Bigger versus Similar: Selecting a Background Corpus for First Story Detection Based on Distributional Similarity

Fei Wang and **Robert J. Ross** and **John D. Kelleher**
Technological University Dublin
ADAPT Research Centre
d13122837@mydit.ie
{robert.ross, john.d.kelleher}@dit.ie

Abstract

The current state of the art for First Story Detection (FSD) are nearest neighbour-based models with traditional term vector representations; however, one challenge faced by FSD models is that the document representation is usually defined by the vocabulary and term frequency from a background corpus. Consequently, the ideal background corpus should arguably be both large-scale to ensure adequate term coverage, and similar to the target domain in terms of the language distribution. However, given these two factors cannot always be mutually satisfied, in this paper we examine whether the distributional similarity of common terms is more important than the scale of common terms for FSD. As a basis for our analysis we propose a set of metrics to quantitatively measure the scale of common terms and the distributional similarity between corpora. Using these metrics we rank different background corpora relative to a target corpus. We also apply models based on different background corpora to the FSD task. Our results show that term distributional similarity is more predictive of good FSD performance than the scale of common terms; and, thus we demonstrate that a smaller recent domain-related corpus will be more suitable than a very large-scale general corpus for FSD.

1 Introduction

Given a stream of documents about news events in a chronological order, the goal of First Story Detection (FSD) is to identify the very first story for each event. Each story is processed in sequence, and a decision is made for a given candidate document on whether or not it discusses an event that has not been seen in previous documents; crucially this decision is made after processing the candidate document but before processing any subsequent documents (Allan et al., 1998; Yang et al., 1998). The decision making process for each incoming document is normally based on a novelty score; namely, if the novelty score of a new document is higher than a given threshold, we say it is a first story.

Hundreds of FSD models have been proposed in prior research, and the nearest neighbour-based models, in which the novelty score is defined as the distance from the new story to the closest existing story, remain the state of the art (Wang et al., 2018). In the implementation of nearest neighbour-based FSD models, the first step is to represent each story with a sound document representation. Even though many deep learning-based document representations have been shown to achieve very good results in a range of NLP (Natural Language Processing) tasks (Goldberg, 2017), the dominant document representation model for FSD remains the traditional term vector models in which each feature represents a term in the vocabulary (Brants et al., 2003; Petrović et al., 2010; Wang et al., 2018; Kannan et al., 2018).

The majority of machine learning research assumes that the data used for building a model and making inference are sampled from the same distribution, i.e., the data generation process is stationary. However, because of its online characteristic, one challenge faced by FSD models is that the system's vocabulary (and hence document representation) cannot be derived from a target corpus, but must instead be defined by the vocabulary of a background corpus. The resultant potential difference between the background and target corpus demonstrates a non-stationary characteristic

Proceedings of Recent Advances in Natural Language Processing, pages 1312–1320,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_150

of FSD. To mitigate for potential differences between background and target data, the ideal background corpus should be both large-scale, so as to ensure an adequate number of common terms between it and the documents in the target stream (i.e., minimize *unknown* words), and similar in the sense of language distribution. In many cases, these two factors cannot be satisfied at the same time, and thus the emphasis has to be placed on the more informative one of the two, which leads to a question of "bigger or similar?". To the best of our knowledge however, there is little research addressing this question empirically, and no metrics have been proposed for the quantitative comparison of the scale and similarity between background corpora relative to a target corpus.

In this paper we examine whether the distributional similarity of common terms between corpora (background and target story stream) is more important than the scale of common terms for FSD. As a basis for our analysis we propose a set of metrics to quantitatively measure the scale of common terms and the distributional similarity between corpora. Using these metrics we rank different background corpora relative to a target FSD corpus. Finally, we apply the models based on different background corpora to the FSD task to determine the relative utility of different assumptions about the background corpus. Our contributions are thus two-fold: an investigation of background corpus similarity versus scale, and a metrics framework for making such an investigation.

2 First Story Detection

FSD as a challenge was initially defined within the Topic Detection and Tracking (TDT) competition series (Yang et al., 1998; Allan et al., 2000b); and was considered to be the most difficult challenge in all five TDT tasks (Allan et al., 2000a). Since then, the need for accurate FSD models has been greatly strengthened by the proliferation of digital content, and social media streams in particular. One of the challenges of FSD is the undefinable characteristic of a first story. We can never know what the next first story will look like; instead, we only know that it must be different to existing stories to some degree. Therefore, we normally consider FSD as an unsupervised learning application, and hence attempt to define and make use of a novelty score in a similar fashion to how novelty-style metrics are defined in other unsupervised learning applications (Wang et al., 2017).

Based on different definitions of novelty scores, it has been proposed that there are three categories of FSD models (Wang et al., 2018): Point-to-Point (P2P) models, Point-to-Cluster (P2C) models, and Point-to-All (P2A) models. P2P models, in which the novelty score is defined as the distance from the incoming story to an existing story, are normally nearest neighbour-based (Yang et al., 1998; Allan et al., 2000b), or approximate nearest neighbour-based (Brants et al., 2003; Petrović et al., 2010, 2012; Moran et al., 2016; Kannan et al., 2018). In P2C models or P2A models, the novelty score is defined respectively as the distance from the new story to a cluster of existing stories (also can be considered as the distance to an existing event), or to all the existing stories. The former is usually clustering-based (Yang et al., 1998; Allan et al., 2000b; Li et al., 2017), and the latter uses all the existing data to build a system, and applies this system to the incoming story to generate a novelty score (Schölkopf et al., 2001; Wurzer et al., 2015). Based on previous literature and research on FSD, it has been shown that nearest neighbour-based P2P models perform the best among all these three categories of FSD models (Wang et al., 2018).

2.1 Term Vector Models for First Story Detection

As presented above, the novelty score in a P2P model is calculated by comparing the incoming story to previous stories and then finding its (approximate) nearest neighbour and the corresponding closest distance. When implementing a P2P model, the first step is to convert the raw stories to document representation vectors that can be fed into the detection model; this is then followed by the quantitative comparisons between these document representations. The state of the art document representation model for P2P FSD models remains the traditional term vector models, due, in part, to their specificity of terms (Wang et al., 2018). In a term vector model, each feature (dimension) represents a term in the vocabulary, so the dimensionality of each vector is generally the same length as the corpus vocabulary.

TF-IDF is the most well-known term vector model and also the most effective one used for FSD (Brants et al., 2003; Petrović et al., 2010; Kannan et al., 2018). A TF-IDF weight is calcu-

lated for each term in a document vector as the product of the TF (term frequency) and IDF (inverse document frequency) components. The TF component captures the number of times a term was encountered in the document, while the IDF component discounts the term weights that are very common in the corpus such that these are judged to have little information relevant to the distinction of documents. A TF-IDF model always stores a vocabulary as well as an IDF dictionary in which the key is each term while the value is the corresponding IDF component for that term. When applying a TF-IDF model, it is necessary to use some corpus to build the vocabulary and the IDF dictionary before calculating the TF-IDF weight for each term in a document using some specific scheme. A widely-applied TF-IDF weighting scheme is shown as follows:

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t) \quad (1)$$

$$idf(t) = log\frac{N}{df(t)} \quad (2)$$

where $tf(t,d)$ represents the TF component, that is just the number of times the term t occurs in document d, and $idf(t)$ represents the IDF component, in which N denotes the total number of documents and $df(t)$ refers to the number of documents that contain the term t.

In the context of FSD, the labelled target corpus is always unavailable before detection because of the online characteristic of FSD, and thus a background corpus is required to build the TF-IDF model, i.e., the vocabulary and the IDF dictionary in the model. As shown in Fig. 1, we assume that a TF-IDF model is built with a background Corpus B and is applied to the FSD task for a target Corpus T. Set 2 is the overlapping term set that contains the terms common to both Corpus B and T, and Set 1 and 3 contain the terms that only exist in Corpus B or T respectively. Consequently, Set 1 and 2 constitute all terms in Corpus B, while Set 2 and 3 constitute all terms in Corpus T.

2.2 Set Overlap and FSD Modelling

In a pre-built TF-IDF model, all the terms in the vocabulary are from the background Corpus B, i.e., the terms used to generate the term vector space are those from Set 1 and 2, while those terms in Set 3 will not appear in the TF-IDF model at all. In other words, the terms in Set 3 are all the

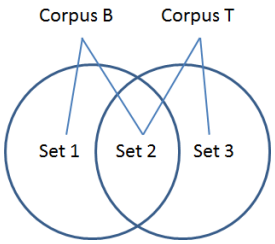


Figure 1: Term Sets within a Background Corpus B and a Target Corpus T

unknown terms with respect to the TF-IDF model. However, when the TF-IDF model is applied to FSD, all the documents to be analysed will be from the target Corpus T, which means that all the terms in Set 1 will not appear at all in the process of FSD; as a result the TF components for these terms are always zero and thus all the final TF-IDF weights of these will always be zero as well. It should be noted that we can look at all the terms of the target corpus here because we are now doing the analysis. However, during the real FSD we will never know whether a term from the background corpus appears in the target corpus or not. Therefore, we have to keep all the terms in Set 1 and 2 that are from the background corpus, even though the weights of all the terms in Set 1 are always zero.

The comparison between TF-IDF representations is usually based on cosine distance calculations for FSD (Allan et al., 2000b; Brants et al., 2003). For such calculations, all representation vectors are normalised so that the cosine distance is not sensitive to the specific weighting schemes (Allan et al., 2000b). To clarify, given an incoming story represented by $\vec{a}$ and an existing story represented by $\vec{b}$, the cosine distance between them is defined as:

$$cosine\_distance(\vec{a},\vec{b}) = 1 - \frac{\vec{a}\cdot\vec{b}}{|\vec{a}|\left|\vec{b}\right|} \quad (3)$$

According to the definition of cosine distance, in each document vector the terms whose weights are always zero do not have any effect on the result of calculation, so they can be ignored when we analyse the calculation. Hence, the valid terms that make sense for FSD are only those in Set 2, which are the common terms in both the background and target corpus. Given this, the effectiveness of the TF-IDF model only depends on Set 2, and specifically, two factors of Set 2: the scale

and the distributional similarity between the background and target corpus. The scale describes the number of common terms between the corpora. The larger the scale of Set 2, the more informative terms are taken into account. The distributional similarity of two corpora refers to similarity of the frequencies of common terms. As the IDF components of these common terms are calculated only based on the background corpus, the more similar the background corpus is to the target corpus in terms of the language distribution, the better the generated weights can represent the common terms for FSD in the target corpus.

Therefore, the ideal background corpus for FSD should be both large-scale and similar in frequency distribution to the assumed target corpus. However, in many cases, these two factors cannot always be satisfied at the same time, and so it can be useful to determine which of these factors is more predictive of good FSD performance.

3 Quantitatively Measuring Background Corpus Suitability

To propose a method for evaluating the relative importance of the quantity of shared terms versus the similarity of language distributions between a background and target corpus, in this section we outline a set of quantitative metrics to make pairwise comparisons between different background corpora relative to the target FSD corpus.

3.1 Measuring the Scale of Common Terms

As shown earlier in Fig. 1, the scale of common terms relative to the target Corpus T can be quantitatively measured using the proportion of common terms of Set 2 relative to all the terms of Corpus T; we refer to this as the overlapping rate of the background Corpus B relative to the target Corpus T. Given any specific target corpus, the bigger the overlapping rate is for a background corpus, the more informative terms are available to be taken into account, and hence the less unknown terms occur in the FSD process.

3.2 Measuring the Distributional Similarity

While measuring the scale of common terms is relatively straightforward, the assessment of distributional similarity is somewhat more involved.

As we focus on the TF-IDF model, the distribution similarity between corpora should be based on the document frequencies. If we order the terms by document frequency for different corpora, each term will likely have a different rank within each corpus. Moreover, if we only look at the ranks of common terms in both corpora (i.e., the terms in Set 2 shown in Fig. 1), it is possible to measure the dissimilarity between two corpora based on their different lists of term ranks.

Before making rank based similarity measurements, some preparation is required. Firstly, the common terms in both corpora (background and target corpus) are extracted as the basis for the comparisons. For each corpus, these common terms are ordered in a descending order based on their document frequencies calculated with only this corpus, and then each term is assigned an index from 1 to n, where n is the number of common terms that are being taken into account. For different corpora, the order of terms will be different, as well as the index of each term. If there are no terms with the same document frequency in an ordered term list, the index of each term can be reasonably considered as its rank in this corpus. However, the fact is that many terms have the same document frequency in a corpus, so they should have the same rank. Instead of assigning different ranks to the neighbouring terms with the same document frequency, we implement some extra operations to make their ranks the same. Specifically, for the terms with the same document frequency, i.e., the terms with indices from i to j, we assign the same average rank $\frac{i+j}{2}$ to all of these, such that this does not affect the rank of any other term. If from the 1^{st} to the 4^{th} terms in the ordered term list have the same document frequency, all of them will be assigned a rank $(1+4)/2 = 2.5$.

After pre-processing, we count the number of inversions and calculate the distance between two ordered same-length term lists to present the dissimilarity between these two corpora:

1 **Inversion count** If the order of two different terms in one corpus is not the same as that in the other corpus, e.g., in one corpus, term X has a rank smaller than term Y, while in the other corpus, term X has a rank larger or equal to term Y, we call this situation an inversion. The inversion count metric is defined as the count of all the inversions between two different ordered rank lists.

2 **Manhattan distance** To calculate the dissimilarity between two same-length rank lists we

subtract the rank of each term in one list from the rank of the same term in the other list and sum the absolute value of each of these differences (Kelleher et al., 2015).

As both these dissimilarity metrics show the degree to which a background corpus is different from the target corpus, we expect that the greater the metric the worse the subsequent model is expected to perform on the FSD task. We only evaluate the distributional similarity based on the frequency ranks of the common terms, rather than the quantitative frequency values, because the comparisons based on the quantitative frequency values usually lead to more emphasis on the terms with high frequency values, which should be avoided. It is worth noting that in real use both of these metrics are normalised to between 0 and 1 by being divided by n^2, where n is the length of the rank lists, i.e, the number of common terms. The calculation of these two metrics requires time complexity of $O(n^2)$ and $O(n)$ respectively.

3.3 Comparison between Two Background Corpora Relative to a Target Corpus

With the metrics proposed above, we can make comparisons between different background corpora relative to a target FSD corpus. For the comparison of the scale of common terms, the overlapping rate can be applied to multiple background corpora to rank them based on their rate values. However, the situation for the comparison of the distributional similarity is more involved.

As explained in their definitions, both two dissimilarity metrics proposed above are calculated based on the common terms of a background corpus and a target corpus. If we want to compare among multiple background corpora relative to a target corpus, the calculation should be based on the common terms of all the background and target corpus to ensure the rank list for each background corpus in the same length[1]. The situation of two background corpora and a target corpus is depicted in Fig. 2, in which the calculation of dissimilarity metrics would be based on Common Set. Generally, the common terms shared by the three corpora will be less than those shared by only any two of them. For each background corpus, the terms used for the comparison (i.e., the terms in Common Set) will be less than those used in FSD (i.e., the terms in Common Set and Set 1 for Corpus B1, and the terms in Common Set and Set 2 for Corpus B2). This will lead to errors in the measures and comparisons, and the more background corpora are being compared, the greater the errors will be. In order to limit this kind of error, we restrict to pairwise comparison between background corpora so that the number of terms used for comparisons are relatively large in comparison to the terms used in FSD.

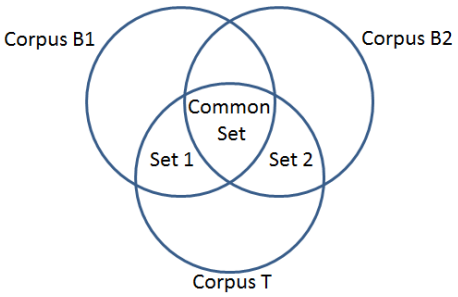


Figure 2: Common Set among two Background Corpora B1 and B2 and a Target Corpus T

4 Experiment Design

In this section, we present our experiments for comparing the scale of common terms and the distributional similarity between different background corpora relative to a target FSD corpus, and apply the models based on different background corpora to the FSD task in an attempt to determine which factor is more predictive of good FSD performance.

4.1 Corpora Used in the Experiments

The target corpus we use for FSD detection is the standard $TDT5$ corpus[2]; the contents of which are newswire stories generated from April to September 2003. The background corpora we are making use of for the current investigation are subsets of $COHA$ (Corpus of Historical American English) (Davies, 2012) and $COCA$ (The Corpus of Contemporary American English) (Davies, 2010). The former covers comprehensive historical English documents from 1810 to 2009 in different domains such as news, fiction, academia

[1] We also tried designing metrics that can be generated based on different terms, i.e., for each background corpus using the terms shared only by the target corpus and itself, rather than common terms shared by all corpora, but we failed because we could not find any valid method to normalise the metrics generated based on different numbers of terms.

[2] https://catalog.ldc.upenn.edu/LDC2006T18

and so on, and the latter is similar to $COHA$ in themes but focuses only on the contemporary contents from 1990 to present. The numbers of documents in $TDT5$, $COHA$ and $COCA$ are about 278,000, 115,000 and 190,000 respectively. As mentioned we make use of subsets of $COHA$ and $COCA$; specifically we mostly include data that predates 2003, i.e., the year of $TDT5$ collection, unless otherwise stated.

In order to answer our underlying research question, i.e., whether bigger or similar background corpora provide the clearer benefit, we carried-out three sets of experiments. In the first set, comparisons are made between $COCA$ and $COHA$ with the assumption that a contemporary corpus will be more similar to the target corpus than a historical one. The second set of experiments supplement the first set and focus on corpus temporality. Comparisons are made between two subsets of the entire COCA corpus - $COCA$ and $COCA\_After\_2003$ that respectively include only the documents before and after 2003, the year when the target corpus was collected. We assume that a corpus with future data is more similar to a target corpus than that with prior data only[3]. The last set of experiments establish comparisons between two subsets of $COCA$ - $COCA\_News$ and $COCA\_Except\_News$, in which $COCA\_News$ contains only the documents in the domain of news, which is the same as the domain of the target TDT5 corpus, while $COCA\_Except\_News$ contains the documents in other domains except news. We also assume that the domain-related corpus is more similar to the target corpus than those in different domains.

4.2 Metric Calculation

In the implementation, we apply all the metrics to each corpus mentioned above, and then make comparisons in each pair of background corpora. In addition, for the comparison of corpus similarity, we examine whether the two proposed metrics, inversion count and Manhattan distance, are consistent with each other in deciding which corpus in each pair is more similar to the target corpus, i.e., whether two metric values for a corpus are both smaller or greater than those for the other corpus in the comparison pair. We also verify whether the results of comparisons correspond with our assumptions about corpus similarity in Sect. 4.1.

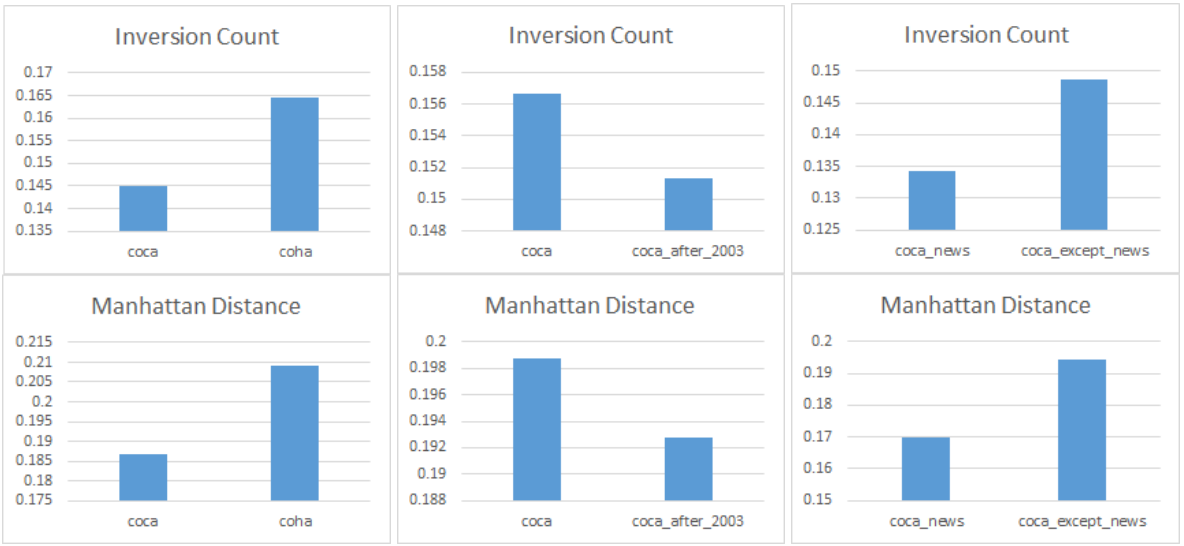
[3]In real FSD, future data is always unavailable. This set of experiments are only for the use of analysis.

4.3 FSD Evaluation

Following background corpus metric calculation, we build TF-IDF models based on the background corpora being compared and apply these models to the FSD task.

The implementation of FSD is based on the nearest neighbour algorithm with the TF-IDF representations we described in Sect. 2.1. We also adopt the cosine distance as the dissimilarity measure between document representations. In order to reduce the effect of useless terms and different term forms, for both the background and target corpus we remove terms with very high and very low document frequency (stop words and typos), and stem all terms. Aligning with previous research (Yang et al., 1998), comparisons are only implemented with the 2000 most recent stories for each incoming story. The output of each FSD model is a list of novelty scores for each document in the target corpus TDT5. Based on these outputs, the standard evaluation method for FSD is to apply multiple thresholds to sweep through all the novelty scores. For each threshold, a missing rate and a false alarm rate are calculated, and then for all thresholds, the missing and false alarm rates are used to generate a DET (Detection Error Tradeoff) curve (Martin et al., 1997), which shows the trade-off between the false alarm error and the missing error in the detection results. The closer the DET curve is to the zero point, the better the FSD model is said to perform. It happens sometimes for the evaluation with DET curves that many curves are in a tangle – making it is difficult to figure out visually which model performs better. Therefore, we calculate Area Under Curve (AUC) for each FSD model, and the model with the lowest AUC is judged to be best.

In order to achieve more comprehensive results for this evaluation, we implement tests for set variants. Specifically, for each set of experiments, we make comparisons not only between the two background corpora being evaluated, but also between each corpus and the union of both corpora; for example for $COCA$ vs. $COHA$, we not only implement the comparison between $COCA$ and $COHA$, but also between $COCA$ and $COCA + COHA$ and between $COHA$ and $COCA + COHA$, where $COCA + COHA$ is the union of $COCA$ and $COHA$. In this way, we



(a) Metric Results for $COCA$ vs. $COHA$

(b) Metric Results for $COCA$ vs. $COCA\_After\_2003$

(c) Metric Results for $COCA\_News$ vs. $COCA\_Except\_News$

Figure 3: Comparisons of Corpus Dissimilarity

have six more comparison results that can be used for the evaluation of the relations between background corpus and model performance for FSD.

5 Results & Analysis

We first look at the comparisons between corpora before looking at FSD performance for different background corpora.

5.1 Results of the Comparisons of Corpus Dissimilarity

We applied the two metrics, inversion count and Manhattan distance, to the three sets of comparisons: $COCA$ vs. $COHA$, $COCA$ vs. $COCA\_After\_2003$ and $COCA\_News$ vs. $COCA\_Except\_News$. The results are shown in Fig. 3. We find firstly that in all comparison sets that the results of the two evaluation metrics are consistent with each other, i.e., the metric values for $COCA$ are both smaller than $COHA$, but greater than $COCA\_After\_2003$, and those for $COCA\_News$ are both smaller than $COCA\_Except\_News$. Secondly, we also find that these comparison results all correspond with our assumptions that more recent domain-related corpora are more similar to the target corpus. Given this, we conclude that both metrics are effective for the comparison of the distributional similarity between background corpora relative to the target corpus, and for the sake of simplicity, we judge Manhattan distance as the most useful metric due to its ease of calculation and interpretation.

5.2 Results of the Relations between Background Corpus and Model Performance for First Story Detection

Results are shown in Table 1, 2 and 3, where the values of the overlapping rates and Manhattan distances are the values for one corresponding background corpus relative to the target corpus. The cells in bold indicate the better results in the comparisons of the scale of common terms and the common term distributional similarity between each pair of background corpora, as well as the better FSD performance. We find that all corpora that are more similar (in terms of term distributions) to the target corpus lead to better performance in FSD, except in the case of very similar performance between $COCA$ and $COCA + COHA$. However, it is worth noting that only six in nine corpora that have a larger scale of common terms correspond with better FSD performance while the other three do not. For example, in Table 3 although the corpus $COCA$ has the much larger scale of common terms, the FSD performance based on it is still worse than that based on $COCA\_News$, because $COCA\_News$ is more similar to the target corpus in terms of language distribution.

Based on these results, it can be argued that term distributional similarity is more predictive of

	coca vs. coha		coca vs. coca+coha		coha vs. coca+coha	
	coca	coha	coca	coca+coha	coha	coca+coha
Overlapping Rate	**0.3771**	0.3255	0.3771	**0.4193**	0.3255	**0.4193**
Manhattan Distance	**0.1869**	0.2090	**0.1996**	0.2068	0.2076	**0.1949**
AUC	**0.1056**	0.1100	**0.1056**	**0.1056**	0.1100	**0.1056**

Table 1: Comparisons between $COCA$ and $COHA$

	coca vs. coca_after_2003		coca vs. coca_all		coca_after_2003 vs. coca_all	
	coca	coca_after_2003	coca	coca_all	coca_after_2003	coca_all
Overlapping Rate	0.3771	**0.4077**	0.3771	**0.4583**	0.4077	**0.4583**
Manhattan Distance	0.1987	**0.1928**	0.1996	**0.1950**	**0.1997**	0.2009
AUC	0.1056	**0.1008**	0.1056	**0.1020**	**0.1008**	0.1020

Table 2: Comparisons between $COCA$ and $COCA\_After\_2003$

	coca_news vs. coca_except_news		coca_news vs. coca		coca_except_news vs. coca	
	coca_news	coca_except_news	coca_news	coca	coca_except_new	coca
Overlapping Rate	0.2932	**0.3184**	0.2932	**0.3771**	0.3184	**0.3771**
Manhattan Distance	**0.1698**	0.1943	**0.1795**	0.1996	0.1986	**0.1880**
AUC	**0.1044**	0.1078	**0.1044**	0.1056	0.1078	**0.1056**

Table 3: Comparisons between $COCA\_News$ and $COCA\_Except\_News$

good FSD performance than the scale of common terms; and, thus we can give general guidance to the selection of background corpus for FSD that a smaller recent domain-related corpus will be more suitable than a very large-scale general corpus for FSD. Of course, our research is directed only at the general situations, as the interpretations do not include extreme situations such as extremely large or small scale of common terms. It is also worth noting that we are purposefully focusing here on the case of a static background corpus and not on the case of updates being made to the TF-IDF model as the FSD process unfolds.

6 Conclusion

We conclude with a highlight of our three main contributions. We proposed a set of metrics for the quantitative evaluation of the scale of common terms and the term distributional similarity of a background corpus relative to a target corpus, and a pairwise comparison scheme between two different background corpora. We also applied the proposed metrics and comparison scheme to the comparisons between background corpora relative to the target FSD corpus, and our results indicate that term distributional similarity is more predictive of good FSD performance than the scale of common terms. Finally, we answered the research question of whether bigger or similar corpus are more useful for FSD by showing that a smaller recent domain-related corpus will be more suitable than a very large-scale general corpus for FSD.

Acknowledgment

The authors wish to acknowledge the support of the ADAPT Research Centre. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Funds.

References

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report .

James Allan, Victor Lavrenko, and Hubert Jin. 2000a. First story detection in tdt is hard. In *Proceedings of the ninth international conference on Information and knowledge management*. ACM, pages 374–381. https://doi.org/10.1145/354756.354843.

James Allan, Victor Lavrenko, Daniella Malin, and Russell Swan. 2000b. Detections, bounds, and timelines: Umass and tdt-3. In *Proceedings of topic detection and tracking workshop*. sn, pages 167–174.

Thorsten Brants, Francine Chen, and Ayman Farahat. 2003. A system for new event detection. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, pages 330–337. https://doi.org/10.1145/860435.860495.

Mark Davies. 2010. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing* 25(4):447–464. https://doi.org/10.1093/llc/fqq018.

Mark Davies. 2012. Expanding horizons in historical linguistics with the 400-million word corpus of historical american english. *Corpora* 7(2):121–157. https://doi.org/10.3366/cor.2012.0024.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies* 10(1):1–309. https://doi.org/10.2200/S00762ED1V01Y201703HLT037.

Jeyakumar Kannan, Ar Md Shanavas, and Sridhar Swaminathan. 2018. Real time event detection adopting incremental tf-idf based lsh and event summary generation. *International Journal of Computer Applications* 975:8887. https://doi.org/10.5120/ijca2018916252.

John D Kelleher, Brian Mac Namee, and Aoife D'arcy. 2015. *Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies*. MIT Press.

Quanzhi Li, Armineh Nourbakhsh, Sameena Shah, and Xiaomo Liu. 2017. Real-time novel event detection from social media. In *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. IEEE, pages 1129–1139. https://doi.org/10.1109/ICDE.2017.157.

Alvin Martin, George Doddington, Terri Kamm, Mark Ordowski, and Mark Przybocki. 1997. The det curve in assessment of detection task performance. Technical report, National Inst of Standards and Technology Gaithersburg MD.

Sean Moran, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. Enhancing first story detection using word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, pages 821–824. https://doi.org/10.1145/2911451.2914719.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. Association for Computational Linguistics, pages 181–189.

Saša Petrović, Miles Osborne, and Victor Lavrenko. 2012. Using paraphrases for improving first story detection in news and twitter. In *Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies*. Association for Computational Linguistics, pages 338–346.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. 2001. Estimating the support of a high-dimensional distribution. *Neural computation* 13(7):1443–1471. https://doi.org/10.1162/089976601750264965.

Fei Wang, Hector-Hugo Franco-Penya, John D Kelleher, John Pugh, and Robert Ross. 2017. An analysis of the application of simplified silhouette to the evaluation of k-means clustering validity. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*. Springer, pages 291–305. https://doi.org/10.1007/978-3-319-62416-7_21.

Fei Wang, Robert J Ross, and John D Kelleher. 2018. Exploring online novelty detection using first story detection models. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, pages 107–116. https://doi.org/10.1007/978-3-030-03493-1_12.

Dominik Wurzer, Victor Lavrenko, and Miles Osborne. 2015. Twitter-scale new event detection via k-term hashing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. pages 2584–2589. https://doi.org/10.18653/v1/D15-1310.

Yiming Yang, Tom Pierce, and Jaime G Carbonell. 1998. A study on retrospective and on-line event detection https://doi.org/10.1145/290941.290953.