Corpora Compared: The Case of the Swedish Gigaword & Wikipedia Corpora

Tosin P. Adewumi **Foteini Liwicki** **Marcus Liwicki**

Machine Learning group, EISLAB
Luleå University of Technology, Sweden.
`firstname.lastname@ltu.se`

Abstract

In this work, we show that the difference in performance of embeddings from differently sourced data for a given language can be due to other factors besides data size. Natural language processing (NLP) tasks usually perform better with embeddings from bigger corpora. However, broadness of covered domain and noise can play important roles. We evaluate embeddings based on two Swedish corpora: The Gigaword and Wikipedia, in analogy (intrinsic) tests and discover that the embeddings from the Wikipedia corpus generally outperform those from the Gigaword corpus, which is a bigger corpus. Downstream tests will be required to have a definite evaluation.

1 Introduction

It is generally observed that more data bring about better performance in Machine Learning (ML) tasks (Adewumi et al., 2019; Stevens et al., 2020). What may not be very clear is the behaviour of variance of homogeneity in datasets. It is always better to have a balanced or broad-based dataset or avoid an overly-represented topic within a dataset (Stevens et al., 2020). Furthermore, noise (or contamination) in data can reduce performance (Hagan et al., 1997). However, not all noise is bad. Indeed, noise may be helpful (Stevens et al., 2020).

In this work, we compare embeddings (in analogy test) from two Swedish corpora: The Gigaword and Wikipedia. The Gigaword corpus by Rødven Eide et al. (2016) contains data from different genre, covering about 7 decades since the 1950s. Meanwhile the Wikipedia is a collection of articles on many, various subjects (Wikipedia, 2019).

Word similarity or analogy tests, despite their weaknesses, have been shown to reveal somewhat meaningful relationships among words in embeddings, given the relationship among words in context (Mikolov et al., 2013; Pennington et al., 2014). It is misleading to assume such intrinsic tests are sufficient in themselves, just as it is misleading to assume one particular extrinsic (downstream) test is sufficient to generalise the performance of embeddings on all NLP tasks (Gatt and Krahmer, 2018; Faruqui et al., 2016; Adewumi et al., 2020b).

The research question being addressed in this work is: does bigger corpus size automatically mean better performance for differently-sourced Swedish corpora? The contribution this work brings is the insight into the differences in the performance of the Swedish embeddings of the Gigaword and Wikipedia corpora, despite the over 40% additional size of the Gigaword corpus. Furthermore, this work will, possibly, enable researchers seek out ways to improve the Gigaword corpus, and indeed similar corpora, if NLP downstream tasks confirm the relative better performance of embeddings from the Wikipedia corpus. The following sections include related work, methodology, results & discussion and conclusion.

2 Related Work

Rødven Eide et al. (2016) created the Swedish corpus with at least one billion words. It covers fiction, government, news, science and social media from the 1950s. The sentences of the first six lines of the content of this Gigaword corpus are:

> 1 knippa dill
> patrik andersson
> TV : Danska Sidse Babett Knudsen har prisats på tv-festivalen i Monte Carlo för rollen
> i dramaserien Borgen .
> Hon sköts med ett skott i huvudet , men tog sig fram till porten och ringde på .
> I början av juni tog hon examen från den tvååriga YH-utbildning , som hon flyt-

tade upp till huvudstaden för att gå .
Det blev kaos , folk sprang fram för att
hjälpa , någon började filma ...

The content of the Wikipedia corpus is a community effort, which began some years ago, and is edited continually. It covers far-reaching topics, including those of the Swedish Gigaword corpus, and in addition, entertainment, art, politics and more. The sentences of the first seven lines of the content of the pre-processed version of the Wikipedia corpus are given below. It would be observed that it contains a bit of English words and the pre-processing script affected non-ascii characters. However, these issues were not serious enough to adversely affect the models generated, in this case, as the embedding system seems fairly robust to handle such noise.

amager r en dansk i resund ns norra och v
stra delar tillh r k penhamn medan vriga
delar upptas av t rnby kommun och drag
rs kommun amager har en yta p nine six
two nine km och befolkningen uppg r
till one nine six zero four seven personer
one one two zero one eight en stor del
av bebyggelsen har f rortspr gel men ven
tskilliga innerstadskvarter finns i k pen-
hamn samt i drag r p den stra delen av n
finns kastrups flygplats amager r delvis
en konstgjord delvis en naturlig s dan n
r mycket l g och vissa delar ligger un-
der havsytan framf r allt det genom f rd
mning.

Adewumi et al. (2020a) created the Swedish analogy test set, which is similar to the Google analogy test set by Mikolov et al. (2013). This was because there was no existing analogy test set to evaluate Swedish embeddings (Fallgren et al., 2016; Précenth, 2019). The analogy set has two main sections and their corresponding subsections: the semantic & syntactic sections. Two native speakers proof-read the analogy set for any possible issues (with percentage agreement of 98.93% between them), after valuable comments from the reviewers of this paper. It is noteworthy that some words can have two or more possible related words. For example, based on the dictionary, the Swedish word *man* can be related to *kvinna* and *dam* in very similar ways. Four examples from the *gram2-opposite* sub-section of the syntactic section are:

medveten omedveten lycklig olycklig
medveten omedveten artig oartig
medveten omedveten härlig ohärlig
medveten omedveten bekväm obekväm

Faruqui et al. (2016) correctly suggest there are problems with word similarity tasks for intrinsic evaluation of embeddings. One of the problems is overfitting, which large datasets (like the analogy set in this work) tend to alleviate (Stevens et al., 2020). In order to have a definite evaluation of embeddings, it's important to conduct experiments on relevant downstream tasks (Faruqui et al., 2016; Faruqui and Dyer, 2014; Lu et al., 2015; Gatt and Krahmer, 2018).

3 Methodology

Table 1 gives the meta-data of the two corpora used. The Gigaword corpus was generated as described by Rødven Eide et al. (2016) while the Wikipedia corpus was pre-processed using the recommended script by (Grave et al., 2018). This script returned all text as lowercase and does not always retain non-asci characters. This created noise in the corpus, which may not necessarily be harmful, as it has been shown in a recent work that diacritics can adversely affect performance of embeddings unlike their normalized versions (Adewumi et al.). A portion of the pre-processed text (given in the previous section) was also tested for coherence on Google Translate and the English translation returned was meaningful, despite the noise. Hence, the noise issue was not serious enough to adversely affect the models generated in this case, as the embedding system seems fairly robust to handle such noise.

Meta-data	Gigaword	Wikipedia
Size	5.9G	4.2G
Tokens	1.08B	767M
Vocabulary	1.91M	1.21M
Year	2016	2019

Table 1: Meta-data for both Swedish Corpora

The authors made use of the fastText C++ library (with default hyper-parameters, except where mentioned) by Grave et al. (2018) to generate 8 word2vec models and 8 subword models from each corpus, based on the optimal hyper-parameter combinations demonstrated by Adewumi et al. (2020b). Each model was intrinsically evaluated using the new Swedish analogy test set by Adewumi et al.

(2020a) in a Python-gensim program (Řehůřek and Sojka, 2010). The hyper-parameters tuned are window size (4 & 8), neural network architecture (skipgram & continuous bag of words(CBoW)) and loss (heirarchical softmax and negative sampling). The subword models used lower & upper character n-gram values of 3 & 6, respectively.

Although each model in the first set of experiments, with default learning rate (LR) of 0.05, was run twice and average analogy score calculated, it would have been more adequate to calculate averages over more runs per model and conduct statistical significance tests. Nonetheless, the statistical significance tests can be conducted for the downstream tasks, which usually are the key tests for the performance of these embeddings. It should also be noted that deviation from the mean of each model performance for their corresponding two runs is minimal. Due to the observation of one model (of Gigaword-CBoW-hierarchical softmax) failing (with *Encountered NaN* error) when using the default LR of 0.05, another set of experiments with the LR of 0.01 was conducted but with single run per model, due to time constraint.

4 Results & Discussion

Table 2 gives mean analogy scores for LR 0.05 of embeddings for the two corpora and table 3 for LR of 0.01. It will be observed that the skipgram-negative sampling combination for both corpora for word2vec and subword models performed best in both tables, except one in table 3, confirming what is known from previous research (Mikolov et al., 2013; Adewumi et al., 2020b,a). From table 2, the highest score is 60.38%, belonging to the word2vec embedding of the Wikipedia corpus. The lowest score is 2.59%, belonging to the CBoW-hierarchical softmax, subword embedding of the Gigaword corpus. The highest score in table 3 also belongs to the Wikipedia word2vec model. Among the 8 embeddings in the word2vec category in table 2, there are 6 Wikipedia embeddings with greater scores than the Gigaword while among the subword, there are 5 Wikipedia embeddings with greater scores. Nearest neighbour qualitative evaluation of the embeddings for a randomly selected word is given in table 4.

We hypothesize that the general performance difference observed between the embeddings of the two corpora may be due to a) the advantage of wider domain coverage (or corpus balance in

	Skipgram (s1)				CBoW (s0)			
	H. S. (h1)		N. S. (h0)		H. S. (h1)		N. S. (h0)	
window (w)	4	8	4	8	4	8	4	8
Word2Vec %								
Wikipedia	47.02	44.09	**60.38**	60.38	29.09	30.09	54.39	56.81
Gigaword	40.26	44.23	**55.79**	55.21	26.23	27.82	55.2	55.81
Subword %								
Wikipedia	46.65	45.8	**56.51**	56.36	28.07	24.95	38.26	35.92
Gigaword	41.37	44.7	**58.31**	56.28	2.59	-	46.81	46.39

Table 2: Mean Analogy Scores for Swedish Gigaword & Wikipedia Corpora with LR=0.05

	Skipgram (s1)				CBoW (s0)			
	H. S. (h1)		N. S. (h0)		H. S. (h1)		N. S. (h0)	
window (w)	4	8	4	8	4	8	4	8
Word2Vec %								
Wikipedia	48.92	49.01	51.71	**53.48**	32.36	33.92	47.05	49.76
Gigaword	39.12	43.06	48.32	**49.96**	28.89	31.19	44.91	48.02
Subword %								
Wikipedia	45.16	**46.82**	35.91	43.26	22.36	21.1	14.31	14.45
Gigaword	39.13	43.65	45.51	**49.1**	31.67	35.07	28.34	28.38

Table 3: Analogy Scores for Swedish Gigaword & Wikipedia Corpora with LR=0.01
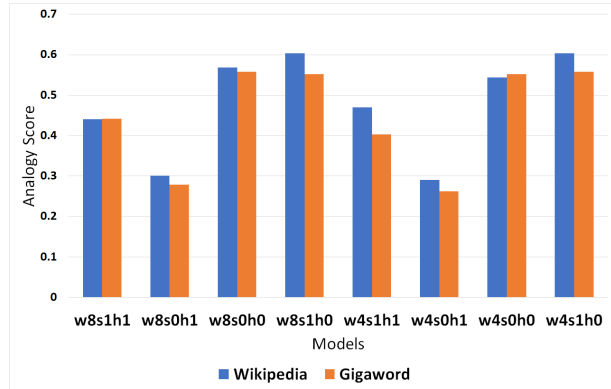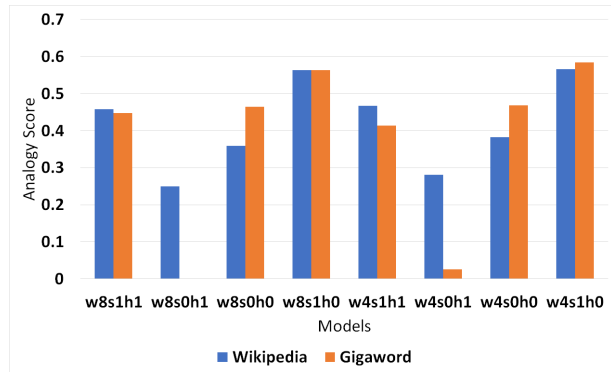

Figure 1: Word2Vec Mean Scores, LR:0.05

Figure 2: Subword Mean Scores, LR:0.05

Nearest Neighbor	Result
Wiki: syster	systerdotter (0.8521), systern (0.8359), ..
Gigaword: syster	systerdotter (0.8321), systerdottern (0.8021), ..

Table 4: Example qualitative assessment of Swedish subword w4s1h0 models

topics) of the Wikipedia corpus - which is the most plausible reason, b) the small noise in the

Wikipedia corpus or c) the combination of both earlier reasons.

Since it's preferable to have more than one criterion for the difference between the two corpora, future work will focus on other tests (particularly downstream tasks) to confirm this (Faruqui et al., 2016; Gatt and Krahmer, 2018). Implementation without using the pre-processing script by (Grave et al., 2018) on the original Wikipedia corpus will also be attempted. It will also be interesting to see the results of merging the two corpora.

5 Conclusion

This work has shown that better performance results from differently sourced corpora of the same language can be based on reasons besides larger data size. Simply relying on larger corpus size for performance may be disappointing. The Wikipedia corpus showed better performance in analogy tests compared to the Gigaword corpus. Broad coverage of topics in a corpus seems important for better embeddings and noise, though generally harmful, may be helpful in certain instances. Future work will include other tests and downstream tasks for confirmation.

6 Acknowledgement

References

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. The challenge of diacritics in yoruba embeddings. In *NeurIPS Machine Learning for Developing Countries Workshop (ML4D), dec 2020*.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2019. Conversational systems in machine learning from the point of view of the philosophy of science—using alime chat and related studies. *Philosophies*, 4(3):41.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2020a. Exploring swedish & english fasttext embeddings with the transformer. *arXiv preprint arXiv:2007.16007*.

Tosin P Adewumi, Foteini Liwicki, and Marcus Liwicki. 2020b. Word2vec: Optimal hyper-parameters and their impact on nlp downstream tasks. *arXiv preprint arXiv:2003.11645*.

Per Fallgren, Jesper Segeblad, and Marco Kuhlmann. 2016. Towards a standard dataset of swedish word vectors. In *Sixth Swedish Language Technology Conference (SLTC), Umeå 17-18 nov 2016*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*.

Martin T Hagan, Howard B Demuth, and Mark Beale. 1997. *Neural network design*. PWS Publishing Co.

Ang Lu, Weiran Wang, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 250–256.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rasmus Précenth. 2019. Word embeddings and gender stereotypes in swedish and english.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Stian Rødven Eide, Nina Tahmasebi, and Lars Borin. 2016. The swedish culturomics gigaword corpus: A one billion word swedish reference dataset for nlp.

Eli Stevens, Luca Antiga, and Thomas Viehmann. 2020. *Deep Learning with PyTorch*. Manning.

Wikipedia. 2019. Swedish wikipedia multistream articles.