Improving Topic Quality by Promoting Named Entities in Topic Modeling

Katsiaryna Krasnashchok and Salim Jouili
EURA NOVA, Mont-Saint-Guibert, Belgium

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 work 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.

Proposed model

Based on modifying the input document-term matrix of standard LDA.

1. Independent Named Entity Promoting.

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

For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1*α	6	0*α
D_2	5	3	2*α	2	1*α
D_3	8	4	0*α	4	2*α

By varying the value of α, we can control the importance of named entities in the corpus.

2. Document Dependent Named Entity Promoting.

$$tf_{dw} = \begin{cases} tf_{dw} + \max_w tf_{dw} & \text{if } w \text{ is NE} \\ tf_{dw} & \text{otherwise} \end{cases} \qquad (2)$$

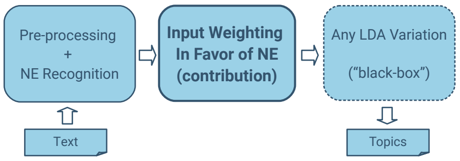
where $\max_w tf_{dw}$ is the most frequent term in the document. For example:

d\w	good	time	ne_nhl	play	ne_espn
D_1	4	2	1+**6**	**6**	0
D_2	**5**	3	2+**5**	2	1+**5**
D_3	**8**	4	0	4	2+**8**

Preferred method, since it does not introduce any new parameters into LDA.

Contribution

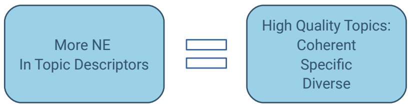
1. Introduced a new weighting model for LDA.


Pre-processing + NE Recognition ⇨ **Input Weighting In Favor of NE (contribution)** ⇨ Any LDA Variation ("black-box")

Text ⇧ Pre-processing

Any LDA Variation ⇩ Topics

2. Demonstrated the competence of named entities as domain-specific terms in news-related content.


More NE In Topic Descriptors = High Quality Topics: Coherent Specific Diverse

Future work: experimenting with different weights for different categories of NE; adding new coherence measures, such as word2vec-based one.

Evaluation and results

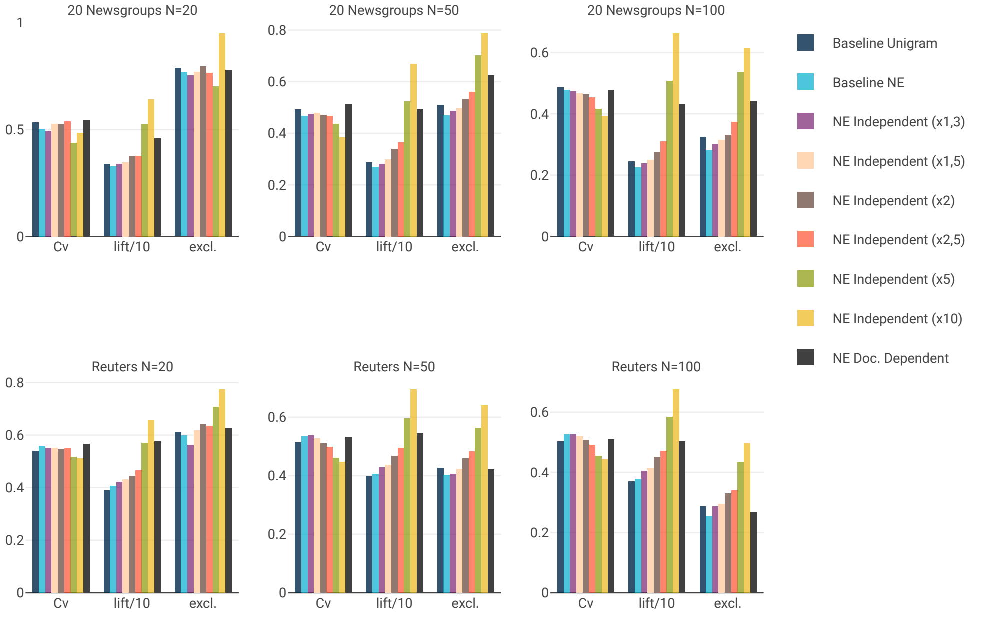


Figure 1. Topic quality results on the corpora.

NE Document Dependent is the optimal model for both datasets: it represents a trade-off between having better or the same coherence and exclusivity, and significantly higher lift, comparing to the baselines.

Produced topics

Topics Baseline Unigram	C_v	Topics 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 1. Comparison of Baseline Unigram and NE Doc. Dependent topics for 20 Newsgroups.

NE Document Dependent produces coherent, diverse and specific topics, containing more important words, such as the organization names, and less common words, such as "like", "use" and "file", resulting in better coherence.

References

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

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.

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.

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.

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*.