Charlet and Geraldine Damnati. “Simbow at SemEval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering”. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. 2017, pp. 315–319. URL: https://www.aclweb.org/anthology/S17-2051.pdf (visited on 03/10/2021).

[6] Vít Novotný et al. “Three is Better than One. Ensembling Math Information Retrieval Systems”. In: *CEUR Workshop Proceedings*. Thessaloniki, Greece, 2020, p. 30. URL: http://ceur-ws.org/Vol-2696/paper_235.pdf (visited on 03/10/2021).

[7] Amir Bakarov. “A survey of word embeddings evaluation methods”. In: *arXiv preprint arXiv:1801.09536v1* (2018). URL: https://arxiv.org/pdf/1801.09536v1.pdf (visited on 03/10/2021).

Bibliography IV

[8] Edouard Grave et al. “Learning word vectors for 157 languages”. In: *arXiv preprint arXiv:1802.06893v2* (2018). URL: `https://arxiv.org/pdf/1802.06893v2.pdf` (visited on 03/10/2021).

[9] Vít Novotný et al. “One Size Does Not Fit All: Finding the Optimal N-gram Sizes for FastText Models across Languages”. In: *arXiv preprint arXiv:2102.02585v1* (2021). URL: `https://arxiv.org/abs/2102.02585v1` (visited on 03/10/2021).

MUNI

FACULTY

OF INFORMATICS