Exploring Swedish & English fastText Embeddings

Tosin Adewumi*, Foteini Liwicki and Marcus Liwicki

ML Group, Luleå University of Technology, Sweden.

Abstract

In this paper, we show that embeddings from relatively smaller corpora sometimes outperform those from larger corpora and we introduce a new Swedish analogy test set and make it publicly available. To achieve good performance in Natural Language Processing (NLP) downstream tasks, several factors play important roles: dataset size, the right hyper-parameters, and well-trained embeddings. We utilize the fastText tool for our experiments. We evaluate both the Swedish and English embeddings that we created using intrinsic evaluation (including analogy & Spearman correlation) and compare them with 2 common, publicly available embeddings. Our English continuous Bag-of-Words (CBoW)-negative sampling embedding shows better performance compared to the publicly available GoogleNews version. We also describe the relationship between NLP and cognitive science. We contribute the embeddings for research or other useful purposes by publicly releasing them.

Keywords

Embeddings, fastText, Analogy set, Swedish

1. Introduction

The embedding layer of neural networks may be initialized randomly or replaced with pre-trained vectors, which act as lookup tables. One of such pre-trained vector tools include fastText, introduced by [1]. The main advantages of fastText are speed and competitive performance to state-of-the-art (SoTA) models. Using pre-trained embeddings in deep networks like the Transformer can improve performance.

Despite the plethora of embeddings in many languages, there's a dearth of analogy test sets to evaluate many of them, including for Swedish [2, 3, 4, 5]. This is because creating labelled or structured datasets can be expensive in terms of time and attention required. [6] created 157 different language embeddings but provided analogy test sets for only 3 languages: French, Hindi and Polish [6]. An analogy test set, such as the one introduced by [7], provides some inclination as to the quality and likely performance of word embeddings in NLP downstream tasks, such as Named Entity Recognition (NER). The analogy evaluation involves prediction of the second value of a pair of two similar words.

Although much research have been putting emphasis on deep, contextual models for generating embeddings, it has recently been shown that the 'non-contextual' embeddings, like word2vec, are still useful in achieving SoTA results in text classification tasks [8]. The two training algorithms (hierarchical softmax and negative sampling) of the two architectures (Skipgram

AIC 2022, Sweden

*Corresponding author.

tosin.adewumi@ltu.se (T. Adewumi); foteini.liwicki@ltu.se (F. Liwicki); marcus.liwicki@ltu.se (M. Liwicki)

and CBoW) of fastText are used in this work. The key contributions of this work are (i) the new Swedish analogy test set publicly made available[1] for the NLP research community and (ii) the produced, relatively optimal English and Swedish embeddings. The quality of the Swedish model by [6] is evaluated for the first time. The embedding hyper-parameters are based on previous research, which used grid search to determine optimal hyper-parameters [9]. The rest of this paper is organised as follows: a brief survey of related work, the methodology used, results and discussion, and the conclusion.

2. Related Work

Distributed representation of words has been in use for some time [10]. fastText, based on the original distributed representation by [7], contains two architectures. The CBoW architecture averages word vectors, fed into a linear classifier, into text representation while the skipgram uses bag of character n-grams for represented words by summing them [11, 1]. The use of subword representations has proven to be helpful when dealing with out-of-vocabulary (OOV) words. Indeed, [12] used word embeddings to guide the parsing of OOV words in their work on meaning representation for robots.

WordSimilarity-353 (WordSim) test set is another analysis tool for word vectors [13]. It is based on human expert-assigned semantic similarity on two sets of English word pairs. It is used to measure an embedding's intrinsic quality. Despite the weaknesses of intrinsic tools (such as weak correlation with downstream evaluation), they have been shown to reveal somewhat meaningful relationships among words in embeddings [7, 14]. It is misleading to assume such intrinsic tests are sufficient in themselves, just as it is misleading to assume one particular extrinsic test is sufficient to generalise the performance of embeddings on all NLP tasks [15, 16, 9]. For Swedish, a common evaluation resource for words is SALDO [17], which is a lexical-semantic resource that links words by their associations. SALDO extends SAL (Svenskt associationslexikon, a set of classified synonyms) with inflectional morphological information [17, 18]. QVEC-CCA may be used as an intrinsic evaluation metric with features from a language resource like SALDO [19, 3].

[1] noted that other implementations of their fastText model could be much slower. Indeed, implementations in Python, an interpreted language, are expected to be slower and will use up more energy resources, compared to the original C++ implementation [20, 21]. The English and Swedish language models by [6] were trained on Common Crawl & Wikipedia datasets, using CBoW of 300 dimensions, with character n-grams of length 5 and window size 5. These are the embeddings we compare with in this work. Common Crawl contains petabytes of data, resulting in 630 billion words after preprocessing in a previous work [22].

Cognitive science & NLP: NLP essentially deals with semantics and knowledge, in general [23]. It overlaps with cognitive science, which mostly deals with how knowledge is acquired and processed in the brain. The overlap occurs in areas such as data or the methods used. [23]. There's been convergence of models (such as the Adaptive Control of Thought–Rational) coming from both fields. The Adaptive Control of Thought–Rational (ACT-R) explains how

[1] spraakbanken.gu.se/en/resources/analogy

the operation of the brain is modular and how the modules interact in making comprehension possible. Development in one of these two fields, therefore has the potential to benefit the other field.

3. Methodology

All the embeddings in English and Swedish were generated using the original C++ implementation [6]. They were run on a shared DGX cluster running Ubuntu 18 with 80 CPUs. Gensim (3.8.1) library was used to evaluate all models against their corresponding analogy test sets in Python (3.6.9). Some of the default hyper-parameter settings were retained [11]. All models are 300 dimensions and trained for 10 epochs. The lower and upper boundaries for the character n-gram were 3 and 6, respectively.

Both the English and Swedish training datasets used are 2019 Wikipedia dumps of 27G (4.86B words) and 4G (767M words), respectively, after pre-processing [24, 25]. They were pre-processed using the recommended script by [6]. It would have been ideal to run each training multiple times to obtain averages but because of the limited time involved, a work-around was adopted, which was to run a few random models twice to ascertain if there were major differences per model. It was established that differences were little enough to accept a single run per model. Besides, each run took hours within the range of about 2 and 36 hours and there were 32 pre-trained models to be generated: 8 English subword and no-subword (word2vec) models each and 8 Swedish subword and no-subword models each. The English embeddings are evaluated using the analogy test set and WordSim while the Swedish embeddings are evaluated using the new Swedish analogy set.

3.1. Swedish analogy test set

The Swedish analogy test set follows the format of the original Google version and many of the samples were drawn from there. The original has been observed to be slightly unbalanced, having 8,869 semantic samples and 10,675 syntactic samples (making a total of 19,544). The new Swedish set is bigger and balanced across the 2 major categories, having a total of 20,637, made up of 10,380 semantic and 10,257 syntactic samples. It is also roughly balanced across the syntactic subsections but the *capital-world* has the largest proportion of samples in the semantic subsection. This is because of the difficulty involved in obtaining world currencies in Swedish and the limited nomenclature of family members. A similar difficulty was experienced by [5], who noted that not all words in the original Google analogy test set can be directly translated to other languages, while creating a much smaller Finnish version. In all, there are 5 semantic subsections and 6 syntactic subsections. Table 1 presents further details on the test set. It was constructed, partly using the samples in the English version, with the help of tools dedicated to Swedish dictionary/translation[2] and was proof-read for corrections by two native speakers (with an inter-annotator agreement score of 98.93%). New, relevant entries were also added. Examples in the family subsection of the semantic section are:
kung drottning man kvinna

[2]https://bab.la & https://en.wiktionary.org/wiki/

pojke flicka man kvinna
bror syster morfar mormor.

Table 1
The new Swedish analogy test set details

Semantic	Syntactic
capital-common-countries (342)	gram2-opposite (2,652)
capital-world (7,832)	gram3-comparative (2,162)
currency (42)	gram4-superlative (1,980)
city-in-state (1,892)	gram6-nationality-adjective (12)
family (272)	gram7-past-tense (1,891)
	gram8-plural (1,560)

4. Results & Discussion

The WordSim result output file from the Gensim library program always has more than one value reported, including the Spearman correlation. An example output for the embedding by [6] is given below:

> *((0.6853162842820049, 2.826381331182216e-50),*
> *SpearmanrResult(correlation=0.70236817646248, pvalue=9.157442621319373e-54), 0.0)*

The first value is reported as WordSim score1 in Table 3. It is a cosine variant. Spearman correlation measures the relationship between the pairs of words in the WordSimilarity-353 dataset [13] by using a monotonic function. Intrinsic results for the pre-trained models are given in Tables 2 and 3. An important trend that can be observed is the higher scores for skipgram-negative sampling in all the cases (English & Swedish), except one. This appears to confirm previous research [7, 9]. It is noteworthy that the released, original pre-trained word2vec model was of the same combination [7]. This English word2vec embedding was trained on GoogleNews dataset of 100 billion words and represented as *'GN'* [7] while the models by [6], trained on the large Common Crawl & Wikipedia datasets, are represented by *'Gr'* in the relevant table. The English subword embeddings have 5 models with higher analogy scores than their word2vec equivalent, out of 8. The WordSim score1 and corresponding Spearman correlation for English word2vec models were higher than their corresponding subword models in all cases, except one. Comparison of the scores of the *'GN'* embeddings to our best scores for the English embeddings in Table 3 indicates that mere increase of the training data does not equate to better performance, as the choice of hyper-parameters or dataset can bring additional improvement. It may not be proper to compare the scores of the English to the Swedish models since both were based on different test sets of varying sizes.

Since [11] observed that using character n-grams led to smaller improvements for English than other languages, we expect that the scores for each language may not follow similar trends.

Table 2
Skipgram English & Swedish intrinsic scores (highest score/row in bold). H.S.: hierarchical softmax; N. S.: negative sampling

	Skipgram (s1)			
	H. S. (h1)		**N. S. (h0)**	
window (w)	**4**	**8**	**4**	**8**
Subword %				
Analogy	62.6	58.8	74.4	69.8
WordSim score1	64.8	66.3	69.9	**70**
Spearman	67.6	69.4	**74.3**	73.6
Word2Vec %				
Analogy	61.3	58.3	73.5	70.4
WordSim score1	66.3	67.3	69.6	**70.1**
Spearman	70	70.9	74.5	**74.7**
Swedish				
Subword %	45.05	39.99	53.53	53.36
Word2Vec %	45.53	41.21	**58.25**	57.30

Table 3
CBoW English & Swedish intrinsic scores (highest score/row in bold). H.S.: hierarchical softmax; N. S.: negative sampling; Gr: [6], GN: Google News [7]

	CBoW (s0)					
	H. S. (h1)		**N. S. (h0)**		**Gr [6]**	**GN [7]**
window (w)	**4**	**8**	**4**	**8**		
Subword %						
Analogy	67.2	68.7	71.6	71	**82.6**	
WordSim score1	62.6	66.2	47.3	51.1	68.5	
Spearman	65.3	70.3	45.3	49.5	70.2	
Word2Vec %						
Analogy	59.7	61.9	**76.2**	75.4		74
WordSim score1	64.1	66.7	65.4	67.5		62.4
Spearman	68.2	71.2	66.9	69.4		65.9
Swedish						
Subword %	26.5	23.93	36.79	35.89	**60.9**	
Word2Vec %	28.02	28.04	52.81	55.64		

In addition, accuracy falls for morphologically complex languages, like German, making analogy predictions difficult [26]. While working on Finnish embeddings, it was observed that fastText (subword) CBoW had lower analogy score than word2vec CBoW while fastText skipgram had higher score than word2vec skipgram, even for zero OOV words [5]. Indeed, determining the best embedding in each category requires the additional step of applying them to downstream tasks[27].

4.1. Assessment of Embedding

Qualitative assessment for a randomly selected input for the Swedish model: subword Skipgram-hierarchical softmax-window 4 (w4s1h1) is given in table 4. The nearest neighbour to syster is halvsyster with a score of 0.8688. And the result of the analogy query of rom - italien + kairo is egypten with the score of 0.4889.

Table 4
Example assessment of the Swedish Skipgram-hierarchical softmax-window 4 (w4s1h1) model

Nearest Neighbor/ Analogy Query	Result
syster	halvsyster (0.8688), systerdotter (0.8599), ...
rom - italien + kairo	egypten (0.4889), norditalien (0.4317), ...

5. Conclusion

This work presents relatively optimal fastText embeddings in Swedish and English from relatively smaller corpora. It also presents the new Swedish analogy test set for intrinsic evaluation of Swedish embeddings. The intrinsic evaluation shows the trend of better performance with skipgram-negative sampling embeddings across the two languages. Merely increasing the training dataset size alone does not equate to better performance and optimal hyper-parameters can bring additional improvement. Future work may involve using these embeddings in downstream tasks, including using CRF-based models for NER. Transfer learning with hyper-parameter tuning may bring SoTA results.

Acknowledgment

The authors wish to thank the anonymous reviewers for their valuable feedback. We are thankful to Carl Borngrund and Karl Ekström for their very useful help in proof-reading the analogy set. The work in this project is partially funded by Vinnova under the project number 2019-02996 ”Språkmodeller för svenska myndigheter”.

References

[1] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of tricks for efficient text classification, arXiv preprint arXiv:1607.01759 (2016).
[2] R. Al-Rfou, B. Perozzi, S. Skiena, Polyglot: Distributed word representations for multilingual nlp, arXiv preprint arXiv:1307.1662 (2013).
[3] P. Fallgren, J. Segeblad, M. Kuhlmann, Towards a standard dataset of swedish word vectors, in: Sixth Swedish Language Technology Conference (SLTC), Umeå 17-18 nov 2016, 2016.
[4] R. Précenth, Word embeddings and gender stereotypes in swedish and english, 2019.
[5] V. Venekoski, J. Vankka, Finnish resources for evaluating language model semantics, in: Proceedings of the 21st Nordic Conference on Computational Linguistics, 2017, pp. 231–236.

[6] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, T. Mikolov, Learning word vectors for 157 languages, arXiv preprint arXiv:1802.06893 (2018).
[7] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).
[8] A. Adhikari, A. Ram, R. Tang, J. Lin, Rethinking complex neural network architectures for document classification, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4046–4051.
[9] T. P. Adewumi, F. Liwicki, M. Liwicki, Word2vec: Optimal hyper-parameters and their impact on nlp downstream tasks, arXiv preprint arXiv:2003.11645 (2020).
[10] G. E. Hinton, et al., Learning distributed representations of concepts, in: Proceedings of the eighth annual conference of the cognitive science society, volume 1, Amherst, MA, 1986, p. 12.
[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.
[12] J. Thomason, A. Padmakumar, J. Sinapov, N. Walker, Y. Jiang, H. Yedidsion, J. Hart, P. Stone, R. Mooney, Jointly improving parsing and perception for natural language commands through human-robot dialog, Journal of Artificial Intelligence Research 67 (2020) 327–374.
[13] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, E. Ruppin, Placing search in context: The concept revisited, ACM Transactions on information systems 20 (2002) 116–131.
[14] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.
[15] A. Gatt, E. Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, Journal of Artificial Intelligence Research 61 (2018) 65–170.
[16] M. Faruqui, Y. Tsvetkov, P. Rastogi, C. Dyer, Problems with evaluation of word embeddings using word similarity tasks, arXiv preprint arXiv:1605.02276 (2016).
[17] L. Borin, M. Forsberg, L. Lönngren, Saldo: a touch of yin to wordnet's yang, Language resources and evaluation 47 (2013) 1191–1211.
[18] S. R. Eide, N. Tahmasebi, L. Borin, The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for nlp, in: Digital Humanities 2016. From Digitization to Knowledge 2016: Resources and Methods for Semantic Processing of Digital Works/Texts, Proceedings of the Workshop, July 11, 2016, Krakow, Poland, 126, Linköping University Electronic Press, 2016, pp. 8–12.
[19] Y. Tsvetkov, M. Faruqui, C. Dyer, Correlation-based intrinsic evaluation of word vector representations, arXiv preprint arXiv:1606.06710 (2016).
[20] T. P. Adewumi, Inner loop program construct: A faster way for program execution, Open Computer Science 8 (2018) 115–122.
[21] T. P. Adewumi, M. Liwicki, Inner for-loop for speeding up blockchain mining, Open Computer Science 10 (2020) 42–47.
[22] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, A. Joulin, Advances in pre-training

distributed word representations, arXiv preprint arXiv:1712.09405 (2017).
[23] T. Poibeau, S. Vasishth, Introduction: Cognitive issues in natural language processing, Language, Cognition, and Computational Models (2018). URL: https://hal.archives-ouvertes.fr/hal-01722353/file/intro-alinev2.pdf.
[24] Wikipedia, English wikipedia multistream articles (2019). URL: https://dumps.wikimedia.org/backup-index.html.
[25] Wikipedia, Swedish wikipedia multistream articles (2019). URL: https://dumps.wikimedia.org/backup-index.html.
[26] M. Köper, C. Scheible, S. S. im Walde, Multilingual reliability and "semantic" structure of continuous word spaces, in: Proceedings of the 11th international conference on computational semantics, 2015, pp. 40–45.
[27] B. Chiu, A. Korhonen, S. Pyysalo, Intrinsic evaluation of word vectors fails to predict extrinsic performance, in: Proceedings of the 1st workshop on evaluating vector-space representations for NLP, 2016, pp. 1–6.