When Topic Models Disagree: Keyphrase Extraction with Multiple Topic Models

Lucas Sterckx, Thomas Demeester, Johannes Deleu, Chris Develder
Ghent University - iMinds
Gaston Crommenlaan 8
Ghent, Belgium
firstname.lastname@intec.ugent.be

ABSTRACT

We explore how the unsupervised extraction of topic-related keywords benefits from combining multiple topic models. We show that averaging multiple topic models, inferred from different corpora, leads to more accurate keyphrases than when using a single topic model and other state-of-the-art techniques. The experiments confirm the intuitive idea that a prerequisite for the significant benefit of combining multiple models is that the models should be sufficiently different, i.e., they should provide distinct contexts in terms of topical word importance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Topical Keyphrase Extraction, Topic Models

1. INTRODUCTION

Keyphrases are defined as a set of terms or noun phrases which concisely summarize the content of a document. Automatic Keyphrase Extraction (AKE) has been beneficial for various applications such as document categorization and contextual advertising on Web pages. A distinction can be made between supervised and unsupervised methods. State-of-the-art unsupervised methods apply a graph-based approach. These methods build a graph from the input documents, each node corresponding to a candidate word and edges connecting two co-occurring candidates. Nodes or vertices are ranked according to their importance using a graph-based ranking method like PageRank. Top-ranked vertices are then combined to generate keyphrases. The inclusion of topical information has been shown to be beneficial for extracting keyphrases from documents. Liu et al. propose Topical PageRank (TPR) [4], a variation of PageRank that incorporates topical information by increasing the importance of highly relevant topical words based on Latent Dirichlet Allocation (LDA) [2]. Each word in the graph gets an additional weight (denoted as $W(w_i)$ for word w_i) in the random-walk algorithm proportional to the cosine distance of the topic-document distribution and word-topic distribution from the LDA topic model. Experimental results showed that TPR outperforms other unsupervised AKE-methods. We assess that topical importance strongly depends on the collection of training documents for LDA and their corresponding context. Specific words can be essential in one context yet only secondary in another. First we show that topical word importance varies with the corpus the topic model is trained on. Then we show that a simple combination of multiple different topic models and word scores leads to more accurate AKE results, a prerequisite being the diversity of the training corpora.

Copyright is held by the author/owner(s).
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742731.

2. DISAGREEMENT BY TOPIC MODELS

We demonstrate how we can improve the accuracy of a single-model TPR by combining information from multiple topic models. We use four different corpora to study the influence of the topic models on AKE: **Wikipedia** (a corpus similar to the one used in the original TPR contribution [4]), **Reuters Corpus Volume I (RCV1)** [3] (800,000 manually categorized newswire stories), **Wikinews** [1] (A free-content news source wiki, maintained through collaborative journalism, from February 2013) and **New-York Times** [1] (a collection of 300,000 NYT news articles). It is known that ensemble methods like model averaging obtain better accuracy than can be obtained from any of the constituent learning algorithms. We assess if and when this is the case for learning algorithms based on topic models for AKE. We first investigate how the topical importance scores from the word-document similarities, which are used in TPR, vary with the corpus the models are trained on. We then use this disagreement to make a combined weight applying several methods for averaging. Large test corpora for AKE, containing a broad set of topics, are hard to find and create. The creation of such a set is in progress, but we wish to report promising results on an existing, smaller set of news articles built by Wan and Xiao [5], that contains 308 news articles from the 2001 Document Understanding Conference (DUC) summarization-track, with 2,488 manually assigned keyphrases. The following experiment is conducted: next to training topic models on the original corpora, we reassign documents from each of the mentioned topic model corpora to one of four new collections randomly, and train a 1,000-topic LDA-model on all collections. As in [4], all of the models' vocabularies are reduced to 20,000 words. This results in four different topical word scores indicated as $W^c(w_i)$ with c denoting the index of the model being used. In Figure 1a, standard deviations of the four weights are shown for the shuffled and for the original corpora for each word in the 308 documents of the test-corpus. We observe that there is a much higher variance in the importance of

[1] http://en.wikinews.org/

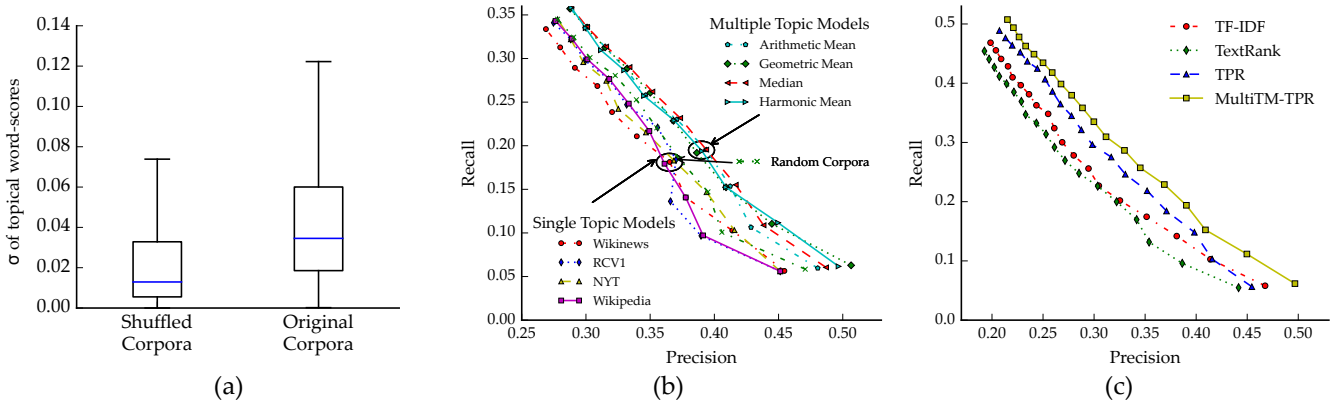

Figure 1: (a) Box plot displaying the average standard deviation for all topical word scores $\{W^c(w_i)\}_{c=1\ldots4}$ for different topic models c, based on the original four collections ('Original Corpora'), versus four topic models based on a random equal share of all data together ('Shuffled Corpora') (b) Precision-recall curve for combinations versus single-model TPR for 1 to 10 extracted keyphrases. (c) MultiTM-TPR versus baselines for 20 extracted keyphrases

the words between models when trained on the specific contexts of documents from the original collections. This means that different topic models trained on corpora with distinct contexts, used in TPR, will produce very different word scores and thus keyphrases, whereas topic models trained on more uniform contexts lead to similar keyphrase rankings.

3. AVERAGING TOPICAL IMPORTANCE

In the previous section the disagreement between models showed the dependence of topical word importance on the corpus the topic model was trained on. We now attempt to leverage this disagreement, composing word scores which reflect a more realistic importance of the words. For this purpose we apply several metrics which combine all weights into a single weight to be used in the PageRank for TPR. For this experiment, all models are trained on the full vocabulary of their respective corpora. We apply four ways of averaging the four weights: the arithmetic mean, the geometric mean, the harmonic mean and the median. We create a mapping between the keyphrases in the gold standard and those in the system output using an exact match. We reduce keyphrases to their stems using the Porter-stemmer and use three standard evaluation metrics for AKE: precision, recall, and F1-measure. Other parameters (for the stemmer, tokenizer and PageRank) are identical to those in [4]. The resulting averaged precision-recall curves for increasing numbers of assigned keyphrases (ranging from 1 to 10 keyphrases) are shown in Figure 1b. The results of all single topic models are approximately equal. When averaging scores generated from topic models from these original corpora, a change in accuracy is noticed. For each combination between the four different topic models some accuracy was obtained. All ways of averaging reach a similar increase in performance with respect to the single models. When looking at the top keywords, a slightly higher precision is observed for those averaging methods that penalize values with more spread, like the harmonic and geometric mean. This increase in accuracy is not observed when randomizing the contexts of the different topic models as demonstrated in Section 2, when there is less variance in the scores topical importance. A topic model was also trained on a single large corpus, consisting of all the single corpora, but this resulted in a similar performance obtained using one of the single topic models trained on a separate smaller corpus. We finally compare our new multi-topic-model method (denoted as 'MultiTM-TPR') to existing baseline methods in Figure 1c and the best single-model TPR. Our MultiTM-TPR outperforms baselines and the original TPR. Also for the highest scored keyphrases, where a single topic model TPR is inferior to the TF-IDF baseline. All improvements of MultiTM-TPR over other methods are verified, using bootstrap resampling, resulting in significance levels of $p < 0.05$.

4. CONCLUSION

In this paper we showed ongoing work demonstrating the benefit of combining multiple topic models for Automatic Keyphrase Extraction. We studied the influence of the corpus the topic model is trained on, and showed disagreement between models which are trained on different corpora. Averaging weights from several topic models leads to an increase in precision of extracted phrases. When training models, an important aspect is the difference in contexts between the corpora, which leads to different topic models and thus disagreement about word importance. We leverage this disagreement by computing a combined topical word importance value which, when used as weight in a Topical PageRank, improves accuracy of extracted keyphrases. Moreover, we show that this benefit of using multiple topic models is attained when the models differ substantially. For future work, we intend to research whether more sophisticated methods for combining or selection of specific models can be applied.

5. REFERENCES

[1] K. Bache and M. Lichman. UCI machine learning repository, 2013.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *JMLR*, 3(4-5):993–1022, 2003.

[3] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397, Dec. 2004.

[4] Z. Liu, W. Huang, Y. Zheng, and M. Sun. Automatic keyphrase extraction via topic decomposition. In *Proceedings of the 2010 Conference on EMNLP*, pages 366–376, 2010.

[5] X. Wan and J. Xiao. CollabRank: towards a collaborative approach to single-document keyphrase extraction. *Coling*, pages 969–976, Aug. 2008.