A

On Collocations and Topic Models

Jey Han Lau[1,2], Timothy Baldwin[1,2] and David Newman[3]
[1] Dept of Computing and Information Systems, The University of Melbourne, Australia
[2] NICTA Victoria Research Laboratory, Australia
[3] Dept of Computer Science, University of California, Irvine, USA
jhlau@csse.unimelb.edu.au, tb@ldwin.net, newman@uci.edu

We investigate the impact of pre-extracting and tokenising bigram collocations on topic models. Using extensive experiments on four different corpora, we show that incorporating bigram collocations in the document representation creates more parsimonious models and improves topic coherence. We point out some problems in interpreting test likelihood and test perplexity to compare model fit, and suggest an alternate measure that penalises model complexity. We show how the Akaike information criterion is a more appropriate measure, which suggests that using a modest number (up to 1000) of top-ranked bigrams is the optimal topic modelling configuration. Using these 1000 bigrams also results in improved topic quality over unigram tokenisation. Further increases in topic quality can be achieved by using up to 10,000 bigrams, but this is at the cost of a more complex model. We also show that multiword (bigram and longer) named entities give consistent results, indicating that they should be represented as single tokens. This is the first work to explicitly study the effect of n-gram tokenisation on LDA topic models, and the first work to make empirical recommendations to topic modelling practitioners, challenging the standard practice of unigram-based tokenisation.

Categories and Subject Descriptors: I.2.7 [**Natural Language Processing**]: Text Analysis; Language Parsing and Understanding

General Terms: Artificial Intelligence

1. INTRODUCTION

Blei et al. [2003] introduced the Latent Dirichlet Allocation (LDA) topic model as an unsupervised method for learning topics from a document collection, where a topic is a multinomial distribution over terms.[1] Topic models have since achieved notable successes in areas such as multi-document summarisation [Haghighi and Vanderwende 2009], word sense discrimination [Brody and Lapata 2009; Lau et al. 2012], sentiment analysis [Titov and McDonald 2008] and information retrieval [Wei and Croft 2006]. The input to the topic model is the bag-of-words representation of a document collection, where word counts are preserved but word order is lost. Despite discarding word order, topic models have a remarkable ability to learn semantically meaningful topics. Since a large amount of semantic information is captured by collocations, we investigate whether we can improve topic models by representing collocations as single tokens.

Following Choueka [1988], we define a collocation to be a sequence of consecutive words that has the characteristics of a syntactic and semantic unit. Examples of collocations are *stock market*, *White House*, and *health care*. In this work we will focus primarily on bigram collocations (and use the terms "bigram" and "collocation" interchangeably).

Usually when topic modelling, words from text documents are tokenised as unigrams, i.e. single words. As word order is lost after tokenisation, bigrams such as *health care* function as two discrete units in the topic model. In a Gibbs-sampled LDA topic model, every single token in the corpus has a topic assigned to it, i.e. *health*

[1] A full introduction to LDA is beyond the scope of this paper, and we refer the reader instead to the excellent introduction by Blei and Lafferty [2009].

Lau, Jey Han, David Newman and Timothy Baldwin (to appear) On Collocations and Topic Models, ACM Transactions on Speech and Language Processing.

and *care* obtain separate topic assignments.[2] If these two tokens appear as a bigram, it would make sense that they receive the same topic assignment, and by encoding bigrams as a single token, we can guarantee this. Therefore one might expect that treating bigrams as single tokens might help topic models. We also expect that bigrams may help some observed problems with topic quality. Mimno et al. [2011] gave an example of a topic whose top three words were *acids*, *fatty* and *nucleic*. This topic is a mix of two very distinct biological concepts: fatty acids are derived from fats and are an important source of fuel; nucleic acids, on the other hand, are building blocks of DNA and RNA, and transmit genetic information. These two very distinct concepts are brought together in a topic via their co-appearance with the term *acid*). Encoding *fatty_acids* and *nucleic_acids* as single tokens would have likely prevented the creation of this low quality mixed-concept topic.

Intuitively, it seems clear that collocations should improve topic modelling, but as any researcher knows, intuitively-appealing predictions aren't always supported by empirical evidence. The core question of this paper, therefore, is:

> *Do collocations empirically enhance topic models, and if so, under what conditions?*

We explore this question by comparing topic models learned from unigram bag-of-words data, with topic models learned from bag-of-words data that includes pre-extracted bigram collocations. We conduct an extensive suite of experiments, using four datasets, three topic settings, multiple seeded models, and four bigram replacement methods. We measure model fit using test likelihood and test perplexity, and also topic coherence. We ultimately find that using a modest amount of bigram replacement (identifying 1000–10,000 highly-ranked bigrams, and representing each one using a single token) improves the quality of topic models.

2. BACKGROUND

The idea of using collocation information in topic models is not a new one, and a number of topic models that use collocations have been proposed. In the Bigram Topic Model [Wallach 2006], word probabilities are conditioned on the previous token (i.e. take the form of bigram probabilities). The LDA Collocation Model [Griffiths et al. 2007] extends the Bigram Topic Model by giving it the flexibility to generate both unigrams and bigrams. The Topical N-grams Model [Wang et al. 2007] adds a layer of complexity to allow the formation of bigrams to be determined by context. Hu et al. [2008] introduced the Topical Word-character Model, challenging the common assumption that the topic of an n-gram is derivable from the topics of its composite words. Johnson [2010] derived the relationship between LDA and Probabilistic Context-Free Grammars, and proposed a model combining LDA and Adaptor Grammars to incorporate collocations while doing topic modelling.

While these models have a theoretically elegant probabilistic treatment of collocations, they do so at higher computational overhead and model size. For example, Wallach's Bigram Topic Model has W^2T parameters, compared to WT for LDA, for number of topics T and size of vocabulary W. As such, these models have not been widely adopted, and are mostly of theoretical interest, and less useful for topic model practitioners.

[2]We note that the sampling of the topic assignments for both words is not completely independent given that they belong to the same document (i.e. they both draw from the same topic distribution that the document has), but as topic assignments are sampled separately, it is possible that the two words receive a different topic assignment.

Document collection	D	N	W	$L = \frac{N}{D}$
ACL	9.8×10^3	17×10^6	165×10^3	1735
BLOGS	660×10^3	55×10^6	616×10^3	83
NEWS	97×10^3	43×10^6	155×10^3	443
PUBMED	74×10^3	3.2×10^6	34×10^3	43
NIH	83×10^3	13×10^6	41×10^3	155

Table I: Statistics of the document collections used in this research (D = documents; N = word tokens; W = word types/vocabulary size; and L = average document length)

The main difference in our methodology is that collocation discovery/extraction is done in pre-processing, and collocations are represented as a single token in the bag-of-words model before learning a standard LDA topic model. Our approach allows flexibility in how one wishes to define a collocation to create a bag-of-words for topic modelling. Learned topics would also contain collocations, and our intuition tells us that these topics would be more coherent and could be better presented to human users. Given the growing trend towards larger document collections and interest in run-time speed, computational efficiency is paramount, and a preprocessing approach to topic modelling with collocations has clear advantages.

3. METHODOLOGY

3.1. Datasets

In our experiments, we use text collections from four distinct sources for our primary experiments. The differences in style and subject matter in these domains provide us with a diverse range of content:

ACL . Conference papers from the ACL Anthology Network[3]
BLOGS . Blog articles dated from August to October 2008 from the Spinn3r blog dataset[4]
NEWS . New York Times news articles dated from July to December 1999, from the English Gigaword corpus
PUBMED . Abstracts from a PubMed search for *traumatic brain injury*[5]

We use a fifth dataset in Section 4.3 for a targeted experiment where we compare automatically-extracted collocations with a set of gold-standard collocations for the document collection:

NIH . NIH grant abstracts dated from 2004 to 2009

Statistics of each document collection are summarised in Table I.

We used very simple tokenisation, removing all punctuation, but not using any stemming or lemmatisation, and only keeping terms longer than 2 characters. As a domain-independent proxy for stopword removal, we simply deleted the 200 most-frequent terms in each corpus, and additionally deleted terms occurring less than once per million words.

We considered four different bigram replacement methods, simply adjusting the extent of bigrams replaced. We first extract bigrams from each document collection using the N-gram Statistics Package [Banerjee and Pedersen 2003], identifying the top

[3] http://clair.eecs.umich.edu/aan/index.php
[4] http://www.icwsm.org/data/
[5] http://www.ncbi.nlm.nih.gov/pubmed

bigrams based on the Student's t-test. Clearly, there is scope to explore other lexical association measures and n-grams of different size/syntactic configuration (c.f., [Pecina 2009]), which we leave for future work. We then used the top-1k, 10k and 100k as the three different bigram replacement methods, comparing these to the unigram bag of words. In this paper, we refer to the unigram and varying-size bigram treatments as $\emptyset$, 1K, 10K, 100K.

We generate the bigram tokenisation variant of each document collection by greedily combining all occurrences of each word pairing in the bigram list into a single token; e.g. if *health care* were extracted as a bigram, all instances of the bigram *health care* would be replaced with the single token *health_care*.[6] We thus have a simple unigram ($\emptyset$) and three bigram (1K, 10K, 100K) variants of each document collection. Using these four alternate document representations, we learned a series of LDA topic models using different settings for the number of topics $T = \{100, 200, 400\}$. For all experiments, we performed the topic modelling with 2 different random seeds. Test data was created by putting aside 10% of each collection before modelling.

3.2. Evaluation Framework

3.2.1. Model Fit and Complexity. We first look at the log likelihood and perplexity of the test data, defined as $\mathcal{L} = \sum_k^N \log p(x_k)$ (where $p(x_k)$ is the probability of term x_k) and Perplexity $= \exp \frac{-\mathcal{L}}{N}$, which is simply the inverse of the geometric mean of per-term likelihood. Although there have been studies that suggest perplexity is not suited to topic model evaluation [Chang et al. 2009], it is still commonly used for comparing different models on the same data in the topic model literature. In our case, we have different treatments of the same collection (effectively producing different input data), but the same LDA algorithm. We will demonstrate the impact of this fact on the evaluation.

When we sample all bigram types from each of our four document collections, we observe that for the vast majority of bigram types (ACL = 96%, BLOGS = 93%, NEWS = 96%, and PUBMED = 99% of cases), the MLE-based probability of the bigram is larger than the product of the unigram probabilities, i.e. $p(x_1)p(x_2) < p(x_1, x_2)$. That is, if we take a random bigram from a given document collection (e.g. *language processing* from ACL), it is highly likely that the probability of that bigram is higher than the probability estimate assuming independence of the two component unigrams *language* and *processing* (i.e. p(language processing) $> p$(language)p(processing)). By definition, these terms are equal only when the occurrence of x_2 is independent of x_1. Replacing the two adjacent tokens with a bigram token improves (increases) likelihood $\mathcal{L}$. However since $\sqrt{p(x_1)p(x_2)} \geq p(x_1, x_2)$ (from positivity), the perplexity measure, which is on a per-term basis, will get worse (increase). This increase in perplexity is consistent with the language's entropy increase. By replacing two frequently-occurring terms with a single bigram term that occurs less frequently (than either unigram component), we reduce the frequency of the two unigram terms, and increase the frequency of a rarer bigram term.

This improvement in likelihood and worsening of perplexity as one replaces top bigrams with bigram-tokens may seem contradictory. However, after recognising that for replaced bigrams $p(x_1)p(x_2) < p(x_1, x_2) \leq \sqrt{p(x_1)p(x_2)}$ we can see this must be true, suggesting that neither likelihood nor perplexity are useful for evaluating which bigram replacement protocol is best.

[6] We also experimented with an n-gram supplementation approach, where the original tokens were preserved and a new n-gram was injected into the document, but in preliminary experiments, found that the n-gram replacement scheme consistently achieved superior results.

By replacing bigrams, we're seeing an improvement in likelihood because our model has more parameters. For every bigram replaced, we add the new bigram-token to the vocabulary, but because these are high-frequency bigrams, in almost all cases the unigram tokens remain in the vocabulary. For example, for the bigram *nuclear weapons* we see 3032 occurrences of *nuclear weapons*, and 14808 occurrences of *nuclear* and 19781 occurrences of *weapons*, clearly not eliminating either of the unigram terms from the vocabulary. Note that creating the bigram token *nuclear_weapons* reduces the topic model's ability to associate the terms *nuclear* and *weapons*. So we expect there to be some limit to the benefit of representing an increasing number of bigrams as single tokens.

The misleading monotonic improvement in likelihood (from the contraction of representing bigrams as single tokens) can be addressed by considering a modified likelihood score that penalises the number of parameters. This is a sensible notion: we desire a model that has good likelihood, but not one that is overparameterised by an excessively large vocabulary (that includes bigrams-as-single-terms). Several methods have been proposed for penalising complexity, with BIC (Bayesian Information Criterion) and AIC (Akaike Information Criterion) being two long-standing approaches [Akaike 1974]. We chose AIC because it is less strict than BIC: AIC's penalty does not grow with the amount of data, as is the case for BIC. This allows us to have some penalty term, but not one that overly dominates the results. Akaike Information Criterion (AIC) is defined as $AIC = -2\mathcal{L} + 2WT$, where WT is the number of parameters in the LDA model.

3.2.2. Topic Coherence. Topic coherence (TC) is a qualitative evaluation of the semantic nature of learned topics, and measures the *interpretability* of a topic based on a human judgment. As topics are typically presented to users via its top-N topic terms, coherence of a topic is judged by whether its top-N terms collectively convey a subject or theme [Chang et al. 2009; Newman et al. 2009; Newman et al. 2010; Lau et al. 2010]. Newman et al. [2010] proposed a pointwise mutual information (TC-PMI) based topic coherence score, while Mimno et al. [2011] presented a variation on that score, using log conditional probability (TC-LCP) instead of PMI:

$$\text{TC-PMI}(\mathbf{w}) = \sum_{j=2}^{10}\sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)P(w_j)}$$

$$\text{TC-LCP}(\mathbf{w}) = \sum_{j=2}^{10}\sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)}$$

where i and j are indices of pairs of word and $\mathbf{w} = \{w_1, w_2, \ldots, w_{10}\}$ are the top-10 most likely terms in a topic. PMI and LCP are based on term co-occurrences $N(w_i, w_j)$—the number of times terms w_i and w_j co-occur together—which is further described below. To calculate the overall coherence of a topic model, we average the topic coherence scores over all topics.

Newman et al. [2010] proposed the use of WIKIPEDIA as an external document source to sample word counts for calculating the PMI scores. Co-occurrence of a pair of words is based on the number of times both words appear in the same sliding window. Mimno et al. [2011], on the other hand, suggested a slightly different approach, using the training topic model document collection for sampling word counts and the full document for deciding co-occurrence of words—essentially co-appearance of terms in a document. However, in establishing the ability of these techniques to model topic coherence, both sets of authors used exclusively unigram tokens. As our topics will potentially contain a mixture of unigrams and bigrams, we need to first establish whether

Topic Type	TC-PMI	TC-LCP$_{\text{DOC}}$	TC-LCP$_{\text{WIKI}}$
Unigram	0.799	0.189	0.604
Collocation	0.865	−0.070	0.743
Combined	0.853	0.057	0.658

Table II: Spearman's ρ for TC-PMI and TC-LCP (based on the topic-modelled document collection and WIKIPEDIA) relative to human ratings

the two measures work equally well over our topics. To this end, we ran a survey where we randomly sampled 40 topics—20 topics that contained only unigram topic words and 20 topics that contained a mixture of unigram and bigram collocation topic words—from the $T = 100$ topic model for NEWS. To maximise comparability, we follow the annotation procedure described in Newman et al. [2010]. In particular, we had 9 annotators rate each of the topics (displayed as a list of 10 words) on a 3-point scale, where a score of 3 indicates a very coherent topic and a score of 1 indicates an incoherent or "junk" topic.[7] Examples were provided to help annotators understand the task and the evaluation measure. The gold-standard coherence score for each topic is calculated by averaging across the human ratings.

We compute TC-PMI scores using WIKIPEDIA and a sliding window of 20 words for sampling word counts, as described in Newman et al. [2010]. We compute TC-LCP scores, on the other hand, based on the document co-occurrence in the collection that we topic modelled (i.e. in-corpus; "TC-LCP$_{\text{DOC}}$"), in line with the methodology proposed in Mimno et al. [2011]. We additionally calculate TC-LCP using WIKIPEDIA (with the 20-word sliding window approach; "TC-LCP$_{\text{WIKI}}$"), for a fairer comparison with TC-PMI. To evaluate the topic coherence scores against human ratings, we calculate the Spearman rank correlation coefficient (ρ); results are presented in Table II. Note that the Spearman's ρ is computed separately for: (a) the unigram topics ($\times 20$; "Unigram"); (b) the collocation topics ($\times 20$; "Collocation"); and (c) the combined set of topics ($\times 40$; "Combined").

From the results we see that the PMI-based topic coherence, TC-PMI, consistently has a stronger correlation with human ratings for both unigram and collocation topics, compared to TC-LCP. We also see that when collocations are included in the topics, TC-LCP performs much better when calculated over WIKIPEDIA, due to data sparseness in the document collection, but that it is still markedly worse than TC-PMI, contrasting with the findings of Mimno et al. [2011] over strictly unigram topics. In light of this, we use TC-PMI to evaluate topic coherence for the remainder of this paper.

4. RESULTS

In this section we present results of our proposed treatment of bigrams, as evaluated by AIC in Section 4.1 and PMI-based topic coherence in Section 4.2. We test the impact of manually-verified bigrams over NIH in Section 4.3. We experiment with collocations that have low compositionality (named entities) in Section 4.4. Lastly, we present an extrinsic evaluation of bigram tokenisation, in the content of document classification using topic model-derived features (Section 4.5)

By way of mention, we noticed that replacing the top-k bigrams with single tokens only has a mild effect on the overall corpus-wide frequency of bigram-tokens. Across our first four datasets, replacing the top-1K bigrams produces a corpus with 2% to 5% bigram-tokens, and replacing the top-100K bigrams produces a corpus with 12%

[7]The annotators are Computer Science postgraduate student working in the area of language technology at The University of Melbourne.

	$\emptyset$	1K	10K	100K
$T = 100$	65.5	**65.0**	66.1	82.8
$T = 200$	98.1	**97.8**	100.7	134.9
$T = 400$	**163.9**	**163.9**	170.4	239.7

(a) ACL

	$\emptyset$	1K	10K	100K
$T = 100$	229.4	**228.8**	229.4	244.7
$T = 200$	350.9	**350.5**	352.8	385.5
$T = 400$	**595.8**	**595.8**	601.8	669.0

(b) BLOGS

	$\emptyset$	1K	10K	100K
$T = 100$	110.7	**109.5**	109.5	124.3
$T = 200$	140.3	**139.2**	141.0	173.1
$T = 400$	200.7	**200.1**	205.4	271.9

(c) NEWS

	$\emptyset$	1K	10K	100K
$T = 100$	**12.0**	12.1	13.7	16.7
$T = 200$	**18.7**	19.0	22.4	28.4
$T = 400$	**32.2**	32.9	39.8	51.9

(d) PUBMED

Table III: AIC results over the different document collections for varying numbers of topics (T) and collocations ($N \in \{\emptyset, 1\text{K}, 10\text{K}, 100\text{K}\}$).

to 16% bigram-tokens. Since topic modelling is driven by token counts, the relative occurrence of bigram-tokens in top-10 topic terms is even less, with no more than 10% of topic terms being bigram-tokens on average (suggesting that the effect on topic coherence scores may be slight). Note that when we replace a bigram, we're generally reducing the counts of two high frequency terms, and increasing the count of the less frequent bigram-term, thereby increasing the entropy of the language.

4.1. Model Fit and Complexity

We show the AIC results in Table III. For AIC, the lower the number, the better the model. From the table we see **(in bold)** that in the majority of cases, replacing the top-1K bigrams generally produces the most parsimonious models, across four datasets and three settings of topics (note that all numbers presented are of magnitude 10^6). PUBMED was the only exception, where unigrams uniformly gave the best AIC scores.

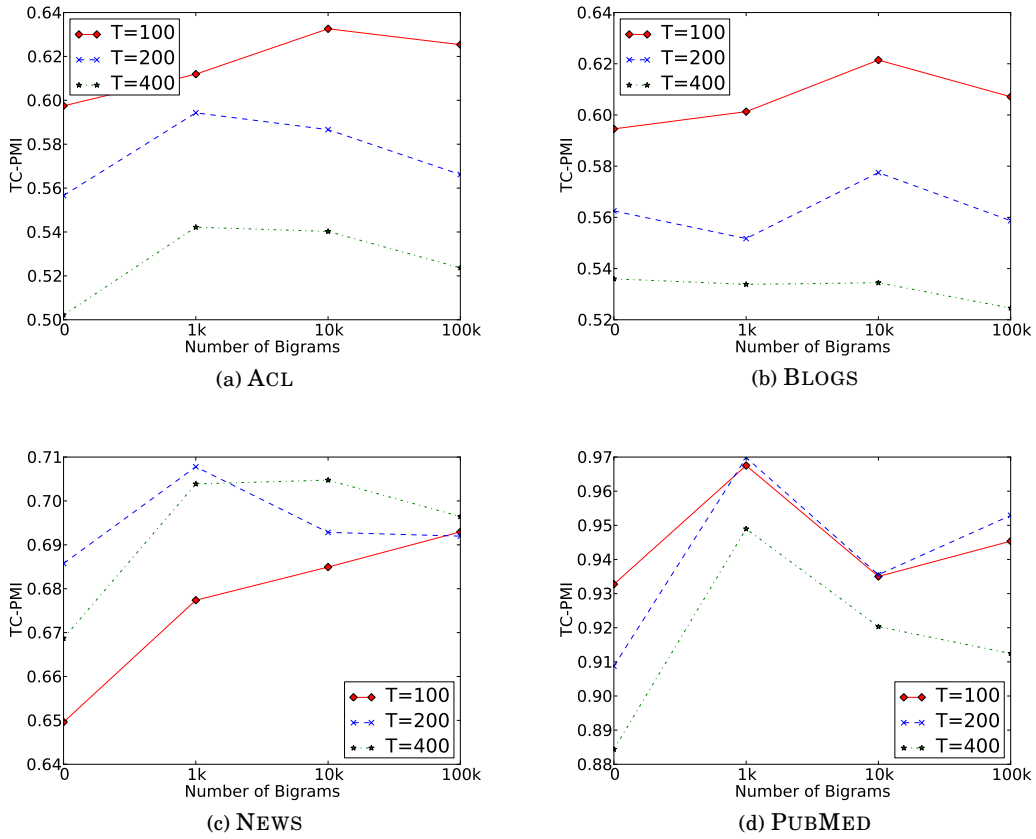
4.2. Topic Coherence

PMI-based topic coherence (TC-PMI) results are presented in Figure 1 for each of ACL, BLOGS, NEWS and PUBMED (note that the overall topic coherence score of a topic model is an average of topic coherence scores over all its topics). We consistently see an improvement in topic coherence as we substitute bigrams, across all four datasets and three settings of topics (T). We generally see the maximum topic coherence occur when using 1K to 10K bigrams, and the topic coherence falls off when using 100K bigrams.

4.3. Gold-Standard Collocations

All results to date have been based on automatically-extracted collocations, the majority of which are highly compositional, and many of which are not true collocations (e.g. *central nervous*). To better explore the impact of the quality of the collocation set on the topic model we ran additional experiments on the NIH dataset where we had access to gold-standard collocations specific to the document collection.

The NIH dataset is a collection of 83,000 grant abstracts from the National Institutes of Health (NIH) in the USA. The gold standard collocations were extracted as follows. First, we created the list of 10,000 most frequent bigrams from the collection. This list was then reviewed by a domain expert who identified "true" collocations. This review yielded a set of 4,254 bigram terms. As the validation was done based on a pre-extracted set of bigrams, we have no way of evaluating the recall of the gold-standard collocation set, and an alternative extraction/validation process may result in a collocation set with better coverage. On the other hand, this approach leads to a collocation



(a) ACL (b) BLOGS

(c) NEWS (d) PUBMED

Fig. 1: PMI-based topic coherence (TC-PMI) for varying numbers of bigrams, under different T settings.

set which is comparable with the collocations used elsewhere in this research (since the candidates were initially automatically extracted), just of higher quality.

For comparison, we additionally automatically extracted the same number (4,254) of bigrams, ranked by the Student's t-test. In summary, we have three tokenisation methods: (1) unigram tokenisation ($\emptyset$); (2) bigram replacement using 4,254 automatically extracted bigrams (AUTO); and (3) bigram replacement using 4,254 manually verified bigrams (GOLD). We evaluated the three variants using AIC scores and topic coherence for three settings of number of topics ($T = 100, 200, 400$). AIC and topic coherence results are presented in Table IV (note that all AIC scores presented are of magnitude 10^6).

Looking at the AIC scores, we see that $\emptyset$ outperforms AUTO and GOLD at $T = 200, 400$. Although somewhat discouraging, we note this is similar to the results we saw for PUBMED in Section 4.1. As NIH and PUBMED are both medical domains and their style of text are very similar, the results we see here are perhaps unsurprising. As a consolation, GOLD at $T = 100$ outperforms $\emptyset$, indicating that, depending on the topic model

	$\emptyset$	AUTO	GOLD
$T = 100$	22.0	22.1	**21.9**
$T = 200$	**29.3**	30.3	30.5
$T = 400$	**44.5**	47.2	48.2

(a) AIC Scores

	$\emptyset$	AUTO	GOLD
$T = 100$	1.223	1.177	**1.351**
$T = 200$	1.204	1.196	**1.346**
$T = 400$	1.184	1.176	**1.325**

(b) Topic coherence

Table IV: AIC and PMI-based topic coherence results for the three different NIH models: $\emptyset$, AUTO and GOLD.

settings, gold-standard collocations can have a positive impact compared to the baseline unigram model.

In terms of topic coherence, GOLD consistently outperforms both $\emptyset$ and AUTO. This observation is enlightening, and it suggests that using "higher quality" collocations causes the model to produce more coherent topics.

4.4. Named Entities

To further explore the impact of compositionality on topic modelling and our hypothesis that low-compositionality collocations benefit the most from our preprocessing method, we are ideally after a dataset with large numbers of non-compositional collocations. In the absence of such a dataset—and also the absence of a reliable method for predicting the compositionality of an arbitrary MWE (despite promising results in the recent work of Reddy et al. [2011] and Hermann et al. [2012], inter alia)—we turn to a class of multiword expressions with a high preponderance of non-compositional items, namely multiword named entities. That is, the named entities such as *Los Angeles* and *John Smith* are non-compositional, as their components do not have a well-defined semantic interpretation in isolation for a compositional interpretation to be based off. While there are certainly significant numbers of compositional multiword named entities (such as *South Melbourne* and, more subtly, *Saif al-Islam Gaddafi*[8]), manual analysis would suggest that the relative level of compositionality among multiword named entities is considerably less than among the more general set of bigram collocations.

Multiword named entities should help reduce the complexity of the topic model and thus improve it, on the grounds that: (1) effective vocabulary size is reduced for "cranberry" expressions (i.e. expressions where one or more components don't exist as tokens outside that named entity, such as *Angeles* in *Los Angeles* for a standard English document collection); and (2) spurious associations between named entities with lexical overlap (e.g. *John Smith* and *John Carpenter*, notwithstanding the existence of named entities where lexical overlap is semantically significant, such as between *Saif al-Islam Gaddafi* and *Muammar Gadaffi*).

To explore the impact of multiword named entities on topic modelling, we automatically identified named entities in the NEWS corpus using the Stanford named entity recogniser [Finkel et al. 2005], and pre-tokenised all instances of multiword named entities of type person, location and organisation. As the named entity recogniser is trained over news articles, we opted to run the named entities experiment only for NEWS for compatibility reasons.[9] AIC and topic coherence results for the unigram

[8]Our claim of compositionality here stems from the fact that the surname *Gaddafi* is so strongly associated with *Muammar Gadaffi* that any person mentioned in Western media with that surname can reliably be assumed to have some association with the former Libyan dictator.

[9]To illustrate the importance of compatibility, over 380,000 named entities were extracted for NEWS but less than 8,000 named entities were extracted for PUBMED, out of which none are protein named entities.

	$\emptyset$	NE
$T = 100$	**20.1**	23.3
$T = 200$	**32.6**	39.1
$T = 400$	**57.7**	71.0

(a) AIC Scores

	$\emptyset$	NE
$T = 100$	0.628	**0.662**
$T = 200$	0.633	**0.659**
$T = 400$	0.654	**0.665**

(b) Topic coherence

Table V: AIC and PMI-based topic coherence results for the unigram model ($\emptyset$) and the model that has named entities incorporated (NE) for NEWS

model ($\emptyset$) and the model with named entities incorporated (NE) are presented in Table V. Note that we did not include a model that uses automatically extracted bigrams, as the comparison would not be meaningful due to the difference in nature of named entities and bigrams (e.g. named entities are generally composed of two or more words, and tend to occur with lower frequency than bigrams).

In terms of AIC scores, $\emptyset$ consistently outperforms NE (bolded). This is somewhat predictable, as there are large numbers of named entities (over 380,000) incorporated in NE. A significant number of them have very low frequency and are unlikely to make a strong impact on the model.[10] The AIC scores are heavily penalised because of the increased number of parameters (from incorporating the large number of named entities), and as such favour $\emptyset$.

Topic coherence, on the other hand, tells another story. Encouragingly, the topics produced by NE have, on average, higher coherence than $\emptyset$, supporting our hypothesis that pre-tokenising low-compositionality collocations improves the topic model.

4.5. Extrinsic Evaluation

Looking back over the intrinsic results to date based on AIC and topic coherence, on the whole, the bigram variants outperform the unigram model. We have yet to answer the burning question, however, of whether these results translate across to improved results using topic modelling with bigram tokenisation in an application.

To this end, we apply our methodology to the task of document classification over the Reuters-21578 dataset. We used Distribution 1.0 of the data,[11] and the standard "ModApte" split to generate the training and test partitions. After performing standard tokenisation and lemmatisation,[12] we removed the 200-most frequent terms as stopwords and pruned any resulting empty documents, producing 8762 training documents and 3009 test documents.

Using the same approach as before, we extracted and ranked bigrams from the dataset using the Student's t-test, and applied the three bigram replacement methods—1K, 10K and 100K—before running the topic models. As a control, we also run a topic model with no bigram replacement ($\emptyset$). Three settings of number of topics ($T = \{100, 200, 400\}$) were used, generating a total of 12 topic models (3 settings of $T \times$ 4 tokenisation settings).

We treated the document classification task as a binary classification problem and tested the 10 most populous classes in the Reuters-21578 dataset separately, in line with past research [Joachims 1998; McCallum 1999; Nigam et al. 2000, inter alia]. For

[10] Approximately 320,000 of the named entities have a frequency less than 5. Note that we did not filter named entities based on frequency, so the frequency could be as low as 1.

[11] `http://archive.ics.uci.edu/ml/datasets/Reuters-21578+Text+Categorization+Collection`

[12] OpenNLP is used for tokenisation, and Morpha for lemmatisation [Minnen et al. 2001]

Class	$\emptyset$	1K	10K	100K
earn	95.01	**97.67**	97.18	96.91
acq	93.75	**95.45**	95.08	94.95
money-fx	97.27	**97.64**	**97.64**	97.34
grain	**97.31**	96.74	96.71	96.91
crude	97.67	**97.94**	97.47	97.21
trade	**97.74**	97.18	97.67	97.41
interest	97.34	**97.87**	97.11	96.94
ship	98.50	**98.64**	**98.64**	98.47
wheat	**98.04**	97.81	97.81	97.84
corn	**98.47**	98.40	98.40	98.40
Macro-Avg	97.11	**97.53**	97.37	97.24

(a) $T = 100$

Class	$\emptyset$	1K	10K	100K
earn	96.31	97.71	**97.74**	96.78
acq	94.18	**95.71**	95.31	94.75
money-fx	96.81	97.51	**97.67**	97.18
grain	97.57	97.71	**97.84**	97.04
crude	97.24	**98.01**	97.97	97.54
trade	97.71	97.54	**97.94**	97.74
interest	97.24	97.47	**97.67**	97.57
ship	**98.90**	98.74	98.67	98.31
wheat	**98.14**	97.71	98.04	97.71
corn	98.37	98.47	**98.54**	98.44
Macro-Avg	97.25	97.66	**97.74**	97.31

(b) $T = 200$

Class	$\emptyset$	1K	10K	100K
earn	96.68	**97.87**	97.54	97.24
acq	94.42	94.82	**95.51**	94.55
money-fx	96.58	**97.14**	96.98	96.91
grain	97.67	**98.21**	97.51	96.81
crude	97.71	**98.07**	97.91	97.54
trade	97.77	**97.84**	97.41	97.61
interest	97.08	**97.77**	97.57	97.67
ship	**98.27**	98.17	98.17	97.77
wheat	97.81	**98.17**	98.01	97.84
corn	98.24	98.37	**98.57**	98.47
Macro-Avg	97.22	**97.64**	97.52	97.24

(c) $T = 400$

Table VI: Classification accuracy over the 10 most populous class in Reuters-21578.

each class, we learnt a linear-kernel SVM using the topic features and evaluated over the test data using classification accuracy.[13] Results are summarised in Table VI.

From the results, we see that the 1K and 10K bigram variants have the highest accuracy in most cases (indicated by boldface). The macro-averaged accuracy of each of the bigram variants (1K, 10K and 100K) is also consistently higher than that of the unigram model ($\emptyset$), indicating that incorporating bigrams into the topic model improves its ability to model the data for document categorisation purposes. For $T = \{200, 400\}$,

[13]We used SVM-Light for all our experiments [Joachims 1999].

the 1K and 10K bigram variants are significantly better than the unigram model ($\emptyset$) based on a one-tailed paired t-test ($p < 0.05$).

5. DISCUSSION AND CONCLUSIONS

We set out in this paper to explore the question of whether collocations empirically enhance topic models, and if so, under what conditions. The answer to the first part of the question would appear to be a definitive yes, in the sense that, for every document collection we ran experiments over, we were able to achieve improvements in mean topic coherence through a simple preprocessing step of identifying collocations and greedily replacing all occurrences of those collocations with a single token. Extrinsic evaluation in the application of document classification supported this claim. The results using Akaike Information Criterion (AIC) were slightly less conclusive, in that: (1) we observed an improvement when we incorporated collocations for three out of the four document collections (all based on 1K collocations), with the one exception being PUBMED where the fit with the topic model worsened with the inclusion of collocation information; and (2) we saw that the unigram model outperformed the bigram models in most settings with the gold-standard collocations and named entities, although we were able to identify reasons for this finding in both cases.

We can hypothesise about how bigrams impact on topic models, partly theoretically and partly based on observation of the output of the topic model. Unsurprisingly, the bigrams that have the greatest positive impact on the topic model are those which have lower compositionality [Baldwin et al. 2003; Baldwin and Kim 2009]. For example, a *melting pot* is most likely to be an environment in which ideas are integrated/assimilated into one, rather than a physical *pot* in which *melting* of substances takes place; rather, the usage is based on analogy with a vessel which is used to combine together compounds at high temperature. We would expect the vast majority of occurrences of the expression *melting pot* to correspond to this metaphorical usage of the term, and a greedy replacement strategy with the token *melting_pot* would: (a) lead to cleaner topic modelling of the remaining (predominantly literal usages) of each of *melting* and *pot*; and (b) better capture the non-compositional semantics of *melting_pot*. All up, therefore, the topic model should be enhanced. In practice, this is very much what we observe, particularly at lower values of T where the topic model is working harder to generalise; at higher values of T, the topic model tends to automatically distinguish between different sub-usages of each token and implicitly tease apart the non-compositional usages of tokens such as *pot* in the process.

For compositional bigrams, on the other hand, the effects are more mixed. Consider *lung cancer* and *breast cancer*, for example, which are specific types of *cancer* affecting the indicated body parts. By representing *lung_cancer* and *breast_cancer* as a single token in the model, the direct connection with *cancer* is lost. The net effect of this appears to be that, at lower values of T, a very generic *cancer* topic is generated, and the effects of the lower-frequency sub-types such as *lung cancer* and *breast cancer* are not self-evident. That is, the ability of the topic model to directly model the association between *cancer* and *lung_cancer* is diminished, but the *cancer* topic is, if anything, more coarse-grained and cohesive as a result (e.g. tokens such as *treatment* and *recovery* are boosted and tokens specific to cancer sub-types such as *lung* and *mamogram* are pushed down in the token ranking for that topic). At higher values of T, on the other hand, the model progressively partitions off specific topics for different cancer subtypes, with or without the collocations, and the impact of the tokenisation method is diminished.

To summarise the overall finding of this paper to topic modelling practitioners: Representing top-ranked bigrams with single tokens is beneficial for LDA topic models.

Acknowledgements

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT centre of Excellence programme. DN is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20155. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

REFERENCES

AKAIKE, H. 1974. A new look at the statistical model identification. In *IEEE Transactions on Automatic Control 19 (6)*. 716–723.

BALDWIN, T., BANNARD, C., TANAKA, T., AND WIDDOWS, D. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Sapporo, Japan, 89–96.

BALDWIN, T. AND KIM, S. N. 2009. Multiword expressions. In *Handbook of Natural Language Processing* 2nd Ed., N. Indurkhya and F. J. Damerau, Eds. CRC Press, Boca Raton, USA.

BANERJEE, S. AND PEDERSEN, T. 2003. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico, 370–381.

BLEI, D. AND LAFFERTY, J. 2009. Topic models. In *Text Mining: Classification, Clustering, and Applications*, A. Srivastava and M. Sahami, Eds. Chapman & Hall/CRC.

BLEI, D. M., NG, A. Y., AND JORDAN, M. I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993–1022.

BRODY, S. AND LAPATA, M. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece, 103–111.

CHANG, J., BOYD-GRABER, J., GERRISH, S., WANG, C., AND BLEI, D. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of Twenty-Fourth Annual Conference on Neural Information Processing Systems (NIPS 2009)*. 288–296.

CHOUEKA, Y. 1988. Looking for needles in a haystack, or locating interesting collocational expressions in large textual databases. In *Proceedings of Recherche d'Informations Assistee par Ordinateur 1988 (RIAO'88)*. Cambridge, USA, 609–623.

FINKEL, J. R., GRENAGER, T., AND MANNING, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*. Ann Arbor, USA, 363–370.

GRIFFITHS, T. L., STEYVERS, M., AND TENENBAUM, J. B. 2007. Topics in semantic representation. *Psychological Review 114,* 2, 211–244.

HAGHIGHI, A. AND VANDERWENDE, L. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2009)*. Boulder, USA, 362–370.

HERMANN, K. M., BLUNSOM, P., AND PULMAN, S. 2012. An unsupervised ranking model for noun-noun compositionality. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics*. Montréal, Canada, 132–141.

HU, W., SHIMIZU, N., NAKAGAWA, H., AND SHENG, H. 2008. Modeling chinese documents with topical word-character models. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK, 345–352.

JOACHIMS, T. 1998. Text categorization with suport vector machines: Learning with many relevant features. In *Proceedings of the 10th European Conference on Machine Learning (ECML-1998)*. London, UK, 137–142.

JOACHIMS, T. 1999. Making large-scale support vector machine learning practical. In *Advances in kernel methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. MIT Press, Cambridge, USA.

JOHNSON, M. 2010. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In *Proceedings of the 48th Annual Meeting of the ACL (ACL 2010)*. Uppsala, Sweden, 1148–1157.

LAU, J., NEWMAN, D., KARIMI, S., AND BALDWIN, T. 2010. Best topic word selection for topic labelling. In *Coling 2010: Posters*. Beijing, China, 605–613.

LAU, J. H., COOK, P., MCCARTHY, D., NEWMAN, D., AND BALDWIN, T. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the EACL (EACL 2012)*. Avignon, France, 591–601.

MCCALLUM, A. 1999. Multi-label text classification with a mixture model trained by EM. In *Proceedings of the AAAI'99 Workshop on Text Learning*. Orland, USA.

MIMNO, D. M., WALLACH, H. M., TALLEY, E. M., LEENDERS, M., AND MCCALLUM, A. 2011. Optimizing semantic coherence in topic models. In *EMNLP*. 262–272.

MINNEN, G., CARROLL, J., AND PEARCE, D. 2001. Applied morphological processing of English. *Natural Language Engineering 7,* 3, 207–223.

NEWMAN, D., KARIMI, S., AND CAVEDON, L. 2009. External evaluation of topic models. In *Proceedings of the Fourteenth Australasian Document Computing Symposium (ADCS 2009)*. Sydney, Australia, 11–18.

NEWMAN, D., LAU, J. H., GRIESER, K., AND BALDWIN, T. 2010. Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*. Los Angeles, USA, 100–108.

NIGAM, K., MCCALLUM, A. THRUN, S., AND MITCHELL, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning 39*, 103–134.

PECINA, P. 2009. Lexical association measures: Collocation extraction. Ph.D. thesis, Charles University.

REDDY, S., MCCARTHY, D., AND MANANDHAR, S. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand, 210–218.

TITOV, I. AND MCDONALD, R. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International World Wide Web Conference (WWW-2008)*. Beijing, China, 111–120.

WALLACH, H. M. 2006. Topic modeling: beyond bag-of-words. In *ICML '06: Proceedings of the 23rd International Conference on Machine Learning*. Pittsburgh, USA, 977–984.

WANG, X., MCCALLUM, A., AND WEI, X. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *ICDM '07: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. Omaha, USA, 697–702.

WEI, X. AND CROFT, W. B. 2006. LDA-based document models for ad-hoc retrieval. In *Proceedings of 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006)*. Seattle, USA, 178–185.