ABSTRACT
We explore how the unsupervised extraction of topic-related keyphrases 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.

2. DISAGREEMENT BY TOPIC MODELS
We demonstrate how we can improve the accuracy of a single-topic 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 used 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

---

1http://en.wikinews.org/
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 harmonic mean, the geometric 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 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.

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