Weak Supervision for Semi-Supervised Topic Modeling via Word Embeddings

Gerald Conheady, Derek Greene

School of Computer Science, University College Dublin, Ireland
gerry.conheady@ucdconnect.ie ✉, derek.greene@ucd.ie

Abstract. Semi-supervised algorithms have been shown to improve the results of topic modeling when applied to unstructured text corpora. However, sufficient supervision is not always available. This paper proposes a new process, Weak+, suitable for use in semi-supervised topic modeling via matrix factorization, when limited supervision is available. This process uses word embeddings to provide additional weakly-labeled data, which can result in improved topic modeling performance.

1 Introduction

Unsupervised algorithms, such as Non-Negative Matrix Factorization (NMF) [4], have been used to uncover the underlying topical structure in unlabeled text corpora [1]. Semi-supervised NMF (SS-NMF) algorithms use background information, in the form of word and document constraints, to produce more accurate topic models [5]. In real-world applications it is reasonably easy to obtain a limited sample of labeled data from domain experts. However, when dealing with large corpora, this may not be enough to obtain improved results. We aim to address this issue by using weak supervision automatically generated from a small amount of input provided by an expert.

Recently, word embeddings have been used in a range of domains, where words are represented by vectors in a multi-dimensional space [6]. Words with related meanings will tend to be close together in the vector space. Based on this idea, in Section 3 we introduce a new method of **Weak+** supervision for topic modeling, which uses word embeddings to generate additional "weakly-labeled" data. This supervision takes the form of a list of candidate words that are semantically related to a small number of "strong" words supplied by an expert to describe a topic. Our initial experiments in Section 4 show that, when this weak supervision is fed to SS-NMF, the results of topic modeling are improved.

2 Related Work

Topic modeling allows for the discovery of themes in an unsupervised manner. While probabilistic approaches have often been used for topic modeling, approaches based on NMF [4] have also been successful [1]. The counts of all the terms in each document serves as an input to NMF in the form of a non-negative

document-term matrix $\mathbf{A}$, corresponding to m documents by n words. NMF seeks to find k topics and starts by randomly initialising an m by k document-topic matrix $\mathbf{W}$ and k by n topic-word matrix $\mathbf{H}$. The algorithm seeks a solution for $\mathbf{W}$ and $\mathbf{H}$ such that $\mathbf{A} \approx \mathbf{WH}$.

SS-NMF typically involves grouping data where limited supervision is available in the form of constraints imposed on pairs of items, provided by a human expert or "oracle" [2]. Methods have been proposed for incorporating constraints into matrix factorization [5]. This paper follows the Utopian approach for weakly-supervised topic modeling [3] which minimises the objective function:

$$||A - WH||_F^2 + ||(W - W_r)M_W||_F^2 + ||(H - H_r D_H)M_H||_F^2$$

Human selected constraints are passed to NMF in the form of reference matrices, $\mathbf{W_r}$ for documents and $\mathbf{H_r}$ for words, whose values are used to initialise the selected documents and words in the $\mathbf{W}$ and $\mathbf{H}$ matrices. The remaining documents and words are initialised randomly. Masking matrices $\mathbf{M_W}$ and $\mathbf{M_H}$ are also used to adjust the effect of the $\mathbf{W_r}$ and $\mathbf{H_r}$ matrices. The diagonal matrix $\mathbf{D_H}$ is used for automatic scaling. The success of such an approach will depend on the availability of useful constraints to populate these matrices [2].

NMF does not directly take into account semantic associations. Related meanings of words, such as between 'computer' and 'data', do not explicitly influence the factorization process. Many applications of word embeddings are based on the original *word2vec* model [6]. Their approach creates distributed representations of words, in the form of dense lower-dimensional vectors, as trained on a large corpus of text using a neural network with one hidden layer. The input and output layers have one entry for each word in the vocabulary n. The hidden layer is considered the dimension layer and has d entries. This allows the output from the hidden layer to be represented by a $n \times d$ matrix. This representation can be used to measure the associations between the words in the corpus vocabulary.

3 Methods

Semi-supervised learning relies on input from domain experts. By definition this input is limited due to the availability of expert time. In this paper, we consider an extreme case, where the expert, referred to as the "oracle", will provide a list of five relevant words and five relevant documents relating to a single topic of interest. An example might be the provision by an auditor of five emails and five words from a health organization's email relating to data privacy breaches.

The Weak+ process uses the Gensim implementation of *word2vec* [7] to produce an embedding representing the corpus being analyzed. It uses this model to extend the list of words provided by the oracle as follows:

1. Construct a skip-gram *word2vec* model for the full corpus.
2. Request the initial list of "strong" supervised words and documents from the oracle for one or more topics.
3. For each of the topic(s) to be supervised:

– For each supervised word, identify the list of top most similar words in the embedding space based on cosine similarity.
– Alternate between the lists, adding words to the extended list until a required number of words have been added (This is done to ensure a good balance of similar words relative to the original list).

We then apply the SS-NMF algorithm to the document-term matrix representation $\mathbf{A}$ of the corpus as described in [3], where the reference matrix $\mathbf{H_r}$ is populated from the extended list of supervised words and the reference matrix $\mathbf{W_r}$ is populated from the list of documents originally provided by the oracle.

4 Experimental Evaluation

4.1 Experimental Setup

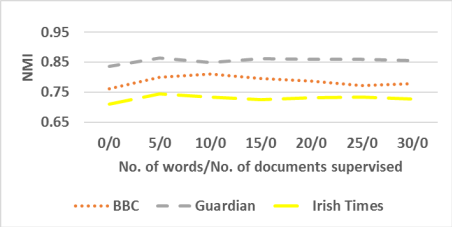
The aim of our experiments is to investigate whether SS-NMF topic modeling can be improved using the additional words generated by Weak+. The *bbc*, *guardian-2013* and *irishtimes-2013* news corpora are used for the evaluation [2], where the documents in these corpora have been assigned a single human-annotated ground truth topic. The top 100 most relevant words per topic are identified based on these annotations. These documents and words are considered as emanating from the oracle for our experiments. We construct *word2vec* embeddings for each corpus, using a skip-gram model with 100 dimensions and a document frequency threshold of 5.

For each corpus, we identify the least coherent ground truth topic (*i.e.* the topic with the lowest mean within-topic to between-topic cosine similarity ratio). These are 'business' (*bbc*), 'music' (*guardian-2013*), and 'politics' (*irishtimes-2013*). We then examine the extent to which we can improve the identification of these difficult topics. Specifically, we produce supervision in the form of a list of five documents and words for these topics based on the oracle, and then apply the Weak+ supervision process to extend the word list, varying the number of words to be supervised from 0 to 30 in steps of 5. Document supervision is restricted to 0 and 5 documents, as we wish to focus on the effect of word supervision. For each level of supervision, we apply SS-NMF for 50 runs, where the entries in the reference matrices are set to 1 for the supervised words and documents, and 0 otherwise. For the purpose of our experiments, we fix the number of topics k to be the number of ground truth topics in each corpus.

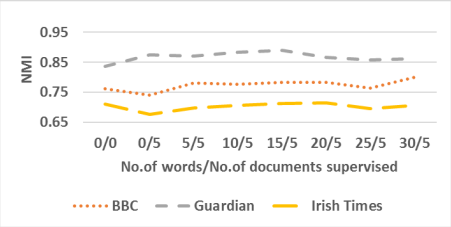
4.2 Results

Firstly, we use Normal Mutual Information (NMI) to measure model accuracy across all topics. For each run of SS-NMF, we compare the disjoint partition produced from the topic-document weights with the ground truth document assignments. The results in Figures 1a and 1b show a small improvement in NMI scores with word supervision alone, and a slightly greater improvement when using both word and document supervision.

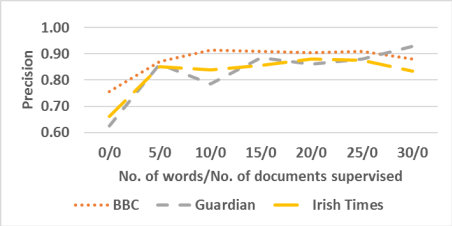
Since our main interest lies in examining the impact on the "difficult" topics in the three corpora, we next consider precision and recall relative to these supervised topics only. Precision measures the proportion of relevant documents found for a supervised topic – we consider the top 200 documents. Initially 151 of the top 200 documents retrieved by SS-NMF were found for the *bbc* 'business' topic giving a precision of 0.76. Precision shows a bigger improvement than NMI increasing from 0.76 to 0.87, 0.63 to 0.86 and 0.66 to 0.85 respectively for the *bbc*, *guardian-2013* and *irishtimes-2013* topics, with supervision based on just the five words provided by the oracle, Figure 1c. The use of the Weak+ supervision words increase these scores to as high as 0.92, 0.93, 0.88. Further improvement is seen when supervision of five documents given by the oracle takes place, resulting in scores as high as 0.99, 1.00 and 0.89, Figure 1d.



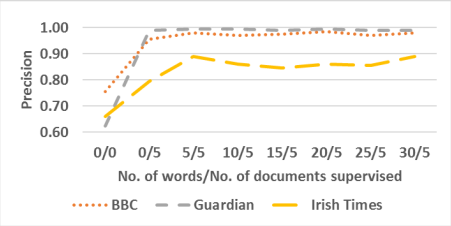
(a) NMI - word supervision only



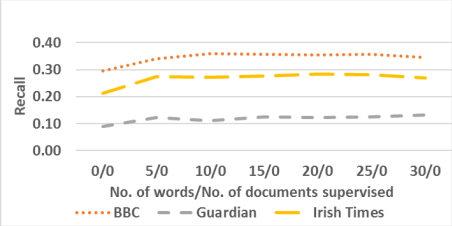
(b) NMI - word/document supervision



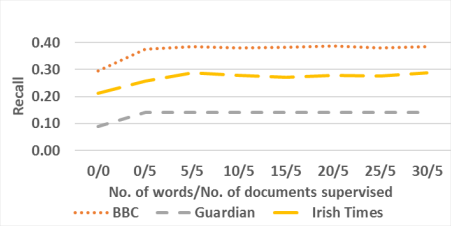
(c) Precision - word supervision only



(d) Precision - word/document supervision



(e) Recall - word supervision only



(f) Recall - word/document supervision

Fig. 1: Plot of NMI, precision, and recall scores for SS-NMF applied on three corpora. The five words and documents provided by the oracle for one topic per corpus are supplemented by words recommended by the Weak+ process.

Recall measures the ratio of the number of documents found and the number of documents in the dataset for a given supervised topic. Initially 151 of 510 possible *bbc* 'business' documents were in the top 200 documents retrieved by SS-NMF for the topic, giving a recall of 0.30. Recall improves from 0.30 to 0.34, 0.09 to 0.12, and 0.21 to 0.27 respectively for the *bbc*, *guardian-2013* and *irishtimes-2013* with supervision based on just the five given words, Figure 1e. The use of the Weak+ supervision words increase these scores slightly to 0.36, 0.13, 0.28. Further improvement is seen when document supervision is added, resulting in scores as high as 0.39, 0.14 and 0.29, Figure 1f.

5 Conclusions and Future Work

In this paper, we have shown that topic modeling can be improved through the use of word and document supervision. Precision and recall for "difficult topics" can be further improved using our new Weak+ approach based on word embeddings. This is a cheap mechanism to increase the number of labelled words with no extra effort required from the human oracle. This suggests that, if an oracle can provide a few good examples of words and documents relating to a topic, a large number of relevant documents can be readily identified. The next step in this work will be to apply these methods in the context of enterprise email corpora. Rather than simply using topic modeling to find the dominant topics in these corpora, we will focus on the identification of niche topics of interest, such as data privacy breaches, which may be difficult to identify using unsupervised topic modeling approaches. We will also extend the Weak+ process to provide document supervision and to operate in an iterative manner, where the user will be able to select from a list of suggested words at each iteration.

Acknowledgement. This research was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

1. Arora, S., Ge, R., Moitra, A.: Learning topic models–going beyond svd. In: 53rd Annual Symposium on Foundations of Computer Science (FOCS). pp. 1–10 (2012)
2. Greene, D.: Constraint Selection by Committee : An Ensemble Approach to Identifying Informative Constraints for Semi-supervised Clustering pp. 140–151 (2007)
3. Kuang, D., Choo, J., Park, H.: Nonnegative Matrix Factorization for Interactive Topic Modeling and Document Clustering. Partitional Clustering Algorithms pp. 1–28 (2015)
4. Lee, D.D., Seung, H.S.: Learning the parts of objects by non-negative matrix factorization. Nature 401(6755), 788–91 (1999)
5. Li, T., Ding, C., Jordan, M.I.: Solving Consensus and Semi-supervised Clustering Problems Using Nonnegative Matrix Factorization 1(2) (2007)
6. Mikolov, T., Corrado, G., Chen, K., Dean, J.: Efficient Estimation of Word Representations in Vector Space. Proc. ICLR 2013 pp. 1–12 (2013)
7. Radim Rehurek: gensim 1.0.0rc1 : Python Package Index, `https://pypi.python.org/pypi/gensim`