# Improving Topic Quality by Promoting Named Entities in Topic Modeling

**Katsiaryna Krasnashchok**
EURA NOVA
Rue Emile Francqui, 4
1435 Mont-Saint-Guibert, Belgium
katherine.krasnoschok@euranova.eu

**Salim Jouili**
EURA NOVA
Rue Emile Francqui, 4
1435 Mont-Saint-Guibert, Belgium
salim.jouili@euranova.eu

## Abstract

News-related content has been extensively studied in both topic modeling research and named entity recognition. However, expressive power of named entities and their potential for improving the quality of discovered topics has not received much attention. In this paper we use named entities as domain-specific terms for news-centric content and present a new weighting model for Latent Dirichlet Allocation. Our experimental results indicate that involving more named entities in topic descriptors positively influences the overall quality of topics, improving their interpretability, specificity and diversity.

## 1 Introduction

News-centric content conveys information about events, individuals and other entities. Analysis of news-related documents includes identifying hidden features for classifying them or summarizing the content. Topic modeling is the standard technique for such purposes, and Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is the most used algorithm, which models the documents as distribution over topics and topics as distribution over words. A good topic model is characterized by its coherence: any coherent topic should contain related words belonging to the same concept. A good topic must also be distinctive enough to include domain-specific content. For news-related texts domain-specific content can be represented by named entities (NE), describing facts, events and people involved in news and discussions. It explains the need to include named entities in topic modeling process.

The main contribution of this work is improving topic quality with LDA by increasing the importance of named entities in the model. The idea is to adapt the topic model to include more domain-specific terms (NE) in the topic descriptors. We designed our model to be flexible, in order to be used in different variations of LDA. We ultimately employ a term-weighting approach for the LDA input. Our results show that: i) named entities can serve as favorable candidates for high-quality topic descriptors, and ii) weighting model based on pseudo term frequencies is able to improve overall topic quality without the need to interfere with LDA's generative process, which makes it adaptable to other LDA variations.

The paper is organized in the following manner: in Section 2 we present the related work; Section 3 describes the proposed solution and is followed by Section 4, where the details of evaluation process and results are outlined. We finish with Section 5, concluding the results and next steps.

## 2 Related Work

This section describes the related work in the area of topic modeling, specifically LDA.

### 2.1 Topic Modeling and Named Entities

Several works explored the relation between LDA and named entities in recent years. The most famous model is CorrLDA2 (Newman et al., 2006). It introduces two types of topics, general and entity, and represents word topics as a mixture of entity topics. Hu et al. (2013) reverses the concept, assuming that entities are critical for news-centric content. Their entity-centered topic model (ECTM) designs entity topics as a mixture of word topics and shows better results in entity prediction than CorrLDA2 (Hu et al., 2013). Both models, however, introduce significant changes to the LDA algorithm. In this paper we strive to incorporate named entities into LDA in a natural way, without affecting the generative algorithm, to keep it



*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 247–253
Melbourne, Australia, July 15 - 20, 2018. 

flexible and adaptable to any LDA variations.

Lau et al. (2013) study the impact of collocations on topic modeling and work with the input of LDA by replacing unigrams with collocations. Adding multiword named entities, as a special type of collocations, enhanced the topic model for the tested dataset (Lau et al., 2013). Our work follows similar tokenization process, but goes further in improving the topic model by promoting named entities in it.

### 2.2 Topic Modeling and Term Weighting

Traditionally, the input of LDA is a document-term matrix of term frequencies (TF), according to the bag-of-words model (BoW). However, Wilson and Chew (2010) showed that point-wise mutual information (PMI) term weighting model can be successfully applied to eliminate stop words from topic descriptors. More weighting schemes were evaluated by Truica et al. (2016) and showed promising results for clustering accuracy. Therefore, term weighting approach in LDA can be beneficial for certain tasks. In this paper we introduce unnormalized TF-based weighting scheme using pseudo frequency as a way of increasing the weight of a term.

## 3 Proposed model

LDA model has been criticized for favoring highly frequent, general words in topic descriptors (O'Callaghan et al., 2015). This problem can be partly solved by eliminating domain-specific stopwords from the corpus. On the other hand, instead of narrowing the corpus, it may be more efficient to promote domain-specific important words, especially if such words can be identified automatically, like named entities. In this paper we deal with the online Variational Bayes version of the LDA algorithm from Hoffman et al. (2010), as alternative to collapsed Gibbs sampling, used by Wilson and Chew (2010) and Truica et al. (2016) to incorporate weights into the LDA model. In Hoffman et al. (2010) the authors demonstrate that the objective of the optimization relies only on the counts of terms in documents, and therefore documents can be summarized by their TF values. Our proposed model takes the TF scores as initial term weights (unnormalized). To increase the weight of a named entity we add a pseudo-frequency to its TF without changing the weights of other terms. This strengthens the chances of NE to appear in a topic descriptor, even if originally it was not mentioned often in the corpus. There are multiple ways of increasing the weights, *e.g.* we can promote all NE in the same proportion, or set their weights separately for each document in the corpus.

### 3.1 Independent Named Entity Promoting

NE Independent model assumes that all named entities in the corpus are $\alpha$ times more important than their initial weights (TF), i.e. they may not be the most important terms in the corpus, but they should weigh $\alpha$ times more than they do now. Therefore, for each column $m_w$ of document-term matrix $M$, we apply scalar multiplication:

$$m_w = \begin{cases} \alpha * m_w & \text{if } w \text{ is NE} \\ m_w & \text{otherwise} \end{cases} \quad (1)$$

By varying $\alpha$, we can set the importance of named entities in the corpus and impact the outcome of topic modeling. The value need not be an integer, since typical LDA implementation can deal with any numbers. In Section 4 we provide results for several tested values of $\alpha$ parameter and discuss our findings.

### 3.2 Document Dependent Named Entity Promoting

While we want the topics produced by LDA to include more named entities as domain-specific words, we may assume that NE, in fact, should be the most important, i.e. the most frequent, terms in each document. In order to set the weights accordingly, the maximum term-frequency per document is calculated and added to each named entity's weight in each document:

$$m_{dw} = \begin{cases} m_{dw} + \max\limits_{w} m_{dw} & \text{if } w \text{ is NE} \\ m_{dw} & \text{otherwise} \end{cases} \quad (2)$$

This weighting scheme obliges named entities to be the "heaviest" terms in each document. At the same time, we do not change the weight of other frequent terms, so eventually they still have a high probability to make the top terms list.

## 4 Evaluation

We designed a series of tests to evaluate our proposed model: a) **Baseline Unigram**: basic model on the corpus consisting of single tokens

(no named entities involved); b) **Baseline NE**: basic model on the corpus with named entities (the strategy of injecting NE in all tests is replacement instead of supplementation, as suggested by Lau et al., 2013); c) **NE Independent**: independent named entity promoting model described in Section 3.1; and d) **NE Document Dependent**: document dependent named entity promoting model described in Section 3.2. We evaluate the tests using the topic quality measures presented below.

### 4.1 Dataset And Preprocessing

Our test corpora consists of news-related publicly-available datasets: 1) 20 Newsgroups[1]: widely studied by NLP research community dataset (Aletras and Stevenson, 2013; Truica et al., 2016; Wallach et al., 2009; Röder et al., 2015; Hu et al., 2013). Contains 18846 documents with messages discussing news, people, events and other entities. 2) Reuters-2013: a set of 14595 news articles from Reuters for year 2013, obtained from Financial News Dataset[2], first compiled and used in (Ding et al., 2014). The documents in Reuters-2013 are generally longer than in 20 Newsgroups. For NE recognition we used NeuroNER[3], a tool designed by Dernoncourt et al. (2016, 2017), trained on CONLL2003 dataset and recognizing four types of NE: person, location, organization and miscellaneous. The further preprocessing pipeline consists of classic steps used in topic modeling.

### 4.2 Topic Coherence

The term "topic coherence" covers a set of measures describing the quality of the topics regarding interpretability by a human. Most widely used measures are based on PMI (or NPMI, normalized) and log conditional probability, both of which rely on the co-occurrence of terms (Lau et al., 2013, 2014; O'Callaghan et al., 2015; Aletras and Stevenson, 2013; Newman et al., 2010; Mimno et al., 2011; Nikolenko, 2016; Nguyen et al., 2015; Syed and Spruit, 2017). Recently a study by Röder et al. (2015) put all known coherence measures into single framework, assessed their correlation with human ratings and discovered the best performing measure - previously unknown $C_v$, based on cosine similarity of word vectors over a sliding window. We inferred the definition from Röder et al. (2015):

$$C_v = \frac{1}{N} \sum_{t=1...N} \frac{1}{N_t} \sum_{i=1...N_t} s_{cos}(\vec{v}_{NPMI}(w_i), \vec{v}_{NPMI}(W_t)) \quad (3)$$

where $N$ is the number of topics, $W_t$ is the set of top $N_t$ terms in topic $t$, the vectors are defined as:

$$\vec{v}_{NPMI}(w_i) = \Big\{ NPMI(w_i, w_j) \Big\}_{j \in W_t} \quad (4)$$

$$\vec{v}_{NPMI}(W_t) = \Big\{ \sum_{w_i \in W_t} NPMI(w_i, w_j) \Big\}_{w_j \in W_t} \quad (5)$$

and the underlying measure is NPMI with probability $P_{sw}$ over a sliding window. $C_v$ with sliding window of 110 words (Röder et al., 2015) is the coherence measure we use in this paper.

Majority of studies also use a reference corpus like Wikipedia for calculating word frequencies and co-occurrences (Aletras and Stevenson, 2013; O'Callaghan et al., 2015; Lau et al., 2014; Röder et al., 2015; Yang et al., 2017). In our case the need for reference corpus is particularly significant, since we change natural frequencies of named entities in the corpus, therefore coherence will definitely decline if calculated on original data. For the tests we have preprocessed the dump of English Wikipedia from 2014/06/15 with the same pipeline as used for the test corpora.

### 4.3 Generality Measures

Coherence measures tend to favor topics with general highly frequent terms. As a result we end up with well understandable but quite generic topics. A good topic should also be specific enough to distinguish documents (O'Callaghan et al., 2015). Moreover, averaging the coherences of all topics may produce very good coherence for a model with many repeating words across topics. For covering these aspects of the topic quality we adopt two other measures.

**Exclusivity:** Represents the degree of overlap between topics, based on the appearance of terms in multiple descriptors (O'Callaghan et al., 2015). We define exclusivity as $\frac{|W_u|}{|W|}$, where $|W_u|$ is the number of unique terms and $|W|$ is the total number of terms in topic descriptors.

**Lift:** Generally used for reranking the terms in descriptors (Taddy, 2012; Sievert and Shirley, 2014), lift is employed here as a topic quality metric. It is defined as $\frac{\beta_{ti}}{b_i}$, where $\beta_{ti}$ is the weight of word $i$ in topic $t$ and $b_i$ is the probability of

[1] http://qwone.com/~jason/20Newsgroups/
[2] https://github.com/philipperemy/financial-news-dataset
[3] https://github.com/Franck-Dernoncourt/NeuroNER

| Topics | Test | 20 Newsgroups | | | Reuters-2013 | | |
|---|---|---|---|---|---|---|---|
| | | $C_v$ | Lift | Excl. | $C_v$ | Lift | Excl. |
| 20 | Baseline Unigram | 0,534 | 3,390 | 0,788 | 0,539 | 3,891 | 0,610 |
| | Baseline NE | 0,503 | 3,273 | 0,767 | 0,559 | 4,059 | 0,598 |
| | NE Independent (x1,3) | 0,494 | 3,394 | 0,755 | 0,551 | 4,209 | 0,563 |
| | NE Independent (x1,5) | 0,527 | 3,464 | 0,770 | 0,552 | 4,308 | 0,618 |
| | NE Independent (x2) | 0,525 | 3,756 | 0,797 | 0,548 | 4,449 | 0,640 |
| | NE Independent (x2,5) | 0,539 | 3,779 | 0,765 | 0,550 | 4,661 | 0,635 |
| | NE Independent (x5) | **0,543** | 5,071 | 0,898 | 0,517 | 5,701 | 0,708 |
| | NE Independent (x10) | 0,486 | **6,416** | **0,950** | 0,511 | **6,560** | **0,773** |
| | NE Doc. Dependent | **0,543** | 4,600 | 0,780 | **0,566** | 5,749 | 0,625 |
| 50 | Baseline Unigram | 0,492 | 2,882 | 0,511 | 0,514 | 3,977 | 0,427 |
| | Baseline NE | 0,467 | 2,704 | 0,469 | 0,534 | 4,064 | 0,402 |
| | NE Independent (x1,3) | 0,476 | 2,825 | 0,487 | **0,538** | 4,291 | 0,406 |
| | NE Independent (x1,5) | 0,479 | 2,987 | 0,497 | 0,527 | 4,370 | 0,423 |
| | NE Independent (x2) | 0,471 | 3,394 | 0,533 | 0,510 | 4,684 | 0,459 |
| | NE Independent (x2,5) | 0,467 | 3,652 | 0,561 | 0,499 | 4,958 | 0,483 |
| | NE Independent (x5) | 0,437 | 5,243 | 0,702 | 0,461 | 5,956 | 0,564 |
| | NE Independent (x10) | 0,385 | **6,693** | **0,787** | 0,447 | **6,943** | **0,641** |
| | NE Doc. Dependent | **0,512** | 4,951 | 0,624 | 0,532 | 5,452 | 0,421 |
| 100 | Baseline Unigram | **0,486** | 2,457 | 0,325 | 0,503 | 3,692 | 0,286 |
| | Baseline NE | 0,478 | 2,248 | 0,282 | 0,525 | 3,775 | 0,253 |
| | NE Independent (x1,3) | 0,473 | 2,391 | 0,300 | **0,527** | 4,041 | 0,286 |
| | NE Independent (x1,5) | 0,467 | 2,499 | 0,315 | 0,520 | 4,126 | 0,295 |
| | NE Independent (x2) | 0,463 | 2,737 | 0,332 | 0,508 | 4,505 | 0,329 |
| | NE Independent (x2,5) | 0,453 | 3,108 | 0,374 | 0,491 | 4,705 | 0,339 |
| | NE Independent (x5) | 0,416 | 5,079 | 0,537 | 0,455 | 5,840 | 0,432 |
| | NE Independent (x10) | 0,394 | **6,622** | **0,614** | 0,444 | **6,747** | **0,498** |
| | NE Doc. Dependent | 0,478 | 4,310 | 0,442 | 0,509 | 5,030 | 0,266 |

Table 1: Topic quality results on the corpora

word $i$ in the reference corpus. The overall model measure is the average of the log-lift of descriptor terms and shows the degree of presence of non-general words in topics.

### 4.4 Results

Table 1 depicts the results of running the experiments[4] with $N = \{20, 50, 100\}$ topics and top 10 words used for the measures. Firstly, we can observe one common outcome: NE Independent (x10) model exhibited the best exclusivity and lift values across all tests, which is logical since this model enforced the biggest number of pseudo-frequent words to be in topic descriptors. However, the same model also showed the lowest coherence in all experiments. This confirms the secondary status of lift and exclusivity: the full performance of the model is decided by the combination of all three measures. From the table we can see that for 20 Newsgroups, Baseline Unigram model resulted in better coherence than Baseline NE. Previously Lau et al. (2013) showed that coherence (NPMI-based) is supposed to improve with NE replacement model. However, the goal of this work goes beyond just including named entities into LDA. We want to demonstrate that our weighting model increases the number of NE in topic descriptors, which makes them more understandable and diverse. For these purposes we use different coherence measure (Röder et al., 2015), and include additional NE type - miscellaneous, which was omitted in (Lau et al., 2013) though it contains some potentially important named entities. Hence, at the moment we do not compare our results with Lau et al. (2013). For each dataset we chose the baseline for comparison depending on

[4]Tests were run with *gensim*: https://radimrehurek.com/gensim/

| Topic Baseline Unigram | $C_v$ | Topic NE Doc. Dependent | $C_v$ |
|---|---|---|---|
| game, good, year, team, player, play, think, get, time, like | 0,507 | game, ne_espn, ne_nhl, player, team, ne_steve, think, run, play, good | **0,565** |
| game, san, espn, chicago, lose, new, won, day, york, road | 0,488 | **ne_nhl**, ne_brown, ne_tor, ne_cal, ne_flyers, team, ne_det, ne_rangers, ne_lindros, ne_edmonton | **0,584** |
| year, ar, know, hockey, league, slave, new, file, list, slip | 0,291 | | |
| space, launch, earth, mission, orbit, satellite, moon, planet, solar, spacecraft | 0,816 | ne_earth, ne_saturn, ne_pluto, ne_jupiter, **ne_nasa**, ne_venus, ne_mars, ne_galileo, ne_uranus, ne_sun | **0,902** |
| gun, file, control, firearm, research, crime, new, information, law, use | 0,424 | **ne_nra**, ne_united states, ne_congress, ne_federal, ne_code, ne_gun control, ne_senate, ne_section, ne_constitution, ne_hci | **0,530** |

Table 2: Comparison of Baseline Unigram and NE Doc. Dependent topics for 20 Newsgroups

coherence: Baseline Unigram for 20 Newsgroups, and Baseline NE for Reuters-2013.

In the majority of cases NE Document Dependent ended up being the optimal model for both datasets: while it did not perform best in terms of lift or exclusivity, it achieved the best or good enough coherence values, better lift and better or the same exclusivity as baseline models. The exceptions are 20 Newsgropus with 20 topics, where NE Independent (x5) became the optimal model, and Reuters-2013 with 100 topics, where NE Independent (x1,3) performed the best for combination of all three measures. The only case where baseline model achieved superior coherence is 20 Newsgroups with $N = 100$, but we note that NE Document Dependent model came close in terms of coherence while having much better lift and exclusivity, therefore it can also be considered optimal. In general, NE Independent model showed improvement in coherence up to a certain value of $\alpha$ (different in each case), followed by a decline, reaching very low values for NE Independent (x10). On the other hand, NE Document Dependent model does not introduce new parameters into LDA and manages to achieve best performance in the majority of settings, thus being more stable and easy to use.

Table 2 demonstrates qualitative analysis on the individual topics from 20 Newsgroups, generated by Baseline Unigram, and their semantically closest counterparts from NE Document Dependent model. As evident from the table, baseline topics describe mostly abstract concepts of ”sport”, ”space” and ”gun control”. From NE Document Dependent topics we get more specific descriptors, resulting in better coherence (as well as lift/exclusivity). It is worth particularly noting the names of the organizations (in bold), crucial to the corresponding topics, that, despite being unigrams, only appear in NE Document Dependent model, because they are not met often enough in the test corpus.

## 5 Conclusion

Presented results indicate that, firstly, our proposed model is capable of improving topic quality by only modifying the TF scores in the input of LDA in favor of named entities. This makes it applicable to any LDA-based models relying on the same input. Secondly, we have shown that named entities are well suited to be used as domain-specific terms and produce high-quality topics in news-related texts. Next steps in our research include experimenting with different weights for different categories of named entities, as well as adding new coherence measures, such as word2vec-based one, used by O’Callaghan et al. (2015).

## Acknowledgments

The elaboration of this scientific paper was supported by the Ministry of Economy, Industry, Research, Innovation, IT, Employment and Education of the Region of Wallonia (Belgium), through the funding of the industrial research project Jericho (convention no. 7717).

## References


Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *IWCS*.

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

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, page ocw156.

Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. 2014. Using structured events to predict stock price movement: An empirical investigation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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.

Linmei Hu, Juanzi Li, Zhihui Li, Chao Shao, and Zhixing Li. 2013. Incorporating entities in news topic modeling. In *Communications in Computer and Information Science*, pages 139–150. Springer Berlin Heidelberg.

Jey Han Lau, Timothy Baldwin, and David Newman. 2013. On collocations and topic models. *ACM Transactions on Speech and Language Processing*, 10(3):1–14.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing*, pages 262–272. Association for Computational Linguistics.

David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. 2006. Statistical entity-topic models. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD06*. ACM Press.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 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*, pages 100–108. Association for Computational Linguistics.

Dat Quoc Nguyen, Kairit Sirts, and Mark Johnson. 2015. Improving topic coherence with latent feature word representations in map estimation for topic modeling. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 116–121.

Sergey I. Nikolenko. 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 - SIGIR16*. ACM Press.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM15*. ACM Press.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*. Association for Computational Linguistics.

Shaheen Syed and Marco Spruit. 2017. Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. In *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE.

Matt Taddy. 2012. On estimation and selection for topic models. In *Artificial Intelligence and Statistics*, pages 1184–1193.

Ciprian-Octavian Truica, Florin Radulescu, and Alexandru Boicea. 2016. Comparing different term weighting schemas for topic modeling. In *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. IEEE.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *Advances in neural information processing systems*, pages 1973–1981.

Andrew T Wilson and Peter A Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *human language technologies: The 2010 annual conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473. Association for Computational Linguistics.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. 2017. Adapting topic models using lexical associations with tree priors. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.