Rethinking Topic Modelling: From Document-Space to Term-Space

Magnus Sahlgren
Research Institutes of Sweden (RISE)
Box 1263, 164 29 Kista, Sweden
magnus.sahlgren@ri.se

Abstract

This paper problematizes the reliance on documents as the basic notion for defining term interactions in standard topic models. As an alternative to this practice, we reformulate topic distributions as latent factors in term similarity space. We exemplify the idea using a number of standard word embeddings built with very wide context windows. The embedding spaces are transformed to sparse similarity spaces, and topics are extracted in standard fashion by factorizing to a lower-dimensional space. We use a number of different factorization techniques, and evaluate the various models using a large set of evaluation metrics, including previously published coherence measures, as well as a number of novel measures that we suggest better correspond to real-world applications of topic models. Our results clearly demonstrate that term-based models outperform standard document-based models by a large margin.

1 Introduction

Topic models are often used in real-world text analysis scenarios as tools for efficient data exploration. The typical modus operandi in such scenarios is to run a topic model with standard parameter settings on the data, and to extract some fixed number n of topics and some fixed number m of words per topic, and then manually interpret, and draw conclusions from, the resulting term lists. A common choice for both n and m is around 10. This means that the human analyst only needs to look at around 100 terms in total instead of reading a text collection consisting of possibly several hundreds of thousands, or even millions, of running words. In terms of efficiency, this is an invaluable tool for content analysis.

Topic models extract topics by uncovering (latent) interactions between terms in document space. This methodology obviously assumes that data arrives with clear and consistent document boundaries, and, in the best case, a fairly even distribution of number of words per document. Unfortunately, this assumption rarely holds in real-world scenarios, where data may arrive in streams, in batches without clear document boundaries, or with very large variations in document lengths. To handle such scenarios, it would be desirable to use a model that is insensitive to the formatting of the input data. In this paper, we discuss and evaluate one such approach, which embeds the topic modelling process entirely in term space. This makes the model less sensitive to document-formatting, and, as it turns out, also more precise.

This work is primarily motivated by the practical usability of topic models in real-world analysis scenarios. In such applications – common in particular in the social sciences, and in security and defence applications – the analyst only cares about the top ranked terms in the resulting term lists. We therefore introduce a number of additional evaluation metrics for topic models, which may correspond better to practical considerations than the commonly used intrinsic (and mostly theoretical) evaluation measures. We also provide an evaluation that casts the topic modelling as a document annotation scenario, and that uses manual annotations as gold standard. Our results – across all evaluation metrics – clearly demonstrate that term-based approaches outperform standard document-based topic models by a large margin.

2 Document-based Topic Models

Topic models are a family of latent-variable methods that attempt to identify interesting patterns in term occurrences over documents. Most topic models take as starting point a standard vector space model (VSM, i.e. a term-document matrix that has been weighted by some suitable term

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2250–2259
November 16 - 20, 2020.

weighting scheme such as TF-IDF). This term-document space is then factorized into a lower-dimensional representation in which the dimensions are interpreted as topics. This allows for both documents and terms to be described as distributions over topics, and conversely for topics to be described as distributions over terms and documents. The choice of factorization technique is the main design choice when it comes to topic modelling. Common approaches include Singular Value Decomposition (SVD; Deerwester et al. (1990)), Non-negative Matrix Factorization (NMF; Lee and Seung (2001)), Latent Dirichlet Allocation (LDA; Blei et al. (2003)), and more recently deep neural networks (Cao et al., 2015; Miao et al., 2017).

Despite the choice of factorization method, all document-based topic models rest on the assumption that latent interactions between terms are due to topical variation in document space. This assumption is neatly summarized by the generative story told by models such as pLSA (Hofmann, 1999) or LDA, which amounts to a subject choosing a (set of) topic(s) to talk about, and for each topic choosing a set of representative terms to utter. This story makes intuitive sense, but note that the notion of a document is completely ad hoc to the story; it only enters the story as the unit of text being output by the subject. We argue that the notion of a document is an unnecessary restriction for topic models that limit the application of such models to data with proper formatting, and that topical term interactions can be better modelled directly in term space.

3 From Document Space to Term Space

We thus suggest to focus entirely on term space, and to remove the dependence on the notion of documents completely. Instead of building term vectors for each *document* in the data, (i.e. a standard VSM), we build word embeddings for all *terms* in the data from large context windows spanning something like 50 tokens.[1] Using such wide context windows ensures that the embeddings have the capacity to encode wider, and thus more topical, contextual information.

There are many ways to build word embeddings. We include four different approaches in this paper:

- Co-occurrence matrix (COOC), a standard term-term matrix that weights co-occurrence counts collected within a sliding context window with Positive Pointwise Mutual Information (Levy et al., 2015).
- Random Indexing (RI), an incremental random projection technique that accumulates embeddings for a word by summing the random index vectors for all words in its context (Sahlgren, 2005).
- Word2Vec (W2V), a shallow neural network that learns embeddings using a language modeling objective (Mikolov et al., 2013).
- Doc2Vec (D2V), a shallow neural network that uses the same architecture as Word2Vec, but that learns to predicts document identifiers instead of words (Le and Mikolov, 2014).

These techniques have their respective merits and drawbacks. The COOC approach is simple and straightforward, but the dimensionality of the embeddings is equivalent to the size of the vocabulary, which can become prohibitive for large data. RI solves the dimensionality problem, since it uses fixed-sized vectors, but at the cost of added noise. W2V is widely acknowledged to be both efficient and precise, but requires sufficient amounts of training data. D2V, on the other hand, is designed primarily for document-processing applications, which might make it an interesting candidate for more topic-oriented applications, such as the present one.

For each of the resulting word embeddings, we compute a similarity matrix that contains the pairwise similarities between all term vectors in the embedding space. We prune the similarity matrix by removing entries with too small values, which gives us a sparser similarity space to operate in. This is beneficial from a computational perspective, and it also removes noise from the representations. To extract topics from the similarity space, we can apply any type of algorithm that identifies clusters or latent variables.[2] In our case, we opt

[1] The size of the context window is of course a parameter that can be tuned and optimized for specific data and analysis scenarios. We default to 50 tokens in these experiments, and we acknowledge that other parameter settings may lead to other results.

[2] Our preliminary experiments included standard clustering methods, such as k-means, agglomerative clustering and density-based methods, but we did not observe any consistent improvements using clustering as compared to factorization.

for a number of simple factorization methods, including:

- Singular Value Decomposition (SVD, Golub and Van Loan (1996)),
- Non-negative Matrix Factorization (NMF, Lee and Seung (2001)),
- Dictionary Learning (DL, Mairal et al. (2009))

For each of these factorization techniques, we extract n components (where n defaults to 10) in the same way as in a standard topic model. In the experiments in Sections 6.3 to 6.5, all results are averaged over 10 runs of the various factorization techniques.

4 Related Work

There have been a few previous studies that explore the use of term-based representations for topic modelling. One example is Arora et al. (2013), who also base their solution on a term-term matrix, but their term-term matrix is not a co-occurrence matrix but a *correlation* matrix produced from a standard term-document matrix. As such, their model still relies on the data being properly formatted in a coherent document structure. By contrast, the models we consider do not put any constraints on the formatting of the data, while at the same time adopting a stricter definition of topical relationships in the form of context windows spanning (in our case) 50 terms, which is typically significantly smaller, and thus more precise, than a whole document.

Another example is Rangarajan Sridhar (2015), who cluster a word embedding (produced with Word2Vec) using a Gaussian Mixture Model (GMM). This is the previous work that comes closest to the approaches we consider, but there are a number of significant differences. We explore a range of word embedding techniques, we use wider context windows (50 terms instead of 11–17), and we use a range of standard factorization techniques instead of GMM to extract term clusters. Despite these differences, we consider Rangarajan Sridhar (2015) to be an important inspiration to our work.

Also similar in spirit to our work is Shi et al. (2018), who incorporate a word similarity matrix with a standard document-based NMF model. The word similarity matrix is built using the Skipgram model from Word2Vec, and is used as an additional term in the block coordinate descent algorithm used to solve the NMF. The approach, aptly named Semantics Assisted NMF (SeaNMF) is primarily designed for data with short documents, in which case the size of the context windows used for the Skipgram embeddings equals the length of the documents in the data. Shi et al. (2018) argue that the sparsity of their word similarity matrix is highly beneficial for the efficiency of the model, and the same advantage consequently applies to our case. The most significant difference between the SeaNMF model and the approaches we consider is that the latter rely *only* on the word similarity matrix, and thus do not use any term-document matrix at all.

In contrast to these previous studies, we focus on the general idea of using word embeddings rather than a VSM as the basis for topic modelling, and we compare a range of different word embeddings using a range of different factorization techniques. We also use a wider range of evaluation methods, and introduce a number of novel measures that better correspond to practical usage of topic models.

5 Evaluation Methods

Since (most) practical use of topic models only focus on the resulting lists of terms, this should also be our focus for evaluation. A challenge here is that determining the quality of term lists can be a notoriously subjective task, comparable (jokingly) to reading tea leaves (Chang et al., 2009). There have been attempts to arrive at more objective evaluation measures for topic models, which usually take the form of using various forms of intrinsic information measures such as entropy, perplexity, or coherence (Wallach et al., 2009; Newman et al., 2010; Mimno et al., 2011; Stevens et al., 2012a). However, such information theoretic measures do not always correlate with semantic interpretability, as noted by Chang et al. (2009), and even if they do, it is not clear why *semantic* interpretability should correlate with *topical* coherency.

In light of these difficulties, it is somewhat remarkable that gold standard topic annotations are not used habitually as standard evaluation metric for topic models. Of course, we will probably not be able to find such annotations for individual terms or term lists, but we might be able find such

annotations at the text level. Even if topic modelling is essentially different from text categorization and text clustering, we can still use text categories as evaluation targets for topic models, given that the categories are topical in nature. That is, if we can find a text collection where the text has been manually labelled with one or more topics, we can simply compare this gold standard topic assignment with that produced by a topic model.

One simple way to do this, which also simulates how a human analyst might use the output of a topic model in a practical analysis scenario, is to collect all the documents covered by each topic (i.e. in which one or more of the terms in the topic list occurs), and then count the overlap between this set of documents and the set of documents labelled by topic categories in the gold standard. Doing so will arrive at a proportion of overlap between the topic model and the gold standard. We argue that this is a simple and straightforward way to evaluate topic models that maps directly to usability in practical application. We refer to this metric as "Truth" in Tables 2 to 5.

A human analyst might also be interested in other factors, such as:

- Overlap: how much do the topics overlap? We quantify this as the proportion of identical terms in the topics; lower is presumably better from an analyst's perspective.

- Coverage: how much of the data do the topics cover? An analyst may prefer a solution that We quantify this as the proportion of texts that contain terms in the topics; whether it is desirable with large or small coverage depends on the analysis scenario.

- Uniqueness: how often do terms from different topics co-occur in the same document? If we want low coverage of the data (i.e. small and focused topics), we should probably strive for high uniqueness of the topics, while if we aim for large coverage of the data, we should expect less unique topics.

- Separation: how much do the embeddings differ between topics? This is measured as the average difference between the cosine similarities between terms *within* topics, and the cosine similarities between terms *between* topics.

- Time: how long does it take to factorize (i.e. infer topics in) the similarity space? For the sake of replicability and comparability, we use the factorization functions from `scikit-learn`[3] with default settings as far as possible.

We also include the UCI (Newman et al., 2010) and UMASS (Mimno et al., 2011) coherence measures as a comparison. These measures sum the PMI values of all pairs of words in the topics; the UMASS measure considers co-occurrences within the entire document, while the UCI measure delimit co-occurrences within a sliding window:

$$\text{UMASS}(w_i, w_j, \epsilon) = \log \frac{D(w_i, w_j) + \epsilon}{D(w_j)} \quad (1)$$

$$\text{UCI}(w_i, w_j, \epsilon) = \log \frac{p(w_i, w_j) + \epsilon}{p(w_i)p(w_j)} \quad (2)$$

Following Stevens et al. (2012b), we set $\epsilon < 1$, in our case to $\epsilon = 0.001$.

6 Experiments

The following experiments use a number of different datasets, built from two different data sources. The first data source is Swedish news, which have been manually collected and annotated with topics by human experts. As such, this dataset corresponds well to a real-world analysis scenario. However, since the Swedish dataset is relatively small, and not publicly available,[4] we also generate a number of artificially annotated English datasets based on the English Wikipedia. The various datasets are detailed in Table 1, and in the following sections.

All experiments in this paper are run on a machine with Intel Xeon E5-2620 2.40GHz CPUs and 192 GB of RAM. All factorization techniques are run using standard settings and the implementations in `scikit-learn` version 0.20.0. We use the `Gensim`[5] implementations of Word2Vec and Doc2Vec with standard parameter settings, and in-house Python implementations of COOC and RI. The RI implementation is available at: `https://ghetto.sics.se/mange/ri`.

[3] `https://scikit-learn.org/`

[4] The data may be attainable by contacting the authors of the Swedish data study, Johansson and Strömbäck (2019).

[5] `https://radimrehurek.com/gensim/`

Data	# Texts	# Tokens	# Types	# Topics	Min. Freq.
Swedish News	895	366,456	33,358	34	5
English Wikipedia	100,000	14,784,214	269,741	40,109	10
English Wikipedia (small topics)	213,656	30,873,801	273,056	125,397	20
English Wikipedia (medium topics)	112,653	16,316,965	173,509	11,194	10
English Wikipedia (large topics)	1,273	196,378	13,398	20	5

Table 1: Datasets used in the experiments.

6.1 Data based on Swedish News

The Swedish dataset consists of news articles collected from the major Swedish newspapers (Svenska dagbladet, Dagens nyheter, Aftonbladet, Expressen) by Johansson and Strömbäck (2019). Each news article has been manually annotated with several different categories by experts at the Department of Journalism, Media and Communication at the University of Gothenburg. We use the category `HuvudÄmne` (eng. `main topic`) as gold standard label, since it explicitly represents the main topic of the news article. A practically useful topic model should be able to minimally identify these 34 different main topics from the data. The data contains 895 news articles with a total amount of 366,456 tokens. The average document length is around 400 terms, with very high variance. We ignore terms with a frequency less than 5 for the Swedish data.

6.2 Data based on Wikipedia

Since the Swedish news data set is comparatively small and not publicly available, we also include a number of larger datasets based on random samples of English Wikipedia articles. The samples are produced by randomly sampling text paragraphs from Wikipedia, and using the title of the Wikipedia entry as topic label for the text. Two examples of such topics are "Climate change in Finland" and "Mike Tyson vs. Michael Spinks". We use a probabilistic sampling strategy that produces on average 20 text samples per topic, with standard deviation around 10, and a minimum number of samples at around 5.

As seen in Table 1, we produce four different datasets based on this strategy.[6] The first contains 100,000 texts with a total of 14,784,119 unigram terms. The average document length is around 150 terms, with a standard deviation of around 50 (the longest document contains approximately 1,000 terms, and the shortest approximately 50). In order to be able to study the effect of topic size on the topic models, we also produce three different datasets with varying numbers of texts per topic. We produce data for small, medium-sized, and large topics, where small topics are those with 5 or less texts per topics, big topics are those with 50 or more texts per topic, and those in between are counted as medium-sized. This leads to 125,397 small topics spanning 30,873,748 tokens, 11,194 medium-sized topics spanning 16,316,954 tokens, and 20 big topics containing 196,378 tokens. We use a minimum frequency threshold of 10 occurrences for the English data, with the exception of the small topics data, where we instead set the minimum frequency threshold to 20 occurrences, and the big topics data, where we use 5 occurrences as threshold.

6.3 Documents vs. Terms

In the first set of experiments, we compare document-based topic models with term-based models. We include two different document-based models; NMF and LDA,[7] both applied to a standard VSM with TF-IDF weighting. We compare these baseline models with four different term-based models that use NMF as factorization;[8] a standard co-occurrence matrix weighted with PPMI (COOC), Random Indexing (RI), Word2Vec (W2V), and Doc2Vec (D2V).

Table 2 shows the results on the Swedish data. The baseline document-based models get higher scores on the document-based UMASS measure, but significantly lower scores on the word-based UCI measures. The document-based models have a higher overlap between topics, and they also cover more of the data, but at the expense of less unique topic assignments. The term-based models

[6] The Wikipedia datasets can be downloaded from: https://bit.ly/33hhyiQ

[7] We use NMF and LDA since they are the most common factorization methods used by standard topic models.

[8] We use NMF here because it is comparably robust. A comparison of different factorization techniques for term-based models is provided in Table 4.

Embedding	Model	UMASS	UCI	Overl.	Cover.	Uniq.	Sep.	Truth	Time
VSM	NMF	-7.72	58.73	0.17	1.00	0.11	0.21	0.18	123.20
	LDA	**-7.24**	53.37	0.36	1.00	0.11	0.17	0.19	202.44
COOC	NMF	-9.92	95.50	0.08	1.00	0.28	0.35	0.28	356.54
RI	NMF	-8.46	145.74	0.03	0.74	0.91	0.31	**0.34**	361.56
W2V	NMF	-10.41	**159.60**	0.00	0.26	0.92	0.37	0.31	468.50
D2V	NMF	-15.54	146.11	0.00	0.55	0.79	**0.45**	0.23	477.92

Table 2: Results on the Swedish data for different embeddings (VSM, COOC, RI, Word2Vec, Doc2Vec) over 7 different evaluation metrics, including the UMASS and UCI coherence scores, topic overlap, topic coverage, uniqueness, separation, and overlap with truth. We also give the processing time (in seconds) for each factorization. All scores are the average over 10 runs.

Embedding	Model	UMASS	UCI	Overl.	Cover.	Uniq.	Sep.	Truth	Time
VSM	NMF	-9.07	56.44	0.16	1.00	0.10	0.13	0.00	14,462.20
	LDA	-2.31	58.55	0.34	1.00	0.12	0.15	0.01	12,399.78
COOC	NMF	1.62	152.08	0.00	1.00	0.95	0.27	0.03	8,895.03
RI	NMF	**11.55**	178.86	0.00	0.74	0.95	0.11	**0.06**	10,169.69
W2V	NMF	-5.67	141.64	0.00	0.57	0.86	**0.50**	0.01	12,098.17
D2V	NMF	9.63	**185.88**	0.00	0.14	0.97	0.44	0.02	9,571.91

Table 3: Results on the English data for different embeddings (VSM, COOC, RI, Word2Vec, Doc2Vec) over 7 different evaluation metrics, including the UMASS and UCI coherence scores, topic overlap, topic coverage, uniqueness, separation, and overlap with truth. We also give the processing time (in seconds) for each factorization. All scores are the average over 10 runs.

have a higher average separation between terms within vs. across topics, and they correspond better to manual topic assessment; the best model with respect to overlap with truth labels is RI, which overlaps to 34% with the gold standard.

Table 3 shows the results on the English data. We note that in this case, the term-based models significantly outperform the document-based models not only on the UCI measure, but also on the UMASS measure, with the exception of W2V, which has a lower score than the baseline VSM+LDA model. We again note that the document-based models have higher overlap between topics, where the term-based models have no overlap at all for the English data. Note also that document-based models tend to cover more of the data than term-based models, and that term-based models have more unique topic assignments. The term-based models also have a higher average separation between terms within vs. across topics, and they also tend to correspond better to the gold standard annotations – but we note the very low overlap for all models on the English data; the best model in this case is again RI, which has en overlap of only 6% with the human annotations.

6.4 Factorization Methods

Turning to the effects of using different factorization techniques for the various representations. Tables 2 and 3 shows that the difference between NMF and LDA for the document-based model is more pronounced for the larger English data, where LDA performs slightly better than NMF. For the smaller Swedish data, there is not consistent difference.

Table 4 shows the effects of using different factorization techniques using term-based models. We include three different factorization techniques for two different embeddings (COOC and W2V) in these results. Note that NMF leads to the best results for both embeddings using the Swedish data, but that the results are more mixed for the English data. For both the COOC and W2V embeddings, Dictionary Learning leads to the best UMASS, UCI measures. SVD leads to the best separation within and across topics for the COOC embeddings, but NMF leads to the best separation for the W2V embeddings. Dictionary Learning leads to the best overlap with the human topic annotations for the COOC embeddings, while there is no difference in overlap between the different factorization techniques for the W2V em-

Swedish									
Embedding	Model	UMASS	UCI	Overl.	Cover.	Uniq.	Sep.	Truth	Time
COOC	NMF	**-9.08**	**121.04**	0.00	1.00	0.72	**0.31**	**0.34**	204.13
	SVD	-11.24	101.69	0.03	1.00	0.30	0.28	0.30	**100.34**
	DL	-13.60	117.97	0.05	0.95	0.57	0.26	0.24	222.05
W2V	NMF	**-8.80**	**150.04**	0.00	0.28	0.97	**0.46**	**0.47**	244.76
	SVD	-9.03	139.70	0.00	1.00	0.88	0.38	0.29	**98.82**
	DL	-12.83	149.31	0.00	0.40	0.88	0.37	0.37	221.82
English									
Embedding	Model	UMASS	UCI	Overl.	Cover.	Uniq.	Sep.	Truth	Time
COOC	NMF	2.16	152.68	0.00	1.00	0.93	0.26	0.02	8,167.09
	SVD	-4.89	117.62	0.02	1.00	0.31	**0.30**	0.01	**3,529.35**
	DL	**11.41**	**162.32**	0.07	0.55	0.95	0.19	**0.05**	3,818.89
W2V	NMF	-7.51	137.2	0.00	0.60	0.85	**0.49**	0.01	8,892.46
	SVD	-2.35	150.28	0.00	0.82	0.76	0.39	0.01	**4,965.02**
	DL	**1.13**	**162.53**	0.00	0.44	0.88	0.36	0.01	5,973.62

Table 4: Results using different factorization techniques (NMF, SVD and Dictionary Learning) for the COOC and Word2Vec embeddings, on the Swedish data (top) and English data (bottom). The processing times are in seconds (using the implementations in `scikit-learn`, and all scores are the average over 10 runs.

Topic size	Embedding	Model	UMASS	UCI	Overl.	Cover.	Uniq.	Sep.	Truth
Small	VSM	LDA	4.93	102.12	0.13	1.00	0.18	0.27	0.00
	W2V	NMF	0.63	161.66	0.00	0.17	0.92	**0.53**	0.00
	RI	SVD	**16.24**	**188.14**	0.00	0.38	0.93	0.09	**0.01**
Medium	VSM	LDA	4.76	107.27	0.10	1.00	0.20	0.27	0.04
	W2V	NMF	2.84	168.03	0.00	0.25	0.89	**0.49**	0.01
	RI	SVD	**15.29**	**186.31**	0.00	0.31	0.85	0.11	**0.06**
Large	VSM	LDA	-10.77	90.33	0.06	1.00	0.23	0.28	0.31
	W2V	NMF	**-9.93**	**136.60**	0.00	0.30	0.97	**0.54**	**0.88**
	RI	SVD	-10.15	104.26	0.01	1.00	0.56	0.28	0.41

Table 5: Performance of the document-based LDA model, NMF-based W2V and SVD-based RI on data with different topic sizes. As before, all scores are the average over 10 runs.

beddings. SVD is the fastest technique using the `scikit-learn` implementations.

6.5 Topic Size

Since topics normally arrive in different sizes, it is a relevant question how the various models handle different sizes of topics. As described in Section 6.2, we use three datasets with topics of different sizes; small topics covering at most 5 texts each, large topics covering at least 50 texts each, and medium-sized topics, covering between 5 and 50 texts each. Table 5 shows the results of the document-based LDA model, the NMF-based W2V embeddings, as well as SVD-based RI embeddings. We include RI in this example, since it performs remarkable well on the small and medium-sized topics.

The most notable aspect of the results in Table 5 is that none of the models perform well, with respect to the overlap with the gold standard labels, on the small and medium-sized topics. The RI embeddings with SVD factorization gets surprisingly high UMASS and UCI scores, and is the only model with any discernible overlap with the truth labels for the small topics (a meager 1% overlap), and also has the most overlap for the medium-sized topics (6%). For the large topics, all models work significantly better with respect to the overlap with the truth labels; the document-based model has an overlap of 31%, RI has an overlap of 41%, and W2V has a very high overlap of 88%.

7 Discussion

As is obvious from our experiments in this paper, different topic models have different properties, and the proper choice of topic model depends on the specific information need of a particular analysis scenario. Even if term-based models in general outperform standard document-based models across all data and metrics used in this paper, there may still be situations where a document-based model would be suitable to use. One such scenario would be if the analyst requires a solution with large coverage of the data; document-based models tend to lead to higher coverage of the data, but there tends to be overlap between topics, and the topic assignments (counted as the occurrence of topic terms in documents) are less unique compared to term-based models.

Term-based models, on the other hand, produce more unique topics with less overlap, and better separation, between topics. The term-based models also reach higher scores on all evaluation metrics (UMASS and UCI coherence, representation separation, and overlap with truth labels) – with the exception of the Swedish data, where document-based models lead to higher UMASS coherence. In general, the difference between document-based and term-based models is lower when considering the UMASS measure than when looking at the UCI measure, which may be explained by the fact that the former uses documents as units for counting co-occurrences, while the latter uses words.

We note that there is a high variance between runs, which makes it difficult to draw any definite conclusions regarding the optimal design choice for a term-based topic model. Certain factorization techniques seem to be more suitable for certain representations and certain data. NMF in general seems to work best in these experiments for most word embeddings on the Swedish data, but Dictionary Learning works best in these experiments for the English data. On the other hand, if the topics are small, SVD seems to work better, in particular for the RI embeddings.

With regards to the different types of word embeddings, we note that the COOC model typically leads to the highest coverage of the data, followed by RI, which also tends to have the best overlap with human gold standard annotations, except for the case of large topics where Word2Vec is significantly better. We note that Word2Vec and Doc2Vec both have high average separation of terms within vs. across topics, but that the addition of document information in Doc2Vec does not seem to be useful for topic inference.

Note that the data used in these experiments contain only one topic per document, whereas many other topic modelling scenarios operate with multiple topics per document. We do not consider this restriction to have any effect on the generality of our results, since term-based models are eminently applicable to multi-topic scenarios. The proposed gold standard comparison is also directly applicable to multi-topic data.

8 Conclusion

This paper has demonstrated the usefulness of casting the topic inference in topic models as pursuit of latent factors in term-space rather than document-space. We have proposed a simple term-based model that uses standard word embeddings with standard factorization techniques. Despite their simplicity, such term-based models outperform all tested document-based models on all evaluation metrics used in this paper. We have also proposed a topic categorization task that utilizes gold standard topic annotations, as well as a range of other metrics that may correspond more closely to a real-world analysis scenario than the type of intrinsic measures commonly used in literature on topic models. The use of these additional measures enables us to characterize the different properties of topic models, and to make informed choices of topic model design for specific information needs.

Our experiments have demonstrated that the optimal model is likely to be data- and task specific, and that the optimal choice of specific representation and factorization technique will likely be different from case to case. However, as a robust baseline, we suggest to use Word2Vec representations with NMF factorization.

We conclude that term-based models are competitive, if not superior, in comparison with traditional document-based models, with a number of added benefits that include independence of document-formatting, and relative robustness to topic size. Although the models investigated in this paper outperform document-based models on all metrics, we consider our term-based approach to be a simple baseline model with a large potential for improvement.

Acknowledgements

This work was partly supported by the Swedish Defense Research Agency (FOI) and partly by the Swedish Research Council under grant 2017-02429 (Linguistic Explorations of Societies). This paper has benefited from discussions with Magnus Rosell (FOI), Johannes Johansson (Department of Journalism, Media and Communication, University of Gothenburg), and Stefan Dahlberg (Department of Humanities and Social Sciences, Mid Sweden University). Special thanks to Bengt Johansson (Department of Journalism, Media and Communication, University of Gothenburg) for providing access to the annotated Swedish data.

References

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML'13)*, pages II–280–II–288. JMLR.org.

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

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *AAAI Conference on Artificial Intelligence*.

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, NIPS'09, pages 288–296, USA. Curran Associates Inc.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Gene H. Golub and Charles F. Van Loan. 1996. *Matrix Computations*, third edition. The Johns Hopkins University Press.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Bengt Johansson and Jesper Strömbäck. 2019. *Kampen om mediebilden: nyhetsjournalistik i valrörelsen 2018*. Institutet för Mediestudier.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Daniel D. Lee and H. Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. 2009. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 689–696, New York, NY, USA. Association for Computing Machinery.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 2410–2419. JMLR.org.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of International Conference on Learning Representations (ICLR 2013)*.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA. Association for Computational Linguistics.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries*, JCDL '10, page 215–224, New York, NY, USA. Association for Computing Machinery.

Vivek Kumar Rangarajan Sridhar. 2015. Unsupervised topic modeling for short texts using distributed representations of words. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 192–200. Association for Computational Linguistics.

Magnus Sahlgren. 2005. An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering (TKE)*, Copenhagen, Denmark.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference (WWW'18)*, pages 1105–1114, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012a. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 952–961, Stroudsburg, PA, USA. Association for Computational Linguistics.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012b. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 1105–1112, New York, NY, USA. ACM.