UCL DEPARTMENT OF
COMPUTER SCIENCE

Research Note
RN/12/16

Automatic Correction of Topic Coherence

26 November 2012

William Martin

John Shawe-Taylor

Abstract

A set of texts is often a poor representation of the language it is written in, and resultantly topics can seem nonsensical to domain experts. This can be for several reasons: misspellings or ‘accidental words’ can be given statistical significance in the case that too many topics are learned; words can appear related or unrelated in the text, even though the opposite is true in the language; too few topics or too many topics are used.
In this position paper we present a novel approach by applying biases derived from external sources during the training process, in order to improve the coherence of topics. This has the effect of improving topic coherence [Newman et al., 2009, 2010], ironing out many of the issues that a sub-optimal number of topics can cause, and imbuing resultant models with real-world word-relationships.

Automatic Correction of Topic Coherence

Position Paper

William Martin
University College London
w.martin@ucl.ac.uk

John Shawe-Taylor
University College London
jst@cs.ucl.ac.uk

ABSTRACT

A set of texts is often a poor representation of the language it is written in, and resultantly topics can seem nonsensical to domain experts. This can be for several reasons: misspellings or 'accidental words' can be given statistical significance in the case that too many topics are learned; words can appear related or unrelated in the text, even though the opposite is true in the language; too few topics or too many topics are used.

In this position paper we present a novel approach by applying biases derived from external sources during the training process, in order to improve the coherence of topics. This has the effect of improving topic coherence [Newman et al., 2009, 2010], ironing out many of the issues that a suboptimal number of topics can cause, and imbuing resultant models with real-world word-relationships.

1. INTRODUCTION

Topic modelling is a method used to classify a corpus into a set of topics, each which are unobserved entities represented by a set of probabilities for each of the terms in the corpus alphabet. Latent Dirichlet allocation (LDA) [Blei et al., 2003] is a fast and well known topic model which, for these reasons, can often be the starting point for projects involving unsupervised text classification.

Due to factors such as irregular words, and the sparse-topic properties of LDA, topics can be found to contain top words which seem out of place. Such topics are referred to as incoherent. Newman et al. [2009, 2010] revealed that automated evaluation of topic coherence is an effective measure of topic performance, as the results correlate well with human expert rankings.

Mimno et al. [2011] presented both a fast, well performing coherence metric that uses document frequency scores, and a revision of the popular LDA Gibbs sampling algorithm [Griffiths, 2002] that incorporates a bias into the training process. The aim of this bias is to enhance topic coherence by discouraging word groupings that are not observed in the corpus, which helps to prevent the effect of terms being forced into topics simply because they do not fit elsewhere. This approach maintains the simplicity of working wholly with the source corpus, and provides a modest improvement in topic coherence over LDA.

In this paper we introduce a novel approach with the externally weighted topic model (EWTM), that builds upon the work done by Newman et al. [2010] and Mimno et al. [2011], and effectively includes biases from external sources using the generalised Pòlya Urn model. This has the effect of imbuing the model with 'real world information', and serves to provide even better results than the model created by Mimno et al. [2011].

In section 2 we give an overview of the field of topic modelling, and in section 3 we present a summary of the recent work on topic coherence. The generalised Pòlya Urn model is introduced in section 4, and the modified inference step we use is detailed in section 5. We present the our revised model in section 6, and include a discussion of its aims and implications for researchers in section 7.

2. TOPIC MODELLING

Topic modelling is an unsupervised machine learning technique which generates a set of topics. Each topic is defined as a set of word probabilities, where a probability value is present for each word in the source alphabet. Topics are typically represented by the top n (often 10) words which have the highest probability values. Topic model generative algorithms encourage sparsity, and discourage topic similarity (topics having a similar set of word probabilities), therefore a top set of words is usually adequately descriptive of a topic.

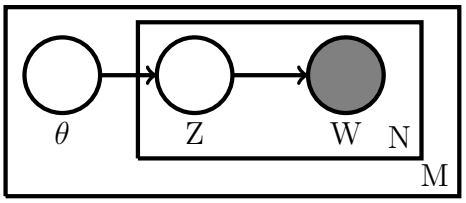
Also generated are a set of topic-document probabilities: each document has a probability for each generated topic, that defines its topic mixture. We discuss what are known as mixture models, in which documents are a mixture of several topics. Generative models run an expectation maximization algorithm to ensure that the topics and topic-document probabilities that are inferred will maximize the likelihood of the source corpus being generated, if we were to randomly generate documents from the model.

StuConOS 2012 University College London, United Kingdom

Table 1: Term definitions for figs. 1 and 2.

M	The number of documents.
θ	The distribution of topics in document d.
N	The number of words in document d.
Z	Topic identity vector for all words in the corpus.
W	Word identity vector for all words in the corpus.
α	Dirichlet prior on the per-document topic distribution.
β	Dirichlet prior on the per-topic word distribution.

Figure 1: Graphical representation of pLSI in plate notation.



2.1 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (pLSI) was created by Hofmann [1999], and was essentially the first probabilistic topic model. It builds upon Latent Semantic Analysis by Deerwester et al. [1990] by incorporating the probabilistic model shown in fig. 1. Terms are defined in table 1.

pLSI is a bag of words model; that is it ignores the ordering of individual words in a document, making them exchangeable.

2.2 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) by Blei et al. [2003] builds on pLSI by incorporating the Dirichlet priors α and β, as fig. 2 shows. Terms are defined in table 1. Dirichlet priors are selected before inference, and are typically small (each value < 1.0) and uniform vectors.

Figure 2: Graphical representation of LDA in plate notation.

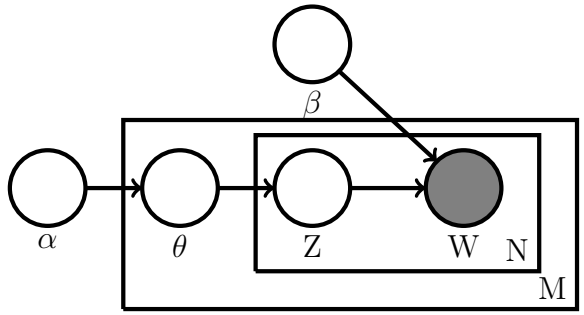


Table 2: Unit definitions for algorithms 1 to 3.

d	A document.
D	The set of all documents in the dataset.
$w^{(d)}$	The set of all terms in the document.
z_i	The topic assignment for term w_i.
$N_{z_i\|d_i}$	The count of terms from document d_i that are assigned to topic z_i.
$N_{w_i\|z_i}$	The number of times the unique term w_i is assigned to topic z_i.
α_z	The alpha Dirichlet parameter for topic z.
β	The beta Dirichlet parameter.
V	The set of similar terms.

2.3 Inference

Blei et al. [2003] defined the variational inference algorithm for training topics. This algorithm works well, but the Gibbs Sampling Monte Carlo algorithm [Griffiths, 2002] is generally favoured for its speed and simplicity.

Algorithm 1 Gibbs Sampling Monte Carlo algorithm for Latent Dirichlet allocation.

Initialise: Randomly select a topic for each observed term.
for all iterations **do**
  for all $d \in D$ **do**
    for all $w_n \in w^{(d)}$ **do**
      Sample: Select the topic z_i for term w_i that w_i is most likely to be generated by.
    end for
  end for
end for

The **Sample** step selects the topic z_i that term w_i is most likely to be generated by, as determined by a probabilistic equation. To perform this equation, we must ignore the term's current involvement with a topic by removing it from the term-topic assignment counts. We then select the topic z_i, and add it back into the term-topic assignment counts under the new topic.

Algorithm 2 Algorithm for a single LDA Gibbs sample.

for all $d \in D$ **do**
  for all $w_n \in w^{(d)}$ **do**
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} - 1$
    $N_{w_i|z_i} \leftarrow N_{w_i|z_i} - 1$
    $z_i \propto (N_{z|d_i} + \alpha_z) \dfrac{N_{w_i|z} + \beta}{\sum_{z'} (N_{w_i|z'} + \beta)}$
    $N_{z_i|d_i} \leftarrow N_{z_i|d_i} + 1$
    $N_{w_i|z_i} \leftarrow N_{w_i|z_i} + 1$
  end for
end for

3. TOPIC COHERENCE

Newman et al. [2010] proved that automated evaluation of topic coherence is a good measure of performance by comparing rankings made by a team of domain experts, in accordance with their human expert ranking system.

Mimno et al. [2011] introduced a metric which does not use

Table 3: Term definitions for eq. (1) and eq. (2).

t	A single topic.
D_{w_i}	Number of documents containing word w_i.
$D_{w_i w_j}$	Number of documents containing both word w_i and word w_j.

any external sources to score topics, and consequently runs very quickly. We dub this metric document frequency information (DFI) due to its use of document frequency statistics in order to find results quickly and efficiently. Mimno et al. [2011] found that DFI shows a reasonably strong correlation with the experts' analysis, and can be used to automatically score a topic model.

Each topic is scored as defined in eq. (1). Terms are defined in table 3.

$$\text{DFI-score}(t) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \log \frac{D_{w_i w_j} + 1}{D_{w_i}} \tag{1}$$

The extent to which DFI scores correlate with human expert rankings depends, of course, on the make-up of the source corpus. Therefore, as an extension, we can run this metric using any reference corpus to score a topic model, not necessarily the source corpus.

In the case of EWTM using an external reference corpus, the metric is likely to find that some words are not present in the reference corpus, and so we modify the algorithm to ensure that the denominator is not 0, as defined in eq. (2). Terms are defined in table 3.

$$\text{DFI-score}(t) = \sum_{j=2}^{n} \sum_{i=1}^{j-1} \log \frac{D_{w_i w_j} + 1}{D_{w_i} + 1} \tag{2}$$

4. PÓLYA URN MODEL

The Pólya Urn model was presented by Mahmoud [2008], and demonstrates how term relatedness can be used to affect the sampling process. The normal sampling process can be imagined as drawing a random, i.i.d. sample w_i from an urn where the number of a particular sample is proportional to its probability, then replacing w_i along with another identical sample.

Using the generalised Pólya Urn model, we replace the sample along with an identical sample and A_{vw_i} additional samples for each $w_i \in 1, \ldots, N$, where A is a $N \times N$ matrix known as the schema. Setting A equal to the identity matrix is analogous to the simple Pólya Urn model, but would not provide a meaningful difference in the sampling process.

Instead, the schema is generated before the model is run according to eqs. (3) to (5).

λ_{w_i} is the standard inverse document frequency (IDF) metric [Jones, 1972] for word w_i, with terms D_{w_i} and $D_{w_i w_j}$ as defined in table 3.

$$A_{w_i w_j} \propto \lambda_{w_i} D_{w_i w_j} \tag{3}$$

$$A_{w_i w_i} \propto \lambda_{w_i} D_{w_i} \tag{4}$$

$$\lambda_{w_i} = \log(\frac{|D|}{D_{w_i}}) \tag{5}$$

Mimno et al. [2011] found that it was empirically helpful to remove off-diagonal elements for the common types occurring in more than 5% of documents. A is proportional to the identities stated above because it is normalised by column to sum to 1.

5. MODIFIED GIBBS SAMPLING

In order to improve the coherence of LDA, we want to encourage words which co-occur in the corpus, and discourage those which do not co-occur. We must take into account that most words are rare due to the power-law characteristic of language.

For these reasons, Mimno et al. [2011] found that the generalised Pólya Urn model is ideal for an adaptation of the training algorithm. This adaptation has the effect that the occurrence of a word w_i increases both the probability of seeing word w_i again, and seeing related words that co-occur.

As stated by Mimno et al. [2011], such a model retains LDA characteristics, but replaces the Pólya Urn topic-word component with a generalised Pólya Urn framework, as described by Mahmoud [2008].

Algorithm 3 Modified generalised Pólya Gibbs sample.

```
for all d ∈ D do
  for all w_n ∈ w^(d) do
    N_{z_i|d_i} ← N_{z_i|d_i} − 1
    for all v ∈ V do
      N_{v|z_i} ← N_{v|z_i} − A_{vw_i}
    end for
    z_i ∝ (N_{z|d_i} + α_z) (N_{w_i|z} + β) / (Σ_{z'} (N_{w_i|z'} + β))
    N_{z_i|d_i} ← N_{z_i|d_i} + 1
    for all v ∈ V do
      N_{v|z_i} ← N_{v|z_i} + A_{vw_i}
    end for
  end for
end for
```

If we compare algorithm 3 to algorithm 2, the lines

$$N_{w_i|z_i} \leftarrow N_{w_i|z_i} - 1$$

and

$$N_{w_i|z_i} \leftarrow N_{w_i|z_i} + 1$$

have each been replaced with loops for all similar terms $v \in V$. Terms are defined in table 2.

6. EXTERNALLY WEIGHTED TOPIC MODEL

We present the externally weighted topic model (EWTM), which uses pairwise term relationships derived from an external validation corpus in order to enhance topic coherence.

Term pair statistics are first gathered from the validation corpus before the topic model is trained using eq. (2). The modified Gibbs sample step discussed in section 5 is then used to imbue the standard model with the gathered external information.

Term pair statistics could be derived in a number of ways. Pointwise mutual information (PMI) [Newman et al., 2010] is an attractive option due to its effectiveness in assessing topic coherence, but for simplicity and speed, we use the document frequency metric defined by Mimno et al. [2011]. This is because we found the PMI algorithm to be too computationally expensive to provide the speed expected from the topic model training process.

Our solution is extensible as the external corpus can be tailored to suit the desired result. For example, the EWTM can be run with the English Wikipedia as the validation corpus to imbue the model with a good representation of the English language. If a more scientific topic model is desired, the EWTM could be run with a validation corpus consisting of scientific conference abstracts, such as the NIPS corpus [Neural Information Processing Systems]. Consider a multilingual corpus. The EWTM could be run with a validation corpus consisting only of one language, to train a topic model consisting predominantly of that language; or alternatively, the model could be run with a validation corpus formed from smaller corpora in different languages, to improve coherence.

7. CONCLUSION

Initial results have shown that topic models have a high coherence score with a relatively low number of topics, but when the number of topics is increased to model more of the corpus, the coherence drops rapidly. This is not the case with the EWTM, which sees only a minimal drop in coherence as the number of topics is increased. EWTM achieves a peak **30%** topic coherence improvement over LDA and the model defined by Mimno et al. [2011].

It is apparent that LDA has a significant speed advantage over its weighted counterparts: in initial results LDA ran on average in 3% of the time taken by the model defined by Mimno et al. [2011], and in 1% of the time taken by EWTM. Hence if presented with a problem involving topic modelling, we would urge the reader to assess whether speed is an important factor in the system: if it is, LDA may be the best option.

EWTM is a no-holds-barred approach to achieve better topic coherence, and does so. The time taken is dependent on the sizes of the validation corpus and the source corpus, and should taken into consideration when selecting a model. Our unique combination of the Pólya Urn model and an external dataset gives rise to excellent topic coherence, and the ability to tailor results. With free selection of the validation corpus, EWTM could prove to be an extremely valuable tool for researchers.

References

David M. Blei, Andrew Ng, and Michael Jordan. Latent Dirichlet allocation. *JMLR*, 3:993–1022, 2003. URL `http://www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf`.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990. URL `http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf`.

Tom Griffiths. Gibbs Sampling in the Generative Model of Latent Dirichlet Allocation, 2002. URL `http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.7.8022`.

Thomas Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM. URL `http://doi.acm.org/10.1145/312624.312649`.

Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

Hosam Mahmoud. *Polya Urn Models*. Chapman & Hall/CRC, 1 edition, 2008. ISBN 9781420059830.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology-new/D/D11/D11-1024.pdf`.

Neural Information Processing Systems. NIPS Conference Abstracts 1987-1999 Corpus. URL `http://nips.cc/`. Accessed: May 2012.

David Newman, Sarvnaz Karimi, and Lawrence Cavedon. External Evaluation of Topic Models. In *Australasian Document Computing Symposium*, pages 11–18, Sydney, December 2009. ISBN 978-1-74210-171-2. URL `http://www.ics.uci.edu/~newman/pubs/Newman-ADCS-2009.pdf`.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 100–108, Stroudsburg, PA, USA, 2010. ISBN 1-932432-65-5. URL `http://www.aclweb.org/anthology-new/N/N10/N10-1012.pdf`.