A Topic-Based Approach to Multiple Corpus Comparison

Jinghui Lu, Maeve Henchion, Brian Mac Namee

Insight Centre for Data Analytics, University College Dublin, Ireland
Tegasc the Agriculture and Food Development Authority, Dublin, Ireland
Jinghui.Lu@ucdconnect.ie, Maeve.Henchion@teagasc.ie,
Brian.MacNamee@ucd.ie

Abstract. *Corpus comparison techniques* are often used to compare different types of online media, for example social media posts and news articles. Most corpus comparison algorithms operate at a word-level and results are shown as lists of individual discriminating words which makes identifying larger underlying differences between corpora challenging. Most corpus comparison techniques also work on pairs of corpora and do need easily extend to multiple corpora. To counter these issues, we introduce *Multi-corpus Topic-based Corpus Comparison* (MTCC) a corpus comparison approach that works at a topic level and that can compare multiple corpora at once. Experiments on multiple real-world datasets are carried demonstrate the effectiveness of MTCC and compare the usefulness of different statistical discrimination metrics - the χ^2 and Jensen-Shannon Divergence metrics are shown to work well. Finally we demonstrate the usefulness of reporting corpus comparison results via topics rather than individual words. Overall we show that the topic-level MTCC approach can capture the difference between multiple corpora, and show the results in a more meaningful and interpretable way than approaches that operate at a word-level.

Keywords: Corpus Comparison, Topic Modelling, Jensen-shannon Divergence

1 Introduction

Many *corpus comparison* techniques are proposed in the literature to reveal the divergence between corpora [8, 16], especially corpora of web-based content such as online news, social media posts, and blog posts [5, 10]. Although these approaches have been shown to be effective, almost all of them are limited by comparing corpus at a word level. Consequently, the results are communicated to a user as a list of unrelated words that are divergent across two corpora, which makes identifying larger underlying differences between the corpora challenging. Additionally, these studies focus on comparing pairs of corpora instead of multiple corpora.

In recent years, *topic modelling* [2] has become a widely used method for revealing thematic information, which is described by a series of high related words called *topic descriptors*, in a collection of documents. We assume the combination of topic modelling techniques and corpus comparison methods has the potential to eliminate the problem described above that arise with word-based corpus comparison approaches. In

other words, a approach uses topic modelling as a basis for corpus comparison to offer users the divergence that can be explained at a topic-level rather than an individual word-level.

This paper describes the *Multi-corpus Topic-based Corpus Comparison* (MTCC) approach to corpus comparison, that leverages topic modelling and statistical discrimination metrics to conduct a topic-based comparison. We describe the approach and demonstrate it on 8 real-world datasets. In a series of experiments we demonstrate that the topics extracted by our models contain divergence information, as well as comparing the effectiveness of different statistical discrimination metrics applied in the algorithm. We also compare the output of MTCC with a word-based corpus comparison method to show that the results output by MTCC are more meaningful and more interpretable than those produced by the word-based method. The contributions of this paper area:

- Multi-corpus topic-based Corpus Comparison, a new corpus comparison technique that leverages topic modelling and statistical discrimination metrics.
- An experiment to show that the topics extracted by the statistical discrimination metrics are capturing divergence.
- An experiment that investigates the effectiveness of different statistical discrimination metrics applied in MTCC.
- A demonstration of the usefulness of topic-based divergence explanations over multiple real-world datasets.

The rest of this paper is organized as follows: Section 2 presents related work; Section 3 describes the MTCC approach; Section 4 describes experiments to measure the divergent information contained in topics found and the effectiveness of different statistical divergence metrics; Section 5 compares the output of the MTCC approach with a word-based method; and, finally, Section 6 summarizes the work and suggests future directions.

2 Related Work

Corpus comparison approaches extract the distinct content from two corpora [16]. Typically, researchers attempted to compute the contribution to the divergence of individual words by applying many statistical discriminating metrics over word frequencies calculated from different corpora, and those words which highly contribute to the difference are selected to be presented as the divergence [5, 8, 16]. Therefore, it is very intuitive to show the comparison results using words.

There is a variety of statistical discrimination metrics in the literature. For instance, Leech and Fallon [11] use χ^2 to measure the discriminative power of a word; log-likelihood [16] and relevance-frequency (RF) [9] were utilized to extract divergent information over corpora of different domains. Besides, Information Gain (IG), Gain Ratio (GR) [17], Kullback-Leibler divergence [4], and Jensen-Shannon divergence (JSD) [5, 13] have been widely employed in corpus comparison.

Latent Dirichlet Allocation[2], as a widely used strategy for exploiting topic information from texts, can automatically infer the distribution of membership to set of

topics in a large collection of documents. It has been shown to have a great ability to find latent topics and cluster documents [1, 6].

As far as we know, Zhao et al. [19] were the first to use topic modelling, specifically LDA, in conjunction with statistical discrimination to find topics specific to a corpus in a pair of corpora being compared, and to use these to explain the differences between the corpora. Zhao et al. first performed independent topic modelling on the corpora being compared, and then applied Jensen-Shannon divergence (JSD) [12] over the topic descriptors to measure the pair-wise similarities between the sets of topics found in the two corpora. If the similarity of the nearest match to a topic in one corpus to a topic in the other corpus was below a specific threshold then that topic was said to be discriminatory. The set of discriminatory topics was then used to explain the differences between the two corpora. Zhao et al. demonstrated this approach by comparing corpora from Twitter and the New York Times. Similarly, Murdock et al. [14] and Sievert et al. [18] used JSD applied to the distributions of word frequencies in topics to measure the distance between topics, but this was not done in a corpus comparison scenario.

Zhao et al's method trains independent LDA model for each corpus which increases the instability of the whole system due to the non-deterministic nature of LDA. Besides, the output highly relies on the setting of thresholds, namely, the small thresholds tend to result in too many similar topics as divergence, but the large thresholds will lose some discriminating topics. The MTCC approach proposed in this paper differs from the work of Zhao et al in two key ways. First, in MTCC a global topic model is trained across a combined corpus that contains all documents from all corpora being compared. Second, because a single topic model is built, intuitively, JSD can be applied directly to topic membership vectors rather than applying JSD to word distributions between topics. As compared to matching similar topics, MTCC skips the empirical setting of thresholds to further automate the comparison process and, since one global topic model is trained, the topic proportion of each corpus can be inferred on which multiple corpus comparison can be based.

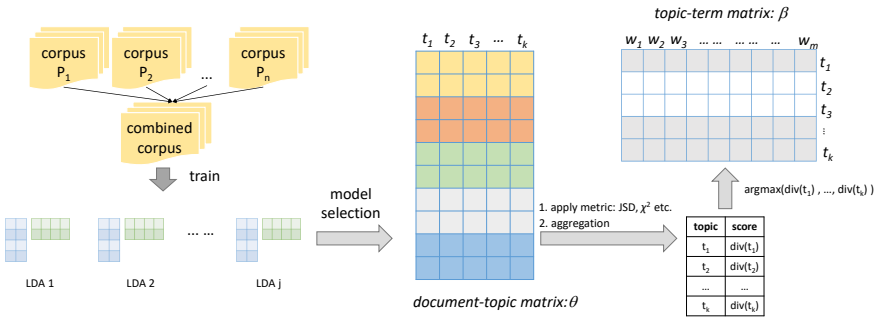


Fig. 1. An overview of MTCC approach. θ is a document-topic matrix where each row denotes a topical representation of a document. The vectors shaded by different colors denote the topical representation of documents from different corpora. β is a topic-term matrix where each row represents the word distribution of the corresponding topic. The rows shaded in grey in β imply the discriminating topics selected to be presented.

3 The Multi-corpus Topic-based Corpus Comparison (MTCC) Approach

In this section, we will first provide a brief description of the use of LDA for topic modelling, then describe the Multi-corpus Topic-based Corpus Comparison (MTCC) approach in detail.

3.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [2] is proposed to infer topic distribution in a collection of documents. The model generates a document-topic matrix θ, and a topic-term matrix β (see Figure 1). Specifically, each row of the document-topic matrix is a topic-based representation of a document where the ith entry determines the degree of association between the ith topic and the document. Each row of the topic-term matrix represents the word distribution of the corresponding topic. Usually, several most common words in one topic will be chosen to be presented as the *topic descriptors*. We use the LDA implemented in Python Gensim package [1].

Properly setting the number of latent topics to be found plays a vital role in the performance of LDA as well as other topic modelling algorithms. There are many approaches for seeking the appropriate choice of the number of latent topics, k, in the literature. O'Callaghan et al [15] proposed a topic coherence metric via word embeddings reflecting the semantic relatedness of topic descriptors, which has been previously used in [1] to determine the number of topics to be found. We also adopt this method, in the experiments, we use word embeddings built across our own test corpora using the FastText algorithm [3].[2]

3.2 Infer Topic Proportion for Corpus

Figure 1 shows the overview of applying the MTCC approach to comparing multiple corpora. All corpora are first pre-processed, including tokenisation, conversion to lower case, and lemmatisation using Python NLTK package [3].

At the outset, all corpora are combined into a single corpus. A set of topic models are trained over this combined corpus with different values of k which is the number of latent topics to be found. The topic coherence score for each topic model is calculated following [15]. The topic model which have the highest averaged topic coherence score is selected as the input for the next stage. The chosen topic model produces two matrices, $\theta \in \mathbb{R}_{+}^{d \times k}$ and $\beta \in \mathbb{R}_{+}^{k \times m}$, where m is the number of unique terms in vocabulary, k is the number of topics and d is the number of documents in the combined corpus.

Assuming the first p documents in θ are from the same text set P_i which is one of the original corpus composing the combined corpus. The membership of the tth topic to the individual corpus P_i, $mem_t(P_i)$, can be given by:

$$mem_t(P_i) = \frac{\sum_{i=1}^{p} \theta[i,t]}{p} \tag{1}$$

[1] https://radimrehurek.com/gensim/models/ldamodel.html

[2] https://radimrehurek.com/gensim/models/fasttext.html

[3] https://www.nltk.org/api/nltk.html

where $\theta[i,t]$ denotes the the proportion of tth topic in the ith documents in the combined corpus and p is the number of documents from corpus P_i.

3.3 Strategies for Comparing Multiple Corpora

Since most of the corpus comparison approaches listed in Section 2 is focused on comparing pairs of corpora, we adopt a one-versus-all strategy in order to conduct a multiple corpus comparison.

The algorithm is quite simple, a specific corpus P_i and a mixture corpus which is a concatenation of all corpora except corpus P_i are considered two corpora being compared. Hence, the proportion of any topic for corpus P_i and the mixture corpus can be given by Equation 1, on which the statistical metrics are applied. Then we can derive the divergence score of tth topic in terms of corpus P_i following [5, 9, 11, 17], which is denoted by $d_t(P_i)$ in this paper.

However, in order to rank the discrimination power of each topic, t, across the full set of corpora to be compared, a single divergence score per topic is required necessitating an aggregation strategy. We define three aggregation strategy which is given as follows:

- sum: $div_t(P_1 \,||\, P_2 \,||\, ... \,||\, P_n) = \sum_{i=1}^{n} d_t(P_i)$
- weighted sum: $div_t(P_1 \,||\, P_2 \,||\, ... \,||\, P_n) = \sum_{i=1}^{n} \pi(P_i) d_t(P_i)$
- maximum: $div_t(P_1 \,||\, P_2 \,||\, ... \,||\, P_n) = max_{i=1}^{n} d_t(P_i)$

where n is the number of corpora for comparison and $\pi(P_i)$ is the proportion of corpus P_i in the combined corpus, $div_t(P_1 \,||\, P_2 \,||\, ... \,||\, P_n)$ denotes the global divergence score of topic t.

There is a modification to the Jensen-Shannon divergence metric, *extended JSD* [13] that can be used directly across multiple corpora at a word-level without the need for the one-versus-all approach. In extended JSD divergence score for the tth word over multiple corpora $P_1, P_2, ..., P_n$ is defined as:

$$div_{JS,t}(P_1 \,||\, P_2 \,||\, ... \,||\, P_n) = -m_t \log m_t + \frac{1}{n} \sum_{i=1}^{n} p_{it} \log p_{it} \tag{2}$$

where p_{it} is the probability of seeing word t in corpus P_i, and m_t is the probability of seeing word t in M. Here, M is a mixed distribution of n corpora where $M = \frac{1}{n} \sum_{i=1}^{n} P_i$. In extended JSD, the global divergence score of the tth topic could be derived from Equation 2 by replacing p_{it} with $mem_t(P_i)$ calculated by Equation 1.

Subsequently, the topics can be ranked by their global divergence score in descending order and the top n topics along with their topic descriptors can be selected to represent the difference between multiple corpora.

A github repository containing the code to implement the MTCC approach and all experiments described in the following section is publicly available.[4]

[4] *https://github.com/GeorgeLuImmortal/topic-based_corpus_comparison*

4 Comparing Topic-based Discrimination Metrics

In this set of experiments we compare the ability of different statistical discrimination metrics to identify discriminative topics across corpora. Corpus comparison is an unsupervised procedure which makes this evaluation somewhat challenging. To overcome these challenges, following [17], we reframe the corpus comparison evaluation as a document classification task. This section describes the experimental approach and the datasets used, and discusses the results of these experiments.

4.1 Datasets

Our experiments are carried out on 8 real-world datasets described in [6]: 6 news article datasets *bbc*, *bbc-sport*, *guardian-2013*, *irishtimes-2013*, *nytimes-1999*, *nytimes-2003* and 2 Wikipedia datasets *wikipedia-high* and *wikipedia-low*. Each dataset has different sections, for example, bbc includes sections: *business, politics, entertainment, sports, tech.* Table 1 shows the total number of documents, the size of vocabulary, the total number of terms, and the number of sections and the value of k in each dataset. We divide each dataset into corpora following these sections.

Table 1. Summary statistics of the datasets used in the experiments. The rightmost column is the number of latent topics to be found for each dataset.

Dataset	No. Docs	Vocab. Size	No. Terms	No. Sections	Best k
bbc	2,225	3,125	484,600	5	150
bbc-sport	737	969	138,699	5	150
guardian-2013	5,414	10,801	2,349,911	6	300
irishtimes-2013	3,093	4,832	916,592	6	150
nytimes-1999	9,551	12,987	3,595,075	4	150
nytimes-2003	5,283	15,001	4,426,313	5	100
wikipedia-high	5,738	17,311	6,638,780	6	350
wikipedia-low	4,986	15,441	5,934,766	10	150

4.2 Measuring the Discriminativeness of Topics Found

Following the approach used in [17] we base our evaluation on the assumption that discriminative topics should be useful features for classifying documents as belonging to different corpora. Thus, we reframe the evaluation of corpus comparison to a document classification task.

The procedure for estimating discriminativeness of a set of topics found by MTCC is as follows:

1. extract the top n most informative topics (MTCC is run and the n most discriminative topics are extracted)
2. construct vector representations of documents based on the selected n topics.
3. build a classification model using the vector representation and measure its performance.

(a) search for a set of optimal hyper-parameters for the classifier using 10-fold cross-validation.
(b) run another 10 cross-validation shuffled by a different random seed using the obtained optimal hyper-parameters and report the result of the second 10 cross-validation based on micro-averaged f1 score [7].

The performance of the second 10 cross-validation is used to assess the discriminativeness of the n topics.

We repeat this procedure for values of $n \in [\frac{k}{10}, k]$, where k is the number of topics used within MTCC and n increases by $\frac{k}{10}$ every time. In other words, given a statistical discrimination metric and a dataset, we will run the above procedure 10 times. Also, we should note here, we rank the topics according to their divergence score from MTCC model in descending order which means the first n topics are the most distinctive topics whereas the remaining topics are not so informative. We also use a baseline, in which n topics are selected randomly.

4.3 Experimental Configuration

We compare the performance of MTCC models utilizing the different statistical discrimination metrics in Section 2, i.e. IG, GR, χ^2, RF, JSD and extended JSD (ext-JSD) across 8 real-world datasets. For each dataset, we treat a section as an individual corpus—for instance, the bbc dataset has 5 corpora. Therefore, this is a multiple corpora comparison task. After extensive preliminary experiments, we choose to use Linear-SVM [7] classifier when measuring performance.[5] Also, the maximum aggregation strategy (see Section 3.3) has been adopted as it was shown to perform better than the other strategies in preliminary experiments. For RF, we set the threshold to 0.01. The baseline method classifiers are trained with the document representations in terms of n topics chosen arbitrarily. To reduce the effect of randomness, for a given n, we run baseline methods 10 times with different random seeds and report the averaged micro-averaged f1 score.

We set the random seed for the LDA model to 1984; α = "$auto$" which indicates the automatically tuning the hyperparameter α;[6] and the random seed for twice cross-validation to 2018 and 0 respectively to make sure the results are consistent and independent of random initial states. The best choice for k found for each dataset using the approach described in Section 3.1 is also reported in Table 1.

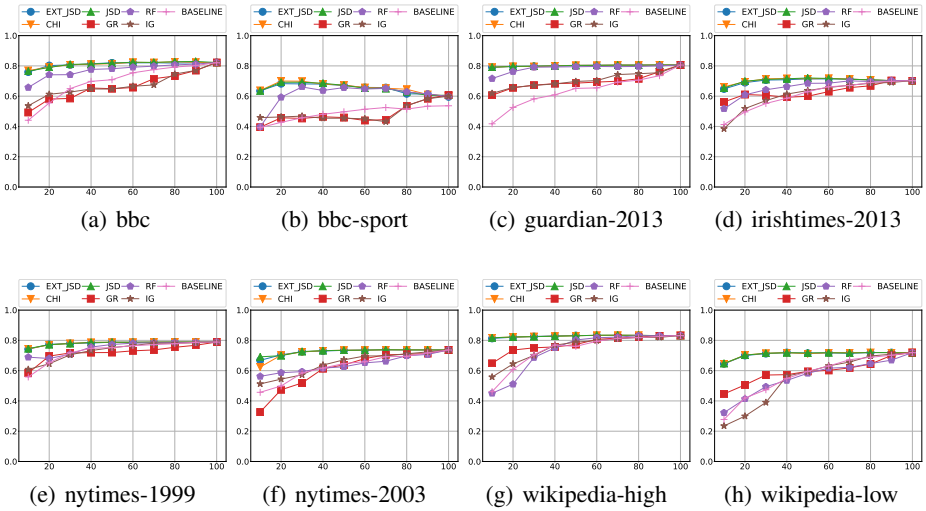
4.4 Results

We present the document classification results of MTCC models with respect to different statistical discrimination metrics in different datasets in Figure 2. The x-axis denotes n, the percentage of total topics used for training the classifier. The y-axis represents the performance which is the micro-averaged f1 score in this case.

We can observe that, almost in all situations, JSD, extended JSD or χ^2 achieves high micro-averaged f1 scores that outperform the baseline (random selection) by a

[5] http://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html
[6] https://rare-technologies.com/python-lda-in-gensim-christmas-edition/

Fig. 2. Micro-f1 score of different statistical discrimination metrics for 8 datasets, the horizontal axis denotes the percentage of total topics used for training the classifier.



(a) bbc (b) bbc-sport (c) guardian-2013 (d) irishtimes-2013

(e) nytimes-1999 (f) nytimes-2003 (g) wikipedia-high (h) wikipedia-low

large margin. This demonstrates that the discrimination metrics can identify the topics that contain divergence information. We can also see from Figure 2 that the best results are achieved by JSD, extended JSD or χ^2 across 8 datasets with very low values for n which denotes the percentage of the total topics used for training (usually $<= 30$). Moreover at these low numbers of topics the performance from the other metrics and from the random baseline is very low. For example, in the guardian-2013 (Figure 2 (c)) and wikipedia-high (Figure 2 (g)) datasets, classifiers nearly reach the best performance (above 0.8) using only the top 10% of topics, meanwhile the performance of baseline is only a little higher than 0.4. This implies that the top 10% topics selected by JSD, extended JSD and χ^2 carry almost all of the divergence information in these two datasets. Similarly, as we can see in bbcsport (Figure 2 (b)), the classifiers based on topics selected using JSD, extended JSD and χ^2 achieve the best results at $n = 20$ and these results even surpass the result where all topics are used ($n = 100$). This indicates that the top 20% of topics are very good at distinguishing between corpora while the remaining topics not only do not contain divergence information but even produce noise for classification.

It is also interesting to note that JSD, extended JSD and χ^2 result in almost the same performance across all datasets. These three metrics are reasonably similar to each other so this is not too surprising. It is interesting to note, however, that although all three will usually select the same topics at the very highest ranks, they do select different topics further down the ordering.

To conclude, JSD, extended JSD and χ^2 can effectively select the discriminative topics and outperform the random baseline and other metrics by a large margin.

5 Demonstrating Topic-based Corpus Comparison

Since we are interested in whether the results of corpus comparison at topic-level are more meaningful and interpretable, in this section, we compare the output of a corpus comparison based on the MTCC topic-level approach to a word-level approach over 4 real-world datasets. We describe the datasets used and an analysis of the results produced by each approach.

5.1 Setup

In this demonstration we perform a corpus comparison that compares the bbc dataset, guardian-2013 dataset, irishtimes-2013 and nytimes-2003 dataset described in [1]. Each dataset is treated as one corpus which is different from the settings in Section 4 and extended JSD is selected as the discrimination metric since its effectiveness shown in Section 4. We therefore conduct a comparison over 4 corpora at one time. These four datasets are all news articles dataset. Also there is both overlap and difference in the sections present in those four datasets (summarised in Table 2). This makes these four corpora an interesting test case for corpus comparison approaches as there are differences that we can expect a corpus comparison approach to discover—for example sections in one corpus that are not in the others (e.g. music section in guardian-2013).

Extended JSD, which is commonly used in extracting keywords from different corpora and has proven its effectiveness in [5, 13], is adopted as the word-level corpus comparison method. The contributions to divergence of individual words across all corpora are computed by Equation 2.

We set the random seed for the LDA model to 1984; α = "$auto$" following the previous experiments; the best choice k are tested in preliminary experiments varies from 100 to 300 in steps of 10. After pre-experiments, 300 is the optimal number for k according to the topic coherence measure described in Section 3.1.

Table 2. The number of documents in each section of the four corpora. Potentially comparable sections are aligned and the potentially distinct sections are highlighted. *Entertainment* is abbreviated to *entmt*.

Sections	bbc	Sections	guardian	Sections	irishtimes	Sections	nytimes
		books	**1,107**				
business	510	business	1,292	economy	364	business	1,024
				crime-law	**699**		
						education	**942**
		fashion	**816**				
				health	273	health	1,046
						movies	**1,247**
		music	**1,403**				
entmt	**386**						
politics	417	politics	844	politics	620		
		football	1,059	soccer	655		
sport	511					sports	1,024
tech	**401**						
				rugby	**482**		

5.2 Results

Table 3 shows the top 3 discriminative topics for each dataset found by MTCC (using the extended JSD metric as the high performance reported in Section 4). The number in front of the text denotes the index of the topic, and is followed by the guess of the meaning of the topics as well as 10 topic descriptors.

Table 3. Three most discriminative topics extracted from the each corpus. The number in front of the text denotes the index of a topic, followed by the 10 most common words to describe the topic. Guesses of the implication of each topic are shown in square brackets.

bbc	guardian
(289) [tech] *people new technology also mobile would could make say many*	(42) [music] *music song band street night album sound love art young*
(114) [sport] *cup england world think like dont want know play football*	(25) [fashion] *fashion woman designer wear dress collection style clothes brand men*
(31) [entmt] *award best oscar nomination prize actor actress category academy ceremony*	(236) [books] *book novel story read author writer writing reader fiction reading*
irishtimes	**nytimes**
(109) [crime-law] *court garda judge case justice man criminal charge law prison*	(299) [health] *health study plan bill government issue change official problem system*
(146) [politics] *european minister government exit euro bailout decision country finance programme*	(13) [education] *school student education parent high district teacher child city program*
(275) [rugby] *rugby australia zealand black leinster coach schmidt test saturday odriscoll*	(102) [movies] *film movie director directed minute character life picture story actor*

The topics extracted using MTCC can be compared to the most discriminative words for each dataset found by the word-based JSD approach which are shown in Table 4.

The first thing that is apparent from comparing Tables 3 and 4 is that the topic-level approach has the advantage of presenting the user with coherent sets of terms in the topic descriptors rather than a long list of disconnected words. This makes it much easier for the user to understand the differences between different corpora.

Looking more deeply at Table 3 we can see that the MTCC approach has successfully identified the divergence information of each dataset, which is evidenced by the presence of distinct topics. For example, the topic descriptors of *topic 109* suggest that this is a topic relating to articles describing court proceedings. We can see from this that MTCC has identified the fact that the irishtimes-2013 corpus has a crime-law section not present in other corpora. Similarly, the presence of other discriminative topics (e.g. *topic 289*, *topic 31*, *topic 42* etc.) has demonstrated the effectiveness of the algorithm in detecting the difference between multiple corpora.

When we look at Table 4, we can find that though word-level corpus comparison properly define the divergence information in some datasets such as guardian-2013 and

Table 4. 30 most distinct words from a comparison for each corpus using word-level extended JSD.

Rank	bbc	Rank	guardian	Rank	irishtimes	Rank	nytimes
1	would	1	game	1	cent	1	school
2	also	2	team	2	garda	2	film
3	people	3	book	3	per	3	student
4	new	4	fashion	4	think	4	center
5	could	5	business	5	player	5	movie
6	say	6	little	6	point	6	program
7	make	7	novel	7	court	7	theater
8	world	8	long	8	end	8	yesterday
9	get	9	side	9	million	9	life
10	take	10	today	10	minister	10	college
11	made	11	album	11	even	11	medicare
12	month	12	song	12	home	12	yankee
13	way	13	cameron	13	know	13	tonight
14	like	14	story	14	hse	14	university
15	back	15	thatcher	15	thing	15	manhattan
16	week	16	club	16	another	16	patient
17	next	17	bst	17	dont	17	never
18	well	18	miliband	18	state	18	including
19	three	19	premier	19	play	19	might
20	many	20	osborne	20	win	20	drug
21	good	21	band	21	hospital	21	street
22	day	22	love	22	health	22	tomorrow
23	may	23	bank	23	need	23	without
24	right	24	guardian	24	leinster	24	district
25	come	25	york	25	four	25	medical
26	work	26	manchester	26	seanad	26	teacher
27	want	27	collection	27	taoiseach	27	ticket
28	going	28	dress	28	rugby	28	doctor
29	still	29	writer	29	put	29	night
30	since	30	form	30	kenny	30	study

nytimes-2013. This approach fails to depict the distinctive content in bbc corpus where a set of unrelated words are presented.

6 Conclusions

In this paper, we introduce the Multi-Corpus Topic-based Corpus Comparison (TMCC) approach to discover distinctive topics across multiple corpora for corpus comparison tasks. We compared the performance of different discrimination metrics on 8 real-world datasets. The results showing that using JSD, extended JSD, or χ^2 can extract topics that contain the most divergence information. We also demonstrated TMCC and compared its output to a word-level approach. Overall we believe that the example presented demonstrates the advantages of using topics for corpus comparison rather than using word-level approaches.

However, because we apply topic modelling on multiple corpora, we are likely to take the risk of losing or blurring topics compared to applying topic modelling over a single corpus (as done by Zhao et al. [19]). Hence, a further exploration that investigates whether or not we are losing or blurring topics together in the multi-corpus topic-based corpus comparison is scheduled in future work.

Acknowledgement. This research was kindly supported by a Teagasc Walsh Fellowship award (2016053) and Science Foundation Ireland (12/RC/2289_P2).

References

1. Belford, M., Mac Namee, B., Greene, D.: Stability of topic modeling via matrix factorization. Expert Systems with Applications 91, 159–169 (2018)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research 3(Jan), 993–1022 (2003)
3. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
4. Degaetano-Ortlieb, S., Kermes, H., Khamis, A., Teich, E.: An information-theoretic approach to modeling diachronic change in scientific english. Selected papers from Varieng–From data to evidence (d2e) (2016)
5. Gallagher, R.J., Reagan, A.J., Danforth, C.M., Dodds, P.S.: Divergent discourse between protests and counter-protests:# blacklivesmatter and# alllivesmatter. PloS one 13(4), e0195644 (2018)
6. Greene, D., O'Callaghan, D., Cunningham, P.: How many topics? stability analysis for topic models. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 498–513. Springer (2014)
7. Kelleher, J.D., Mac Namee, B., D'Arcy, A.: Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. MIT Press (2015)
8. Kilgarriff, A.: Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved lob-brown comparison. In: ALLC-ACH Conference (1996)
9. Lan, M., Tan, C.L., Low, H.B.: Proposing a new term weighting scheme for text categorization. In: AAAI. vol. 6, pp. 763–768 (2006)
10. Lawrence, E., Sides, J., Farrell, H.: Self-segregation or deliberation? blog readership, participation, and polarization in american politics. Perspectives on Politics 8(1), 141–157 (2010)
11. Leech, G., Fallon, R.: Computer corpora–what do they tell us about culture. ICAME journal 16 (1992)
12. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information theory 37(1), 145–151 (1991)
13. Lu, J., Henchion, M., MacNamee, B.: Extending jensen shannon divergence to compare multiple corpora. In: 25th Irish Conference on Artificial Intelligence and Cognitive Science, Dublin, Ireland, 7-8 December 2017. CEUR-WS. org (2017)
14. Murdock, J., Allen, C.: Visualization techniques for topic model checking. In: Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)
15. O' Callaghan, D., Greene, D., Carthy, J., Cunningham, P.: An analysis of the coherence of descriptors in topic modeling. Expert Systems with Applications 42(13), 5645–5657 (2015)
16. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing corpora-Volume 9. pp. 1–6. Association for Computational Linguistics (2000)
17. Sajgalik, M., Barla, M., Bielikova, M.: Searching for discriminative words in multidimensional continuous feature space. Computer Speech & Language 53, 276–301 (2019)
18. Sievert, C., Shirley, K.: Ldavis: A method for visualizing and interpreting topics. In: Proceedings of the workshop on interactive language learning, visualization, and interfaces. pp. 63–70 (2014)
19. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.P., Yan, H., Li, X.: Comparing twitter and traditional media using topic models. In: European conference on information retrieval. pp. 338–349. Springer (2011)