Position Aware Topic Modelling of Text

Keywords: natural language processing, topic modelling, hidden Markov models

Modern text processing algorithms typically model documents as a bag of words or with hidden Markov chains. Intuitively, the bag of words model loses valuable information by ignoring the position of words in a document. Specifically, words in the same sentence or paragraph are more closely related than are two random words appearing in the same document. Further, paragraphs and documents often have complex structures where the beginning sets a topic, a digression occurs in the middle, and the end relates the digression to the topic. This aspect is not captured by hidden Markov models, which assume locality of related text. Finally, position indicates importance i.e. the opening and concluding paragraphs of an essay are usually the most important understanding the document as a whole. I hypothesize that machine learning can model how and where position indicates relatedness and importance of text, ultimately leading to improved topic models.

Topic modelling is a popular and useful technique for automatically identifying the subjects of large numbers of documents. For example, it has been successfully applied to group news articles [2] and to search for academic papers [6]. As the quantity of documents available increases, especially on the Internet, it is crucial for machines to assist users in categorizing and finding information. Doing so vastly improves the ability of people to discover and process relevant information. As search engines have shown, public adoption of such algorithms depends crucially on the quality of results, which this project aims to improve.

In the literature, topic models such as probabilistic latent semantic analysis (pLSA) [6] and latent Dirichlet allocation (LDA) [2] are usually introduced with a bag of words model that simply counts word occurrences in documents. However, other work [5] illustrates the importance of position to the related problem of segmenting documents based on topic. The very notion of segmenting documents by topic suggests that topic models could be improved by accounting for the positions of words. Indeed, both pLSA [3] and LDA [8] were later extended with a Markov model for topic transitions. Learning cue words for topic transitions has also been tried [1] successfully. I believe that natural language topic structures are more complex than captured by these models and propose a plan to make a better one.

Making a new model proceeds in several steps: looking at the data available, developing an intuition about patterns and useful signals in the data, formalizing that intuition in a model, and controlled testing on another data set. I propose to go about making a generative model for text by analyzing how people actually write documents. Revision histories have been used to measure software quality [4] and trust in Wikipedia [7] but I am not aware of any use in document understanding. The idea is that looking at revisions will reveal patterns in the structure and organization of documents. Possible sample data could come from web site revisions, academic paper review cycles, and Wikipedia. Gaining an intuitive understanding of revision scenarios and processes from these data will help me select or create a machine learning algorithm to learn rules from a larger corpus than I can examine. I will then combine this model of text structure with topic modelling to create a new generative model for text. Results will be compared with existing models by using a standard corpus, such as categorized Associated Press articles. Performance is measured by agreement with the subject labels and perplexity, a standard measure of agreement between the model and observed text. The road does not end here–I expect to repeat this process as I learn more

about the problem.

This proposal is well suited to study at Princeton and to my past research. Princeton professor David Blei is the top researcher behind LDA [2] and other topic models. His group has the knowledge, connections, and training data to support advances in topic modelling and segmentation. I first came across his work and used LDA while performing research into source code reorganization. Although I used a bag of words at the time, source code has more structure than text and revisions are usually kept in version control. This is what gave me the original idea to study structure and revision of text in more detail. With a strong group and background in the area, I expect to significantly improve topic modelling and make it more useful to the public.

References

[1] Beeferman, D., Berger, A., and Lafferty, J. Statistical models for text segmentation. *Machine Learning* (1999).

[2] Blei, D., Ng, A., and Jordan, M. Latent dirichlet allocation. *Journal of Machine Learning Research 3* (2003), 993–1022.

[3] Brants, T., Chen, F., and Tsochantaridis, I. Topic-based document segmentation with probabilistic latent semantic analysis. In *CIKM '02: Proceedings of the eleventh international conference on Information and knowledge management* (2002), pp. 211–218.

[4] Canfora, G., and Cerulo, L. Impact analysis by mining software and change request repositories. In *METRICS '05: Proceedings of the 11th IEEE International Software Metrics Symposium (METRICS'05)* (2005), p. 29.

[5] Hearst, M. A. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist. 23*, 1 (1997), 33–64.

[6] Hofmann, T. Probabilistic latent semantic analysis. In *UAI '99: Uncertainty in Artificial Intelligence* (1999).

[7] McGuinness, D. L., Zeng, H., da Silva, P. P., Ding, L., Narayanan, D., and Bhaowal, M. Investigations into trust for collaborative information repositories: A wikipedia case study. In *Proceedings of the Workshop on Models of Trust for the Web* (2006).

[8] Purver, M., Griffiths, T. L., Körding, K. P., and Tenenbaum, J. B. Unsupervised topic modelling for multi-party spoken discourse. In *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL* (2006), pp. 17–24.