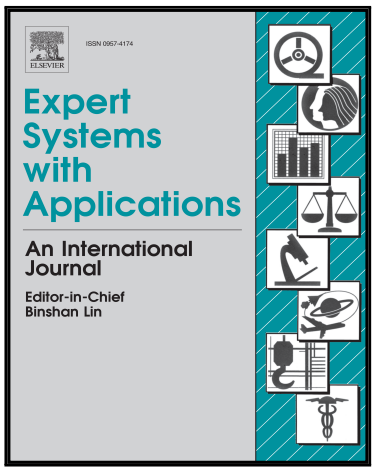

Document-based Topic Coherence Measures for News Media Text

Damir Korenčić, Strahil Ristov, Jan Šnajder

PII: S0957-4174(18)30488-3
DOI: 10.1016/j.eswa.2018.07.063
Reference: ESWA 12117

To appear in: *Expert Systems With Applications*

Received date: 20 April 2018
Revised date: 7 July 2018
Accepted date: 29 July 2018

Please cite this article as: Damir Korenčić, Strahil Ristov, Jan Šnajder, Document-based Topic Coherence Measures for News Media Text, *Expert Systems With Applications* (2018), doi: 10.1016/j.eswa.2018.07.063

Highlights

- Novel class of document-based coherence measures for news topics proposed and evaluated
- High-performing document-based coherence measure identified
- Document-based coherence measures contrasted with state-of-art word-based measures
- Application of document-based measures for semi-automated topic discovery

Document-based Topic Coherence Measures for News Media Text

Damir Korenčić[a,b,**], Strahil Ristov[a], Jan Šnajder[b,*]

[a] *Department of Electronics, Rudjer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia*
[b] *University of Zagreb, Faculty of Electrical Engineering and Computing, Unska 3, 10000 Zagreb, Croatia*

Abstract

There is a rising need for automated analysis of news text, and topic models have proven to be useful tools for this task. However, as the quality of the topics induced by topic models greatly varies, much research effort has been devoted to their automated evaluation. Recent research has focused on *topic coherence* as a measure of a topic's quality. Existing topic coherence measures work by considering the semantic similarity of topic words. This makes them unfit to detect the coherence of transient topics with semantically unrelated topic words, which abound in news media texts. In this paper, we introduce the notion of document-based topic coherence and propose novel topic coherence measures that estimate topic coherence based on topic documents rather than topic words. We evaluate the proposed measures on two datasets containing topics manually labeled for document-based coherence, on which the proposed measures outperform a strong baseline as well as word-based coherence measures. We also demonstrate the usefulness of document-based coherence measures for automated topic discovery from news media texts.

Keywords: Topic models, Topic coherence, Topic model evaluation, Text analysis, News text, Exploratory analysis

1. Introduction

News media, including broadcast, the press, and online news, in many ways mold our perception of the world and influence our decisions. According to a recent study by Newman et al. (2016), online news, including online news sites, news aggregators, search engines, social media, and increasingly also messaging apps are now the predominant source of news, while the majority of consumers

*Principal corresponding author
**Corresponding author
Email addresses: damir.korencic@irb.hr (Damir Korenčić), strahil.ristov@irb.hr (Strahil Ristov), jan.snajder@fer.hr (Jan Šnajder)

discover news stories through algorithms rather than editors or journalists. The increased consumption of online news in textual form has paralleled a growing interest in the use of natural language processing (NLP) and machine learning for automated analysis of news texts. These technologies enable users to derive information from large amounts of text data and target diverse components of the news media ecosystem, from providing end-consumers with more efficient and personalized access to news (Steinberger et al., 2013; Vossen et al., 2014; Li et al., 2011) to support for news production and dissemination (Clerwall, 2014; Popescu & Strapparava, 2017) to content analysis (Flaounas et al., 2013; Neuendorf, 2016).

One important NLP task in the context of news text analysis is the unsupervised discovery of topics from large volumes of texts. *Topic models* (Blei et al., 2003; Blei, 2012) have proven to be extremely useful tool for this task. A topic model is a probabilistic model of text that, given a set of text documents as input, produces word-topic and document-topic probability distributions. Topics are expected to correspond to concepts and can be used as topical summaries or features for various downstream NLP tasks. Table 1 shows an example of topics produced by running a topic model on a corpus of US news from 2015. The main advantage of topic models is that they are unsupervised and require a minimum of linguistic processing. However, the downside is that the quality of the topics may greatly vary – an issue compounded by the stochastic and approximate nature of the topic model inference.

There are a number of ways to characterize the quality of model-derived topics (Boyd-Graber et al., 2014). Recent research has focused on the notion of *topic coherence*, loosely defined in terms of topic's correspondence to a concept (Newman et al., 2010). The existing approaches to topic coherence are word-based, assuming that topic coherence correlates with the coherence of the words assigned to that topic. Consequently, the coherence of model topics is assessed and measured based on top-ranked topic words. As an example, consider the model topics in Table 1. The first two topics would likely be considered coherent, as their top-ranked words correspond to the concepts of "Economy" and "Sport", respectively. Topics 3 and 4 appear less coherent when judging by their top-ranked words, while the last topic, labeled "noise", is an incoherent topic composed of unrelated words.

While the assumption that topic coherence correlates well with the coherence of the top-ranked topic words is intuitive and certainly also true in many cases, we argue that it nonetheless provides a partial view of the notion of topic coherence. In particular, we note that word-topic distribution constitutes just a subset of the topic-related information contained in the model, and that not all topics will lend thesmelves to a semantic interperetation based on topic words alone. A case in point are topics 3 and 4 in Table 1. Unlike "Economy" and "Sport", which are general and enduring topics, topics 3 and 4 are contingent and transient – a trait typical of topics from a newspaper corpus. At the level of topic words, the topics appear incoherent, as the words are semantically dissimilar. However, another important piece of information, thus far mostly overlooked in topic coherence analysis, is the document-topic distribution. In

	Topic label	Top-10 topic words
1.	Economy	rate, economy, growth, fed, dropped, low, market, reserve, price, unemployment
2.	Sport	team, game, players, season, sports, league, fans, football, bowl, pick
3.	US DHS shutdown	boehner, homeland, block, dhs, mcconnell, pass, illegal, speaker, border, deportation
4.	ISIS war authorization	ground, veto, resolution, corker, latino, bob, draft, review, capitol, pass
5.	*(noise)*	paper george animals richard dog pledge era nothing sometimes cooperation

Table 1: Example topics derived from a corpus of news texts compiled by Korenčić et al. (2015), comprising about 24 thousand US news articles from 2015. Each topic is characterized by ten words with the highest word-topic probability. The topic labels were assigned manually based on inspection of documents with highest document-topic probability.

particular, if a human annotator inspects the documents associated with the two topics, she will likely recognize the topics as coherent and semantically interpret them as "US DHS shutdown" and "ISIS war authorization", respectively. This example illustrates that, for the transient topics corresponding to news stories, the topic words provide insufficient information to assess the coherence. In such cases, the topic coherence can often be more easily assessed by inspecting the documents associated with the topic.

Motivated by the above observations, we propose *document-based* topic coherence as an alternative to word-based topic coherence. Document-based topic coherence can better capture topics' semantic interpretability in cases when the topics are transient and contingent, as is often the case with news topics. The main result of our work is a novel method for calulating document-based topic coherence. The method consists of three steps: (1) the selection of topic-related documents, (2) document vectorization, and (3) computation of a coherence score from the document vectors using either distance-based, graph-based, or density-based methods. We consider a number of options for each of the three steps, obtaining different *document-based topic coherence measures*. We experimentally evaluate the measures on two datasets with topics obtained using a standard Latent Dirichlet Allocation topic model, manually annotated for document-based coherence. We show that a graph-based coherence measure performs the best, outperforming a strong baseline document-based method. Furthermore, we experiment with measuring document-based coherence using state-of-art word-based coherence measures, demonstrating that these measures fail to estimate document-based coherence. A qualitative analysis of topics based on word- and document-based coherence reveals that document- and word-based measures can complement each other and that it therefore may be beneficial to combine these two types of measures. Lastly, in a proof-of-concept

study, we demonstrate the usefulness of document-based coherence measures for the task of semi-automated topic discovery.

In summary, the contribution of our work is threefold: (1) we introduce the notion of document-based topic coherence and demonstrate its adequacy for news media texts, (2) we propose novel, document-based topic coherence measures, and (3) we compile and make available two datasets[1] of topics manually annotated with document-based topic coherence scores, as well as the code[2] and the resources necessary to replicate our experiments.

The remainder of the paper is set out as follows. The next section provides background on topic models and an overview of the related work, including applications on news texts and topic model evaluation. In Section 3, we elaborate the notion of the document-based coherence and propose methods for computing document-based coherence measures. In Section 4, we evaluate and analyze the document-based coherence measures, while in Section 5, we compare them against state-of-art word-based coherence measures. In Section 6, we describe the proof-of-concept for the application of document-based coherence measures to semi-automated topic discovery. In Section 7, we conclude the paper and outline future work.

2. Background and Related Work

In this section, we give a brief description of topic models, followed by an overview of related work. There are three threads of research relevant to our work: applications of topic models for news text analysis, evaluation of topic models, and topic coherence evaluation.

2.1. Topic Models

Topic models (Blei et al., 2003) are generative probabilistic models of text with numerous text analysis applications, including exploratory analysis of text collections (Grimmer, 2009; Chuang et al., 2012), information retrieval (Wei & Croft, 2006), feature extraction (Chen et al., 2011), and natural language processing tasks, such as word sense disambiguation (Boyd-Graber et al., 2007b) and sentiment analysis (Lin & He, 2009). The structure of a topic model is defined by a set of random variables and relationships among them, which together define the probabilistic process of text generation. Typically, the variables of interest are topics, defined as probability distributions over words in the dictionary, and document-topic distributions, defining topic salience within each of the documents.

The most widely used topic model is Latent Dirichlet Allocation (LDA), proposed by Blei et al. (2003). This model posits a fixed number of topics, K, with each topic being a probability distribution over words in the dictionary. Topics are represented by a word-topic probability matrix $\boldsymbol{\phi}$, with $\boldsymbol{\phi}_{ij}$ being

[1] `https://rebrand.ly/doc-coh-dataset`
[2] `https://rebrand.ly/doc-coh-code`

the probability of word j in topic i. Similarly, documents are represented by a document-topic probability matrix θ, with θ_{ij} being the probability of topic j in the i-th document. The process of text generation unfolds as follows. First, each topic $\boldsymbol{\phi}_i$ is sampled as a multinomial distribution from a Dirichlet prior distribution with parameter $\vec{\beta}$. Then, for each document D_i, θ_i is sampled as a multinomial distribution from a Dirichlet prior distribution with parameter $\vec{\alpha}$. Lastly, for each word within the i-th document, a topic z_{ij} is sampled from θ_i, and then the word is sampled from the topic $\boldsymbol{\phi}_{z_{ij}}$. This generative process is summarized by the following probabilities for the word-topic matrix and the document collection:

$$p(\boldsymbol{\phi}) = \prod_{i=1}^{K} \mathrm{Dir}(\boldsymbol{\phi}_i|\vec{\beta})$$

$$p(D_i) = \mathrm{Dir}(\theta_i|\vec{\alpha}) \prod_{j} \mathrm{Mult}(z_{ij}|\theta_i)\mathrm{Mult}(w_{ij}|\phi_{z_{ij}})$$

Text documents $D = \{D_i\}$ are the observed variables of the model, and inference algorithms are used to estimate the word-topic distributions $p(\boldsymbol{\phi}|D)$, document-topic distributions $p(\theta|D)$, and assignments of topics to words $p(z|D)$. Typically, the inference is performed using approximate inference methods, such as Gibbs sampling (Griffiths & Steyvers, 2004) and variational inference (Blei et al., 2003; Hoffman et al., 2010). Various extensions of the basic LDA model with a richer structure have been proposed in the literature, including those that model text document metadata (Mimno & McCallum, 2012) and relationships between topics (Blei & Lafferty, 2007), as well as models with a variable number of topics (Blei, 2012).

All the coherence measures considered in this article use either the word-topic probability matrix $\boldsymbol{\phi}$ or the document-topic probability matrix θ to represent topics as lists of topic-related words or documents, respectively, and use this representation to compute the coherence of topics.

While LDA and its variants are certainly the most widely used topic models today, it should be noted that generative models are not the only approach to topic modeling. One alternative is matrix factorization models, such as latent semantic analysis (LSA) (Deerwester et al., 1990) and non-negative matrix factorization (NMF) (Lee & Seung, 1999). These models derive a set of latent factors by approximating the document-word matrix as a product of document-factor and factor-word matrices. The latent factors can be viewed as corresponding to topics, with the semantics of factors defined by document-factor or factor-word weights. Relevant for the work presented in this paper is the fact that coherence measures studied in this paper can also be applied to such factor-based topics, represented by either document-factor or factor-word weights.

2.2. Topic Models for News Text Analysis

Topic models have been applied to diverse news text processing tasks, ranging from exploratory analysis and scientific news analysis to commercially viable applications such as news recommendation, summarization, and retrieval.

Exploratory analysis. These methods use topic models to create visualizations and browsing interfaces that enable users to gain insights into collections of news text. One approach to topic-based exploratory news analysis is to represent objects such as news outlets (Chuang et al., 2014) or storylines (Ahmed et al., 2011) in terms of topic weights. Such representations can be used to create informative topic-based object descriptions or to search for topically similar objects. Topic-document probabilities can be used to visualize temporal salience of topics, as a means of news corpus exploration (Newman et al., 2006) or for detecting and visualizing events (Dou et al., 2012). Furthemore, the probabilistic topic modeling framework can be used to extract relationships between modeled objects; for instance, Newman et al. (2006) treat extracted named entities as words, calculates topic-entity and entity-entity relatedness from conditional probabilities and creates graph-based visualizations. A recent survey of text visualization techniques by Kucher & Kerren (2015) reports a large interest for methods based on topic modeling.

Content analysis. Topic models have established themselves as a useful tool for quantitative content analysis within computational social science (Jacobi et al., 2016), owing to the fact that they can scale up the analysis to large document collections and alleviate the labor-intensive process of manual document categorization (document coding). In the context of news media, two typical use cases are media agenda (McCombs & Shaw, 1972) and news framing (Entman, 1993) analyses. These applications exploit the fact that models' topics often correspond to news issues. For instance, Kim et al. (2014) demonstrated the use of topic models for a comparative analysis of media agenda and public agenda, by comparing the salience of topics from news text against topics from user-generated texts. Similarly, Jacobi et al. (2016) used topic models to first identify the issues of interest and then analyzed how the framing of these issues has changed over time. However, as the correspondence between topics and news issues might in some cases be weak, Korenčić et al. (2015) proposed to address this problem by a semi-supervised method for media agenda analysis consisting of news issue discovery and measurement of issue salience. Because topic models are receiving increased interest in the social science community, Grimmer & Stewart (2013) pointed to a need for new methods for validating these tools before they can be adopted as standard. We see measures of topic coherence proposed in this paper as an important step toward that goal.

Other applications. Topic models have been used for various other news text analysis tasks, either by using topic-document or topic-word probabilities to construct or enrich features of the modeled objects or by using the model to directly calculate probabilistic relationships between them. For example, topic models have been used for news recommendation, by constructing topical features of both user preferences and news texts (Garcin et al., 2013; Li et al., 2014). Gao et al. (2012) proposed an event summarization method that uses a cross-collection topic model of news articles and tweets to produce summaries by ranking article sentences and tweets based on conditional probabilities derived

from the model. Shahaf & Guestrin (2012) proposed a system for interactive discovery of news storylines, represented as temporal sequences of news articles connecting two endpoint articles, where each news article is represented using features constructed from topic models. Yi & Allan (2009) experimented with enhancing news information retrieval using topic models for query expansion and the construction of a document language model. The above studies show that topic models outperform or perform on par with the state-of-art systems across a variety of use cases.

2.3. Evaluation of Topic Models

Topic models are only as useful as the quality of the topics they produce. As noted in the introduction, the downside of topic models is that the quality of the topics depends on a range of factors. First, topic modeling involves a number of design decisions, including which topic model structure and inference algorithm to use, how to set the model hyperparameters, and how to preprocess the text. Second, due to the stochastic nature of the inference process, the quality of the topics can greatly vary even for a single model. Automated topic model evaluation help in addressing both issues: it can be used to narrow down the set of possible design options and to identify high-quality models from among several runs.

The approaches to topic model evaluation may be divided into *extrinsic* (task-dependent) and *intrinsic* (task-independent). The former approach is used to evaluate the quality of a model in terms of how much it improves the performance on a downstream NLP task, such as information retrieval (Wei & Croft, 2006), word sense disambiguation (Boyd-Graber et al., 2007a), sentiment analysis (Titov & McDonald, 2008), or word similarity and document classification (Stevens et al., 2012). In contrast, the intrinsic approach evaluates the quality of the produced topics irrespective of an application. In this paper, we focus on automated intrinsic evaluation, which is more generally applicable than extrinsic evaluation.

Intrinsic evaluation methods may further be divided into four main categories: measures of fit, measures of stability, measures of match with a ground truth, and measures of topic quality. Measures of fit rely on the probabilistic structure of topic models to compute discrepancy between the model and the data. The most commonly used method from this category is to measure the perplexity of held-out text data with respect to the inferred model (Blei et al., 2003; Wallach et al., 2009). A more sophisticated method was proposed by Mimno & Blei (2011), who measure the discrepancy between empirical properties of the learned latent variables and properties expected from the probabilistic model structure.

Measures of stability are motivated by the variability in the inferred topics and the fact that model stability is a desirable property in a number of applications, most notably in social sciences. Stability of a set of models is calculated by averaging the similarity across pairs of models. Similarity can be computed by aligning similar topics of the two models (Waal & Barnard, 2008; Koltcov

et al., 2014; Belford et al., 2018), or by representing the models in terms of words or documents and comparing such representations (Belford et al., 2018).

Topic model evaluation can also be framed as a matching problem, in which model-produced topics are compared against ground-truth labels. In particular, (Ramirez et al., 2012) proposed the evaluation of topic models as clustering algorithms, by treating the topics as soft clusters. In contrast, (Chuang et al., 2013) proposed a framework for matching model topics directly to human-compiled concepts.

The measures of topic quality focus on computing quality scores of individual topics, which can then be aggregated to obtain model quality scores. AlSumait et al. (2009) defined a score of topic quality in terms of the distances between topic-document or topic-word distributions on one side and uninformative distributions (uniform and "vacuous") on the other. Chang et al. (2009) framed the evaluation of topic quality as a *word intrusion* task: human judges were asked to identify intruder words randomly inserted into a set of top topic words, with the idea that, for interpretable topics, the intruders will be easier to spot. The work of Lau et al. (2014) proposes a method to fully automate the original word intrusion task of Chang et al. (2009). Musat et al. (2011) calculated the "conceptual relevance" of topics, defined to measure the ease of attributing a concept to a topic. This is achieved by mapping top-ranked topic words to WordNet concepts and finding WordNet concepts that encompass these words, while being as specific as possible. A special class of topic quality measures are the topic coherence measures, described next.

2.4. Topic Coherence

Relevant to the work described in this paper is the line of research that uses *topic coherence* as a measure of topic quality. This line of research was motivated by the work of Chang et al. (2009), who proposed to measure the quality of model topics in terms of their interpretability. Chang et al. (2009) showed that models that fare better in predictive perplexity often have less interpretable topics, suggesting that evaluation should consider the internal representation of topic models and aim to quantify their interpretability.

The idea soon gave rise to a new family of methods that evaluate the semantic interpretability by measuring the topic coherence (Newman et al., 2010). These methods are word-based: top-ranked topic words are used as input for automatic coherence calculation methods and shown to human annotators that label topics with coherence scores. The majority of these methods calculate topic coherence by averaging pairwise semantic similarities of the top-ranked topic words (typically 5 or 10) or subsets thereof (Röder et al., 2015). The experimental evaluation is usually carried out using a ranking measure or a correlation coefficient (typically: AUC, Kendall's τ, Spearman's ρ, or Pearson's r) to assess the agreement between the calculated coherence scores and human-annotated coherence judgments (either binary coherent/non-coherent judgments or graded judgments on a Likert scale).

The main design decision for a coherence method that works by averaging pairwise similarities of top-ranked topic words is the choice of the word simi-

larity measure. Existing work uses WordNet- and Wikipedia-based word similarity (Newman et al., 2010), pointwise mutual information (Newman et al., 2010; Aletras & Stevenson, 2013; Lau et al., 2014), conditional word probability (Mimno et al., 2011), distributional vectors (Aletras & Stevenson, 2013), tf-idf (Nikolenko et al., 2015), and word embeddings (O'Callaghan et al., 2015; Nikolenko, 2016). Instead of calculating pairwise word similarities, Rosner et al. (2014) proposed the partitioning of the set of top-ranked topic words into subsets, and then averaged the similarities of subset pairs. Röder et al. (2015) proposed a generic framework for topic coherence measures based on aggregation of similarities of either word or subset pairs. By searching through the framework-induced space of measures, the authors derived several novel and efficient topic coherence measures. Two coherence methods from the literature fall outside of the described similarity-based framework: the measure of Ramrakhiyani et al. (2017), which clusters the embeddings of top-ranked topic words and approximates the coherence with the size of the largest cluster, and a set of measures proposed by Newman et al. (2010), which work by analyzing the results of querying a web search engine with top-ranked topic words.

The work presented in this paper falls under the category of topic coherence methods. However, our work differs from all of the above in that we measure topic coherence by reference to documents rather than words associated with a topic. In other words, the coherence measures we propose make use of document-topic distributions rather than word-topic distributions. As argued in the introduction, we hypothesize that characterizing topic coherence in terms of documents associated with the topics may be more adequate in some cases, especially for news media, which abounds with contingent and transient topics. Therefore, unlike all the prior work on topic coherence, we evaluate the proposed document-based coherence measures on datasets annotated specifically for document-based coherence, with topic coherence scores obtained upon inspection of topic-related documents rather than topic-related words.

One point that deserves additional comment is the definition of topic coherence. Existing work relies on operational definitions, by means of providing the annotators with instructions on how to assign coherence scores to topics. These instructions are descriptive and informal, based on topic's correspondence to a concept (Mimno et al., 2011), topic's interpretability (Aletras & Stevenson, 2013; Rosner et al., 2014), or both (Newman et al., 2010; Nikolenko et al., 2015). Other descriptors such as "meaningful", "easy-to-label", and "coherent" are often used to compound the definitions. It can be argued that the definitions that refer to topic correspondence to a concept are essentially equivalent to those that refer to topic interpretability, since interpretation necessarily involves conceptualization. With this in mind, we define a model topic as coherent if a human annotator can recognize its correspondence to a concept. Furthemore, we note that not all concepts can be represented by a model topic; e.g., non-topical concepts such as "breaking news", "good news", or "bad news" cannot be expected to emerge as topics of a topic model trained on a corpus of news articles. We will refer to the concepts that can be represented using topic models as *semantic topics*. In practice, not all topics produced by a topic model will correspond

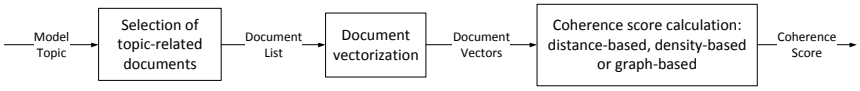

Figure 1: Three steps of the proposed document-based topic coherence measures.

to semantic topics. Two common types of erroneous topics are "mixed" and "noise" topics (Boyd-Graber et al., 2014). Mixed topics are a fusion of two or more otherwise coherent topics – the words and the documents associated with a mixed topic are a union of words and documents of the coherent sub-topics. Noise topics consist of random unrelated words and documents.

Technically, the work most similar to ours is that of AlSumait et al. (2009), who proposed a measure of topic quality based on document-topic distribution and combined it with a word-based measure to produce the final topic quality score. However, unlike AlSumait et al. (2009), we perform a quantitative evaluation of document-based measures using the document-based measure from AlSumait et al. (2009) as a baseline. Our evaluation procedure bears similarities to that of Ramirez et al. (2012), as both evaluation approaches use document-topic distributions, but whereas Ramirez et al. (2012) evaluated complete topic models using manually labeled documents, we evaluate the individual topics by computing their coherence scores.

3. Document-based Topic Coherence Measures

In this section, we describe the proposed document-based topic coherence measures. The calculation of each measure comprises three main steps: (1) selection of topic-related documents, (2) vectorization of the documents, and (3) calculation of a coherence score based on document vectors (Figure 1).

The first step takes as input a topic and document-topic probability matrix θ and outputs a list of documents. Top documents are selected by taking a fixed number of documents with the highest document-topic probabilities. The document vectorization step takes as input a list of text documents and outputs a list of vectors. Vectorization of documents' text is performed using standard bag-of-words or tf-idf vectorization or by aggregating the embeddings of document words. Finally, the vectorized documents are given as input to one of three types of methods for coherence scoring: distance-based, density-based, or graph-based methods. Distance-based methods aggregate pairwise document distances, while density-based methods use a vector probability density to approximate mutual closeness of documents. Graph-based methods derive a graph from document distances and calculate the coherence using one of several graph property measures. We next describe in detail each of the three steps of coherence measure calculation.

3.1. Selection of Topic-Related Documents

The aim of this step is to construct a list of text documents that are representative of a model topic – the documents that are associated with the topic in the context of the topic model. Selecting too many documents (in the extreme case, all the documents) will render the document list incoherent. Alternately, selecting too few documents (in the extreme case, a single top document for the topic) will likely make the list highly coherent.

We opt for a simple and model-independent strategy: given a topic, we select *TopDocs* documents with the highest document-topic probabilities. *TopDocs* is the parameter of the selection step and we optimize its value empirically. A similar strategy has proven effective in the case of word-based coherence measures, where selecting the top 10 words has shown to yield good results.

Recall from Section 2.1 that the document-topic distributions are represented by the document-topic probability matrix θ, where θ_{ij} is the probability of topic j in the i-th document. These probabilities represent the strength of association between the topics and the documents. Formally, for a topic j, the first *TopDocs* documents are chosen from the list of all documents $D_{i_1}, \ldots, D_{i_N}$ ordered by the probability of topic j in descending order ($\theta_{i_1 j} \geq \theta_{i_2 j} \geq \ldots \geq \theta_{i_N j}$).

3.2. Document Vectorization

The purpose of the vectorization step is to transform information contained in documents' texts into vectors that will be input to coherence scoring methods. These vectors must enable the scoring methods to approximate the degree in which a set of documents share a topic. Vector representations useful for clustering or clasifying documents into topical categories as well as for retrieving documents via topical queries are expected to also work well for calculating topical coherence of documents.

We experiment with two standard text vectorization methods commonly used in document classification and retrieval: word probabilities and tf-idf scoring (Schütze et al., 2008). The word probability and tf-idf vectors are derived from the news corpus used to build the topics under evaluation and are thus domain-specific. In addition, we experiment with generic (i.e., mixed-domain) vectors constructed via per-document aggregation of two types of word embeddings derived from a sizable external corpus: CBOW (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). The CBOW and GloVe word embedding vectors are widely used in NLP, while representing documents as vector aggregations has been found to work well in many taks, including clustering documents into topical categories (Zhang et al., 2018) and document retrieval based on topical queries (Galke et al., 2017), which are related to our task.

Word count vectorization. These vectorization methods rely on counting occurrences of words in text documents. This is preceded by document preprocessing, which at the minimum includes tokenization and case-folding, but may also include morphological normalization, such as stemming or lemmatization, and stop-word removal.

Let N denote the total number of documents in the corpus, c_{ij} the number of occurrences of word j in the i-th document, d_i the size of the i-th document, and dc_j the number of documents the word j occurs in. Probability vector $prob_i$ of the i-th document is a vector of empirical word-in-document probabilities, obtained as maximum likelihood estimates, $prob_{i,j} = c_{ij}/d_i$. Tf-idf representation (Salton & Buckley, 1988) combines word-in-document probabilities with frequencies of word occurrence in other corpus documents. We use the tf-idf variant in which the tf-idf vector $tfidf_i$ of the i-th document is defined as $tfidf_{i,j} = tf_{i,j} \times idf_j$ where $tf_{i,j} = \log(c_{ij}) + 1$ and $idf_j = \log((N+1)/(dc_j+1)) + 1$. In addition, we normalize document tf-idf vectors to a unit L2-norm.

Word embedding aggregation. This method of aggregation relies on pre-constructed word embeddings (Turian et al., 2010): low-dimensional, continuous-valued vectorial representations of words' meanings derived from word co-occurrences in a large text corpus. We experiment with the two most commonly used word embeddings: CBOW (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). CBOW embedding vectors are obtained by optimizing the log-linear prediction of words based on their context words. On the English dataset, we use pre-trained 300-dimensional CBOW embeddings, derived from a 100-billion-word Google News corpus. On the Croatian dataset, we use the word2vec tool to train 300-dimensional CBOW embeddings on the hrWaC corpus – a web corpus of Croatian texts (Ljubešić & Erjavec, 2011) totaling 2.8 billion words.[3] GloVe embedding vectors are obtained by approximating global word co-occurrence probabilities using a weighted least squares regression model. On the English dataset, we use pre-trained 300-dimensional GloVe embeddings, derived from Wikipedia and Gigaword corpora. On the Croatian dataset, we train 300-dimensional GloVe embeddings on the hrWaC corpus using the glove tool.[4] We compute the vector representation of a document text by adding up the word embedding vectors of all its content words. Optionally, we average the resulting vector to account for the differences in document lengths.

3.3. Coherence Scoring Method

After documents representative of a topic have been selected and vectorized, a list of document vectors is fed to a coherence scoring method. We experiment with three types of methods: (1) distance-based methods, which aggregate distances between document vectors, (2) density-based methods, which approximate coherence using a multivariate normal distribution as a model of document vectors, and (3) graph-based methods, which first construct a connectivity graph from document vectors and then compute a suitable graph measure. In total, nine scoring methods are proposed: two distance-based, two density-based, and five graph-based methods. While there are other types of measures we could

[3] CBOW pre-trained vectors for English and the word2vec tool are available from `https://code.google.com/archive/p/word2vec/`

[4] GloVe pre-trained vectors for English and the tool are available from `https://nlp.stanford.edu/projects/glove/`

have considered, we have chosen these tree types as they make different assumptions as to what contributes to coherence of a set of document vectors: mutual closeness (distance-based methods), compactness (density-based methods), or connectivity (graph-based methods).

3.3.1. Distance-based coherence

Distance-based methods rely on a measure of distance between vectors, i.e., a function $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^+$ that assigns a positive number to a pair of vectors. The distance measure does not have to be a metric in the mathematical sense, as long as it gives a useful notion of distance, such as the cosine distance.

We consider two simple distance-based methods: (1) average distance, which calculates an average of distances between all pairs of document vectors, and (2) distance variance, which calculates the average of distances between the document vectors and the center (mean) vector. In both cases, the final coherence score is produced by negating the average to convert a measure of dispersion into a measure of coherence.

3.3.2. Density-based coherence

The density-based coherence method works by first fitting a multivariate normal probability density function to the set of document vectors and then approximating coherence as the average log-density of the vectors under the model. The intuition is that, the higher the density, the tighter the grouping around the mode of the probability density function, and the higher the coherence of the document vectors.

The parameters of the multivariate density function are a mean vector, $\mu \in \mathbb{R}^n$, and a covariance matrix, $\Sigma \in \mathbb{R}^{n \times n}$. To reduce the number of parameters, and in turn prevent overfitting, we restrict the covariance matrix to either a diagonal matrix ($\Sigma = \text{diag}(\sigma_i^2)$, where σ_i^2 is the variance of the i-th vector component) or an isotropic matrix ($\Sigma = \sigma^2\mathbf{I}$). This is equivalent to assuming uncorrelated noise and isotropic noise, respectively.

We fit the probability density function to the document vectors using maximum likelihood estimate. Before fitting, we optionally perform dimensionality reduction on the document vectors using the standard principal component analysis (PCA).

3.3.3. Graph-based coherence

Graph-based methods first construct a graph of selected topic-related documents and then calculate a coherence score using a suitable graph measure. The nodes of the graph correspond to the documents, while graph edges are constructed using a measure of distance between document vectors. Graph measures we experiment with correspond to the notion of graph compactness or connectivity: closeness centrality, communicability centrality, clustering coefficient, number of connected components, and the size of the minimum spanning

tree.[5]

Graph construction. We experiment with two methods to construct edges between document nodes. The first method constructs a fully connected weighted graph with edge weights set to the distance between document vectors. The second method uses a distance threshold to connect only those pairs of documents whose distance falls below a given threshold. After thresholding, the remaining edges can retain their weights, or the graph can be converted into an unweighted graph.

Closeness centrality. The first graph measure we consider is closeness centrality (Freeman, 1978). Closeness centrality of a graph node v is defined as the inverse of the average shortest path distance between the node and all other reachable nodes:

$$cc(v) = \frac{|C(v)| - 1}{\sum_{w \in C(v)} d(v, w)} \tag{1}$$

where $C(v)$ is the set of all nodes reachable from the node v (the nodes in the connected component containing the node v). Closeness centrality of an isolated node ($C(v) = v$) is 0.

To avoid assigning high closeness centrality to nodes of a fragmented graph (a graph with many small connected components), the closeness centrality is normalized by the relative size of the node's connected component

$$cc_{\text{norm}}(v) = \frac{|C(v)| - 1}{N - 1} \frac{|C(v)| - 1}{\sum_{w \in C(v)} d(v, w)} \tag{2}$$

where N is the number of nodes in the graph.

We calculate the coherence score as the average normalized closeness centrality of all the graph nodes:

$$CC(G) = \frac{1}{N} \sum_{v \in G} cc_{\text{norm}}(v) \tag{3}$$

Subgraph centrality. Subgraph centrality (Estrada & Rodriguez-Velazquez, 2005) is a measure of node centrality correlated with the number of closed walks (cycles allowing node repetition) that start and end in a node. Let $\mu_k(v)$ denote the number of closed walks of length k originating in the node v. Subgraph centrality of a node v is defined as:

$$sc(v) = \sum_{k=1}^{\infty} \frac{\mu_k(v)}{k!} \tag{4}$$

[5] All of the measures considered here are available as part of the NetworkX (Schult & Swart, 2008) library available at `http://networkx.readthedocs.io`.

The number of closed walks $\mu_k(v)$ is scaled by $k!$ to ensure the convergence of the series. Subgraph centrality of a node can be efficiently computed via spectral decomposition of the graph's adjacency matrix. Edge weights are irrelevant for subgraph centrality because the measure is based on the number of walks, not the weights of the walks. For this reason, subgraph centrality is not applied to complete graphs, as all the graphs with the same number of nodes would be assigned the same centrality score.

We calculate the coherence score by averaging subgraph centralities of all the nodes:

$$SC(G) = \frac{1}{N} \sum_{v \in G} sc(v) \tag{5}$$

Clustering coefficient. The clustering coefficient of a node v is the number of actual triangles that go through the node v, denoted as $T(v)$, divided by the number of all possible triangles that could go through that node. A triangle through a node v corresponds to a set of three distinct nodes – v, u_1, and u_2 – such that edges vu_1, u_1u_2, and u_2v exist. The clustering coefficient is defined as:

$$cc(v) = \frac{T(v)}{\frac{deg(v)(deg(v)-1)}{2}} = \frac{2T(v)}{deg(v)\big(deg(v)-1\big)} \tag{6}$$

A weighted version of the clustering coefficient, which we apply to weighted graphs, is defined as (Saramäki et al., 2007):

$$cc(v) = \frac{1}{deg(v)\big(deg(v)-1\big)} \sum_{u_1,u_2} \big(w'(v,u_1)w'(u_1,u_2)w'(u_2,v)\big)^{1/3} \tag{7}$$

Here, the sum is over all pairs of nodes that close a triangle with v, and $w'(u,v)$ is the weight of the edge between nodes u and v divided by the maximum edge weight in the graph.

We calculate the coherence score by averaging the clustering coefficients of all the graph nodes:

$$CC(G) = \frac{1}{N} \sum_{v \in G} cc(v) \tag{8}$$

Connected components and spanning tree. We experiment with two additional measures based on the connectivity structure of the graph. The first measure is the inverse of the number of connected components in the graph. We use this measure only in combination with thresholded graph construction, as for complete graphs, the number of connected components is always one. The second measure is the negative weight of the minimum spanning tree of the graph. We use this measure only in combination with a non-thresholded, fully connected weighted graph.

4. Experiments with Document-Based Coherence

We now proceed with evaluating and analyzing the proposed document-based coherence measures. To this end, we use two datasets of topics manually annotated for document-based coherence. After defining a set of coherence measures that correspond to sensible parameter values, we select the best measures on a development set and evaluate these on two test sets, using area under the ROC curve (AUC) as the performance metric. Lastly, we analyze the best-performing measures. We begin by describing the construction of the datasets.

4.1. Datasets

As argued in the introduction, document-based topic coherence has the potential to better capture the semantic interpretability of news topics than word-based coherence. The standard way of evaluating the proposed document-based coherence measures is to compare their coherence scores against coherence judgments obtained from human annotators. As we are specifically interested in evaluating document-based coherence measures, the human judgments should also be based on topic-related documents rather than just topic-related words.

To the best of our knowledge, no dataset that meets the above desiderata is publicly available. Hence, we created our own dataset of news text topics manually annotated with document-based topic coherence judgments. One can obtain the coherence judgments directly, be asking the annotators to judge the coherence of each topic, or indirectly, by first asking the annotators to semantically interpret and label each topic, and then use these labels to derive the coherence judgments. We chose the latter approach as it allowed us to reuse existing datasets with manually labeled topics.

As a starting point, we used the dataset of Korenčić et al. (2015), which contains LDA topics derived from a corpus od 24k political news articles from mainstream US news outlets. Korenčić et al. ran five LDA models initialized with different random seeds – three models with 50 topics and two models with 100 topics. The so-obtained topics were then pooled together, similarly as in (Lau et al., 2014), to obtain a final set of 350 topics. Having a pooled set of topics allows us to evaluate the coherence measures on a more diverse set of topics.

The topics were then manually labeled by two annotators, with the annotation procedure set up as follows. The annotators were instructed to inspect a model topic represented as a list of article titles and words and to infer, if possible, the corresponding semantic topics. A semantic topic was described as either an abstract concept or a concept corresponding to an entity, event, or a story. After inspecting a model topic, the annotators consulted a shared list of semantic topics discovered so far and either updated the list with new topics or re-used the existing ones. The model topic was then labeled with semantic topics and, if the topic contained unrelated random documents or words, with an additional "noise" label. For calibration, annotators labeled a shared set of 50 topics and updated the labeling conventions. The top-ranked documents were

presented to the annotators as a list of titles sorted by document-topic probabilities, with the full texts of articles available for inspection. The annotators were instructed to inspect the documents in sorted order and to stop when the document-topic probability fell below 10%. Topic-related words were presented as a list of top 20 topic words. For labeling, the annotators relied primarily on documents, as the words proved either seemingly unrelated, vague, or too abstract whereas lists of well-formed document titles provided more accurate and specific information. Consequently, the decisions on the semantic topics and their correspondence to model topics, as well as the decision on existence of noise, were made based on topic-related documents, while the words at best served to confirm this decision. In summary, 52% of the model topics were labeled with one semantic topic, 15% were labeled with one semantic topic and the noise label, 17% were labeled with two semantic topics, 4% were labeled with two semantic topics plus noise, while 12% were labeled as noise.

Using the above-described labeled dataset as input, we defined as *coherent* all topics that have been annotated with a single semantic topic, possibly with the addition of noise, and as *incoherent* otherwise.[6] In other words, we consider a topic as coherent as long as a human annotator can recognize that the topic corresponds to a single semantic topic, which is in line with the definitions of topic coherence used by Newman et al. (2010) and Mimno et al. (2011), whereas an incoherent topic corresponds either to noise or to a mixture of two or more semantic topics. This resulted in 235 topics (67%) being labeled as coherent and 115 topics (33%) labeled as incoherent. On a sample of 50 topics annotated by both annotators, the annotators agreed on 88% of the topics, while the chance-corrected kappa agreement coefficient (Landis & Koch, 1977) is 0.674.

We randomly split the described set of 350 topics into two subsets: the *development set*, containing 120 topics, and the *test set*, containing 230 topics. We use the development set to optimize the parameters of the coherence measures and the test set for final evaluation of these measures. To ensure that both the development and the test sets are representative of the entire dataset, the split was stratified across the five labels: single semantic topic, semantic topic plus noise, two semantic topics, two semantic topics plus noise, and noise.

In addition to the US news topics dataset described above, we introduce a second dataset that serves as an additional test set for assessing the robustness of the results. We refer to this second dataset as *test-cro*. The test-cro dataset consists of topics derived from a corpus of news texts in Croatian language, originally compiled for a media agenda setting study in (Korenčić et al., 2016). The topics were pooled from four LDA models – three models with 50 topics and one model with 100 topics, which resulted in a total of 250 topics.[7] Coherence

[6] Following (Nikolenko et al., 2015; Nikolenko, 2016), we work with binary judgments of coherence, since obtaining these from existing labels is rather straightforward. The alternative would have been to convert the existing labels into graded judgments of topic coherence, but it is not clear how one would proceed about this.

[7] Korenčić et al. (2016) originally used 200 topics, however, to make the size of the test-cro dataset comparable to that of the test set, we built and labeled an additional model with 50

labels were derived by the same procedure used for the US topics – model topics were first annotated with semantic topics and the "noise" label, after which the coherence labels were derived from the annotations. Of the 250 topics, 166 topics (66%) were labeled coherent, whereas 84 topics (34%) were labeled incoherent.

We make available both the US news topics dataset and the Croatian news topics dataset.[8]

4.2. Evaluation of Coherence Measures

We use the Area Under the ROC Curve (AUC) metric (Ling et al., 2003) to evaluate the performance of topic coherence methods. AUC is a metric employed for both classification and ranking evaluation. As a ranking measure, it has been used to evaluate word-based topic coherence (Nikolenko et al., 2015; Nikolenko, 2016) by comparing binary topic labels (coherent or incoherent) against the numerical coherence scores produced by the coherence measures. We use the AUC metric because it is well suited for comparing binary judgments of coherence, obtained by our topic labeling method, with real-valued coherence scores.

Generally, given a model M producing confidence scores and data points $x \in D$ labeled with binary class labels, AUC corresponds to the probability that, for two points x and x' such that x belongs to the positive and x' belongs to the negative class, the positive one receives a higher score, i.e., $M(x) > M(x')$ (Nikolenko et al., 2015). In the case of coherence measures, with topics labeled as either coherent (positive class) or incoherent (negative class), AUC of a coherence measure Coh is the probability that, for a coherent topic t and an incoherent topic t', the coherent topic gets a higher coherence score, i.e., $Coh(t) > Coh(t')$. The AUC scores are confined to the $[0, 1]$ interval, with 0 and 1 being the worst and the best score, respectively, and 0.5 being the expected score of an uninformative random measure.

The alternative interpretation of AUC rests on the idea that a model M producing confidence scores can be converted into a binary classifier by thresholding its output. The false positive rate (fall-out) and the true positive rate (recall) of the classifier for different threshold values define the receiver operating characteristics (ROC) curve. The performance of a perfect classifier corresponds to point $(0, 1)$ (no fall-out, complete recall). In contrast, the performance of a random classifier for different threshold values corresponds to a straight line from $(0, 0)$ to $(1, 1)$. Given an ROC curve, AUC is defined as the area under the ROC curve. For the case of a binary classifier for topic coherence based on thresholding a coherence measure, the recall corresponds to the proportion of coherent topics detected by the classifier, while fallout corresponds to the proportion of incoherent topics falsely detected as coherent.

topics.

[8] `https://rebrand.ly/doc-coh-dataset`

Scoring	Vectorization	# Measures
Distance	Cnt	48
Distance	Embd	80
Density	Cnt	96
Density	Embd	128
Graph	Cnt	936
Graph	Embd	1560

Table 2: Six categories of coherence measures, each corresponding to one combination of the coherence scoring and document vectorization methods, along with the number of distinct measures considered in the evaluation.

4.3. Baseline Method

We use as the baseline the document-based measure of "topic significance," proposed by AlSumait et al. (2009). To the best of our knowledge, this measure is the only document-based measure of topic quality. The measure represents each topic as a probability distribution over the set of corpus documents obtained by normalizing topic-document probabilities. Topic significance is then calculated as the distance between the described distribution and the uninformative uniform distribution, using either cosine or KL-divergence as the distance measure. We use the variant based on the cosine distance, since it performs better on both datasets. In (AlSumait et al., 2009) this measure is not evaluated on its own but instead combined with similarly defined word-based measures in a composite measure of topic quality, which is then evaluated qualitatively by an inspection of high- and low-scored topics.

4.4. Model Selection

Our goal is to identify the well-performing document-based coherence measures from Section 3: those that have a high correlation with human-provided scores of document-based coherence of model topics, as measured by the AUC score. To this end, we first introduce a set of parameters that describe the structure of these measures. We then proceed to define a set of sensible parameter values corresponding to a set of coherence measures that will be considered in further evaluation.

Coherence measure categories. To ease the analysis, we group the coherence measures into six categories, as shown in Table 2. Each category corresponds to a pairing of two attributes: the coherence scoring method and the document vectorization method. We consider these two attributes to be the most distinguishing properties of a coherence measure.

The first attribute is the coherence scoring method, which essentially determines how the documents are viewed (as points in a vector space or as nodes in a graph) and how to estimate the coherence score from a set of documents (cf. Section 3.3). Distance-based methods (Distance) rely on a measure of vector distance and aggregate the distances directly, while density-based methods

(Density) rely on a probability density defined on the space of vectors and calculate the dissipation of vectors around the center of distribution. In contrast, graph-based methods (Graph) add structure to the set of vectorized documents by constructing a document graph using a measure of vector distance for edge definition and calculate the coherence using measures describing various graph properties.

The second attribute, the document vectorization method (cf. Section 3.2), determines the representation of the top-ranked topic documents that are given as input to the coherence scoring method. Here, we distinguish between two types of vectorization methods: Cnt and Embd. The former methods are based on word-in-document counts (normalized bag-of-words and tf-idf) derived from the same corpus that was used to build the topic model under evaluation. Document preprocessing we use to obtain word-in-document counts consists of stemming and stop-word removal. In contrast, Embd vectorization methods refer to representations based on the aggregation of word embeddings (CBOW and GloVe), which have been derived from a large, external corpus. Apart from the difference in how the vectors are constructed in these two cases, an important difference is that the Cnt vectors are domain-specific (in our case: the domains of US and Croatian political news), whereas Embd vectors are generic (i.e., mixed-domain). This difference is likely to have an influence on the coherence measure calculation: in contrast to domain-specific vectors, the generic vectors will generally be more ambiguous and might not correspond to the domain-specific senses of some words.[9] Alternately, generic vectors might better capture the meaning of rare words, and might generally be statistically more reliable because they are derived from a larger corpus.

Coherence measure parameters. To allow for a systematic analysis of the document-based coherence measures proposed in Section 3, we parametrize these measures with respect to the selection of topic-representative documents, the vectorization method, and the coherence scoring method. Table 3 outlines the parameters and their values considered in subsequent experiments. The parameters are broken down by coherence scoring method except for the first three parameters, which are shared by all methods. The shared parameters together define how a model's topic is transformed into a set of document vectors, while the method-specific parameters define the details of the coherence score calculation. For a more precise description of the parameters, the reader is referred to Section 3.

Note that not all parameter-value combinations are sensible. More specifically, *EmbeddingAgg* is applicable only if *DocVect* is `cbow` or `glove`. For Density scoring method, if *DocVect* equals `bow` or `tfidf` (high-dimensional vectors), the value of *DimReduce* is varied among the full set of values (5, 10, 20, 50, 100), while if *DocVect* equals `cbow` or `glove` (vectors of lower dimen-

[9] As many polysemous words have domain-specific senses, restricting the domain from which the representations are derived will typically decrease the word-level ambiguity. This sense-domain relation has also been leveraged for improving word sense disambiguation, e.g., (Magnini et al., 2002).

Scoring	Parameter	Values
(All)	*TopDocs*	10, 25, 50, 100
	DocVect	`bow`, `tfidf`, `cbow`, `glove`
	EmbeddingAgg	`average`, `sum`
DISTANCE	*DistanceMeasure*	`l1`, `l2`, `cosine`
	DistanceAgg	`average`, `variance`
DENSITY	*CovMatrix*	`isotropic`, `diagonal`
	DimReduce	`None`, 5, 10, 20, 50, 100
GRAPH	*DistanceMeasure*	`l1`, `l2`, `cosine`
	GraphAlgo	`closeness-centrality`, `subgraph-centrality`, `clustering`, `num-connected`, `min-spanning-tree`
	DistanceThresh	`None`, 0.02, 0.05, 0.1, 0.25, 0.5, 0.75
	Weighted	`True`, `False`

Table 3: Coherence measure parameters and values for the different coherence scoring methods. The first three parameters are shared by all three scoring methods.

sion), the value of *DimReduce* is varied among the values 5, 10, and 20. The *DimReduce* value of `None` is combined with all the vectorization methods, resulting in a vector size of approximately 24k dimensions for `bow` and `tfidf`, and 300 dimensions (the original size of the word embeddings) for `cbow` and `glove`. Similarly, not all parameter combinations make sense for the GRAPH scoring method: for fully connected weighted graphs (*DistanceThresh* set to `None`), only the `closeness-centrality`, `clustering`, and `min-spanning-tree` methods are sensible. For thresholded graphs (*DistanceThresh* set to a positive real number), `subgraph-centrality` and `num-connected` are used only for unweighted graphs (*Weighted* set to `False`). Table 2 lists the numbers of sensible parameter combinations for each of the measure categories. In total, we consider 2,848 distinct coherence measures.

Another point worth mentioning is the treatment of the *DistanceThresh* parameter for the GRAPH scoring method. To account for the fact that different distance measures and vectorization methods generally yield distances at different scales, we proceed as follows. For each combination of *DistanceMeasure* and *DocVect*, we estimate the distribution of distances from a random sample of 100k document-vector pairs from the corpus. We then treat *DistanceThresh* as a percentile rank from the so-obtained distribution, varying the threshold among the values 0.02, 0.05, 0.1, 0.25, 0.5, and 0.75.

4.5. Results

Table 4 shows the AUC scores on the test and test-cro sets for each category of coherence measures. For each category, we report the AUC score of the best-performing coherence measure from that category on the development set. The best-performing measures are chosen from the set of measures defined by

Measure Category		test		test-cro	
Scoring	Vectorization	AUC	p-value	AUC	p-value
Graph	Cnt	**0.804**	–	**0.812**	–
Distance	Cnt	0.754	0.001	0.785	0.028
Density	Cnt	0.745	0.000	0.774	0.009
Distance	Embd	0.732	0.001	0.746	0.029
doc-dist-cosine	–	0.730	0.006	0.748	0.025
Density	Embd	0.728	0.001	0.725	0.005
Graph	Embd	0.694	0.003	0.671	0.000

Table 4: Coherence measures' AUC scores for each of the six categories and the baseline, ordered by the score on the test set. The p-values are derived by comparing the AUC score from the first row with the AUC score from the other six rows.

parameters described in Section 4.4. Note that we carry out no such optimization on the test-cro dataset, as we want to test the cross-dataset robustness of the measures' parameters. The *doc-dist-cosine* on both datasets is the baseline method (cf. Section 4.3).

As seen from Table 4, the best-performing measure comes from the Graph-Cnt category, which achieves an AUC score of over 0.8 on both test and test-cro sets and outperforms other measures by at least 0.027 AUC. We use the DeLonge's test (DeLong et al., 1988)[10] to test the statistical significance of differences between the AUC score of the Graph-Cnt measure and that of the other seven measures, including the baseline; the p-values are shown in Table 4 next to the corresponding AUC scores. Two observations follow from Table 4. The first is that the ordering of the measures by AUC scores is almost perfectly consistent for the two datasets. The second observation is that on the test-cro set the Cnt-based measures and the baseline achieve higher AUC scores than on the test set.

Figure 2 shows the ROC curves of the best-performing coherence measures on the test dataset. As described in Section 4.2, an ROC curve measures the classification performance of a coherence measure: each point on a curve corresponds to a classifier based on a coherence measure paired with a coherence threshold used to decide whether a topic is coherent or incoherent. Figure 2 contrasts the best-performing coherence measures against the baseline measure (green curve) and the globally best Graph-Cnt measure (red curve). The ROC curves show that all the measures perform better than random classifiers. Furthermore, the ROC curves complement Table 4 in showing the performance gap between the Graph-Cnt and other measures. For the Graph-Cnt measure, the recall of 0.8 or above may be achieved with fall-out of at least 0.33. This

[10]DeLong's test is designed to compare the AUCs of two correlated ROC curves (ROC curves derived from different measures applied to the same data points). We use the implementation from the pROC R package (Robin et al., 2011).

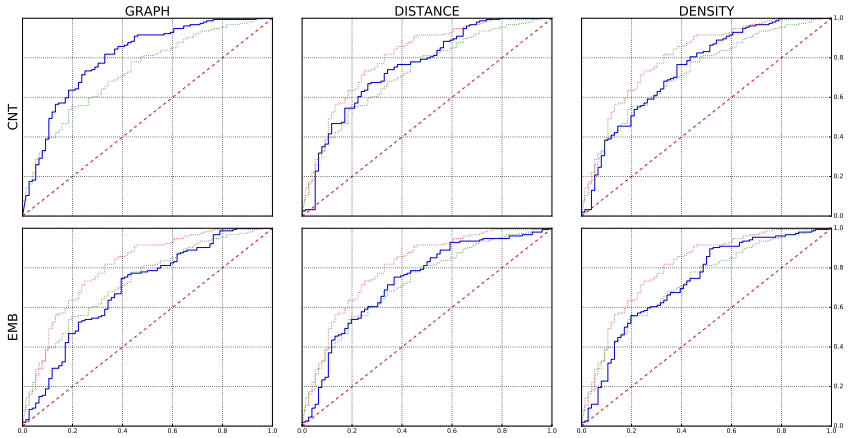

Figure 2: ROC curves of the best-performing coherence measures from the Table 4. The curve of the baseline measure is plotted in green on each plot. The curve of the measure from the top-performing category (top left) is plotted in red on other measures' plots.

means that if one wants a Graph-Cnt-based classifier to detect 80% of the coherent topics, the tradeoff is to have at least 33% of the incoherent topics classified as coherent. The other measures can achieve the recall of 0.8 with fall-out rates close to 0.5 (with the exception of Density-Cnt yielding 0.43 fall-out). On the other hand, if one wishes to achieve fall-out rates below 0.2 (reliable detection of incoherent topics), the Graph-Cnt measure can achieve this with a recall of 0.64 or below, while for the other measure the recall would be at best 0.56.

As seen from both Table 4 and Figure 2, the baseline measure described in Section 4.3 is a strong baseline – its performance is only slightly worse than all the proposed measures, except for the Graph-Cnt measure, which markedly outperforms the baseline.

In the above evaluation, each of the measure categories was represented by a single best-performing measure on the development set. This rises the question whether the chosen measures are indeed representative as being the best measures within their respective categories. To answer this question, from each of the six categories we selected the top 10 or top 5% (depending on the category size) measures with the best performance on the development set (instead of a single best-performing measure per category) and evaluated these on the test and test-cro sets. Our analysis showed that the best-performing measures from Table 4 are indeed, with one exception, representative of their categories, achieving performance comparable to the measures with the best test and test-cro performance within the same category. An exception is the Graph-Embd category, where the measure that performs the best on the development set does not perform well on either the test or the test-cro set. However, the selected Graph-Embd measures with the top test and test-cro performances

Category	Parameters	AUC score Dev	Test
Graph-Cnt	*DocVect*=`tfidf`, *TopDocs*=50, *DistanceMeasure*=`l2`, *DistanceThresh*=0.05, *Weighted*=`False`, *GraphAlgo*=`subgraph`	0.778	0.812
	DocVect=`tfidf`, *TopDocs*=50, *DistanceMeasure*=`cosine`, *DistanceThresh*=0.02, *Weighted*=`False`, *GraphAlgo*=`subgraph`	0.782	0.804
Graph-Embd	*DocVect*=`cbow`, *TopDocs*=50, *DistanceMeasure*=`cosine`, *DistanceThresh*=0.25, *Weighted*=`False`, *GraphAlgo*=`subgraph`	0.730	0.766
	DocVect=`glove`, *TopDocs*=50, *DistanceMeasure*=`l1`, *EmbeddingAgg*=`average`, *DistanceThresh*=0.25 *Weighted*=`True`, *GraphAlgo*=`clustering`	0.792	0.694
Distance-Cnt	*DocVect*=`bow`, *TopDocs*=50, *DistanceMeasure*=`cosine`, *DistanceAgg*=`average`	0.735	0.754
	DocVect=`bow`, *TopDocs*=50, *DistanceMeasure*=`cosine`, *DistanceAgg*=`variance`	0.739	0.754
Distance-Embd	*DocVect*=`cbow`, *TopDocs*=50, *DistanceMeasure*=`cosine`, *DistanceAgg*=`average`	0.711	0.746
	DocVect=`glove`, *TopDocs*=25, *DistanceMeasure*=`cosine`, *DistanceAgg*=`variance`	0.719	0.732
Density-Cnt	*DocVect*=`tfidf`, *TopDocs*=50, *DimReduce*=100, *CovMatrix*=`isotropic`	0.704	0.745
	DocVect=`tfidf`, *TopDocs*=50, *DimReduce*=`None`, *CovMatrix*=`isotropic`	0.704	0.745
Density-Embd	*DocVect*=`cbow`, *TopDocs*=25, *DimReduce*=5, *EmbeddingAgg*=`average`, *CovMatrix*=`isotropic`	0.701	0.734
	DocVect=`cbow`, *TopDocs*=25, *DimReduce*=10, *EmbeddingAgg*=`average`, *CovMatrix*=`isotropic`	0.708	0.728

Table 5: Parameter values of the best-performing coherence measures. For each of the six categories two best-performing measures are shown: one best performing on the development set and the other best-performing on the test set.

perform markedly better. A detailed analysis of the representativeness of the best development set measures is provided in the supplementary material.[11]

We now turn to a question of practical importance: which parameter values yield the best measures, i.e., which measures should one apply in practice? In Table 5, we give an indicative summary which shows the parameter values for two best-performing measures from each of the categories – one with the best performance on the development set and another with the best performance on the test set. A detailed analysis of the parameters of the best measures is provided in the supplementary material,[11] while here we summarize only the most interesting findings. Interestingly, all of the best-performing Graph-Cnt mea-

[11] `https://rebrand.ly/doc-coh-supplementary`

sures use thresholding to filter the edges, but ultimately use an unweighted graph. Of the five proposed graph-based coherence algorithms (cf. Section 3.3.3), three algorithms emerge as components of the best-performing measures: `subgraph-centrality`, `closeness-centrality`, and `clustering-coefficient`. Common to all three algorithms is that they calculate local connectivity scores of graph nodes and average them to obtain the graph score, unlike the other two algorithms (`num-connected` and `min-spanning-tree`), which calculate global graph connectivity properties. As regards the edge threshold value, `subgraph` and `closeness` centrality algorithms generally prefer smaller thresholds yielding sparse graphs (percentile ranks 0.02, 0.05, and 0.10), while the `clustering` algorithm seems to prefer higher thresholds (percentile ranks 0.25 and 0.5). The best Graph-Embd measures have the same structure, but fail to achieve the performance of the best Graph-Cnt measures. All the best measures that use Embd-based document vectorization in combination with `l1` and `l2` distances perform the aggregation of document vectors by averaging instead of summing the word embeddings. This is expected because the `l1` and `l2` distance measures, unlike the cosine distance, are not invariant to vector length. Generally, representing documents with Cnt vectors seems preferable over Embd vectors. More specifically, for graph-based measures the choice of Cnt vectors over Embd vectors leads to a large increase in performance, yielding globally best performance.

4.6. Conclusions

The main finding from the above experiments is that the best overall performance is achieved by coherence measures from the Graph-Cnt category, i.e., measures that rely on count-based document vectorization derived from an in-domain corpus and combined with graph-based algorithms for coherence scoring. This result was confirmed on both the US news topics dataset and the Croatian news topics dataset. Measures from the Graph-Cnt category construct graphs of top topic documents by removing all edges above a small distance threshold and calculate the coherence score by aggregating local node connectivity information, suggesting that the most effective way to estimate coherence of a set of documents is to use vectors of word-in-document counts to represent the documents and average local similarities in the neighborhoods of each document. In contrast, the measures from the Graph-Embd category, which fail to reach the performance of best Graph-Cnt measures, also average local similarities but use embedding-based document vectorization. Measures from other four non-Graph categories calculate the global similarity of the entire set of documents.

Another observation arising from the experiments is that, not surprisingly, the choice of document vectorization method has a strong influence on the performance of the coherence measures. These observations indicate that improvements to existing methods could be achieved by means of some other vectorization method.

Finally, the choice of top 50 topic-related documents as input for calculation of a topic's coherence score emerged as the best option that results in

measures achieving best or nearly-best results for all the measure categories, as demonstrated by the results in Section 4.5 (and additional analyses provided the supplementary material).

5. Experiments with Word-based Coherence

The previous experiment has investigated the efficacy of document-based coherence measures. In this section, we turn to the question of how document-based coherence measures, which we propose in this work, relate to the word-based coherence. While in the introduction we argued that document-based coherence measures may be better suited for topics derived from news media texts, the objective of the experiments described in this section is to quantify the magnitude and consistency of these gains. To this end, we measure how well the state-of-art word-based coherence measures approximate document-based coherence of topics derived from news media texts. Lastly, to gain additional insight into differences between the two types of measures, we perform a qualitative analysis of topics coming from the high and low ends of the document- and word-based coherence spectrum.

5.1. Word-Coherence Measures

As a reference point for selecting the word-based coherence measures, we use the study of Röder et al. (2015), who carried out a detailed and systematic analysis of a large number of word-based coherence measures on six publicly available datatasets. For our experiments, we select five top-performing measures from this study, designed to predict coherence scores formed by inspection of top-ranked topic words: (1) the C_{UCI} measure of Newman et al. (2010), (2) the C_{NPMI} measure of Aletras & Stevenson (2013), (3) the C_A measure of Aletras & Stevenson (2013), and the (4) C_V and (5) C_P measures, both discovered by an exhaustive search of the parameter space in (Röder et al., 2015).[12]

In the generic framework of Röder et al. (2015), the above measures are defined by an appropriate partitioning of the set of topic words into subsets, followed by computing the average of an appropriate similarity measure[13] between all pairs of so-obtained subsets. This includes the calculation of similarity of word pairs as a special case. The pairwise similarities between word subsets are calculated from probabilities based on word co-occurrences derived from a corpus, which are affected by the choice of the corpus and the choice of the corpus preprocessing method (tokenization, stopword removal, and word normalization).

The measures C_{UCI} and C_{NPMI} average the similarity of word pairs computed using pointwise mutual information (PMI) and its normalized version (NPMI),

[12] All five considered measures are implemented in the open-source software package Palmetto, available at `https://github.com/dice-group/Palmetto`.

[13] Röder et al. (2015) use the term "confirmation measure" to refer to a similarity measure between subsets of topic words.

respectively. A sliding window is used to derive word co-occurrence for both measures. The C_A measure also averages the similarity of word pairs, with the difference that the words in the pair are first represented as vectors of NPMI-based similarities with other top-ranked topic words and the similarity of the pair is calculated as the similarity of these two vectors. The C_V measure averages the similarities between pairs each comprising a topic word and its complement set that consists of all other topic words. Similarly, as for C_A, indirect vector similarity is used to compare a word with its complement set. Lastly, the C_P measure averages similarities of pairs each comprising a topic word and all topic words ranked above it by the word-topic probabilities, using as similarity a measure based on conditional probability of a single word given a word set (Fitelson, 2003).

5.2. Estimating Document-Based Coherence with Word-Based Coherence

In this experiment, we examine how well the top-performing word-based coherence measures approximate document-based coherence scores. Namely, word-based coherence measures from Section 5.1 are designed to predict, using as input a set of top-ranked topic words, coherence scores of topics assigned by human annotators based on the inspection of top-ranked topic *words*. In contrast, document-based coherence scores are assigned by annotators who inspected the top-ranked topic *documents*, and accordingly document-based measures considered in Section 4 use as input the top-ranked topic documents. We evaluate word-based coherence measures in the same way as we have evaluated the document-based measures, namely, on the test and test-cro sets of manually annotated topics using AUC score as the performance measure (cf. Sections 4.1 and 4.2).

We use the measures with the parameters optimized for word-based coherence estimation (Röder et al., 2015), all of which use top-10 topic words as input and derive the co-occurrence counts from Wikipedia. For the US news topics dataset, we derive the counts from the English Wikipedia (dump from June 2016), while for the Croatian news topics dataset, we derive the counts from the Croatian Wikipedia (dump from November 2017). In both cases the counts are derived with the preprocessing we used when we built our topic models.[14] Before preprocessing, we removed the redirection, disambiguation, category, and portal pages from both Wikipedia datasets.

Table 6 shows the performance of state-of-art word-based coherence measures compared against the baseline document-based coherence method *doc-dist-cosine* (cf. Section 4.3). The accompanying p-values are obtained using the DeLonge's test (cf. Section 4.5) with the null hypothesis of no difference between the AUC scores of a word-based measure and that of the baseline. As is evident from the results, document-based coherence measures outperform word-based

[14] For the US topics we also tried using the original counts used by Röder et al. (2015), available online, but the counts obtained with our preprocessing turned out to give better AUC scores for all the measures.

	test		test-cro	
Measure	AUC	p-value	AUC	p-value
doc-dist-cosine	0.730	–	0.748	–
C_V	**0.607**	0.002	0.508	0.000
C_A	0.579	0.001	0.442	0.000
C_P	0.548	0.000	**0.614**	0.009
C_{NPMI}	0.498	0.000	0.595	0.002
C_{UCI}	0.482	0.000	0.571	0.001

Table 6: Performance of word-based coherence methods in estimating document-based coherence, compared against the document-based baseline coherence measure.

measures by a considerable margin – the best word-based measures achieve AUC scores slightly above 0.6, while the document-based baseline achieves scores of at least 0.73. Partitioning the set of top-ranked topic words into pairs of words and word sets, as implemented by C_V and C_P measures, seems to yield somewhat better AUC scores than partitioning into word pairs, as implemented by all other word-based coherence measures.

5.3. Qualitative Analysis

The previous experiment has shown that, when comparing to the proposed document-based coherence measures, state-of-art word-based coherence measures fall short of estimating document-based coherence. This raises the question of what is the relation between these two approaches to measuring coherence: is document-based coherence simply a better model of topic coherence, or are word- and document-based coherence two different, although correlated and possibly complementary views on topic coherence?

To investigate which of the two is the case, we carried out a qualitative analysis of the topics from our US news topics dataset. The analysis is done along two dimensions: document-based coherence and word-based coherence. In each of the two dimensions, we select topics of high coherence and topics of low coherence, giving us four categories of topic coherence. The topics are selected from a sample of 230 topics constituting the test set (cf. Section 4.1), based on scores produced by the document- and word-based coherence measures. More specifically, for document-based coherence we select from the top 30% and bottom 30% of topics ranked by the best-performing measure from the GRAPH-CNT category (cf. Section 4.5), while for word-based coherence we do the same using the C_P measure (this measure performs well on word-based coherence but poorly on document-based coherence; cf. Table 6). Furthermore, for the purpose of this analysis, we manually categorize each semantic topic as either *concrete* (topics pertaining to entities, events, or news stories) or *abstract* (topics pertaining to general issues or abstract semantic categories).

High document/low word coherence. In this category, we find 23 topics from the test set, the majority (21) of which are concrete. Table 7 gives five examples,

Topic label	Top-10 topic words
Chicago mayoral election	mayor chicago emanuel de giuliani garcia love blasio rudy runoff
Restoration of ties with Cuba	foundation cuba list malley summit rubio cuban donations trump island
Ted Cruz	cruz ted liberty tea imagine evangelical r-texas declared candidacy obamacare
Iran negotiations	nuclear agreement sanctions iranian weapons kerry framework tehran cotton corker
Vaccination	vaccines parents science kids choice huffpost carson measles research believes

Table 7: Topics with high document coherence and low word coherence.

where all but the "vaccination" topic are concrete. The low word-based coherence scores are in line with the observation that the top-ranked topic words are, as a whole, semantically unrelated. However, high document-based coherence scores are a consequence of the topic-related documents being highly similar news articles describing the same entity, event, or story. Note that, since a state-of-art word-based coherence measure is unable to recognize the coherence of these topics, they would likely be discarded as low-quality topics. Likewise, a human annotator – unless familiar with the news corpus – would judge the set of top-ranked words as unrelated and the topics as incoherent. Hence, high document/low word coherence topics are a paradigmatic case for the use of document-based coherence measures.

High document/high word coherence. Among the 20 topics in this category, 14 are abstract and 6 are concrete. This suggests that topics scored with high word-based coherence tend to be abstract. The top-ranked topic words of such abstract topics are semantically related. Table 8 shows five example topics from this category. Among these, two are concrete topics pertaining to entities, which shows that concrete topics can also have high word coherence if they are characterized by words relating to a common concept. On the other hand, the fact that 14 abstract topics are also scored as coherent by the document-based coherence measure suggests that document-based coherence can detect the coherence of not only concrete but also abstract semantic topics.

Low document/high word coherence. From the 12 topics in this category, three are incoherent, while the rest are abstract topics. This again confirms the correlation between topic abstractness and word-based coherence. The abstract topics fall into two groups. The first is a group of four topics (lawsuits, journalism, social media, and radio & television) whose low document coherence is due to topics being unrepresented in the documents, i.e., they are mentioned in a relatively small portion of document text otherwise dominated by other topics.

Topic label	Top-10 topic words
Environment	climate energy global science environmental warming fuel scientists emissions plants
Budget	billion domestic fiscal balance deficit medicare repeal priorities ryan trillion
Consumer debt crisis	debt loans dollars contract fees payments taxpayers borrowers treasury consumers
Robert Menendez	attorney menendez lawyer criminal allegations file sentence prosecutors convicted prison
Yemen	saudi strike target al yemen intelligence arabia houthis pakistan qaeda

Table 8: Topics with high document coherence and high word coherence.

This makes the topic-related documents heterogeneous and incoherent, which in turn lowers the document-based coherence score. The second group are five topics that are coherent but for which the document-based coherence score is either misestimated or low because a topic is highly abstract and associated with a set of documents that are less semantically related. Table 9 shows all the topics from the first group and one example topic from the second group. Taken together, this category of topics shows that it might be beneficial to combine document- and word-based coherence measures: the word coherence could be used as a fallback in cases when a document-based coherence measure fails to detect coherence, for instance because the topic is being underrepresented in the documents.

Low document/low word coherence. Among the 26 topics in this category, the majority of topics (18) are incoherent (mixture of topics or noise), while the remaining 8 topics pertain to a single semantic topic but with the addition of noise. Additionally, the majority of the remaining topics (7 out of 8) are concrete – a property that correlates with low word-based coherence. The remaining set of 8 topics in this category represents a hard case whose coherence is difficult to detect even with a combination of a document- and word-based coherence measure – these topics are specific in that they contain both noise (which decreases document-based coherence) and are concrete (which decreases word-based coherence). This category of topics complements the one containing topics with low document and high word coherence in demonstrating the usefulness of combining word-based and document-based coherence measures. Among the topics with low document coherence, word coherence correlates with true coherence: most (18 out of 26) low word coherence topics are indeed incoherent, while most (10 out of 12) high word coherence topics are indeed coherent.

Topic label	Top-10 topic words
Lawsuits	file board lawsuit complaint suit violated georgia damage settled accused
Radio and television	morning host watch night radio television tv network update station
Journalism	writing published article piece journalists paper journal newspaper quoted editor
Social media	fox twitter host night tweeted morning facebook com watch remarks
Crime	prison criminal crime sentence convicted attorney prosecutors trial lawyer judge
(noise)	video someone thought probably maybe else guy anything everyone yes

Table 9: Topics with low document coherence and high word coherence.

5.4. Conclusions

We motivated the need for document-based coherence by essentially claiming that word-based coherence in some cases is not informative enough to gauge the coherence of a topic derived by a topic model, especially for transient and contingent topics typical of the news domain. Results of the experiments in Section 5.2 confirm this by showing that when state-of-art word-based coherence measures proposed in the literature are used for estimating document-based coherence, their performance is markedly below the document-based coherence baseline. However, some word-based measures outperform a random baseline, indicating a degree of correlation between document-based and word-based coherence.

The qualitative analysis of model topics with high and low document- and word-based coherence indicates that word- and document-based coherence measures are complementary to each other and that they may be combined to detect coherent topics more accurately – there exist coherent topics (model topics corresponding to semantic topics) that would be discarded as incoherent if the detection were based on word- or document-based coherence alone. Interestingly, high word-based coherence correlates with abstract topics, whereas low word-based coherence correlates with concrete topics.

6. Semi-Automated Topic Discovery

As mentioned in the introduction, the increased availability and consumption of online news in textual form brought about increased interest in automated content analysis of news texts. Most content analysis techniques rely on an inventory of topics, which ideally corresponds to the topics covered by the news texts under analysis. However, large quantities of news data available on one hand, and the transience of topics reported about in the news on the other,

make it impossible to establish a fixed and comprehensive topic inventory for the news domain. The alternative is to rely on techniques for (semi)automated topic discovery from text corpus, for which topic models have proven to be an extremely useful tool.

However, although topic models make topic discovery much easier, for the reasons discussed in the introduction the results are of varying quality – more precisely, not all topics induced by a topic model will be semantically interpretable. Hence, to be usable for content analysis, the results of the topic model typically need to be examined by a human expert, which requires considerable effort. Moreover, to increase the coverage of the topics, one would typically want to run several differently parametrized topic models (e.g., models with varying number of topics) and examine the topics collected from all these models, which further increases the complexity of the task.

In this section we demonstrate how our proposed document-based coherence measures can be used to improve the efficiency of semi-automated topic discovery based on topics collected from several models. The idea is to use document-based topic coherence scores as a heuristic for which topics should be examined first by the human expert. The assumption is that, if the human expert examines the topics in the order of their coherence score rather than at random, and if care is taken to avoid re-discovering duplicate topics, this may substantially improve the discovery rate of the semantic topics (the number of discovered topics per number of topics examined), thus reducing the overall human effort.

6.1. Experimental Setup

We simulate the described topic discovery scenario using our US news topics dataset of model topics annotated with semantic topics (cf. Section 4.1) by traversing the topics sorted in descending order of coherence score and keeping a count of the number of distinct semantic topics.

We simulate the semantic topic discovery process using the entire US news topics dataset consisting of 350 model topics. Each coherent topic is labeled with a single corresponding semantic topic, while both fused topics and noise topics are labeled with zero semantic topics. Discovery is simulated by traversing the list of model topics sorted either at random or by scores assigned by a coherence measure. Each model topic is compared with the topics from the list of already discovered topics and discarded if it is found to be a duplicate. If the topic is not a duplicate, it is counted as a topic examined by a human expert, and if the topic corresponds to a semantic topic, the semantic topic is added to the list of discovered topics. We measure the effectiveness of the discovery process by examining the topic discovery rate: the number of discovered semantic topics per number of topics examined.

Two points deserve to be mentioned: topic duplication and fused topics. While using more than a single topic model will generally improve the coverage of the semantic topics, it will also result in some topics being duplicated, especially those that are more salient in the corpus. Because incoherent topics are random, being the result of either random noise or of the fusing of two random

semantic topics, they are very unlikely to be duplicated. Alternately, coherent topics will match semantic topics, which are limited in number, and therefore topics that get duplicated are most likely coherent. As a consequence, sorting the topics by coherence will push the duplicates toward the start of the list, thereby lowering the topic discovery rate, as each duplicate topic needs to be examined but yields no new semantic topics. To prevent this, we remove all duplicates from the pooled list of topics. We consider two topics to be duplicates if the cosine distance between the topics' probability vectors is less than 0.5 – a conservative choice, as topics have to be almost identical to meet this threshold. When simulating the topic discovery process, we apply duplicate removal to both topics sorted by coherence and randomly ordered topics.

Another point worth mentioning concerns the fused topics: model topics corresponding to two or more semantic topics. As described in Section 4.1, 21% of model topics in our dataset were labeled with two semantic topics. In this experiment, we treat fused topics as noise, simulating the topic discovery scenario in which the fused topics are recognized by the human expert as noise and discarded – a task that can be performed efficiently. Some semantic topics that only occur as fused may be missed, but using a pool of topic models rather than a single model makes it more likely that the fused topic will eventually emerge as separate and coherent topic in one of the models. Alternatively, the human expert could attempt to split up fused topics into individual semantic topics but this would be a time-consuming and potentially error-prone task. In this approach to topic discovery the benefit of using a coherence measure to improve the discovery rate would be reduced, since the fused topics generally have low coherence. To remedy this, the coherence measures could be compounded with measures for detection of fused topics, but we leave this for future work.

6.2. Results

We tested four coherence measures: the baseline document-based measure (cf. Section 4.3), the C_P word-based measure (cf. Section 5.1), and two top-performing document-based measures from the Graph-Cnt category with the highest AUC score on the development set and the test set (cf. Table 5), subsequently denoted `graph1` and `graph2`, respectively. We compare the four coherence measures against a random baseline: we run the simulation for 20 random topic orderings and calculate the average, minimum, and maximum of discovered topics per the number of examined topics.

The results are shown in Figure 3. The curves on the left plot show the number of topics discovered as a function of the number of topics examined, while the right plot shows the difference between the number of topics discovered and the average number of topics discovered when examined in random order. The main observation is that topic discovery guided by the document-based coherence measures markedly outperforms both the random baseline and the state-of-art word-based coherence measure. Furthermore, the two Graph-Cnt coherence measures outperform the document-based coherence baseline, giving the overall best performance. Another observation is that these two coherence measures, albeit they score nearly identical in terms of AUC (cf. Table 5), yield

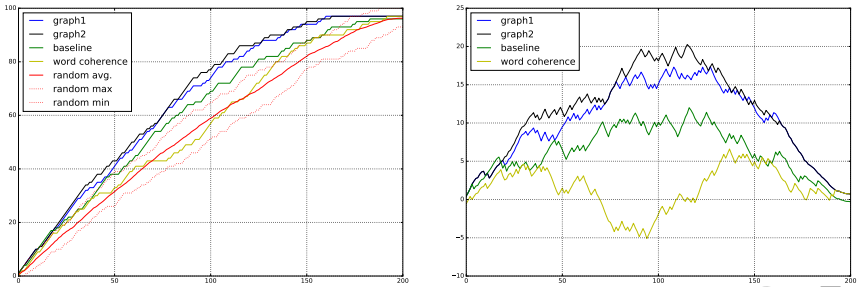

Figure 3: Semantic topics discovered (y-axis) per model topics examined (x-axis): (a) absolute number of topics discovered and (b) the average difference between the number of topics discovered when using ordering based on a topic coherence measure and random ordering.

different topic discovery curves. This suggests that for evaluating coherence measures for specific applications it might be a good idea to complement rank-based metrics such as AUC with an application-oriented evaluation metric.

In more concrete terms, the results on this dataset show that by sorting the topics by coherence-based coherence it is possible to discover all semantic topics after examining 160 model topics – this is in contrast to an average of 200 model topics which would have to be examined if the order were random. Assuming that topic examination takes 6 minutes on average (an estimate based on the annotation of the dataset), the coherence measure would save the annotator four hours. If the aim is to examine only 100 model topics, perhaps to get an overview of semantic topics covered by the corpus, examination in random order would on average discover only 59 semantic topics, whereas examination based on document-coherence would discover 77 topics – an increase of 30%.

7. Conclusion

Topic models are a popular tool for unsupervised discovery of topics from text corpora, including texts from the news domain. A well-known problem with topic models, however, is that the quality of the generated topics typically varies. This motivated the development of a number of model evaluation techniques, most notably those based on the calculation of topics' semantic coherence. The existing topic coherence methods estimate the coherence based on the semantic relatedness of the topic words. This approach, however, is inadequate for news media texts, where topics are often contingent and transient and therefore associated with semantically unrelated words. To solve this problem, we proposed a novel class of topic coherence methods that estimate topic coherence based on topic documents rather than topic words. The underlying assumption is that, because documents contain more information than words, document-based topic coherence can better capture topics' semantic interpretability.

The proposed document-based methods calculate the coherence of a topic in three steps: selection of topic-related documents, vectorization of the docu-

ments, and calculation of a coherence score from document vectors. We proposed a number of different methods, including distance-based, density-based, and graph-based methods, and evaluated them on two datasets of topics manually labeled with coherence scores. The method that uses tf-idf or bag-of-words document representations, builds an unweighted similarity graph, and estimates the coherence score by aggregating node connectivity scores was found to outperform all other considered methods in terms of coherence ranking performance, including a strong baseline. Furthermore, we have shown that the method can be used to speed up the otherwise tedious task of semi-automated topic discovery from a corpus of news media texts.

To investigate the relationship between document-based and word-based topic coherence, we evaluated state-of-art word-based coherence methods on our datasets of topics labeled with document-based coherence scores. We found that word-based coherence methods, which are optimized for word-based coherence, perform poorly on our datasets. An examination of model topics with estimated high and low document- and word-based coherence demonstrated the potential merit in combining word- and document-based coherence measures to detect coherent topics more accurately.

There exist a number of interesting directions for future work. On the technical side, the methods we propose could probably be improved by using a more effective document vectorization method. Document-level neural embeddings (Lau & Baldwin, 2016) and kernel-based aggregations of word-level embeddings (Zhang et al., 2018) might be a good starting point. The experiments in Section 4.5 show that the baseline document-based method exhibits fairly good performance, so one possibility might be to try to improve the baseline or combine it with the measures we proposed. Another promising direction for future work, indicated by the experiments in Section 5, would be to combine document- and word-based coherence measures.

At a conceptual level, we believe that further investigating the possible applications of both word- and document-based coherence measures in exploratory analysis might prove these measures useful for a task separate from topic model evaluation. Such applications might also help us gain a better understanding of the concept of topic coherence. Document-based coherence measures can be used, in line with similar word-based experiments (Stevens et al., 2012; O'Callaghan et al., 2015), for a systematic comparative analyses of various topic models as well as other clustering models. Ideally, such an analysis would also consider the application of document-based coherence to other genres of texts besides news texts, especially genres with shorter and topically focused texts, such as social media posts and other user-generated content, as well as the topically more heterogeneous scientific articles.

8. Acknowledgments

This research has been supported by the European Regional Development Fund under the grant KK.01.1.1.01.0009 (DATACROSS). We would like to

thank the anonymous reviewers, as well as Domagoj Alagić, Mladen Karan, and Maria Pia Di Buono for the helpfull feedback.

References

Ahmed, A., Ho, Q., Eisenstein, J., Xing, E., Smola, A. J., & Teo, C. H. (2011). Unified analysis of streaming news. In *Proceedings of the 20th international conference on World wide web* (pp. 267–276). ACM.

Aletras, N., & Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)* (pp. 13–22).

AlSumait, L., Barbará, D., Gentle, J., & Domeniconi, C. (2009). Topic significance ranking of lda generative models. *Machine Learning and Knowledge Discovery in Databases*, (pp. 67–82).

Belford, M., Namee, B. M., & Greene, D. (2018). Stability of topic modeling via matrix factorization. *Expert Systems with Applications*, *91*, 159 – 169. doi:https://doi.org/10.1016/j.eswa.2017.08.047.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*, 77–84.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*, 17–35.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models: Problems, diagnostics, and improvementes. In *Handbook of Mixed Membership Models and their Applications* (pp. 225–254). CRC Press.

Boyd-Graber, J. L., Blei, D. M., & Zhu, J. (2007a). Probabalistic walks in semantic hierarchies as a topic model for WSD. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Boyd-Graber, J. L., Blei, D. M., & Zhu, X. (2007b). A topic model for word sense disambiguation. In *EMNLP-CoNLL* (pp. 1024–1033).

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Nips* (pp. 1–9). volume 31.

Chen, M., Jin, X., & Shen, D. (2011). Short text classification improved by learning multi-granularity topics. In *IJCAI* (pp. 1776–1781).

Chuang, J., Fish, S., Larochelle, D., Li, W. P., & Weiss, R. (2014). Large-scale topical analysis of multiple online news sources with media cloud. *NewsKDD: Data Science for News Publishing, at KDD*, .

Chuang, J., Gupta, S., Manning, C., & Heer, J. (2013). Topic model diagnostics: Assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on machine learning (ICML-13)* (pp. 612–620).

Chuang, J., Ramage, D., Manning, C., & Heer, J. (2012). Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 443–452). ACM.

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, *8*, 519–531.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*, 391.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, (pp. 837–845).

Dou, W., Wang, X., Skau, D., Ribarsky, W., & Zhou, M. X. (2012). Leadline: Interactive visual analysis of text data through event identification and exploration. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on* (pp. 93–102). IEEE.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of communication*, *43*, 51–58.

Estrada, E., & Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, *71*, 056103.

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*, 194–199.

Flaounas, I., Ali, O., Lansdall-Welfare, T., De Bie, T., Mosdell, N., Lewis, J., & Cristianini, N. (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-contenttopics, style and gender. *Digital Journalism*, *1*, 102–116.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks*, *1*, 215–239.

Galke, L., Saleh, A., & Scherp, A. (2017). Word embeddings for practical information retrieval. In M. Eibl, & M. Gaedke (Eds.), *INFORMATIK 2017* (pp. 2155–2167). Gesellschaft fr Informatik, Bonn.

Gao, W., Li, P., & Darwish, K. (2012). Joint topic modeling for event summarization across news and social media streams. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (pp. 1173–1182). ACM.

Garcin, F., Dimitrakakis, C., & Faltings, B. (2013). Personalized news recommendation with context trees. In *Proceedings of the 7th ACM conference on Recommender systems* (pp. 105–112). ACM.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, *101*, 5228–5235.

Grimmer, J. (2009). A bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, *18*, 1–35.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, (pp. 1–31).

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent Dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems 23* (pp. 856–864). Curran Associates, Inc.

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, *4*, 89–106.

Kim, Y., Kim, S., Jaimes, A., & Oh, A. (2014). A computational analysis of agenda setting. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion* (pp. 323–324). ACM.

Koltcov, S., Koltsova, O., & Nikolenko, S. (2014). Latent dirichlet allocation: stability and applications to studies of user-generated content. In *Proceedings of the 2014 ACM conference on Web science* (pp. 161–165). ACM.

Korenčić, D., Ristov, S., & Šnajder, J. (2015). Getting the agenda right: measuring media agenda using topic models. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications* (pp. 61–66). ACM.

Korenčić, D., Grbeša-Zenzerović, M., & Šnajder, J. (2016). Topics and their salience in the 2015 parliamentary election in Croatia: A topic model based analysis of the media agenda. In *Proceedings of the International Conference on the Advances in Computational Analysis of Political Text - PolText 2016*.

Kucher, K., & Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)* (pp. 117–121). doi:10.1109/PACIFICVIS.2015.7156366.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Lau, J. H., & Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. *CoRR*, *abs/1607.05368*. URL: `http://arxiv.org/abs/1607.05368`.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL* (pp. 530–539).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*, 788.

Li, L., Wang, D.-D., Zhu, S.-Z., & Li, T. (2011). Personalized news recommendation: a review and an experimental investigation. *Journal of computer science and technology*, *26*, 754–766.

Li, L., Zheng, L., Yang, F., & Li, T. (2014). Modeling and broadening temporal user interest in personalized news recommendation. *Expert Systems with Applications*, *41*, 3168–3177.

Lin, C., & He, Y. (2009). Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 375–384). ACM.

Ling, C. X., Huang, J., & Zhang, H. (2003). Auc: a statistically consistent and more discriminating measure than accuracy. In *IJCAI* (pp. 519–524). volume 3.

Ljubešić, N., & Erjavec, T. (2011). hrWaC and slWac: Compiling web corpora for Croatian and Slovene. In *Text, Speech and Dialogue - 14th International Conference* Lecture Notes in Computer Science (pp. 395–402). Springer.

Magnini, B., Strapparava, C., Pezzulo, G., & Gliozzo, A. (2002). The role of domain information in word sense disambiguation. *Natural Language Engineering*, *8*, 359–373.

McCombs, M. E., & Shaw, D. L. (1972). The agenda-setting function of mass media. *Public Opinion Quarterly*, *36*, 176–187.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, *abs/1301.3781*. URL: `http://arxiv.org/abs/1301.3781`. `arXiv:1301.3781`.

Mimno, D., & Blei, D. (2011). Bayesian checking for topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 227–237). Association for Computational Linguistics.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Association for Computational Linguistics.

Mimno, D. M., & McCallum, A. (2012). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. *CoRR*, *abs/1206.3278*.

Musat, C., Velcin, J., Trausan-Matu, S., & Rizoiu, M.-A. (2011). Improving topic evaluation using conceptual knowledge. In *22nd International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 1866–1871). volume 3.

Neuendorf, K. A. (2016). *The content analysis guidebook*. Sage.

Newman, D., Chemudugunta, C., Smyth, P., & Steyvers, M. (2006). Analyzing entities and topics in news articles using statistical topic models. In *ISI* (pp. 93–104). Springer.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 100–108). Association for Computational Linguistics.

Newman, N., Fletcher, R., Kalogeropoulos, A., Levy, D., & Kleis Nielsen, R. (2016). Reuters institute digital news report 2017.

Nikolenko, S. I. (2016). Topic quality metrics based on distributed word representations. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval* (pp. 1029–1032). ACM.

Nikolenko, S. I., Koltcov, S., & Koltsova, O. (2015). Topic modelling for qualitative studies. *Journal of Information Science*, *43*, 88–102.

O'Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, *42*, 5645–5657.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP* (pp. 1532–1543). volume 14.

Popescu, O., & Strapparava, C. (Eds.) (2017). *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics. URL: `http://www.aclweb.org/anthology/W17-42`.

Ramirez, E. H., Brena, R., Magatti, D., & Stella, F. (2012). Topic model validation. *Neurocomputing*, *76*, 125–133.

Ramrakhiyani, N., Pawar, S., Hingmire, S., & Palshikar, G. K. (2017). Measuring topic coherence through optimal word buckets. *EACL 2017*, (p. 437).

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, *12*, 77.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining* (pp. 399–408). ACM.

Rosner, F., Hinneburg, A., Röder, M., Nettling, M., & Both, A. (2014). Evaluating topic coherence measures. *CoRR*, *abs/1403.6397*. URL: `http://arxiv.org/abs/1403.6397`. `arXiv:1403.6397`.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, *24*, 513–523.

Saramäki, J., Kivelä, M., Onnela, J.-P., Kaski, K., & Kertesz, J. (2007). Generalizations of the clustering coefficient to weighted complex networks. *Physical Review E*, *75*.

Schult, D. A., & Swart, P. (2008). Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conferences (SciPy 2008)* (pp. 11–16). volume 2008.

Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval* volume 39. Cambridge University Press.

Shahaf, D., & Guestrin, C. (2012). Connecting two (or less) dots: Discovering structure in news articles. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*, 24.

Steinberger, R., Pouliquen, B., & der Goot, E. V. (2013). An introduction to the europe media monitor family of applications. *CoRR*, *abs/1309.5290*. URL: `http://arxiv.org/abs/1309.5290`. `arXiv:1309.5290`.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., & Buttler, D. (2012). 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* (pp. 952–961). Association for Computational Linguistics.

Titov, I., & McDonald, R. T. (2008). A joint model of text and aspect ratings for sentiment summarization. In *ACL* (pp. 308–316).

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics* (pp. 384–394). Association for Computational Linguistics.

Vossen, P., Rigau, G., Serafini, L., Stouten, P., Irving, F., & Van Hage, W. R. (2014). NewsReader: recording history from daily news streams. In *LREC* (pp. 2000–2007).

Waal, A. D., & Barnard, E. (2008). Evaluating topic models with stability.

Wallach, H. M., Murray, I., Salakhutdinov, R., & Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th annual international conference on machine learning* (pp. 1105–1112). ACM.

Wei, X., & Croft, W. B. (2006). Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 178–185). ACM.

Yi, X., & Allan, J. (2009). A comparative study of utilizing topic models for information retrieval. In *Advances in Information Retrieval* (pp. 29–41). Springer Berlin Heidelberg.

Zhang, R., Guo, J., Lan, Y., Xu, J., & Cheng, X. (2018). Aggregating neural word embeddings for document representation. In *Advances in Information Retrieval* (pp. 303–315). Cham: Springer International Publishing.