Evaluating a Topic Modelling Approach to Measuring Corpus Similarity

Richard Fothergill,[♢] **Paul Cook**[♣] **and Timothy Baldwin**[♢]
♢ Department of Computing and Information Systems
The University of Melbourne
Victoria 3010, Australia
♣ Faculty of Computer Science
University of New Brunswick
Fredericton, NB E3B 5A3, Canada
rjfo@unimelb.edu.au, paul.cook@unb.ca, tb@ldwin.net

Abstract

Web corpora are often constructed automatically, and their contents are therefore often not well understood. One technique for assessing the composition of such a web corpus is to empirically measure its similarity to a reference corpus whose composition is known. In this paper we evaluate a number of measures of corpus similarity, including a method based on topic modelling which has not been previously evaluated for this task. To evaluate these methods we use known-similarity corpora that have been previously used for this purpose, as well as a number of newly-constructed known-similarity corpora targeting differences in genre, topic, time, and region. Our findings indicate that, overall, the topic modelling approach did not improve on a chi-square method that had previously been found to work well for measuring corpus similarity.

Keywords: Analyzing corpora, corpus similarity, topic modelling

1. Introduction

In constructing traditional corpora, such as the British National Corpus (Burnard, 2000, BNC), documents are chosen based on particular selection criteria such as domain, genre, and time period. The composition of such a corpus in terms of these factors is therefore understood. On the other hand, many web corpora are constructed automatically on the basis of web crawls (Ferraresi et al., 2008), or the results of search engine queries (Baroni and Bernardini, 2004). We therefore don't have the same understanding of the composition of the corpora that are built, and the size of such corpora generally precludes manual analysis of their composition. This provided the underlying motivation for this paper: can we develop automatic analytic methods to help gain a better understanding of the composition of automatically-constructed corpora, such as web corpora?

One way to analyze a corpus is to measure the extent to which it is similar to other corpora, in particular reference corpora which we do know the composition of. Unfortunately, other than Kilgarriff (2001), there has been very little work to-date on this topic. Kilgarriff (2001) analyzed a number of empirical methods for measuring corpus similarity and found a method based on the chi-square statistic ("χ^2") to perform best. In this paper we consider an alternative approach to measuring corpus similarity based on topic modelling (Blei et al., 2003).

Kilgarriff (2001) cautioned that corpus similarity is complex, noting that two corpora can be similar in some ways and different in others, and that a single measure of corpus similarity is therefore limited. However, because of the lack of prior work in this area, Kilgarriff (2001) argued a measure of corpus similarity to be a useful starting point. In this paper we further explore the types of difference between corpora that a variety of corpus similarity measures are able to detect.

To evaluate methods for measuring corpus similarity, Kilgarriff (2001) constructed corpora with known similarity from the BNC. Here we too consider known-similarity corpora from the BNC. We additionally construct corpora that are known to differ specifically with respect to genre, topic, time, and region, to examine the extent to which measures of corpus similarity can detect these types of differences between corpora. We further consider known-similarity corpora that are much larger than those used by Kilgarriff (2001) to consider the effect of corpus size on measures of corpus similarity.

Our findings are somewhat surprising. Although topic modelling has been successfully applied to a wide range of NLP tasks (Brody and Lapata, 2009; Haghighi and Vanderwende, 2009; Hardisty et al., 2010; Lau et al., 2014), we find that, overall, our topic modelling-based approach to measuring corpus similarity is not an improvement over the χ^2 method of Kilgarriff (2001). The current best approach to measuring corpus similarity thus remains χ^2.

2. Related Work

2.1. Comparing Corpora

Corpora can be compared in a variety of ways. In perhaps the simplest case, the most frequent words, possibly restricted to a particular part-of-speech, can be compared for two corpora (Schäfer and Bildhauer, 2013). Lists of keywords — i.e., words that are marked with respect to frequency in one corpus compared to another — computed through any of a variety of methods such as ratio of relative frequency (Kilgarriff, 2009) or the χ^2 statistic, can also be compared. These approaches, however, give an impressionistic view of corpus similarity, as opposed to a quantitative measure, which is the focus of this work.

Kilgarriff (2001) considered a number of measures of corpus similarity based on the χ^2 statistic, Spearman rank correlation co-efficient, and perplexity of language models. He found the χ^2 method to perform best. In this method, the χ^2 statistic is calculated for the N most frequent words in the union of two corpora; this statistic is then taken as the similarity between those corpora. Kilgarriff (2001) found $N = 500$ to work well. This method is attractive in that it is inexpensive to compute and is based on only the words in a corpus (i.e., it does not require any processing such as part-of-speech tagging which could influence a corpus similarity measure).

Lippincott et al. (2010) use a topic modelling approach to measure variation in biomedical subdomains. They use latent Dirichlet allocation (Blei et al., 2003) to build a topic model for a corpus of biomedical articles. For each article in the corpus, this gives a distribution over topics. Articles in their corpus are associated with subdomains. A topic distribution for each subdomain is produced by combining the distributions for the documents (weighted by document size in tokens) in a given subdomain. The similarity between two subdomains is then measured as the Jensen-Shannon divergence between their topic distributions. Lippincott et al. (2010) used this approach to explore the differences between subdomains, but did not evaluate it as a (sub-)corpus similarity measure. In this paper we implement and evaluate a similar topic modelling approach to measuring corpus similarity, although we do not specifically consider biomedical subdomains.

2.2. Known Similarity Corpora

Kilgarriff (2001) constructs known-similarity corpora ("KSC") from two known-different source corpora, A and B, by mixing different ratios of the source corpora. A KSC collection is made up of N corpora each containing $N-1$ partitions, with each partition drawn from one of the two source corpora. For the first corpus in the KSC collection all partitions are drawn from A. For the second corpus, one partition is drawn from B and the remaining $N-2$ partitions are drawn from A. Thus the third corpus is a $2 : N-3$ mixture, and so on to the last corpus which is drawn entirely from B.

Although the true similarities between KSC are not in fact known, we can still assume that certain inter-KSC similarities will be greater than others. For example, the similarity between the second and third corpora must be greater than the similarity between the first and fourth corpora, by virtue of the fact that the first corpus contains less of A than the second corpus, and the fourth corpus contains more of A than the third corpus. Wherever the interval of one corpus pair contains the interval of another, we say that the inner pair must have a higher similarity. This gives a gold-standard partial order on the similarities between corpora in a KSC collection. A similarity measure is evaluated on how many of these gold standard similarity comparisons it reproduces.

3. Method

We implemented a selection of corpus similarity measures based on n-gram language models and topic models of the corpora, for comparison with the benchmark χ^2 similarity set by Kilgarriff (2001). To evaluate our selection of corpus similarity measures, we assembled a suite of KSC collections, including KSC collections provided by Kilgarriff (2001), and new KSC collections constructed using the same method.

3.1. Similarity Measures

χ^2 similarity is a statistic that compares the corpus frequencies of words directly. Kilgarriff (2001) justifies this choice on the grounds that "reliable statistics depend on features that are reliably countable". In basing our additional similarity measures on generative language models, we too have a foundation in reliably countable phenomena, but aim to better capture syntactic and semantic differences between corpora. In this section we detail the similarity measures we studied and their metaparameters.

3.1.1. χ^2 Similarity

We implemented χ^2 similarity as defined in Kilgarriff (2001), however we varied the cap on the lexicon size to the top N words for $N \in \{200, 500, 1000, 2000, 4000\}$. We also tested χ^2 similarity with an uncapped lexicon (that is, all word types in the corpus contribute to the statistic).

We did not discard words outside the top N completely. Instead, we counted them all as tokens of a single wordform *__OTHER__*. This ensures the χ^2 similarity is calculated as a sum across the contingency table of an entire event space.

3.1.2. Perplexity Similarity

We implemented perplexity similarity using the SRILM language modelling toolkit (Stolcke et al., 2011). To calculate the similarity between two corpora A and B, our perplexity similarity measure first builds an n-gram language model of each: M_{A} and M_{B} respectively. The final similarity is:

$$-\frac{P(\mathrm{B}, M_{\mathrm{A}}) + P(\mathrm{A}, M_{\mathrm{B}})}{2}$$

where $P(C, M)$ is the perplexity of model M with respect to corpus C. The score is negated because high perplexity is indicative of difference, not similarity.

Note that the perplexity similarity measure implemented by Kilgarriff (2001) had a much more complicated algorithm, for the sake of symmetry with the paired n-fold cross-validation based homogeneity measure he used. We do not require a measure of corpus homogeneity for the known-similarity corpora we consider here.

Rather than just use trigram language models as Kilgarriff (2001) did, we tested the perplexity similarity measure using n-gram models for $n \in \{1, 2, 3, 4, 5\}$. We applied SRILM in its default configuration which produces models with Good-Turing discounted estimates and uses the Katz backoff method (Stolcke, 2002).

3.1.3. Topic Similarity

Our final measure, topic similarity, combines the documents in the two corpora to be compared, and builds a topic model of the complete set of contained documents. It then builds a vector representation of each corpus and compares the resulting vectors to derive the similarity between the corpora.

The vector representation $\vec{a}$ of corpus A has a dimension for each topic in the topic model. The value of $\vec{a}_i$ is the number of tokens in A assigned topic i by the topic model. We used three vector similarity measures to compare corpus topic vectors.

1. Euclidean similarity

$$-\|(\vec{a} - \vec{b})\|$$

The distance is negated to give more similar vectors a "greater" similarity.

2. Cosine similarity

$$\frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$

3. Jensen-Shannon similarity

$$1 - D_{\text{JS}}(\frac{\vec{a}}{\|\vec{a}\|_1}, \frac{\vec{b}}{\|\vec{b}\|_1})$$

where $\|\cdot\|_1$ is the ℓ_1 norm (sum of absolute values) and D_{JS} is the Jensen-Shannon divergence:

$$D_{\text{JS}}(\vec{a}, \vec{b}) = \frac{D_{\text{KL}}(\vec{a}, \vec{m}) + D_{\text{KL}}(\vec{b}, \vec{m})}{2}$$

where $m = (\vec{a}+\vec{b})/2$ and D_{KL} is the Kullback-Leibler divergence, or relative entropy:

$$D_{\text{KL}}(\vec{a}, \vec{b}) = \sum_i \vec{a}_i \log \frac{\vec{a}_i}{\vec{b}_i}$$

We subtract D_{JS} from 1 to make the measure a positive value that increases with similarity.

In our experiments, we used topic models with T topics, for $T \in \{10, 50, 100, 500, 1000\}$.

3.2. KSC Construction

Here, we give a short description of our implementation of the known-similarity corpus construction method. A full treatment of the method for constructing KSC can be found in Kilgarriff (2001).

When constructing KSC from source corpora A and B, we construct $N = 11$ KSC in all cases,[1] meaning that the percentages of B in individual KSC are exactly $0\%, 10\%, 20\%, ..., 100\%$ (and similarly the percentages of A are $100\%, 90\%, 80\%, ..., 0\%$).

We split the source corpora at the token level, assigning the same number of tokens to each KSC. However, we do preserve sentence and document boundaries for the purpose of the topic similarity measure, introducing artificial boundaries when splits occur mid-sentence or mid-document.

Except where otherwise stated, text is assigned to KSC in contiguous chunks from the source corpora in a size appropriate to the position of the KSC in the KSC set. For example, if $60k$ words are assigned from corpus A to KSC$_4$ then the next $50k$ words of A will be assigned to KSC$_5$. When a source corpus consists of multiple files, the order is determined by the lexicographical sort order of the source file names.

KSC set	Number of corpora in set
acc_gua	10
art_gua	11
bmj_gua	9
env_gua	9
gua_tod	11

Table 1: KSC based on the BNC from Kilgarriff (2001).

3.2.1. KILGARRIFF KSC

We evaluated each of our measures on a subset of the KSC used in Kilgarriff (2001), referred to as KILGARRIFF, comprising the text type pairs indicated in Table 1. Within each KSC set, the number of words in the corpora varies between $111k$ and $114k$. The three letter codes refer to subsets of the BNC, and are described in Kilgarriff (2001).

3.2.2. WDC KSC

The WeSearch Data Collection ("WDC") is a collection of user-generated text designed to capture differences in both subject matter and writing style (Read et al., 2012). It contains text on the separate topics of NLP and the Linux operating system taken from blogs, Wikipedia, software reviews and forums.

Using Linux as a fixed topic and varying the writing style by varying the source through blogs, reviews and forums, we created three KSC with differences in genre. Then, using forums as a fixed source, we created additional KSC by mixing the topics of NLP and Linux. Table 2 shows details of each WDC KSC we constructed.

3.2.3. GIGAWORD Corpus KSC

The Gigaword corpus (Parker et al., 2009, "GIGAWORD") is a collection of date-stamped newswire text. We used the L.A. Times/Washington Post ("ltw") subset and the New York Times ("nyt") subset to create large KSC sets with regional differences for the same time period and to compare time differences for the same region.

Details of the GIGAWORD KSC we created can be found in Table 3. nyt_jun consists of texts from the nyt subset for June 2005 and 2006. This time-differentiated KSC consists of corpora that are roughly an order of magnitude larger than those used by Kilgarriff (2001). nyt_5678 is similar, but consists of texts from May–August, and as such provides even larger corpora. ltw_nyt consists of texts from the ltw and nyt subsets for June 2006, while ltw_nyt_long consists of texts from the same subsets for May–July 2006. These corpora allow us to compare region-differentiated corpora at two different sizes, both of which are again much larger than those used by Kilgarriff (2001).

Our standard implementation of the KSC construction method takes samples from the source corpora from the start of the dataset. If one of the source corpora is larger, it is effectively truncated by this selection policy. Since GIGAWORD source files are sorted chronologically and the NYT portion is larger than the LTW portion, a shorter timespan would be selected from the NYT portion for the KSC. To alleviate this, for GIGAWORD we alter the sampling method slightly: after each KSC has received its allo-

[1] This does not apply to the KSC sets based on the BNC provided by Kilgarriff (2001), where the number of KSC varies between 9 and 11 (as shown in Table 1).

KSC set	Source A	Source B	Words per corpus	Difference
wlb_wlr	Blogs	Reviews	48250	genre
wlb_wlf	Blogs	Forums	147740	genre
wlr_wlf	Reviews	Forums	48250	genre
wnb_wlb	NLP	Linux	124930	topic

Table 2: KSC constructed from the WDC. Each KSC set contains 11 corpora.

KSC set	Source A	Source B	Words per corpus	Difference
nyt_jun	NYT 2005 Jun	NYT 2006 Jun	1247550	time
nyt_5678	NYT 2005 May–Aug	NYT 2006 May–Aug	4897340	time
ltw_nyt	LTW 2006 Jun	NYT 2006 Jun	443630	region
ltw_nyt_long	LTW 2006 May–Jul	NYT 2006 May–Jul	1338460	region

Table 3: KSC constructed from the GIGAWORD corpus. Each KSC set contains 11 corpora.

cation from a source corpus, we skip forward in that corpus to a position proportional to the amount taken so far. This ensures that the final samples come from near the end of the time range.

Note that although our method ensures that mixed LTW/ NYT KSC contain text drawn from aligned timespans, it will still be the case that the earlier KSC in the set come from earlier time periods than later KSC in the set. This means that our location-differentiated KSC are also somewhat time differentiated. We mitigate this by limiting the total timespan from which samples are drawn for location-differentiated KSC to three months, whereas time-differentiated KSC are separated by one year.

4. Results

To evaluate the methods for measuring corpus similarity, we apply each method to each pair of corpora in each KSC set. We then calculate the accuracy for a method on a KSC set as the proportion of correct corpus similarity judgements — according to the gold-standard known corpus similarities — for that set.

The average accuracy for each method on each KSC set is shown in Table 4. The χ^2 methods for all values of n, except for the lowest value of $n = 200$, outperform all other methods. The best accuracy of 93.9% is for $n = 4000$, and is substantially higher than the accuracy for $n = 500$, the parameterization suggested by Kilgarriff (2001). The best topic modelling approach ($T = 1000$, JS) achieves 91.4% accuracy. For the different vector similarity measures for the topic modelling methods, Jensen-Shannon divergence scores higher than Euclidean distance or Cosine similarity for many values of t. When using Jensen-Shannon divergence we further see that larger values of t give higher accuracy. The perplexity similarity approaches are remarkably poor, achieving accuracies lower than those of all other methods considered.

These initial results indicate that the proposed topic modelling approach to measuring corpus similarity is not an improvement over χ^2, which remains the best method overall for this task. A further advantage of the χ^2 method is that it can be computed relatively quickly compared to training a topic model or a language model. The relatively strong performance of χ^2— which relies on the (square of the absolute value of the) differences between the observed and expected frequencies of words — compared to the other approaches suggests that high frequency words are very informative of corpus differences.

Similarity measure		Accuracy
χ^2	$n = 4000$	0.939
χ^2	$n = \infty$	0.937
χ^2	$n = 2000$	0.933
χ^2	$n = 1000$	0.930
χ^2	$n = 500$	0.918
Topic	$T = 1000$, JS	0.914
Topic	$T = 500$, JS	0.910
χ^2	$n = 200$	0.907
Topic	$T = 100$, JS	0.902
Topic	$T = 50$, JS	0.892
Topic	$T = 100$, Euclidean	0.881
Topic	$T = 10$, JS	0.880
Topic	$T = 50$, Euclidean	0.879
Topic	$T = 10$, Euclidean	0.872
Topic	$T = 50$, Cosine	0.872
Topic	$T = 10$, Cosine	0.869
Topic	$T = 500$, Euclidean	0.864
Topic	$T = 100$, Cosine	0.863
Topic	$T = 1000$, Euclidean	0.863
Topic	$T = 1000$, Cosine	0.842
Topic	$T = 500$, Cosine	0.840
Perplexity	$n = 3$	0.832
Perplexity	$n = 4$	0.832
Perplexity	$n = 5$	0.832
Perplexity	$n = 2$	0.828
Perplexity	$n = 1$	0.499

Table 4: The accuracy of each similarity measure averaged over all KSC sets.

Table 5 shows the top-5 methods for each group of KSC sets — KILGARRIFF, WDC, and GIGAWORD. Here we see that the χ^2 method also performs best, on average, for each group, although for the GIGAWORD KSC the best accuracy is obtained when the vocabulary is not restricted (i.e., $n = \infty$). Furthermore, although the perplexity methods performed relatively poorly overall (i.e., averaged over all KSC sets, as shown in Table 4) they do perform well on the GIGAWORD KSC.

Dataset	KSC set	Similarity measure		Accuracy
KILGARRIFF	acc_gua	χ^2	$n = \infty$	0.971
KILGARRIFF	art_gua	χ^2	$n = 4000$	0.983
KILGARRIFF	bmj_gua	χ^2	$n = 4000$	0.980
KILGARRIFF	bmj_gua	χ^2	$n = 2000$	0.980
KILGARRIFF	bmj_gua	χ^2	$n = 500$	0.980
KILGARRIFF	bmj_gua	χ^2	$n = 200$	0.980
KILGARRIFF	env_gua	χ^2	$n = 1000$	0.997
KILGARRIFF	gua_tod	χ^2	$n = 4000$	0.971
KILGARRIFF	gua_tod	χ^2	$n = 2000$	0.971
KILGARRIFF	gua_tod	Topic	$T = 500$, Euclidean	0.971
GIGAWORD	ltw_nyt	χ^2	$n = 2000$	0.986
GIGAWORD	ltw_nyt_long	χ^2	$n = 200$	0.998
GIGAWORD	ltw_nyt_5678	Perplexity	$n = 3$	0.909
GIGAWORD	nyt_jun	Topic	$T = 1000$, JS	0.986
WDC	wlb_wlf	χ^2	$n = 2000$	0.945
WDC	wlb_wlr	χ^2	$n = 2000$	0.880
WDC	wlr_wlf	Topic	$T = 500$, JS	0.952
WDC	wnb_wlb	Topic	$T = 1000$, JS	0.976

Table 6: The best-performing method, and corresponding accuracy, on each individual KSC set. In cases where multiple methods tied for the best accuracy on a KSC set, all of these methods are shown.

KILGARRIFF

Similarity measure		Accuracy
χ^2	$n = 4000$	0.964
χ^2	$n = 2000$	0.962
χ^2	$n = 1000$	0.962
χ^2	$n = 500$	0.957
χ^2	$n = \infty$	0.954

WDC

Method		Accuracy
χ^2	$n = 4000$	0.927
χ^2	$n = 2000$	0.925
χ^2	$n = 1000$	0.921
χ^2	$n = 500$	0.921
χ^2	$n = \infty$	0.917

GIGAWORD

Method		Accuracy
χ^2	$n = \infty$	0.936
χ^2	$n = 4000$	0.920
Perplexity	$n = 3$	0.907
Perplexity	$n = 4$	0.907
Perplexity	$n = 5$	0.907

Table 5: The top-5 similarity measures, and their average accuracies, for each of the KILGARRIFF, WDC, and GIGAWORD KSC sets.

We now examine the best-performing methods for each individual KSC set. Results are shown in Table 6. For each KILGARRIFF KSC set, a χ^2 method gives the best results (or is tied for the best), although there is some variation as to which specific value of n gives the best accuracy.

Turning to the GIGAWORD KSC, for ltw_nyt and ltw_nyt_long — the region-differentiated KSC — the best results are again obtained with a χ^2 method. Here the best results for ltw_nyt and ltw_nyt_long are obtained with n set to 2000 and 200, respectively, suggesting that the best parameterization of the χ^2 method might depend on corpus size.

For the time-differentiated GIGAWORD KSC, however, we see a different pattern. For nyt_5678 the best results are achieved with perplexity ($n = 3$), a method that performs relatively poorly for many other KSC sets. For nyt_june, topic modelling ($T = 1000$, JS) gives the best results. In each case the accuracy for the best χ^2 method (not shown in Table 6) is 2–3 percentage points lower than that of the best method. These findings suggest that the χ^2 method might not be as well-suited to identifying similarities between corpora from different time periods, but that are otherwise comparable. In future work we plan to construct additional time-differentiated corpora from other sources (such as social media) to investigate this further.

For the topic-differentiated WDC KSC (i.e., wnb_wlb), topic modelling ($T = 1000$, JS) gives the best results with an accuracy of 97.6%, although χ^2 ($n = 4000$) is close behind at 97.3%. That topic modelling does well at identifying differences in topic is perhaps not so surprising, and this could be seen as consistent with the good performance of topic modelling on the nyt_june KSC, where news articles from different time periods would be expected to be on somewhat different topics. (However, this does not provide an account for why perplexity does so well for nyt_5678 where we would also expect differences in topic.)

Overall our topic modelling approach is most competitive on KSC with differences in subject matter: NLP vs Linux or news from different years. For the KSC with other differentiating features it is possible that training topic models on the union of the two corpora risks representing the commonalities well and losing the less common unshared phenomena. An alternate approach would be to train topic models separately on either side of a corpus comparison. As with the n-gram perplexity method, similarity could then be measured using the perplexity of one corpus relative to the model of the other. Topic model perplexity can be computed using methods developed for evaluating topic models on held out documents such as the Chib-style estimator or left-to-right method of Wallach et al. (2009).

The genre-differentiated WDC KSC don't show a clear pattern. The χ^2 method gives the best accuracy for wlb_wlf and wlb_wlr ($n = 2000$ in each case), while topic modelling ($T = 500$, JS) is best for wlr_wlf.

5. Conclusions

We evaluated a number of approaches to measuring corpus similarity, including an approach based on topic modelling that had not been previously evaluated for this task. The evaluation was carried out using known-similarity corpora based on the BNC from an earlier corpus similarity study (Kilgarriff, 2001), as well as newly-constructed known-similarity corpora specifically targeting differences in genre, topic, time, and region. Overall, the topic modelling method did not perform better than the χ^2 approach that Kilgarriff (2001) had previously found to perform best, although there was some variation for certain known-similarity corpora, particularly those differentiated by time and topic.

In future work we intend to explore further approaches to measuring corpus similarity. The method based on the perplexity of n-gram language models performed relatively poorly. There have, however, been recent advances in language modelling through neural network-based approaches (Mikolov et al., 2010; Pennington et al., 2014, for example). Such methods could lead to improved language modelling approaches to measuring corpus similarity. More-recent approaches to topic modelling are tailored specifically to learning topics from multiple corpora (Wang et al., 2009; Buntine and Mishra, 2014, for example) and therefore might be particularly well-suited to measuring corpus similarity. In future work we also intend to consider such topic models.

Finally, although the topic modelling approach we considered here did not perform as well as the χ^2 method, topic modelling could nevertheless still be a useful tool for comparing corpora. Potential pitfalls of our topic model approach might be avoided by adapting the n-gram perplexity method to topic model perplexity using the probability estimation techniques of Wallach et al. (2009). Alternatively, Kilgarriff (2012) presents a method for "getting to know your corpus" in which he suggests manually clustering the top-100 keywords for a focus corpus with respect to a reference corpus, to determine the major differences between the corpora in terms of, for example, topic, formality, and language variety. By computing keyness for topics, as opposed to words, and then examining the highest probability words for those key-topics, it might be possible to produce a similar summary of the differences between corpora with less manual intervention. We plan to explore this possibility in future work.

6. Acknowledgements

This research was supported in part by a Learning and Teaching Initiative grant from The University of Melbourne.

7. Bibliographical References

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*.

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 EACL (EACL 2009)*, pages 103–111, Athens, Greece.

Buntine, W. and Mishra, S. (2014). Experiments with non-parametric topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'14)*, pages 881–890, New York, USA.

Burnard, L. (2000). *User Reference Guide for the British National Corpus*. Technical report, Oxford University Computing Services.

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54, Marrakech, Morocco.

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies 2009 (NAACL HLT 2009)*, pages 362–370, Boulder, USA.

Hardisty, E., Boyd-Graber, J., and Resnik, P. (2010). Modeling perspective using adaptor grammars. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 284–292, Cambridge, USA.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, 6(1):97–133.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference*, Liverpool, UK.

Kilgarriff, A. (2012). Getting to know your corpus. In *Proceedings of the 15th International Conference on Text, Speech and Dialogue (TSD 2012)*, pages 3–15, Brno, Czech Republic.

Lau, J. H., Cook, P., McCarthy, D., Gella, S., and Baldwin, T. (2014). Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014)*, pages 259–270, Baltimore, USA.

Lippincott, T., Ó Séaghdha, D., Sun, L., and Korhonen, A. (2010). Exploring variation across biomedical subdomains. In *Proceedings of the 23rd International Confer-*

ence on Computational Linguistics (Coling 2010), pages 689–697, Beijing, China.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, pages 1045–1048, Makuhari, Japan.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2009). *English Gigaword Fourth Edition*. Linguistic Data Consortium, Philadelphia, USA.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1532–1543, Doha, Qatar.

Read, J., Flickinger, D., Dridan, R., Oepen, S., and Øvrelid, L. (2012). The WeSearch corpus, treebank, and treecache. a comprehensive sample of user-generated content. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation*, Istanbul, Turkey.

Schäfer, R. and Bildhauer, F. (2013). *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, San Francisco, USA.

Stolcke, A., Zheng, J., Wang, W., and Abrash, V. (2011). SRILM at sixteen: Update and outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, page 5, Waikoloa, USA.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *7th International Conference on Spoken Language Processing*, Denver, USA.

Wallach, H., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112, Montreal, Canada.

Wang, C., Thiesson, B., Meek, C., and Blei, D. (2009). Markov topic models. In *Proceedings of Artificial Intelligence and Statistics*, Clearwater Beach, USA.