Labeled Topics for News Corpora Using Word Embeddings and Keyword Identification

Conference Paper · July 2016

CITATION
1

READS
126

3 authors, including:

Abdulkareem Alsudais
Claremont Graduate University

11 PUBLICATIONS **4** CITATIONS

SEE PROFILE

Labeled Topics for News Corpora Using Word Embeddings and Keyword Identification

Abdulkareem Alsudais, Hovig Tchalian, Brian Hilton
Claremont Graduate University
{abdulkareem.alsudais, hovig.tchalian, brian.hilton}@cgu.edu

Abstract

In this paper, a simple pipeline for identifying labeled topics for temporally ordered and topic-specific news corpora using word embeddings and keyword identification is proposed. The steps in the pipeline rely on NLP techniques to identify keywords, corpus periodization, and word embeddings. The proposed method is evaluated by applying it on a dataset consisting of TV and Radio transcripts on "Donald Trump." The results demonstrated that the topics captured using this pipeline are more coherent and informative than ones generated using Latent Dirichlet Allocation. Findings from this preliminary experiment suggest that word embeddings models and common NLP keyword identification techniques can be used to identify coherent and labeled topics for a temporally ordered news corpus.

1 Introduction

"Word embeddings" refer to models that create dense vector representation for words or phrases in a text corpus by utilizing their immediate syntactic context, defined by a window with proximate terms. Recently, these models have gained popularity, partially due to advances in computing powers that have made creating vectors for large corpora more feasible [Goth, 2016]. Successful models for generating vector representations for words and phrases such as word2vec and its SKIPGRAM model [Mikolov *et al.*, 2013a], GloVe [Pnnington *et al.*, 2014], Swivel [Shazeer *et al.*, 2016] and others [Levy and Goldberg, 2014b; Turian *et al.*, 2010] have proven to be successful in performing various language-related tasks. These tasks include solving analogies equations and generating word similarities.

Researchers and scientists in Natural Language Processing (NLP) have leveraged these word embeddings models to solve various research problems related to text corpora. For example, Mikolov *et al.*, [2013b] used them to translate texts between English and Spanish; Alemi and Ginsparg [2015] used them to segment text documents; Lebret *et al.*, [2015] used them to generate image captions; and Leeuwenberga *et al*,.[2016] used them to find synonyms for words. These examples demonstrate the effectiveness of solutions that rely on word and phrase vectors as produced by the aforementioned word embeddings models. The success of these approaches also highlights the potential of word embeddings models for solving common and current NLP and text mining related research problems. In this paper, the ability to utilize word embeddings models and keyword identification techniques to generate labeled topics for news corpora is investigated.

One common feature of word embeddings models is a function that identifies a list of words or phrases that are most similar to a given word. This function simply locates the words or phrases that have similar vectors to a given word. According to Mikolov *et al.*, [2013c], the types of similarities returned by the function in SKIPGRAM varies, while Levy and Goldberg [2014a] suggest that the similarities are mostly topical.

While representing a significant step forward, current word embeddings tools and models nonetheless still require customization, modification, and enhancement in order to accomplish domain-specific and NLP-related tasks. Therefore, researchers studying and examining text corpora might benefit from learning how other researches created pipelines or blueprints that utilize word embeddings models to solve and complete specific problems. Additionally, new solutions that leverage recent advances in NLP to create novel applications of word embeddings tools have the potential to advance the NLP field even further.

Most word embeddings methods, for instance, rely only on the linear context of the text, which results in losing additional contextual information present in the text. Levy and Goldberg [2014a] proposed a new word embeddings method that generalizes SKIPGRAM by preserving dependencies such as "direct object" or "nominal subject" that each word in the text represents. They argued that incorporating these dependencies improved the similarity results generated by SKIPGRAM. We build on Levy and Goldberg research by demonstrating how adding additional contextual markers can enhance the performance of specific tasks popular when using word embeddings models.

One particularly appropriate application of such enhanced solutions is discovering topics of a text corpus, and specifically in the area of topic modeling, an algorithmic approach that generates thematic clusters in a corpus based on the distribution of co-occurrence probabilities across word vectors. Research in topic modeling was pioneered by the work of Blei *et al.*,[2003] and their LDA (Latent Dirichlet Allocation)

method. There are a number of solutions that have extended the LDA method, the most successful and widely accepted of which is Labeled-LDA [Ramage *et al.*, 2009]. Labeled-LDA extend LDA and generate not only topics for the corpus, but also labels that define the topics. The algorithm leverages the labels or tags linked to each document in the corpus. For instance, if the examined corpus consists of academic papers on text mining, utilizing the authors' keywords for each paper, the algorithm generates labeled topics for the corpus, such as "entity resolution," and "deep learning."

The growing work in topic modeling represents precisely the kind of opportunity for the enhanced NLP applications we discussed above. When analyzing text corpora, researchers commonly use topic modeling algorithms to generate topics latent in a text corpus. These topics can be used to establish context for the corpus, and to identify, efficiently and effectively, the most important themes in the corpus. To the best of our knowledge, there has not been any work on capturing labeled topics for a corpus using word embeddings.

In this paper, the topical similarities as identified by SKIPGRAM are leveraged to generate corpus-wide labeled topics. We find that utilizing the similarities as generated by Mikolov *et al.*, [2013a] to generate labeled topics offers a significant improvement over approaches that do not incorporate word embeddings to generate topics. The proposed pipeline relies on recent advances in temporal text mining an widely accepted Natural Language Processing techniques such as tokenizing, lemmatizing, Part of Speech (POS) tagging, and keyword identification to generate "context" for the corpus. We argue that by providing "context" for the corpus, SKIPGRAM can be used to generate labeled topics that are more informative than ones generated by LDA, in particular for news corpora. Our application demonstrates that word embeddings can be effective and informative in generating labeled topics for text corpora when used to supplement common NLP techniques performed on the corpus.

2 Methodology

In this section, the steps in the proposed pipeline are briefly described. Figure 1 illustrates the entire pipeline, including the results achieved after each step.

2.1 Periodization

Corpus periodization is the process of segmenting a corpus into a set of smaller and discursively coherent periods while retaining the chronological order of the corpus. Social scientists in fields such as sociology and history commonly use periodization to study various changes in a specific discourse across time by fragmenting a corpus into a set of focal periods [Morley and Bayley, 2009] . For example, when Ruef [1999] examined thirty years of textual news articles on market reform in the U.S. healthcare sector, he identified historical events and political acts in this time frame, and used them to segment the corpus into periods. Corpus periodization is incorporated in the proposed pipeline and utilized to identify the most important keyowords or noun phrases of the corpus. We explain the periodization method and the justification for using it in another paper [Alsudais and Tchalian, 2016].

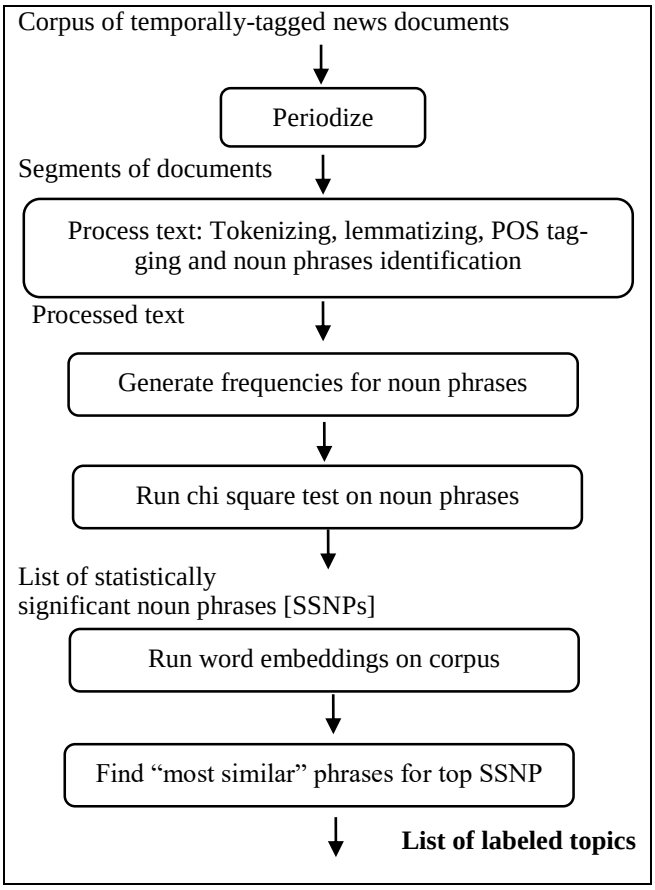

Figure 1. The steps in the propsed pipeline

2.2 Text Preprocesing

In this step, a number of common text processing techniques are applied on the corpus. The purpose of this step is to refine the corpus and only retain information relevant to the process of topic labeling and generation. First, all the news articles in each period are tokenized into a set of sentences. Then, all words in the sentences are tokenized, transformed to lower case, and subsequently lemmatized, the accepted approach within the NLP field. By changing words to lower case and lemmatizing them, natural synonyms such as "Companies" and "company" get combined and counted as the same entity.

2.3 Noun Phrases Extraction

The purpose of this step is to identify the noun phrases most central to the corpus, a critical intermediate step in this pipeline. To complete this step, all the noun phrases in each period are captured and their frequencies are counted. Afterwards, a chi square goodness of fit test is computed with the frequency counts of the noun phrases in each period as the columns. The result of this step is a list of chi square values for each unique noun phrase in the corpus. A list of statically significant noun phrases is then created according to a selected significance level or threshold. The list includes all the noun phrases that pass the defined threshold level.

2.4 Word Embeddings

After preprocessing the corpus, removing all non-noun phrases, and saving noun phrases in the corpus as single units, in this step, a word embeddings model is run on the modified corpus. The result of this step is a dense vector representation for the noun phrases in the text. SKIPGRAM is used in this paper to generate the vector representations for the noun phrases. Since various word embeddings models can produce different vectors for the same word or phrase in a corpus, using a different word embeddings model may not result in generating topical similarities identical to the ones SKIPGRAM generates.

2.5 Labeled Topics Generation

The final step in this pipeline is to produce labeled topics for the corpus. Using the "most similar" function in word2vec, lists of the most similar noun phrases are generated for all the statistically significant noun phrases (SSNP). This function simply calculates and detects the noun phrases that have vectors that are most similar to a given word or phrase. The SSNPs are then used as labels for the generated topics. While only topics for the top statically significant noun phrases were captured for the purposes of this paper, further extensions of this work will also identify topics for all SSNPs and combine similar and overlapping ones to create multi-labeled topics that fully capture the all the most popular themes in the corpus.

3 Experiment

In this section, the results of applying the proposed method on a corpus of temporally tagged news corpora are demonstrated.

3.1 Dataset

In this paper, a dataset consisting of transcripts of news-related television and radio shows in which the name "Donald Trump" appeared is used. This dataset is selected to demonstrate the effectiveness of the method proposed in this paper in generating labeled topics for temporally ordered and topic-specific news corpora.

Donald Trump announced his candidacy for President of the United States on June 16, 2015. Ever since, a considerable public dialogue has been taking place around the various 'hot button' topics his candidacy has elicited. Whatever the merit of the dialogue itself, its time-bound nature, considerable volume and focus on a limited number of topics makes it an ideal dataset for purposes of evaluating the method proposed in this paper. In particular, the method can be evaluated by measuring its accuracy in capturing and labeling the topics latent in the news corpus representing the political dialogue.

Two sources were used for the dataset: Fox News Network and National Public Radio (NPR), which together provide a substantial sample of the range of political discussions surrounding the candidate. The dataset includes transcripts of shows that aired from June 1st, 2015 to March 31st, 2016. The total number of news transcripts is 2,271 documents. Just over a third of the corpus, 1,528 documents, aired on Fox News Network while the rest, 743 documents, aired on NPR.

3.2 Periodization

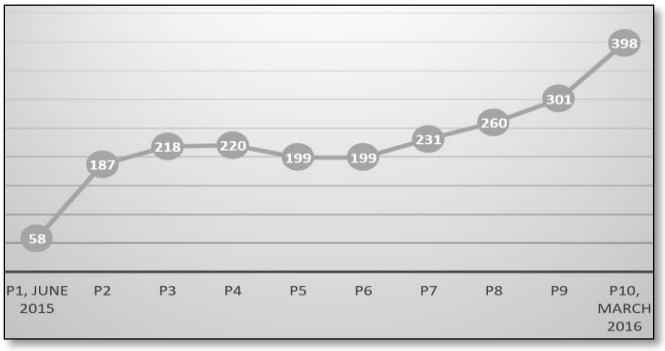

Figure 2. The number of articles in each period.

When the number of the initial temporal break points in the corpus is small, using a corpus periodization method to segment the corpus is not necessary. In this dataset, the initial points were the individual months between June, 2015 and March 2016. Due to the small number of the initial points, natural break points were used in this experiment and articles were grouped according to the month and year combination of when they were published. Accordingly, a set of ten periods was created. Figure 2 shows the breakdown of the periods, and the number of articles in each one.

Rank	Keyword	Rank	Keyword
1	Delegate	13	Cleveland
2	Scott walker	14	Cruz
3	Convention	15	Caucus
4	Paris	16	Iowa
5	Biden	17	South carolina
6	Carly Fiorina	18	Assad
7	Muslims	19	Nevada
8	Iowa caucuse	20	Iowa and new hampshire
9	Illegal immigrant	21	Committee
10	Carson	22	Planned Parenthood
11	Server	23	Putin
12	Home state	24	Iran
25	Caucuse	37	Gun
26	New Hampshire	38	Primary
27	Kasich	39	Syria
28	Murder	40	Paul ryan
29	Mexico	41	Shooting
30	Terrorism	42	Nuclear Weapon
31	Sanders	43	Russia
32	Refugee	44	Afghanistan
33	Illegal immigration	45	Agreement
34	Jeb Bush	46	Nuclear deal
35	Michigan	47	Sanction
36	Benghazi	48	Outsider

Table 1. Top SSNPs, as captured by periodization, the chi-square test and traditional bag-of-words approaches

3.3 Noun Phrases / Keywords

After preprocessing the corpus using the techniques described in section 2.2, all noun phrases were identified along with the number of times they were used in each period. Accordingly, the chi square values were computed for each noun phrase. Finally, a list of Statistically Significant Noun Phrases (SSNP) was generated based on the chi square values of the noun phrases at a significance level of 95%. The list contained over 700 noun phrases. In table 1, the top 48 SSNPs are listed. These words and phrases are later used as labels that define the topics. This list alone provides some contexts and summary of the corpus and the main themes discussed in the processed 2,271 documents across the ten periods.

3.4 Labeled topics

After identifying the list of statistically significant noun phrases, all the documents in the dataset were processed using word2vec to create dense vector representations for all the unique phrases in the corpus. The final step in the pipeline was completed by simply querying word2vec to retrieve the lists of the most similar word or phrases to each of the words and phrases in the SSNPs list. These lists are used as topics, and the SSNP used to create them are used as the labels for the topics.

We observed that labeled topics are not as informative when the label or noun phrase is a proper noun, such as "Biden" or "Cruz." For example, for the label "Cruz," the topic contained the names of a number of other presidential candidates such Rubio and Bush. Thus, more rigorous testing is needed to implement methodical changes that systematically, and perhaps iteratively, refine the topics. Additionally, there are instances where the labeled topics overlap and share the same underlying terms. Therefore, identifying and combining these topics is needed.

In table 2, a sample of the generated topics is displayed with more emphasis on topics that are not labeled with a proper noun. The effectiveness of the performance of the method proposed in this paper is demonstrated by comparing the generated labeled topics to topics generated by LDA on the same corpus. Some topics such "Planned parenthood" and "Terrorism" are more coherent than others.

Selected Labeled Topics Generated by the Method Proposed in this Paper		**Topics as Generaetd by LDA**
Label	**Topic**	
Delegate	['ballot', 'convention', 'primary', 'first ballot', '1,237 delegate', '30 percent', '50 percent', 'republican primary', '20 percent', 'contested convention']	'-- ', 'bolling', 'williams', 'guilfoyle ', 'gutfeld ', 'perino: ', 'that's', 'video', 'right', 'clip'
Scott walker	['rick santorum', 'john kasich', 'rick perry', 'mike huckabee', 'chris christie', 'rand paul', 'lindsey graham', 'mitt romney', 'marco rubio', 'jeb bush']	'trump ', 'donald ', 'cruz ', 'ted ', 'republican ', "he's ", 'hillary ', 'rubio ', 'win ', 'campaign'
Paris	['brussels', 'belgium', 'paris attack', 'terror attack', 'france', 'san bernardino', 'mali', 'turkey', 'isis terrorist', 'bombing']	'kelly', '-- ', 'so,', 'unidentified', 'video', 'clip', 'know', 'male', 'well, ', "he's"
Muslims	['complete shutdown', 'more muslims', 'temporary ban', 'southern border', 'christians', 'refugee', 'proposed ban', 'jihad', 'total and complete shutdown', 'fear']	'carlson', 'planned ', 'abortion', 'parenthood', 'pro-life', 'body', 'fields', 'parenthood', 'rose ', 'abortion'
Iowa caucuse	['super tuesday', 'new hampshire primary', 'iowa caucus', 'new poll', 'caucuse', 'republican race', 'national poll', 'south carolina primary', 'next week', 'tuesday']	"o'reilly", '-- ', 'unidentified', 'right', 'so, ', 'yes', "that's", 'watters:', 'male', 'clip'
Illegal immigrant	['illegal alien', 'san francisco', 'deportation', 'five time', 'criminal', 'sanctuary city', 'immigrant', 'criminal alien', 'felon', 'citizen']	'-- ', 'wallace', "he's", 'well', "that's", 'debate', 'know', 'republican', 'trump', 'lot'
Server	['e-mail', 'classified information', 'email', 'private server', 'mail', 'top secret', 'state department', 'e', 'private e-mail', 'benghazi']	'siegel', 'robert', 'greene', 'know', 'well', "you're", 'kind', "that's", 'mean', 'young'
Home state	['double digit', 'third place', 'florida', 'second place', 'other state', 'latest poll', 'win', 'ohio and florida', 'winner', 'wisconsin']	'baier', 'president', 'fox', 'news', 'video', '-- ', 'end', 'begin ', 'clip', 'u.s.'
Caucus	['early state', 'caucuse', 'ground game', 'turnout', 'republican primary', 'voting', 'polling', 'iowa and new hampshire', 'primary', 'evangelical']	'trump', 'kurtz', 'recording', 'media', 'donald', 'ari', "he's", 'unidentified', 'press ', 'news\n'
Terrorism	['threat', 'terror', 'homeland', 'radical islam', 'initial response', 'middle east', 'al qaeda', 'islamic state', 'isil', 'region']	'pope ', 'religious', 'carson', 'ben', 'faith', 'christian', 'church', 'catholic', 'joe', 'pope.'
Assad	['vacuum', 'putin', 'coalition', 'red line', 'ukraine', 'vladimir putin', 'force', 'no-fly zone', 'iraq', 'ally']	'trump', '-- ', 'know', 'donald', "we're ", '[applause] ', 'great ', 'it.', 'trump', 'lot'
Planned parenthood	['abortion', 'body part', 'federal funding', 'abortion part', 'entire federal government', 'funding', 'taxpayer funding', 'abortion practice', 'reimbursement', 'aborted fetuse']	'audie', 'cornish', 'scott', 'scott:', 'horsley', 'e.', 'meyers', 'e.j.', 'as', 'dionne'
Committee	['hearing', 'inspector general', 'benghazi committee', 'classified information', 'justice department', 'document', 'email', 'testimony', 'state department', 'mrs. clinton']	'cavuto: ', 'it`s ', '-- ', 'don`t ', 'well, ', 'i`m ', 'that`s ', 'he`s ', 'we`re ', 'so,\n'
Iran	['sanction', 'agreement', 'nuclear weapon', 'nuclear deal', 'north korea', 'iranians', 'deal', 'u.n.', 'nuclear program', 'regime']	'- ', '(soundbite ', 'david ', 'archived ', 'steve ', 'gonyea: ', 'inskeep: ', 'sarah ', "npr's ", 'donald\n'

Table 2. Summary of selected labeled topics as generated by proposed pipeline (left) and topics as generated by LDA.

4. Conclusion and Future Work

In this paper, a simple pipeline for identifying labeled topics for temporally ordered and topic-specific news corpora is proposed. The steps in this pipeline rely on 1) widely accepted Natural Language Processing techniques and approaches that preprocess and clean documents in the corpus, 2) a periodization method that utilizes the temporal features of the corpus to periodize it and create comprehensive textual keywords, and 3) a word embeddings model that creates dense vector representation for all unique words and phrases in the corpus.

The main contributions of this paper are 1) a demonstration that word embeddings models can be used to identify coherent topics for a corpus and 2) a pipeline that can be replicated to produce a list of labeled topics for a temporally ordered news corpus.

The effectiveness of this proposed method was demonstrated by applying it to a corpus consisting of 2,271 television and radio transcripts containing the term "Donald Trump." We demonstrated that the labeled topics generated by the proposed method were more informative than ones generated by Latent Dirichlet Allocation (LDA). Our method captured more coherent topics than ones generated by LDA. Furthermore, this method more accurately captured semantic and syntactic context, as confirmed in the topics labeled "Planned Parenthood" and "Terrorism."

Additional and more rigorous testing and evaluation of this method is necessary. For instance, it is common to preprocess a corpus before generating topics with LDA. We acknowledge that, in this paper, LDA was run on the corpus without applying any preprocessing on the text. Preprocessing the corpus might produce topics that are more concise and relevant. Thus, it is important to run additional tests, in order to determine whether such preprocessing can improve the accuracy, coherence and relevance of the standard LDA approach, possibly helping explain a proportion of the contribution difference between that method and the one proposed in this paper. Very minor discrepancies in the topics generated using the processed and un-processed runs of this method proposed here strongly suggest that the difference would be negligible. Additionally, topic coherence measures such as C_V and C_P [Röder *et al.*, 2015] should be used to evaluate the results and quantitatively asses the coherence of each individual topic.

This work can also be refined and extended in several ways. For example, examining the labeled topics reveals clear overlap between some of the topics. There are opportunities, therefore, to further leverage the similarities between the terms and topics as identified by the word embeddings model, in order to merge and collapse certain topics. The results of this should be a set of topics that provide a comprehensive and complete summary for the corpus being studied. Moreover, examining the labeled topics also reveals different types and classes of topics. Thus, future work includes examining these types and creating a method that systematically classify them.

References

[Alemi and Ginsparg, 2015] Alexander A Alemi and Paul Ginsparg. Text Segmentation based on Semantic Word Embeddings. *arXiv Preprint*. arXiv:1503.05543.

[Alsudais and Tchalian, 2016] Abdulkareem Alsudais and Hovig Tchalian, H. Corpus Periodization Framework to Periodize a Temporally Ordered Text Corpus. In *Proceedings of the 22nd Americas Conference on Information Systems (AMCIS)*.

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

[Goth, 2016] Gregory Goth. Deep or Shallow, NLP Is Breaking Out. *Communication of the ACM*, 59(3), 13–16.

[Lebret *et al.*, 2015] Remi Lebret, Pedro O. Pinheiro, and Ronan Collobert. Phrase-based Image Captioning. *arXiv Preprint.* arXiv:1502.03671.

[Leeuwenberga *et al.*, 2016] Artuur Leeuwenberga, Mihaela Velab, Jon Dehdaribc, and Josef van Genabithb. A Minimally Supervised Approach for Synonym Extraction with Word Embeddings. *The Prague Bulletin of Mathematical Linguistics*, (105), 111–142.

[Levy and Goldberg, 2014a] Omer Levy and Yoav Goldberg. Dependency-Based Word Embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*. pp. 302–308. Baltimore, Maryland, USA: Association for Computational Linguistics.

[Levy and Goldberg, 2014b] Omer Levy and Yoav Goldberg. Neural Word Embedding as Implicit Matrix Factorization. A*dvances in Neural Information Processing Systems*, 2177–2185.

[Mikolov *et al.*, 2013a] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26, 3111–3119.

[Mikolov *et al.*, 2013b] Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *arXiv Preprint*. arXiv:1301.3781.

[Mikolov *et al.*, 2013c] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. Exploiting Similarities among Languages for Machine Translation. *arXiv Preprint*.

[Morley and Bayley, 2009] John Morley and Paul Bayley (Eds). Corpus-assisted discourse studies on the Iraq conflict: Wording the war. Routledge.

[Pnnington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe : Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532–1543. Association for Computational Linguistics.

[Ramage *et al.*, 2009] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. pp. 248–256. Association for Computational Linguistics.

[Röder *et al.*, 2015] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*.

[Ruef, 1999] Martin Ruef. Social ontology and the dynamics of organizational forms: Creating market actors in the healthcare feld, 1966-1994. *Social Forces*, 77(4), 1403–1432. 1999.

[Shazeer *et al.*, 2016] Noam Shazeer, Ryan Doherty, Colin Evans, and Chris Waterson. Swivel: Improving Embeddings by Noticing What's Missing. *arXiv Preprint*. 2016.

[Turian *et al.*, 2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. 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.