Topic Modelling with Word Embeddings

Fabrizio Esposito
Dept. of Humanities
Univ. of Napoli Federico II
fabrizio.esposito3
@unina.it

Anna Corazza, Francesco Cutugno
DIETI
Univ. of Napoli Federico II
anna.corazza|francesco.cutugno
@unina.it

Abstract

English. This work aims at evaluating and comparing two different frameworks for the unsupervised topic modelling of the CompWHoB Corpus, namely our political-linguistic dataset. The first approach is represented by the application of the latent DirichLet Allocation (henceforth LDA), defining the evaluation of this model as baseline of comparison. The second framework employs Word2Vec technique to learn the word vector representations to be later used to topic-model our data. Compared to the previously defined LDA baseline, results show that the use of Word2Vec word embeddings significantly improves topic modelling performance but only when an accurate and task-oriented linguistic pre-processing step is carried out.

Italiano. *L'obiettivo di questo contributo è di valutare e confrontare due differenti framework per l'apprendimento automatico del topic sul CompWHoB Corpus, la nostra risorsa testuale. Dopo aver implementato il modello della latent DirichLet Allocation, abbiamo definito come standard di riferimento la valutazione di questo stesso approccio. Come secondo framework, abbiamo utilizzato il modello Word2Vec per apprendere le rappresentazioni vettoriali dei termini successivamente impiegati come input per la fase di apprendimento automatico del topic. I risulati mostrano che utilizzando i 'word embeddings' generati da Word2Vec, le prestazioni del modello aumentano significativamente ma solo se supportati da una accurata fase di 'pre-processing' linguistico.*

1 Introduction

Over recent years, the development of political corpora (Guerini et al., 2013; Osenova and Simov, 2012) has represented one of the major trends in the fields of corpus and computational linguistics. Being carriers of specific content features, these textual resources have met the interest of researchers and practitioners in the study of topic detection. Unfortunately, not only has this task turned out to be hard and challenging even for human evaluators but it must be borne in mind that manual annotation often comes with a price. Hence, the aid provided by unsupervised machine learning techniques proves to be fundamental in addressing the topic detection issue.

Topic models are a family of algorithms that allow to analyse unlabelled large collections of documents in order to discover and identify hidden topic patterns in the form of cluster of words. While LDA (Blei et al., 2003) has become the most influential topic model (Hall et al., 2008), different extensions have been proposed so far: Rosen-Zvi et al. (Rosen-Zvi et al., 2004) developed an author-topic generative model to include also authorship information; Chang et al. (Chang et al., 2009a) presented a probabilist topic model to infer descriptions of entities from corpora identifying also the relationships between them; Yi Yang et al. (Yang et al., 2015) proposed a factor graph framework for incorporating prior knowledge into LDA.

In the present paper we aim at topic modelling the CompWHoB Corpus (Esposito et al., 2015), a political corpus collecting the transcripts of the White House Press Briefings. The main characteristic of our dataset is represented by its dialogical structure: since the briefing consists of a question-answer sequence between the US press secretary and the news media, the topic under discussion may change from one answer to the fol-

lowing question, and vice versa. Our purpose was to address this main feature of the CompWHoB Corpus associating at each answer/question only one topic. In order to reach our goal, we propose an evaluative comparison of two different frameworks: in the first one, we employed the LDA approach by extracting from each answer/question document only the topic with the highest probability; in the second framework, we applied the word embeddings generated from the Word2Vec model (Mikolov and Dean, 2013) to our data in order to test how dense high-quality vectors represent our data, finally comparing this approach with the previously defined LDA baseline. The evaluation was performed using a set of gold-standard annotations developed by human experts in political science and linguistics. In Section 2 we present the dataset used in this work. In Section 3, the linguistic pre-processing is detailed. Section 4 shows the methodology employed to topic-model our data. In Section 5 we present the results of our work.

2 The dataset

2.1 The CompWHoB Corpus

The textual resource used in the present contribution is the CompWHoB (Computational White House press Briefings) Corpus, a political corpus collecting the transcripts of the White House Press Briefings extracted from the American Presidency Project website, annotated and formatted into XML encoding according to TEI Guidelines (Consortium et al., 2008). The CompWHoB Corpus spans from January 27, 1993 to December 18, 2014. Each briefing is characterised by a turn-taking between the podium and the journalists, signalled in the XML files by the use of a *u* tag for each utterance. At the time of writing, 5,239 briefings have been collected, comprising 25,251,572 tokens and a total number of 512,651 utterances (from now on, utterances will be referred to as 'documents'). The document average length has been measured to 49.25 tokens, while its length variability is comprised within a range of a minimum of 0 and a maximum of 4724 tokens. The dataset used in the present contribution was built and divided into training and test set by randomly selecting documents from the CompWHoB Corpus in order to vary as much as possible the topics dealt with by US administration.

2.2 Gold-Standard Annotation

Two hundred documents of the test set were manually annotated by scholars with expertise in linguistics and political science using a set of thirteen categories. Seven macro-categories were created taking into account the US major federal executive departments so as not to excessively narrow the topic representation, accounting for 28.5% of the labelled documents. Six more categories were designed in order to take into account the informal nature of the press briefings that makes them an atypical political-media genre (Venuti and Spinzi, 2013), accounting for the remaining 71.5% (Table 1). The labelled documents represent the gold-standard to be used in the evaluation stage. This choice is motivated by the fact that even if metrics such as perplexity or held-out likelihood prove to be useful in the evaluation of topic models, they often fail in qualitatively measuring the coherence of the generated topics (Chang et al., 2009b). Thus, more formally our gold-standard can be defined as the set $G = \{g_1, g_2, ..., g_S\}$ where g_i is the ith category in a range $\{1, S\}$ with $S = 13$ as the total number of categories.

Crime and justice	Culture and Education
Economy and welfare	Foreign Affairs
Greetings	Health
Internal Politics	Legislation & Reforms
Military & Defense	President Updates
Presidential News	Press issues
Unknown topic	

Table 1: Gold-Standard Topics

3 Linguistic Pre-Processing

In order to improve the quality of our textual data, special attention was paid to the linguistic pre-processing step. In particular, since LDA represents documents as mixtures of topics in forms of words probability, we wanted these topics to make sense also to human judges. Being press briefings actual conversations where the talk moves from one social register to another (e.g. switch from the reading of an official statement to an informal interaction between the podium and the journalists) (Partington, 2003), the first step was to design an *ad-hoc* stoplist able to take into account the main features of this linguistic genre. Indeed, not only were words with a low frequency discarded,

but also high frequency ones were removed in order not to overpower the rest of the documents. More importantly, we included in our stoplist all the personal and indefinite pronouns as well as the most commonly used honorifics (e.g. Mr., Ms., etc.), given their predominant role in addressing the speakers in both informal and formal settings (e.g. "Mr. Secretary, you said oil production is up, [...]"). Moreover, the list of the first names of the press secretaries in office during the years covered by the CompWHoB Corpus was extracted from Wikipedia and added to the stoplist, since most of the time used only as nouns of address (Brown et al., 1960). As regards the proper NLP pipeline implemented in this work, the Natural Language ToolKit[1] (NLTK) platform (Bird et al., 2009) was employed: word tokenization, POS-tagging, using the Penn Treebank tag set (Marcus et al., 1993) and lemmatization were carried out to refine our data. When pre-processing is not applied to the dataset, only punctuation is removed from the documents.

4 Methodology

This section deals with the two techniques employed in this work to topic-model our data. We first discuss the LDA approach and then focus on the use of the word embeddings learnt employing Word2Vec model. Both the techniques were implemented in Python (version 3.4) using the Gensim[2] library (Rehurek and Sojka, 2010).

4.1 Latent DirichLet Allocation

In our first experiment we ran LDA, a generative probabilistic model that allows to infer latent topics in a collection of documents. In this unsupervised machine learning technique the topic structure represents the underlying *hidden* variable (Blei, 2012) to be discovered given the *observed* variables, i.e. documents' items from a fixed vocabulary, be them textual or not. More formally, LDA describes each document d as multinomial distribution θ_d over topics, while each topic t is defined as a multinomial distribution ϕ_t over words in a fixed vocabulary where $i_{d,n}$ is the nth item in the document d.

4.1.1 Topic modelling with LDA

Data were linguistically pre-processed prior to training LDA model and only words pos-tagged as nouns ('NN') were kept in both the training and test sets' documents. This choice was motivated by the necessity of generating topics that could be semantically meaningful. After having carried out the pre-processing step, we trained LDA model on our training corpus by employing the online variational Bayes (VB) algorithm (Hoffman et al., 2010) provided by the Gensim library. Based on online stochastic optimization with a natural gradient step, LDA online proves to converge to a local optimum of the VB objective function. It can be applied to large streaming document collections being able to make better predictions and find better topic models with respect to those found with batch VB. As parameters of our model, we set the k number of topics to thirteen as the numbers of classes in our gold-standard, updating the model every 150 documents and giving two passes over the corpus in order to generate accurate data. Once the model was trained, we inferred topic distributions on the unseen documents of the test set. For each document d_i, the topic $t_{max(i)}$ with the highest probability in the multinomial distribution was selected and associated to it. The cluster ω_k corresponds then to the set of documents associated to the topic t_k. Due to the presence of a gold-standard, the *external criterion* of *purity* was chosen as evaluation measure of this approach. *Purity* is formally defined as:

$$\textit{purity}(\Omega, G) = \frac{1}{N}\sum_{k}\max_{j}|w_k \cup g_j|$$

$\Omega = \{\omega_1, \omega_2, ..., \omega_K\}$ is the set of clusters and $G = \{g_1, g_2, ..., g_S\}$ is the set of gold-standard classes. The *purity* computed for the LDA approach is:

$$purity \approx 0.46$$

This measure constituted the baseline of comparison with the Word2Vec word embeddings approach.

4.2 Word2Vec

Word2Vec (Mikolov et al., 2013a) is probably the most popular software providing learning models for the generation of dense embeddings. Based on Zelig Harris' *Ditributional Hypothesis* (Harris, 1954) stating that words occurring in similar contexts tend to have similar meanings, Word2Vec model allows to learn vector representations of words referred to as *word embeddings*. Differently from techniques such as LSA (Dumais, 2004),

[1] http://www.nltk.org

[2] https://radimrehurek.com/gensim/

LDA and other topic models that use *documents* as context, Word2Vec learns the distributed representation for each target word by defining the context as the terms surrounding it. The main advantage of this model is that each dimension of the embedding represents a latent feature of the word (Turian et al., 2010), encoding in each word vector essential syntactic and semantic properties (Mikolov et al., 2013c). In this way, simple vector similarity operations can be computed using *cosine similarity*. Moreover, it must not be forgotten that one of Word2Vec's secrets lies in its efficient implementation that allows a very robust and fast training.

4.2.1 Topic modelling with Word2Vec

Training data were linguistically pre-processed beforehand according to the *ad-hoc* pipeline implemented in this work. The model was initialised setting a minimum count for the input words: terms whose frequency was lower than 20 were discarded. In addition, we set the default threshold at $1 \exp -3$ for configuring the high-frequency words to be randomly downsampled in order to improve word embeddings quality (Mikolov and Dean, 2013). Moreover, as highlighted by Goldberg and Levy (Goldberg and Levy, 2014), both sub-sampling and rare-pruning seem to increase the effective size of the window making the similarities more topical. Finally, based on the recommendation of Mikolov et al. (Mikolov et al., 2013b) and Baroni et al. (Baroni et al., 2014), in this work we trained our model using the CBOW algorithm since more suitable for larger datasets. The dimensionality of our feature vectors was fixed at 200. Once constructed the vocabulary and trained the input data, we used the learnt word vector representations on our unseen test set documents. Then, we calculated the centroid c for each document d, where $e_{d,i}$ is the ith embedding in d, so as to obtain a meaningful topic representation for each document (Mikolov and Dean, 2013). Finally, we clustered our data using the k-means algorithm. In order to compare our approach with the baseline previously defined, the external criterion of *purity* was computed also in this experiment to evaluate how well the k-means clustering matched the gold-standard classes:

$$purity \approx 0.54$$

This technique proved to outperform the LDA topic model approach presented in this work. Surprisingly, notwithstanding the fact that Word2Vec relies on a broad context to produce high-quality embeddings, this framework showed to perform better using a linguistically pre-processed dataset where only nouns are kept. Table 2 shows the results obtained in the two experiments.

Topic Models Results	
Framework	Results
LDA without pre-processing	**0.45**
LDA with pre-processing	**0.46**
Word2Vec without pre-processing	**0.44**
Word2Vec with pre-processing	**0.54**

Table 2: Results of the two frameworks. When pre-preprocessing is not applied, only punctuation is removed.

5 Conclusions

In this contribution we have presented a comparative evaluation of two unsupervised learning approaches to topic modelling. Two experiments were carried out: in the first one, we applied a classical LDA model to our dataset; in the second one, we trained our model using Word2Vec so as to generate the word embeddings for topic-modelling our test set. After clustering the output of the two approaches, we evaluated them using the external criterion of purity. Results show that the use of word embeddings outperforms the LDA approach but only if a linguistic task-oriented pre-processing stage is carried out. As at the moment no comprehensive explanation can be provided, we can only suggest that the main reason for these results may lie in the fluctuating length of each document in our dataset. In fact, we hypothesise that the use of word embeddings may prove to be the boosting factor of Word2Vec topic model since encoding information about the close context of the target term. As part of future work, we aim to further investigate this aspect and design a topic model framework that could take into account the main structural and linguistic features of the CompWHoB Corpus.

Acknowledgments

The authors would like to thank Antonio Origlia for the useful and thoughtful discussions and insights.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

David M Blei. 2012. Probabilistic Topic Models. *Communications of the ACM*, 55(4):77–84.

Roger Brown, Albert Gilman, et al. 1960. The Pronouns of Power and Solidarity. *Style in language*, pages 253–276.

Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. 2009a. Connections between the Lines: Augmenting Social Networks with Text. In *Knowledge Discovery and Data Mining*.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009b. Reading Tea Leaves: How Humans Interpret Topic Models. In *Advances in neural information processing systems*, pages 288–296.

TEI Consortium, Lou Burnard, Syd Bauman, et al. 2008. *TEI P5: Guidelines for electronic text encoding and interchange*. TEI Consortium.

Susan T Dumais. 2004. Latent semantic analysis. *Annual review of information science and technology*, 38(1):188–230.

Fabrizio Esposito, Pierpaolo Basile, Francesco Cutugno, and Marco Venuti. 2015. The CompWHoB Corpus: Computational Construction, Annotation and Linguistic Analysis of the White House Press Briefings Corpus. *CLiC it*, page 120.

Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Marco Guerini, Danilo Giampiccolo, Giovanni Moretti, Rachele Sprugnoli, and Carlo Strapparava. 2013. The New Release of CORPS: A Corpus of Political Speeches Annotated with Audience Reactions. In *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 86–98. Springer.

David Hall, Daniel Jurafsky, and Christopher D Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 363–371. Association for Computational Linguistics.

Zellig S Harris. 1954. Distributional Structure. *Word*, 10(2-3).

Matthew Hoffman, Francis R Bach, and David M Blei. 2010. Online Learning for latent Dirichlet Allocation. In *advances in neural information processing systems*, pages 856–864.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.

T Mikolov and J Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in neural information processing systems*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013b. Exploiting Similarities among Languages for Machine Translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, volume 13, pages 746–751.

Petya Osenova and Kiril Simov. 2012. The Political Speech Corpus of Bulgarian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Alan Partington. 2003. *The Linguistics of Political Argument: The Spin-Doctor and the Wolf-Pack at the White House*. Routledge.

Radim Rehurek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-Topic Model for Authors and Documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 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*, pages 384–394. Association for Computational Linguistics.

M Venuti and C Spinzi. 2013. Tracking the change in institutional genre: a diachronic corpus-based study of White House Press Briefings. *The three waves of globalization: winds of change in Professional, Institutional and Academic Genres*.

Yi Yang, Doug Downey, Jordan Boyd-Graber, and Jordan Boyd Graber. 2015. Efficient Methods for Incorporating Knowledge into Topic Models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.