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

Zhiqiang Cai[1], Haiying Li[2], Xiangen Hu[1,3], and Arthur C. Graesser[1]

[1]University of Memphis, [2]Rutgers University, & [3]Central China Normal University

Methods for Document Representation

- Conventional Method: Topic Proportion
- Our Method: Word Probability Sum

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

For a document d: $p_k(w_i)$ is the probability of the word w_i in topic k and $f(w_i, d)$ is the frequency of the word w_i in the document d.

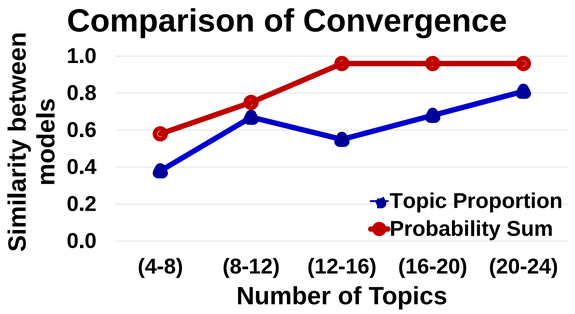
The method of word probability sum is simpler, but does it perform better than the method of topic proportion for any language processing task?

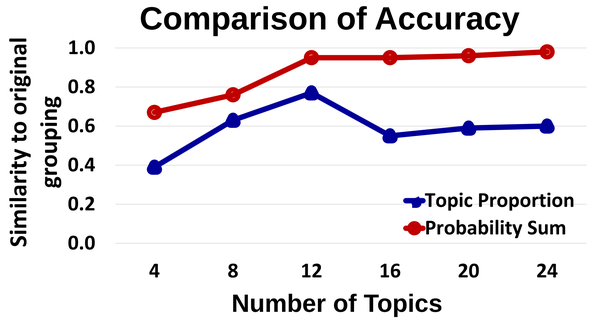
Using two Representations on a Classification Task

- Corpus
  - 8 expository texts
  - Over 180 students' summaries for each text
- Representing each summary by topic proportion and word probability sum, respectively
- Using these two representations as features to classify the summaries
- Comparing the accuracy of the summary classification task

Title	FKGL*	Words	Summary
Butterfly and Moth	8.6	255	183
Diabetes	11.7	241	182
Effects of Exercising	9.1	195	189
Floods	9.2	230	186
Hurricane	9.4	222	184
Job Market	10.9	240	181
Kobe and Jordan	9.2	299	188
Walking and Running	8.9	399	187
Total			1480

Convergence





- Methods: topic proportion scores and probability sum scores
- 6 Topic models
- Number of topics 4, 8, 12, 16, 20, and 24
- For each method, the clustering using 4 topics is compared to that using 8 topics, 8 to 12, and so on.
- Clusters converge when number of topics increases.

- Each clustering compared with original grouping
- Highest clustering similarity:
  - Topic proportion
    0.77 for topic 12 and then dropped
  - Word probability sum
    Close to 1 for topic ≥ 12

$$Clustering\ similarity = \frac{\sum number\ of\ shared\ documents\ in\ an\ aligned\ cluster\ pair}{Total\ documents}$$

Classification Accuracy

	1	2	3	4	5	6	7	8
	Topic Proportion Based Clusters							
Butterfly and Moth	**160**	0	0	0	0	20	1	2
Diabetes	6	5	**101**	1	0	69	0	0
Effects of Exercising	0	1	**186**	0	1	1	0	0
Floods	11	7	21	1	1	**139**	5	1
Hurricane	1	0	1	1	**173**	3	5	0
Job Market	0	0	1	0	0	**179**	0	1
Kobe and Jordan	0	0	0	0	1	1	**185**	1
Working and Running	1	0	**164**	0	1	20	0	1
	Word Probability Sum Based Clusters							
Butterfly and Moth	**180**	0	0	1	0	1	1	0
Diabetes	0	**176**	0	0	0	6	0	0
Effects of Exercising	0	1	**182**	0	0	5	1	0
Floods	0	0	0	**179**	1	6	0	0
Hurricane	0	0	0	0	**180**	4	0	0
Job Market	0	1	0	0	0	**179**	1	0
Kobe and Jordan	0	0	0	0	1	1	**186**	0
Working and Running	0	0	2	0	0	4	0	**181**

- Topic proportion based clustering (24 topics)
  - Unable to discriminate "*Diabetes*", "*Effect of Exercising*" and "*Walking and Running*"
  - Unable to discriminate "*Flood*" and "*Job Market*"
  - "*Hurricane*" and "*Kobe and Jordan*" were well clustered.

  The cluster similarity to the original summary grouping is 0.60.
- Probability sum based clustering (24 topics)
  - ***Almost perfectly clustered all summaries.***

  The cluster similarity to the original summary grouping is 0.98.

References

1. Blei, D. M., Ng A. Y., and Jordan M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.
2. 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.
3. Li, H. (2015). *The impact of pedagogical agents' conversational formality on learning and learner impressions* (Unpublished doctoral dissertation). University of Memphis, Memphis.
4. 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.

Acknowledgements

This research was supported by the National Science Foundation (DRK-12-0918409, 1108845), the Institute of Education Sciences (R305C120001), Army Research Lab (W911INF-12-2-0030), and the Office of Naval Research (N00014-00-1-0600, N00014-12-C-0643).

*** FKGL: Fesch-Kincaid Grade Level (Measuring reading ability)**