# Evaluating Topic Coherence Using Distributional Semantics

Nikolaos Aletras Mark Stevenson
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP, UK
{n.aletras, m.stevenson}@dcs.shef.ac.uk

**Abstract**

This paper introduces distributional semantic similarity methods for automatically measuring the coherence of a set of words generated by a topic model. We construct a semantic space to represent each topic word by making use of Wikipedia as a reference corpus to identify context features and collect frequencies. Relatedness between topic words and context features is measured using variants of Pointwise Mutual Information (PMI). Topic coherence is determined by measuring the distance between these vectors computed using a variety of metrics. Evaluation on three data sets shows that the distributional-based measures outperform the state-of-the-art approach for this task.

## 1 Introduction

Topic modelling is a popular statistical method for (soft) clustering documents (Blei et al., 2003; Deerwester et al., 1990; Hofmann, 1999). Latent Dirichlet Allocation (LDA) (Blei et al., 2003), one type of topic model, has been widely used in NLP and applied to a range of tasks including word sense disambiguation (Boyd-Graber et al., 2007), multi-document summarisation (Haghighi and Vanderwende, 2009) and generation of comparable corpora (Preiss, 2012).

A variety of approaches has been proposed to evaluate the topics generated by these models. The first to be explored were extrinsic methods, measuring the performance achieved by a model in a specific task or using statistical methods. For example, topic models have been evaluated by measuring their accuracy for information retrieval (Wei and Croft, 2006). Statistical methods have also been applied to measure the predictive likelihood of a topic model in held-out documents by computing their perplexity. Wallach et al. (2009) gives a detailed description of such statistical metrics.

However, these approaches do not provide any information about how interpretable the topics are to humans. Figure 1 shows some example topics generated by a topic model. The first three topics appear quite coherent, all the terms in each topic are associated with a common theme. On the other hand, it is difficult to identify a coherent theme connecting all of the words in topics 4 and 5. These topics are difficult to interpret and could be considered as "junk" topics. Interpretable topics are important in applications such as visualisation of document collections (Chaney and Blei, 2012; Newman et al., 2010a), where automatically generated topics are used to provide an overview of the collection and the top-$n$ words in each topic used to represent it.

Chang et al. (2009) showed that humans find topics generated by models with high predictive likelihood to be less coherent than topics generated from others with lower predictive likelihood. Following Chang's findings, recent work on evaluation of topic models has been focused on automatically measuring the coherence of generated topics by comparing them against human judgements (Mimno et al., 2011; Newman et al., 2010b). Newman et al. (2010b) define topic coherence as the average semantic relatedness between topic words and report the best correlation with humans using the Pointwise Mutual Information (PMI) between topic words in Wikipedia.

| |
|---|
| 1: oil, louisiana, coast, gulf, orleans, spill, state, fisherman, fishing, seafood |
| 2: north, kim, korea, korean, jong, south, il, official, party, son |
| 3: model, wheel, engine, system, drive, front, vehicle, rear, speed, power |
| 4: drink, alcohol, indonesia, drinking, indonesian, four, nokia, beverage, mcdonald, caffeine |
| 5: privacy, andrews, elli, alexander, burke, zoo, information, chung, user, regan |

Figure 1: A sample of topics generated by a topic model over a corpus of news articles. Topics are represented by top-$n$ most probable words.

Following this direction, we explore methods for automatically determining the coherence of topics. We propose a novel approach for measuring topic coherence based on the distributional hypothesis which states that words with similar meanings tend to occur in similar context (Harris, 1954). Wikipedia is used as a reference corpus to create a distributional semantic model (Padó and Lapata, 2003; Turney and Pantel, 2010). Each topic word is represented as a bag of highly co-occurring context words that are weighted using either PMI or a normalised version of PMI (NPMI). We also explore creating the vector space using differing numbers of context terms. All methods are evaluated by measuring correlation with humans on three different sets of topics. Results indicating that measures on the fuller vector space are comparable to the state-of-the-art proposed by Newman et al. (2010b), while performance consistently improves using a reduced vector space.

The remainder of this article is organised as follows. Section 2 presents background work related to topic coherence evaluation. Section 3 describes the distributional methods for measuring topic coherence. Section 4 explains the experimental set-up used for evaluation. Our results are described in Section 5 and the conclusions in Section 6.

## 2 Related work

Andrzejewski et al. (2009) proposed a method for generating coherent topics which used a mixture of Dirichlet distributions to incorporate domain knowledge. Their approach prefers words that have similar probability (high or low) within all topics and rejects words that have different probabilities across topics.

AlSumait et al. (2009) describe the first attempt to automatically evaluate topics inferred from topic models. Three criteria are applied to identify junk or insignificant topics. Those criteria are in the form of probability distributions over the highest probability words. For example, topics in which the probability mass is distributed approximately equally across all words are considered likely to be difficult to interpret.

Newman et al. (2010b) also focused on methods for measuring the semantic coherence of topics. The main contribution of this work is to propose a measure for the automatic evaluation of topic semantic coherence which has been shown to be highly correlated with human evaluation. It is assumed that a topic is coherent if all or the most of its words are related. Results showed that word relatedness is better predicted using the distribution-based Pointwise Mutual Information (PMI) of words rather than knowledge-based measures.

The method using PMI proposed by Newman et al. (2010b) relies on co-occurrences of words in an external reference source such as Wikipedia for automatic evaluation of topic quality. Mimno et al. (2011) showed that available co-document frequency of words in the training corpus can be used to measure semantic coherence. Topic coherence is defined as the sum of the log ratio between co-document frequency and the document frequency for the N most probable words in a topic. The intuition behind this metric is that the co-occurrence of words within documents in the corpus can indicate semantic relatedness.

Musat et al. (2011) associated words in a topic with WordNet concepts thereby creating topical subtrees. They rely on WordNet's hierarchical structure to find a common concept that best describes as many words as possible. It is assumed that the higher the coverage and specificity of a topical subtree,

the more semantically coherent the topic. Experimental results showed high agreement with humans in the word intrusion task, in contrast to Newman et al. (2010b) who concluded that WordNet is not useful for topic evaluation.

Recent work by Ramirez et al. (2012) analyses and evaluates the semantic coherence of the results obtained by topic models rather than the semantic coherence of the inferred topics. Each topic model is treated as a partition of document-topic associations. Results are evaluated using metrics for cluster comparison.

# 3 Measuring Topic Coherence

Let $T = \{w_1, w_2, ..., w_n\}$ be a topic generated from a topic model which is represented by its top-$n$ most probable words. Newman et al. (2010b) assume that the higher the average pairwise similarity between words in $T$, the more coherent the topic. Given a symmetric word similarity measure, $Sim(w_i, w_j)$, they define coherence as follows:

$$Coherence_{Sim}(T) = \frac{\sum_{\substack{1 \leq i \leq n-1 \\ i+1 \leq j \leq n}} Sim(w_i, w_j)}{\binom{n}{2}} \tag{1}$$

where $w_i, w_j \in T$.

## 3.1 Distributional Methods

We propose a novel method for determining topic coherence based on using distributional similarity between the top-n words in the topic. Each topic word is represented as a vector in a semantic space. Let $\vec{w_1}$, $\vec{w_2}$, ..., $\vec{w_n}$ be the vectors which represent the top $n$ most probable words in the topic. Also, assume that each vector consists of $N$ elements and $\vec{w_{ij}}$ is the $j$th element of vector $\vec{w_i}$. Then the similarity between the words, and therefore cohesion of the topic, can be computed using the following measures (Curran, 2003; Grefenstette, 1994):

- The **cosine** of the angles between the vectors:

$$Sim_{cos}(\vec{w_i}, \vec{w_j}) = \frac{\vec{w_i} \cdot \vec{w_j}}{\|\vec{w_i}\| \|\vec{w_j}\|} \tag{2}$$

- The **Dice** coefficient:

$$Sim_{Dice}(w_i, w_j) = \frac{2 \times \sum_{k=1}^{N} min(\vec{w_{ik}}, \vec{w_{jk}})}{\sum_{k=1}^{N} (\vec{w_{ik}} + \vec{w_{jk}})} \tag{3}$$

- The **Jaccard** coefficient:

$$Sim_{Jaccard}(w_i, w_j) = \frac{\sum_{k=1}^{N} min(\vec{w_{ik}}, \vec{w_{jk}})}{\sum_{k=1}^{N} max(\vec{w_{ik}}, \vec{w_{jk}})} \tag{4}$$

Each of these measures estimates the distance between a pair of topic words and can be substituted into equation 1 to produce a topic cohesion measure based on distributional semantics.

Alternatively, the cohesion of a set of topic words can be estimated with a single measure by computing the average distance between each topic word and the centroid:

$$Sim_{centroid} = \frac{\sum_{t \in T} sim_{cos}(T_c, t)}{n} \tag{5}$$

where $T_c$ is the centroid of the vectors for topic $T$. For the experiments reported in this paper the distance of each vector to the centroid is computed using the cosine measure.

### 3.2 Constructing the Semantic Space

Vectors representing the topic words are constructed from a semantic space consisting of information about word co-occurrence. The semantic space was created using Wikipedia[1] as a reference corpus and a window of $\pm$ 5 words[2].

#### 3.2.1 Weighting Vectors

Using the co-occurrence information to generate vectors directly does not produce good results so the vectors are weighted using two approaches.

For the first, **PMI**, the pointwise mutual information for each term in the context is used rather than the raw co-occurrence count. PMI is computed as follows:

$$\text{PMI}(w_i, w_j) = log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{6}$$

Note that this application of PMI for topic cohesion is different from one previously reported by Newman et al. (2010b) since we use PMI to weight vectors rather than to compute a similarity score between pairs of words.

In addition, vectors are also weighted using **NPMI** (Normalised PMI). This is an extension of PMI that has been used for collocation extraction (Bouma, 2009) and is computed as follows:

$$\text{NPMI}(w_i, w_j) = \frac{PMI(w_i, w_j)}{-log(p(w_i, w_j))} \tag{7}$$

Finally, we introduce $\gamma$ which is a parameter to assign more emphasis on context features with high PMI (or NPMI) values with a topic word. Vectors are weighted using $PMI(w_i, f_j)^\gamma$ or $NPMI(w_i, f_j)^\gamma$ where $w_i$ is a topic word and $f_j$ is a context feature. For all of our experiments we set $\gamma = 2$ which was found to produce the best results.

#### 3.2.2 Reducing the Basis

Including all co-occurring terms in the vectors leads to a high dimensional space. We also experimented with two approaches to reducing the number of terms to form a semantic space with smaller basis. Firstly, following Islam and Inkpen (2006), a **Reduced Semantic Space** is created by choosing the $\beta_{w_i}$ most related context features for each topic word $w_i$:

$$\beta_{w_i} = \big(log(c(w_i))\big)^2 \frac{(log_2(m))}{\delta} \tag{8}$$

where $\delta$ is a parameter for adjusting the number of features for each word and $m$ is the size of the corpus. Varying the value of $\delta$ did not effect performance for values above 1. This parameter was set of 3 for the results reported here. In addition a frequency cut-off of 20 was also applied. In addition, a smaller semantic space was created by considering only topic words as context features, leading to $n$ features for each topic word. This is referred to as the **Topic Word Space**.

## 4 Experimental Set-up

### 4.1 Data

To the best of our knowledge, there are no standard data sets for evaluating topic coherence. Therefore we have developed one for this study which we have made publicly available[3]. A total of 300 topics are

[1] http://dumps.wikimedia.org/enwiki/20120104/

[2] We also experimented with different lengths of context windows

[3] The data set can be downloaded from http://staffwww.dcs.shef.ac.uk/people/N.Aletras/resources/TopicCoherence300.tar.gz

generated by running LDA over three different document collections:

- **NYT:** 47,229 New York Times news articles published between May and December 2010 from the GigaWord corpus. We generated 200 topics and randomly selected 100.
- **20NG:** The 20 News Group Data Collection[4] (20NG), a set of 20,000 newsgroup emails organised into 20 different subjects (e.g. sports, computers, politics). Each topic has 1,000 documents associated with it. 100 topics were generated for this data set.
- **Genomics:** 30,000 scientific articles published in 49 journals from MEDLINE, originally used in the TREC-Genomics Track[5]. We generated 200 topics and randomly selected 100.

All document were pre-processed by removing stop words and lemmatising. Topics are generated using *gensim*[6] with hyperparameters $(\alpha, \beta)$ set to $\frac{1}{num\_of\_topics}$. Each topic is represented by its 10 most probable words.

### 4.2 Human Evaluation of Topic Coherence

Human judgements of topic coherence were collected through a crowdsourcing platform, CrowdFlower[7]. Participants were presented with 10 word sets, each of which represents a topic. They asked to judge topic coherence on a 3-point Likert scale from 1-3, where 1 denotes a "Useless" topic (i.e. words appear random and unrelated to each other), 2 denotes "Average" quality (i.e. some of the topic words are coherent and interpretable but others are not), and 3 denotes a "Useful" topic (i.e. one that is semantically coherent, meaningful and interpretable). Each participant was asked to judge up to 100 topics from a single collection. The average response for each topic was calculated as the coherency score for the gold-standard.

To ensure reliability and avoid random answers in the survey, we used a number of questions with predefined answer (either totally random words as topics or obvious topics such as week days). Annotations from participants that failed to answer these questions correctly were removed.

We run three surveys, one for each topic collection of 100 topics. The total number of filtered responses obtained for the NYT dataset was $1,778$ from 26 participants, while for the 20NG dataset we collected $1,707$ answers from $24$ participants. The participants were recruited by a broadcast email sent to all academic staff and graduate students in our institution. For the Genomics dataset the emails were sent only to members of the medical school and biomedical engineering departments. We collected $1,050$ judgements from 12 participants for this data set.

Inter-annotator agreement (IAA) is measured as the average of the Spearman correlation between the set of scores of each survey respondent and the average of the other respondents' scores. The IAA in the three surveys is 0.70, 0.64 and 0.54 for NYT, 20NG and Genomics respectively.

## 5 Results

Table 1 shows the results obtained for all of the methods on the three datasets. Performance of each method is measured as the average Spearman correlation with human judgements. The top row of each table shows the result using the average PMI approach (Newman et al., 2010b) while the next two rows show the results obtained by substituting PMI with NPMI and the method proposed by Mimno et al. (2011). The main part of each table shows performance using the approaches described in Section 3 using various combinations of methods for constructing the semantic space and determining the similarity between vectors.

[4] http://people.csail.mit.edu/jrennie/20Newsgroups
[5] http://ir.ohsu.edu/genomics
[6] http://radimrehurek.com/gensim
[7] http://crowdflower.com

| **NYT** | | |
|---|---|---|
| Newman et al. (2010b) | 0.71 | |
| Average NPMI | 0.74 | |
| Mimno et al. (2011) | -0.39 | |
| Reduced Semantic Space | | |
| | PMI | NPMI |
| Cosine | 0.69 | 0.68 |
| Dice | 0.63 | 0.62 |
| Jaccard | 0.63 | 0.61 |
| Centroid | 0.67 | 0.67 |
| Topic Words Space | | |
| | PMI | NPMI |
| Cosine | **0.76** | 0.75 |
| Dice | 0.68 | 0.71 |
| Jaccard | 0.69 | 0.72 |
| Centroid | **0.76** | 0.75 |

| **20NG** | | |
|---|---|---|
| Newman et al. (2010b) | 0.73 | |
| Average NPMI | 0.76 | |
| Mimno et al. (2011) | 0.34 | |
| Reduced Semantic Space | | |
| | PMI | NPMI |
| Cosine | 0.78 | 0.79 |
| Dice | 0.77 | 0.78 |
| Jaccard | 0.77 | 0.78 |
| Centroid | 0.77 | 0.78 |
| Topic Words Space | | |
| | PMI | NPMI |
| Cosine | 0.79 | **0.8** |
| Dice | 0.79 | **0.8** |
| Jaccard | **0.8** | **0.8** |
| Centroid | 0.78 | 0.79 |

| **Genomics** | | |
|---|---|---|
| Newman et al. (2010b) | 0.73 | |
| Average NPMI | 0.76 | |
| Mimno et al. (2011) | -0.4 | |
| Reduced Semantic Space | | |
| | PMI | NPMI |
| Cosine | 0.74 | 0.73 |
| Dice | 0.69 | 0.68 |
| Jaccard | 0.69 | 0.76 |
| Centroid | 0.73 | 0.71 |
| Topic Words Space | | |
| | PMI | NPMI |
| Cosine | **0.8** | **0.8** |
| Dice | 0.79 | **0.8** |
| Jaccard | **0.8** | **0.8** |
| Centroid | **0.8** | **0.8** |

Table 1: Performance of methods for measuring topic coherence (Spearman Rank correlation with human judgements).

Using the average PMI between topic words correlates well with human judgements, 0.71 for NYT, 0.73 for 20NG and 0.75 for Genomics confirming results reported by Newman et al. (2010b). However, NPMI performs better than PMI, with an improvement in correlation of 0.03 for all datasets. The improvement is down to the fact that NPMI reduces the impact of low frequency counts in word co-occurrences and therefore uses more reliable estimates (Bouma, 2009).

On the other hand, the method proposed by Mimno et al. (2011) does not correlate well with human judgements, (-0.39 for NYT, 0.34 for 20NG and -0.4 for Genomics) which is the lowest performance of all of the methods tested. This demonstrates that while co-document frequency helps to generate more coherent topics (Mimno et al., 2011), it is sensitive to the size of the collection.

Results obtained using the reduced semantic space and PMI are lower than the average PMI and NPMI approaches for the NYT and Genomics data sets. For the 20NG dataset the results are higher then the average PMI and NPMI using these approaches. The difference in relative performance is down to the nature of these corpora. The words found in topics in the NYT and Genomics datasets are often

| Topic Terms | Human Rating |
|---|---|
| **Top-3** | |
| family wife died son father daughter life became mother born | 2.63 |
| election vote voter ballot state candidate voting percent party result | 3 |
| show television tv news network medium fox cable channel series | 2.82 |
| **Bottom-3** | |
| lennon circus rum whiskey lombardi spirits ranch idol make vineyard | 1.93 |
| privacy andrews elli alexander burke zoo information chung user regan | 1.25 |
| twitter board tweet followers conroy halloween kay hands emi post | 1.53 |

Figure 2: Top-3 and bottom-3 ranked topics using Topic Word Space in NYT together with human ratings.

polysemous or collocate with terms which become context features. For example, one of the top context features of the word "coast" is "ivory" (from the country). However, that feature does not exist for terms that are related to "coast", such as "beach" or "sea". The majority of topics generated from 20NG contain meaningless terms due to the noisy nature of the dataset (emails) but these do not suffer from the same problems with ambiguity and prove to be useful for comparing meaning when formed into the semantic space.

Similar results are obtained for the reduced semantic space using NPMI as the association measure. Results in NYT and Genomics are normally 0.01 lower while for 20NG are 0.01 higher for the majority of the methods. This demonstrates that weighting co-occurrence vectors using NPMI produces little improvement over using PMI, despite the fact NPMI has better performance when the average similarity between each pair of topic terms is computed.

When the topic word space is used there is a consistent improvement in performance compared to the average PMI (Newman et al., 2010b) and NPMI approaches. More specifically, cosine similarity using PMI is consistently higher (0.05-0.06) than average PMI for all datasets and 0.02 to 0.04 higher than average NPMI (0.76, 0.79, 0.8 for NYT, 20NG and Genomics respectively). One reason for this improvement in performance is that the noise caused by polysemy and high dimensionality of the context features of the topic words is reduced. Moreover, cosine similarity scores in the reduced semantic space are higher than average PMI and NPMI in all of the datasets, demonstrating that vector-based representation of the topic words is better than computing their average relatedness. Table 2 shows the top-3 and bottom-3 ranked topics in NYT together with human ratings.

Another interesting finding is that the cosine metric produces better estimates of topic coherency compared to Dice and Jaccard in the majority of cases, with the exception of 20NG in reduced semantic space using PMI. Furthermore, similarity to the topic centroid achieves performance comparable to cosine.

# 6 Conclusions

This paper explored distributional semantic similarity methods for automatically measuring the coherence of sets of words generated by topic models. Representing topic words as vectors of context features and then applying similarity metrics on vectors was found to produce reliable estimates of topic coherence. In particular, using a semantic space that consisted of only the topic words as context features produced the best results and consistently outperforms previously proposed methods for the task.

Semantic space representations have appealing characteristics for future work on tasks related to topic models. The vectors used to represent topic words contain co-occurring terms that could be used for topic labelling (Lau et al., 2011). In addition, tasks such as determining topic similarity (e.g. to identify similar topics) could naturally be explored using these representations for topics.

## Acknowledgments

The research leading to these results was carried out as part of the PATHS project (`http://paths-project.eu`) funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 270082.

## References

Loulwah AlSumait, Daniel Barbará, James Gentle, and Carlotta Domeniconi. Topic Significance Ranking of LDA Generative Models. *Machine Learning and Knowledge Discovery in Databases*, pages 67–82, 2009.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors. In *Proceedings of the International Conference on Machine Learning*, pages 25–32, 2009.

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

G. Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology (GSCL '09)*, Potsdam, Germany, 2009.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pages 1024–1033, 2007.

Allison June-Barlow Chaney and David M. Blei. Visualizing topic models. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (IWSCM)*, 2012.

Jonathan Chang, Jordan Boyd-Graber, and Sean Gerrish. Reading Tea Leaves: How Humans Interpret Topic Models. *Neural Information*, pages 1–9, 2009.

James R. Curran. From distributional to semantic similarity. *Ph.D. Thesis, University of Edinburgh*, 2003.

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

Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Springer, 1994.

Aria Haghighi and Lucy Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370, Boulder, Colorado, 2009.

Zellig Sabbettai Harris. Distributional structure. *Word*, 10:146–162, 1954.

Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, pages 50–57, Berkeley, California, United States, 1999.

Aminul Md. Islam and Diana Inkpen. Second Order Co-occurrence PMI for Determining the Semantic Similarity of Words. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, pages 1033–1038, 2006.

Jey Han Lau, Karl Grieser, David Newman, and Timothy Baldwin. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1536–1545, Portland, Oregon, USA, June 2011.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK, 2011.

Claudiu C. Musat, Julien Velcin, Stefan Trausan-Matu, and Marian A. Rizoiu. Improving topic evaluation using conceptual knowledge. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence (IJCAI '11)*, pages 1866–1871, Barcelona, Spain, 2011.

David Newman, Timothy Baldwin, Lawrence Cavedon, Eric Huang, Sarvnaz Karimi, David Martinez, Falk Scholer, and Justin Zobel. Visualizing search results and document collections using topic maps. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(23):169–175, 2010a.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT '10)*, pages 100–108, Los Angeles, California, 2010b.

Sebastian Padó and Mirella Lapata. Constructing semantic space models from parsed corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan, 2003.

Judita Preiss. Identifying comparable corpora using LDA. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 558–562, Montréal, Canada, 2012.

Eduardo H. Ramirez, Ramon Brena, Davide Magatti, and Fabio Stella. Topic Model Validation. *Neurocomputing*, 76(1):125–133, 2012.

Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188, 2010.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation Methods for Topic Models. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09)*, pages 1105–1112, Montreal, Quebec, Canada, 2009.

Xing Wei and W. Bruce Croft. 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 (SIGIR '06)*, pages 178–185, 2006.