More Efficient Topic Modelling Through a Noun Only Approach

Fiona Martin
Department of Computing,
Macquarie University, NSW, 2109,
Australia
fiona.martin@students.mq.edu.au

Mark Johnson
Department of Computing,
Macquarie University, NSW, 2109,
Australia
mark.johnson@mq.edu.au

Abstract

This study compared three topic models trained on three versions of a news corpus. The first model was generated from the raw news corpus, the second was generated from the lemmatised version of the news corpus, and the third model was generated from the lemmatised news corpus reduced to nouns only. We found that the removing all words except nouns improved the topics' semantic coherence. Using the measures developed by Lau et al (2014), the average observed topic coherence improved 6% and the average word intrusion detection improved 8% for the noun only corpus, compared to modelling the raw corpus. Similar improvements on these measures were obtained by simply lemmatising the news corpus, however, the model training times are faster when reducing the articles to the nouns only.

1 Introduction

A challenge when analysing a large collection of text documents is to efficiently summarise the multitude of themes within that collection, and to identify and organise the documents into particular themes. Document collections such as a newspaper corpus contain a wide variety of themes or topics, with each individual article referencing only a very small subset of those topics. Such topics may be broad and coarse grained, such as *politics*, *finance* or *sport*. Alternatively, topics may be more specific, such as articles related to earthquakes in southern California, or to Napa Valley wineries.

Topic modelling is one way to examine the themes in large document collections. Topic modelling considers documents to be a mixture of latent topics. A more formal definition of topics, as provided by Blei (2012), is that a topic is a multinomial distribution over a fixed vocabulary. One of the most prominent algorithms for topic modelling is the Latent Dirichlet Allocation (LDA) algorithm, developed by Blei et al. (2003). Typically the most frequent function words are excluded prior to topic modelling with LDA (termed *stop word removal*). The topics then generated by LDA can be a mixture of nouns, verbs, adjectives, adverbs and any function words not previously excluded. The LDA algorithm treats all word tokens as having equal importance.

It is common to examine the most frequent words associated with the topic, to determine if these words together suggest a particular theme. For example, a topic with the most frequent words {*water plant tree garden flower fruit valley drought*} suggests a possible label of "garden", whereas a topic of {*art good house room style work fashion draw*} seems to combine multiple themes, and is harder to label. Manually assigning a meaning to a topic (e.g. "gardening") is easier for a reviewer if the most frequent words in the topic are semantically coherent. One issue identified with topic modelling is that it can generate 'junk' topics (Mimno et al., 2011), that is, topics lacking coherence (as in the second example above). Such topics are either ambiguous or have no interpretable theme.

While in some instances there may be interest in examining adjectives (say for sentiment analysis), or verbs (if seeking to identify change, for example), often interest centres around entities such as people, places, organisations and events. For articles drawn from all sections of a newspaper (for example, *Sport*, *Business*, *Lifestyle*, *Drive* and so on), it may be useful to organise articles ignoring their section of origin, and instead focus on the subjects of each article, that is, the people, places, organisations and events (e.g. *earthquake* or *Election*). Such information is typically represented in the articles' nouns.

This study builds on the work of Griffiths et al.

Fiona Martin and Mark Johnson. 2015. More Efficient Topic Modelling Through a Noun Only Approach . In *Proceedings of Australasian Language Technology Association Workshop*, pages 111−115.

(2005), Jiang (2009) and Darling et al. (2012), where topics were generated for specific parts of speech. The novelty in this current study is that it is concerned solely with noun topics, and reduces the corpus to nouns prior to topic modelling. As a news corpus tends have a broad and varied vocabulary, that can be time consuming to topic model, limiting articles to only the nouns also offers the advantage of reducing the size of the vocabulary to be modelled.

The question of interest in this current study was whether reducing a news corpus to nouns only would efficiently produce topics that implied coherent themes, which, in turn, may offer more meaningful document clusters. The measures of interest were topic coherence and the time taken to generate the topic model. Previous work by Lau et al. (2014) suggests that lemmatising a corpus improves topic coherence. This study sought to replicate that finding, and then examine if further improvement occurs by limiting the corpus to nouns. The news corpus and the tools applied to that corpus are detailed in the next section. Section 3 provides the results of the topic coherence evaluations, and Section 4 discusses these results in relation to the goal of efficiently generating coherent topics.

2 Data and Methods

2.1 Data and Pre-Processing

Topic models were generated based on a 1991 set of San Jose Mercury News (SJMN) articles, from the Tipster corpus (Harman & Liberman, 1993). The articles in this corpus are in a standard SGML format. The SGML tags of interest were the <HEADLINE>, <LEADPARA>and <TEXT>, where the lead paragraph of the article has been separated from the main text of the article. The SJMN corpus consisted of 90,257 articles, containing 35.8 million words. Part-of-speech (POS) tagging identified 12.9 million nouns, which is just over 36% of the total corpus. The POS tagging meant a single token such as '(text)' was split into three tokens: '(', 'text',')'. Such splits resulted in the lemmatised set of articles being larger, with over 36.2 million tokens. As this split would be done by the topic modelling tool anyway, it made no material difference to the topics generated, but it did increase the number of tokens fed to the topic modeller, slowing the topic generation.

The news articles were pre-processed by part-of-speech (POS) tagging and each word token was lemmatised. POS tagging was done using the Stanford Log-linear Part-of-Speech tagger (StanfordPOS) (Toutanova et al., 2003), v3.3.1 (2014-01-04), using the *wsj-0-18-bidirectional-distsim.tagger* model. The Stanford POS tagger is a maximum-entropy (CMM) part-of-speech (POS) tagger, which assigns Penn Treebank POS tags to word tokens. Following the finding of Lau et al. (2014) that lemmatisation aided topic coherence, the news articles were lemmatised for the second and third versions of the corpus (but not the first set of articles, to be referred to as the *Original* version of the corpus). Lemmatisation was performed using the *morphy* software from NLTK[1], version 2.0.4, and was applied using the POS tag identified for each word. The *morphy* function reduced words to their base form, such as changing 'leveraged' to 'leverage', and 'mice' to 'mouse'. A Python script was used to create a version of the SJMN articles that contained only tokens tagged with the Penn Treebank noun type tags.

Three distinct versions of the articles were formed, to generate three separate series of topic models. The first version was the complete, original SJMN articles. The second version was a lemmatised set of SJMN articles. The third version was a lemmatised set of SJMN articles, reduced to only nouns. Punctuation was removed from the text in all three versions of the news corpus.

2.2 Topic Modelling

Topic modelling was performed using the Mallet software from the University of Massachusetts Amherst (McCallum, 2002). The Mallet software was run to generate topics using the Latent Dirichlet Allocation (LDA) algorithm, configured to convert all text to lowercase, to model individual features (not n-grams), and to remove words predefined in the Mallet English stop-word list prior to topic modelling. The default settings were used for the optimise-interval hyper-parameter (20) and the Gibbs sampling iterations (1,000). The Mallet software uses a random seed, so the resulting topics can vary between models even when generated using the exactly the same settings and corpus. It is expected that, on balance, dominant topics should re-occur each time the topics are generated, but the nature of such unsupervised learning means that this may not al-

[1] http://www.nltk.org/howto/wordnet.html

ways be the case. To account for such variation, topic models were generated ten times for each set of the news articles, and scores averaged across those ten runs.

The Mallet software requires the number of topics to be specified in advance. As there is not yet an agreed best method for determining the number of topics, this study generated separate sets of 20, 50, 100, 200 and 500 topics. All showed similar patterns between the three data sets. The 200 topics produced the highest topic coherence, as assessed by the measures described in the next section, and for brevity, only the results of the 200 topic runs are reported in this paper.

Table 1: *Average Topic Coherence Measures*

Version	Mean (SD)	Median	Range
1. Original	0.162 (0.087)	0.160	0-0.52
2. Lemmatised	0.170 (0.086)	0.165	0-0.49
3. Nouns Only	0.172 (0.081)	0.170	0-0.49

For each version of the articles, OC scores were averaged across the 200 topics, across the ten topic models (n=2,000).

Table 2: *Number of Low Coherence Topics*

Version	OC <0.1	OC = 0
1. Original	409 (20%)	16 (8%)
2. Lemmatised	346 (17%)	9 (5%)
3. Nouns Only	305 (15%)	1 (1%)

Counts are across the ten models of 200 topics (i.e. n=2,000). The figures in brackets are a percent of the 2000 total topics, for each article set.

2.3 Topic Evaluation

The study by Lau et al. (2014)[2] produced two measures found to be well correlated with human evaluations of topic coherence, and those two measures were used in this current study. The first was an observed coherence (OC) measure, that was configured to use normalised point-wise mutual information (NPMI) to determine how frequently words co-occur in a corpus, and then use this to measure the coherence of the top ten most frequent words in each topic. An NPMI OC score closer to 1 reflected greater co-occurrence, whereas a score of 0 indicated the words were independent.

The second measure was an automated word intrusion detection (WI) task. This task required an intruder word to be inserted into a random location in each topic. The intruder words needed to be words common to the corpus, but not related to the themes in the individual topic. The WI software used the word co-occurrence statistics from the reference corpus to choose which word was most likely to be the intruder. The WI software rated accuracy as either detected (1) or not detected (0).

The proportion of topics where the WI software automatically detected the intruder word was calculated per model via a Python script. This result was expressed as a proportion between 0 and 1, with a value of 0.5 indicating that only half of the intruder words were detected across all (200) topics. A proportion of 1 would indicate all intruder words detected, and 0 indicated no intruder words were detected in any topics. The San Jose Mercury corpus was used as the reference corpus for calculating the baseline co-occurrence.

A final check determined the percentage of nouns in the top 19 most frequent words for each topic. This check was done only for topics generated from the original corpus. To be counted as a noun, a word must have been POS tagged as a noun somewhere in the corpus (for example, "burden" might appear as both a verb and a noun at different places in the corpus, but will be counted as a noun for this statistic).

3 Results

The NPMI Observed Coherence (OC) proportions and the Word Intrusion (WI) detection percentages are shown in Table 1 and 3, respectively. These figures suggest an improvement in topic coherence in the second and third models. Table 2 indicates that all three article sets produced substantial numbers of topics with very low coherence scores. The *Nouns Only* articles produced the least number of low and zero OC coherence topics, suggesting lower numbers of 'junk' topics. Additionally, a review of the topics generated from the original (unaltered) article set indicated a clear predominance of nouns, with over 99% of the 19 most frequent words being nouns, for each of the 200 topics.

It must be noted that the OC scores suggest it was a different set of 200 topics generated each of the ten times topic modelling was performed on the same versions of the articles. For a given version of the articles, none of the ten models produced the same average OC scores as another model on that article set. For example, of the ten models for the *Lemmatised* articles, the mean OC scores ranged between 0.1679 and 0.1744, but no two

[2]The software used in the evaluations was downloaded from https://github.com/jhlau/topic_interpretability, on the 1 May 2014.

Table 3: *Average Word Intrusion Detection*

Version	Mean (SD)	Median	Range
1. Original	0.80 (0.03)	0.79	0.77-0.86
2. Lemmatised	0.88 (0.02)	0.89	0.84-0.90
3. Nouns Only	0.87 (0.03)	0.87	0.83-0.91

Average WI scores were calculated for each of the ten 200 topic models, and the averages of these ten are shown here, for each version of the articles topic modelled (n=200).

Table 4: *Average Time to Generate 200 Topics*

	Time (mins)
Version	Mean (SD)
1. Original	92 (1)
2. Lemmatised	104 (2)
3. Nouns Only	75 (3)

were the same. Minimum, maximum and median OC scores showed similar differences across the ten models. These differences indicate that the generated topics were different in each of the ten models generated for a given article set.

Finally, Table 4 shows that the nouns only corpus was faster to topic model than the other two versions of the news corpus. Part-of-speech tagging the articles took, on average, less than one second per article. Memory restrictions encountered with the part-of-speech tagger meant the articles had to be tagged in parallel sets, rather than tagging the complete corpus at once.

4 Discussion

For the two measures evaluated in this study, reducing the SJMN news corpus to only nouns produced equivalent or improved topic semantic coherence, compared to topic modelling the original news articles. Interestingly, even when the original articles contained all words (apart from the stop words), topic modelling still favoured nouns as the most frequent words in the topics. This suggests that reducing the articles to only nouns may be advantageous in that it may remove extra vocabulary items that would not typically be ranked highly among the most frequent words of a topic anyway. The results of this study suggest that for topic coherence, lemmatising the articles could be the most important factor. However, lemmatising alone does not reduce the time taken to generate the topic model.

Drawing conclusions about any performance impacts is more problematic due to the separate, unintegrated nature of the POS tagging and topic modelling used in this study. There was additional time taken for intermediate file operations that could be eliminated in an integrated process (e.g. piping output between tagging and modelling). Future research could look to integrating the POS tagger and the topic model to gain the best efficiency advantage.

The measures of topic coherence used here are based on whether the top ten most frequent words for a topic are words that commonly co-occur. It does not validate whether these words represent a topic which truly reflects one of the top 200 most frequent themes across articles in the corpus. The substantial variability in both the topic coherence and the word intrusion detection indicate it was not the same 200 topics in each of the ten models generated, for each set of articles. This was confirmed by manual reviews of the topics generated, for each of the three sets of the articles. This variability also occurred when more topics were generated (i.e. 500 topic models) and less topics (i.e. 20, 50, 100 topic models). Though variability is not unexpected in an unsupervised method such as topic modelling, such variability indicates the topics may be unreliable, and is of concern if the end-user seeks to draw detailed conclusions about a corpus based on a single topic model. For example, if a topic related to earthquakes occurred in one set of topics, then it cannot be guaranteed that if the model is re-generated, that such an earthquake topic will re-occur. Therefore, caution should be applied when using topics to make inferences about a corpus, and all inferences should be cross checked using alternate means.

5 Conclusion and Future Work

This study replicated the findings of Lau et al. (2014) that lemmatising improves topic coherence, on observed coherence and word intrusion measures. Limiting the lemmatised corpus to nouns only retains this coherence advantage, while reducing model generation time. Therefore, this study found that lemmatising and limiting the news corpus to the nouns offers advantages in topic coherence and speed, compared to topic modelling the raw corpus of SJMN articles, or lemmatising alone. While this study considered topic coherence, future work could seek to improve topic reliability (i.e. topic consistency). This may include new measures of topic reliability, and optimising the number of topics that can be reliably generated for a given corpus.

Acknowledgments

This research is funded by the Capital Markets Co-operative Research Centre.

References

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

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

Darling, W. M., Paul, M. J., & Song, F. (2012). Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the Workshop on Semantic Analysis in Social Media* (pp. 1–9). Stroudsburg, PA, USA: The Association for Computational Linguistics.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. In Saul, L., Weiss, Y., & Bottou, L. (Eds.), *Advances in Neural Information Processing Systems 17*, (pp. 537–544). MIT Press.

Harman, D. & Liberman, M. (1993). TIPSTER Complete LDC93T3A. DVD. `https://catalog.ldc.upenn.edu/LDC93T3D`.

Jiang, J. (2009). Modeling syntactic structures of topics with a nested HMM-LDA. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining* (pp. 824–829). Washington, DC, USA: IEEE Computer Society.

Lau, J. H., Newman, D., & Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)* (pp. 530–539). Gothenburg, Sweden: Association for Computational Linguistics.

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 262–272). Association for Computational Linguistics.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, (pp. 173–180). The Association for Computational Linguistics.