Can Word Probabilities from LDA be Simply Added up to Represent Documents?

Zhiqiang Cai
University of Memphis
Memphis, TN, USA
zcai@memphis.edu

Haiying Li
Rutgers University
New Brunswick, NJ, USA
haiying.li@gse.rugers.edu

Xiangen Hu
University of Memphis
Memphis, TN, USA
xhu@memphis.edu

Art Graesser
University of Memphis
Memphis, TN, USA
agraesser@memphis.edu

ABSTRACT
This paper provides an alternative way of document representation by treating topic probabilities as a vector representation for words and representing a document as a combination of the word vectors. A comparison on summary data shows that this representation is more effective in document classification.

Keywords
Topic modeling, LDA, document clustering, cluster similarity

1. INTRODUCTION
Topic modeling has been one of the most important methods in natural language analysis. It helps to discover underlying topics in a collection of documents. The found topics are used to form topic features for documents. The topic features are then used as input to perform task such as document clustering [11], automated summarization [1], automated essay grading [6], etc. LDA (Latent Dirichlet Allocation) [2, 3] is the most popular way for topic modeling. LDA topic model provides topic proportions as a vector representation of document. We investigated an alternative way of document representation by summing up word probabilities from LDA topic model. The new representation is compared with the topic proportion representation as input of a document clustering task on a summarization data set. The results showed that the simple "probability sum" document representation performs better.

2. LDA and Document Representations
Latent Dirichlet allocation (LDA), first introduced by Blei, Ng and Jordan in 2003 [3], is one of the most popular methods in topic modeling. LDA represents topics by word probabilities. Given a vocabulary with N words, $\{w_1, w_2, \cdots, w_N\}$, the LDA model probabilities $\boldsymbol{P}_k = (p_k(w_1), p_k(w_2), \cdots, p_k(w_N))$ form a representation of the k^{th} topic $(k = 1,2,\cdots,K)$. The words with highest probabilities in each topic usually give a good idea about what the topic is.

In LDA, a document d has an inferred topic proportion which is usually used as topic features to represent the document:

$$\boldsymbol{T}(d) \sim (t_1(d), t_2(d), \cdots, t_K(d)).$$

From the point of view of statistics, topic proportion is probably the only choice for LDA-based document representation. However, if we jump out of the box of statistics, we can simply view the word probabilities across the K topics as a K -dimensional vector representation for each word. Thus, a document can be represented by summing up the word probability vectors:

$$s_k(d) = \sum_{i=1}^{N} p_k(w_i) \log(1 + f(w_i, d)), (k = 1, 2, \cdots, K)$$

In the above formula, $s_k(d)$ is the "probability sum" of the document d on the k^{th} topic, $p_k(w_i)$ is the probability of the word w_i on the k^{th} topic, and $f(w_i, d)$ is the frequency of the word w_i in the document d. The logarithm of word frequency is known as Zipf scale [9].

3. Corpus for Document Clustering
201 participants wrote 1481 summaries for 8 passages, about 185 for each passage [10]. The lengths of the passages ranged from 195 to 399. The Flesch-Kincaid grade level was from 8.6 to 11.7. Some passages had similar topics: *Working and Running*, *Kobe and Jordan*, and *Effects of Exercising* on sports and exercising; and *Floods* and *Hurricane* on disasters.

The summaries were collected from an online experiment. The original goal was to evaluate the effect of an online AutoTutor [5, 9] lesson that teaches summarization. Each subject composed summaries for 2 texts before learning the lesson, 2 after learning, and 4 during learning with a counter-balanced design. The participant wrote each summary immediately after reading a passage. The system automatically controlled summary length (50-100 words) and *plagiarism*. The summary could not be submitted when it was out of range or when it had 10 consecutive words copied from the original passage.

Each summary was treated as a document for topic modeling. The vocabulary size was 4275 after removing stop words. 6 topic models were built for different numbers of topics (4, 8, 12, 16, 20 and 24), respectively. For each model, the topic proportions and the probability sums were computed for each summary. The LDA package used for topic modeling was infer.net from Microsoft [8].

Topic proportions and probability sums were then used as document features for clustering. We used K-Mean clustering method and fixed the number of clusters to 8 for all 6 topic models.

4. Results
We define the similarity of two clustering results by

$$Sim = \frac{\sum_{i=1}^{C} number\ of\ shared\ documents\ in\ cluster\ pair\ i}{total\ number\ of\ documents}.$$

The cluster pairs were best arranged using "Hungarian Algorithm" [7] so that the similarity is the highest under the paring. For each of the two document representations, we first compared the cluster similarity between models with the number of topics 4 and 8, 8 and 12, 12 and 16, 16 and 20, and 20 and 24. We aimed to check whether or not the clusters converge as the number of topics increases.

The results showed that when the number of topics increased, clustering based on probability sum quickly converged. The similarity between 12 topics and 16 topics was 0.96. For topic-proportion-based clustering, the similarity between 8 and 12 topics went close to probability sum. However, it dropped at 12 and 16, and then went up to 0.81 for 20 and 24.

While both representations converged to some clusters, the topic-proportion-based clustering converged to the unevenly distributed clusters. The largest two clusters contained 908 documents out of 1480. In contrast, probability-sum-based clustering converged to clusters of sizes almost the same as the original summary groups.

Table 1 shows the best matched clusters to the original passages for 24-topic model. Topic-proportion-based clusters matches the original passage groups with a similarity of 0.60, whereas probability-sum-based clustering did surprisingly better. The cluster similarity to the original summary grouping was 0.98.

Table 1 Best matched clusters to original passages

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
	Topic Proportion Based Clusters							
BM	160	0	0	0	0	20	1	2
Di	6	5	101	1	0	69	0	0
EE	0	1	186	0	1	1	0	0
Fl	11	7	21	1	1	139	5	1
Hu	1	0	1	1	173	3	5	0
JM	0	0	1	0	0	179	0	1
KJ	0	0	0	0	1	1	185	1
WR	1	0	164	0	1	20	0	1
	Probability Sum Based Clusters							
BM	180	0	0	1	0	1	1	0
Di	0	176	0	0	0	6	0	0
EE	0	1	182	0	0	5	1	0
Fl	0	0	0	179	1	6	0	0
Hu	0	0	0	0	180	4	0	0
JM	0	1	0	0	0	179	1	0
KJ	0	0	0	0	1	1	186	0
WR	0	0	2	0	0	4	0	181

Note: **BM**=Butterfly and Moth, **Di**=Diabetes, **EE**=Effects of Exercising, **Fl**=Floods, **Hu**=Hurricane, **JM**=Job Market, **KJ**=Kobe and Jordan and **WR**=Working and Running.

The cluster similarity changed when the number of topics increased in topic modeling. The topic-proportion-based clustering had its highest cluster similarity 0.77 to the original grouping when the number of topics is 12. It then dropped below 0.60. The probability-sum-based clustering had higher similarities for all models than topic proportion and consistently converged toward 1.

5. ACKNOWLEDGMENTS

This research was supported by the National Science Foundation (DRK-12-0918409, 1108845), the Institute of Education Sciences (R305H050169, R305B070349, R305A080589, R305A080594, R305G020018, R305C120001), Army Research Lab (W911INF-12-2-0030), and the Office of Naval Research (N00014-00-1-0600, N00014-12-C-0643). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, IES, or DoD.

6. REFERENCES

[1] Arora, R. and Ravindran, B. 2008. Latent Dirichelet allocation based multi-document summarization. In *Proceedings of the second workshop on Analytics for noisy unstructured text data* (Singapre, July 24 - 24, 2008). ACM, New York, NY, 91-97.

[2] Blei, D., Griffiths, T., Jordan, M., and Tenenbaum, J. 2004. Hierarchical topic models and the nested Chinese restaurant process. *Advances in Neural Information Processing Systems*, 16 (2004).

[3] Blei, D. M., Ng A. Y., and Jordan M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (March, 2003), 993-1022.

[4] Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*. (2009). 288-296.

[5] Graesser, A. C., D'Mello, S. K., Hu, X., Cai, Z., Olney, A., and Morgan, B. 2012. AutoTutor. In P. M. McCarthy, & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation and resolution*. Hershey, PA: IGI Global. 169-187.

[6] Kakkonen, T., Myller, N., and Sutinen, E. 2006. Applying latent Dirichlet allocation to automatic essay grading. In *Advances in Natural Language Processing*. Springer Berlin Heidelberg, 110-120.

[7] Kuhn, H. W. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*， 2 (1955), 83–97.

[8] Lau, J. H., Grieser, K., Newman, D., & Baldwin, T. 2011. Automatic labelling of topic models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* 1 (June, 2011). Association for Computational Linguistics. 1536-1545.

[9] Li, H. (2015). *The impact of pedagogical agents' conversational formality on learning and learner impressions* (Unpublished doctoral dissertation). University of Memphis, Memphis.

[10] van Heuven, W.J.B., Mandera, P., Keuleers, E., and Brysbaert, M. 2014. Subtlex-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, 67 (2014), 1176-1190.

[11] Xie, P. and Xing, E. 2013. Integrating document clustering and topic modeling. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence* (Bellevue, Washington, USA, July 11 - 15, 2013). UAI 2013. AUAI, Corvallis, Oregon, 694-703.