Comparison of Embedding Techniques for Topic Modeling Coherence Measures

Mark Belford[1]
Insight Centre for Data Analytics, University College Dublin, Ireland
https://www.insight-centre.org/users/mark-belford
mark.belford@insight-centre.org

Derek Greene
Insight Centre for Data Analytics, University College Dublin, Ireland
derek.greene@ucd.ie

Abstract

The quality of topic modeling solutions are often evaluated using topic coherence measures, which attempt to quantify the semantic meaningfulness of the descriptors. One popular approach to evaluate coherence is through the use of word embeddings, where terms are represented as vectors in a semantic space. However, there exist a number of popular embedding methodologies and variants which can be used to construct these vectors. Due to this, questions arise regarding the optimal embedding approach to utilise when calculating the coherence of solutions produced for a given dataset. In this work we evaluate the difference between two popular word embedding algorithms and their variants, using two distinct external reference corpora, to discover if these underlying choices have a substantial impact on the resulting coherence scores.

2012 ACM Subject Classification Information systems → Document topic models

Keywords and phrases Topic Modeling, Coherence, Embeddings

Funding *Mark Belford*: This research was supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289

1 Introduction

Topic modeling facilitates the discovery of the underlying latent themes or topics in a corpus of text documents. These are frequently represented by their top n terms and are referred to as topic descriptors. There are many popular topic modeling approaches including probabilistic techniques such as Latent Dirichlet Allocation (LDA) [2] and those based on matrix factorization such as Non-negative Matrix Factorization (NMF) [5]. Ideally topic modeling solutions should be of high quality and easily interpretable, however this is unfortunately not always the case as poor solutions can be discovered for a number of reasons, such as the stochastic nature of traditional topic modeling algorithms [1]. With this in mind quality metrics are frequently used to evaluate solutions, with *topic coherence* being the most common. These measures typically attempt to evaluate the semantic coherence of a set of topics, relative to a background corpus. While originally a human evaluated task [4], there now exists a variety of automated coherence methodologies [7, 8, 12].

A more recently proposed approach to evaluate coherence utilises word embedding algorithms, such as word2vec [6] and fastText [3]. In both of these approaches, words are represented in a dense, low-dimensional vector space, where words with similar meaning and usage appear to be similar to one another. Both algorithms offer two different model variants to construct these vectors – Continuous Bag-Of-Words (CBOW) and Skip-Gram (SG). The goal of CBOW is to predict a target word while using the surrounding context

[1] Corresponding author

LDK 2019 - Posters Track.
Editors: Thierry Declerck and John P. McCrae

words, based on a sliding window, while SG is the inverse where the goal is to predict the surrounding context words for a given target word. Word embedding models require that they be trained on large external reference corpora to facilitate making these predictions. However, questions arise regarding which of these embedding approaches to utilise when calculating topic coherence for a given dataset, especially as there are many facets which are left to the user to specify and these may have an impact on the results. With this in mind we propose the following research question – how does the choice of embedding algorithm, selected variant, and background reference corpus impact the resulting coherence scores?

2 Methodology

To calculate the coherence of topic descriptors using word embeddings we utilise the approach proposed by [9]. This technique quantifies the intra-topic coherence based on word similarities using their learned vector representations from a given embedding model. However, it is possible that some of these top terms may not have a corresponding vector in the embedding model due to not appearing in the vocabulary of the external reference corpus used for training. To account for this we propose a small modification to this approach in which we construct the list of top terms as the first N terms that appear in a descriptor but are also contained in the embedding vocabulary. By following this approach coherence scores for topics are calculated using the formulation seen in Equation 1. While fastText can generate vectors for terms that are not present in the reference corpus vocabulary we chose not to utilise this feature to ensure a fair comparison with word2vec. Frequently topic coherence is only measured at the individual topic level, such as in Equation 1. However, we can also calculate an overall coherence score at the model level by simply computing the average of these individual topic descriptor coherence scores.

$$TC = \frac{1}{\binom{N}{2}} \sum_{j=2}^{N} \sum_{i=1}^{j-1} similarity(w_i, w_j) \tag{1}$$

For our experiments, we constructed 15 yearly datasets from The Guardian API, where associated article section labels were used as ground truth topics (e.g. “politics”, “technology”). We then built 100-dimensional CBOW/SG word2vec and fastText embeddings on two larger background corpora: (1) 1.6m Guardian news articles published from 2004-2018, (2) 4.9m Wikipedia long abstracts collected in 2016 [10]. These variant and corpus combinations yielded 8 embeddings, as seen in Table 2. For each dataset, we generate 100 runs of randomly-initialized NMF, and compute 100 corresponding model-level coherence scores, before averaging this set to compute a final coherence value, as seen in Equation 2. We repeat this process over a range of topic numbers $k \in [2, 30]$ for each embedding and dataset combination. Table 1 provides a detailed breakdown of these datasets.

$$MeanTC = \frac{1}{r} \sum_{i=1}^{r} TC(model_i) \tag{2}$$

Table 1 Details of the fifteen evaluation corpora and two reference corpora used in our experiments, including the total number of documents n, number of terms m, and number of categories $\hat{k}$ in the associated "ground truth" annotations.

Corpus	n	m	$\hat{k}$
guardian-2004	18,209	20,191	5
guardian-2005	17,311	17,396	4
guardian-2006	24,338	22,491	6
guardian-2007	28,218	27,051	6
guardian-2008	36,774	30,579	8
guardian-2009	30,411	26,825	7
guardian-2010	25,164	25,426	6
guardian-2011	20,840	24,008	5
guardian-2012	28,820	28,783	7
guardian-2013	22,139	24,813	5
guardian-2014	28,774	29,118	7
guardian-2015	32,593	32,098	7
guardian-2016	30,634	31,056	7
guardian-2017	17,918	23,279	5
guardian-2018	15,334	21,520	5
guardian15	1,595,844	557,937	N/A
wikipedia2016	4,899,998	1,333,306	N/A

Table 2 Details of the eight combinations of embedding models with varying embedding approaches, variants and reference corpora.

Combination Name	Embedding	Variant	Reference Corpus
guardian15-ft-cbow-d100	fastText	CBOW	guardian15
guardian15-ft-sg-d100	fastText	SG	guardian15
guardian15-w2v-cbow-d100	word2vec	CBOW	guardian15
guardian15-w2v-sg-d100	word2vec	SG	guardian15
wikipedia2016-ft-cbow-d100	fastText	CBOW	wikipedia2016
wikipedia2016-ft-sg-d100	fastText	SG	wikipedia2016
wikipedia2016-w2v-cbow-d100	word2vec	CBOW	wikipedia2016
wikipedia2016-w2v-sg-d100	word2vec	SG	wikipedia2016

3 Evaluation

3.1 Ranked Correlation

We first investigated whether there was a noticeable difference between the different embedding approaches with respect to their coherence scores by measuring the Spearman rank correlation between the average topic coherence scores produced on each of the 15 Guardian datasets. These results are displayed as a heatmap plot, as seen in Figure 1. It is evident that there is a large difference between embedding models that are trained using different background corpora, with the models having much lower correlation scores with respect to each other. It is also worth noting that, when considering the same background corpora, the different embedding algorithms exhibit relatively high correlation scores. This suggests that they

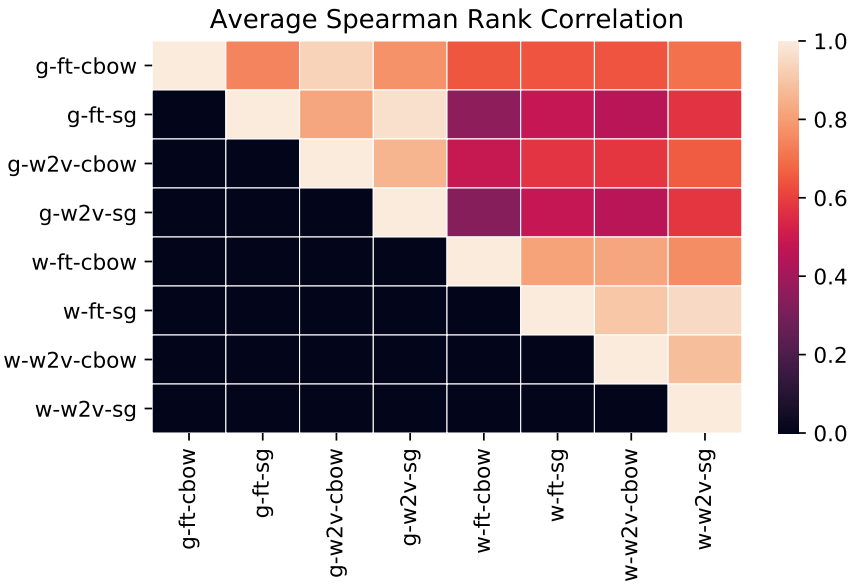

Figure 1 Heatmap of the Pairwise Average Spearman Rank Correlation over all 15 corpora.

may perform similarly when trained on the same data. When exploring this further, it also appears that there is a high level of correlation between the variants of the different embedding algorithms (i.e. CBOW v SG) when utilising the same reference corpora.

3.2 Ground Truth Evaluation

A common application of topic coherence is to select an appropriate number of topics k. Therefore, we further explored the effect of embedding choice as follows. For each dataset and embedding model, we sorted the coherence scores for different k values to identify the top values of k. We then counted the number of times the "ground truth value" of k appears within the top n recommendations, for $n = 1$ to $n = 5$, as seen in Table 3. For example, the wikipedia-w2v-cbow embedding correctly identifies the ground truth number of topics when $n = 5$ for 14 of the 15 datasets. Surprisingly, using the Wikipedia corpus, rather than the domain-specific Guardian corpus produces better embeddings with respect to identifying the "correct" number of topics. This may be due to a temporal effect where The Guardian news articles span over a 15 year duration, while the Wikipedia dump reflects a relatively recent collection of articles. It is also interesting to note that fastText performs considerably worse than the word2vec model in these cases. Across all combinations it is also clear that the CBOW variant performs better than SG, and is likely due to CBOW having to only predict a single target word rather than the context words around it.

4 Conclusion

In this work we have demonstrated that care should be taken when utilising word embeddings in the process of measuring topic coherence. It is clear that the choice of embedding algorithm, model variant, and background corpus has a large impact on the resulting coherence values, which could potentially influence topic model parameter selection choices, and ultimately affect the interpretations made from the topics identified on a given corpus.

Table 3 Results of the number of times the ground truth value of k was identified in the top n elements for each embedding combination.

Combination Name	Top 1	Top 2	Top 3	Top 4	Top 5
guardian15-ft-cbow-d100	0	1	3	3	3
guardian15-ft-sg-d100	0	0	1	1	1
guardian15-w2v-cbow-d100	1	3	3	3	3
guardian15-w2v-sg-d100	1	1	1	1	1
wikipedia2016-ft-cbow-d100	1	1	3	4	4
wikipedia2016-ft-sg-d100	2	4	5	5	5
wikipedia2016-w2v-cbow-d100	**4**	**7**	**7**	**14**	**14**
wikipedia2016-w2v-sg-d100	4	5	5	6	6

References

1 Mark Belford, Brian Mac Namee, and Derek Greene. Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, 91:159–169, 2018.

2 David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

3 Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

4 Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, pages 288–296, 2009.

5 Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

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

7 David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics, 2011.

8 David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

9 Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

10 M. Atif Qureshi and Derek Greene. Eve: Explainable vector based embedding technique using wikipedia. *Journal of Intelligent Information Systems*, Jun 2018.

11 Radim Řehůřek and Petr Sojka. 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, May 2010. ELRA.

12 Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM, 2015.