Using Word Embedding to Evaluate the Coherence of Topics from Twitter Data

Anjie Fang[1], Craig Macdonald[2], Iadh Ounis[2], Philip Habel[2]
[1]a.fang.1@research.gla.ac.uk, [2]{firstname.secondname}@glasgow.ac.uk
University of Glasgow, UK

ABSTRACT

Scholars often seek to understand topics discussed on Twitter using topic modelling approaches. Several coherence metrics have been proposed for evaluating the coherence of the topics generated by these approaches, including the pre-calculated *Pointwise Mutual Information* (PMI) of word pairs and the *Latent Semantic Analysis* (LSA) word representation vectors. As Twitter data contains abbreviations and a number of peculiarities (e.g. hashtags), it can be challenging to train effective PMI data or LSA word representation. Recently, *Word Embedding* (WE) has emerged as a particularly effective approach for capturing the similarity among words. Hence, in this paper, we propose new Word Embedding-based topic coherence metrics. To determine the usefulness of these new metrics, we compare them with the previous PMI/LSA-based metrics. We also conduct a large-scale crowdsourced user study to determine whether the new Word Embedding-based metrics better align with human preferences. Using two Twitter datasets, our results show that the WE-based metrics can capture the coherence of topics in tweets more robustly and efficiently than the PMI/LSA-based ones.

1. INTRODUCTION

Topic modelling approaches can be used by scholars to capture the topics discussed in various corpora, including news articles, books [5] and tweets [4, 15]. Since typically for such scenarios no ground-truth exists to determine how well the topic modelling approach works, a number of topic coherence metrics have been proposed to assess the performance of the topic modelling approaches in extracting comprehensible and coherent topics from corpora. These metrics often capture the semantic similarity of words in a topic using external sources such as Wikipedia or WordNet.

To evaluate the coherence of a topic, a coherence metric averages the semantic similarity of words in topics. Recently, two effective coherence metrics, namely the *Pointwise Mutual Information* (PMI)-based [11] of word pairs and the *Latent Semantic Analysis* (LSA) word representation-based [3] metrics, have been adapted to tweet corpora [3]. Through a large crowdsourcing study, Fang et al. [3] found that a PMI-based metric using a Twitter background dataset aligned best with human preferences of topic coherence. However, some challenges remain, particularly because of the uniqueness of Twitter data, where unlike many other corpora, tweets contain abbreviations, several peculiarities (e.g. hashtags) and a vast vocabulary. For example, the PMI metric leverages the co-occurrence data of approx. 354 million word pairs [3], which is voluminous to store. Recently, *Word Embedding* (WE) has emerged as a more effective word representation than, among others, LSA [8, 9, 10]. Using WE word representation models, scholars have improved the performance of classification [6], machine translation [16], and other tasks. However, it remains to be seen whether Word Embedding can be effectively used to evaluate the coherence of topics in comparison with existing metrics.

In this paper, we propose a new Word Embedding-based metric, which we instantiate using 8 different Word Embedding models (trained using different datasets and different parameters). We also use as baselines two types of existing effective metrics based on PMI and LSA. We conduct a large-scale pairwise user study, comparing human judgements with the 8 WE-based and the 4 PMI/LSA-based baseline metrics. We generate the topic pairs using three topic modelling approaches (i.e. LDA, *Twitter LDA* and *Pachinko Allocation Model*). First, we rank the topic modelling approaches using each of the deployed coherence metrics. Second, we assess the extent to which the topical preferences emanating from the 12 metrics align with human assessments. Using two Twitter datasets, our results show that the new Word Embedding-based metrics outperform the PMI/LSA-based ones in capturing the coherence of topics in terms of robustness and efficientness.

2. BACKGROUND & RELATED WORK

In this paper, we use three topic modelling approaches to generate topics for Twitter data. They have been chosen because of their reasonable computational cost and scalability on high volumes of tweets. First, we use LDA [2], where each of K topics is represented by a term distribution ϕ, while each document has a topic distribution θ. Second, we experiment with an extension of LDA, the *Pachinko Allocation Model* (PAM) [7]. The topic layer in PAM is divided into a super-topic layer (distribution over sub-topics) and a sub-topic layer (distribution over terms). Third, we use a topic modelling approach tailored to tweet corpora, namely *Twitter LDA* (TLDA) [15], where a Bernoulli distribution is estimated and used to control the selection between "real" terms and background terms. PAM and TLDA are known

SIGIR '16, July 17-21, 2016, Pisa, Italy
© 2016 ACM. ISBN 978-1-4503-4069-4/16/07...$15.00
DOI: http://dx.doi.org/10.1145/2911451.2914729

to generate more coherent topics than LDA on news corpora and tweets, respectively [7, 15].

There are two main existing types of effective topic coherence metrics. One metric, *Pointwise Mutual Information* (PMI), developed by Newman et al. [11], captures the semantic similarity of pairs of words in a topic, by examining how the word pairs co-occur in external sources such as Wikipedia. PMI has been tested on news articles and books. Fang et al. [3] showed that the PMI metric deployed using Twitter background datasets was closest to human judgements. Moreover, they also adapted a second type of metric, the LSA word representation metric, to evaluate the coherence of topics. The LSA-based coherence metric proved to be less aligned with human assessments than the PMI-based one on tweet corpora.

Turning to Word Embedding approaches, recently scholars have applied *Feed-Forward Neural Network* (FFNN) for Word Embeddings [1]. In this approach, similar to LSA, a word is represented as a continuous vector in a Word Embedding model. Based on FFNN, Mikolov et al. [8, 9] proposed a *skip-gram* model to generate Word Embeddings from large datasets more efficiently and effectively. Godin et al. [6] improved the performance of document classification via Word Embeddings. Indeed, Neural Networks were shown to generate more effective word representations than LSA [10]. Therefore, in this paper, we propose new Word Embedding-based metrics to capture the coherence of topics. We adopt the *skip-gram* approach to obtain our Word Embedding models. In the next section, we explain how we deploy the PMI, LSA, and WE-based metrics.

3. COHERENCE METRICS

We use three types of coherence metrics for Twitter data based on PMI and LSA, and Word Embeddings (WE), which are instantiated into 12 metrics. We first define the existing PMI & LSA-based metrics before introducing the new Word Embedding-based metric to evaluate the coherence of topics.
PMI & LSA metrics. A topic t can be represented by the top $n = 10$ words ($\{w_1, w_2, ..., w_{10}\}$) (selected by their probabilities ($p(w|z)$) in Φ for this topic). The coherence of a topic can be calculated by averaging the semantic similarity of pairs of words associated with that topic (Equation (1)). Both Newman et al. [11] and Fang et al. [3] showed that the PMI of pairs of words can capture the coherence of topics identified from both standard and tweet corpora. In particular, for the PMI metric, Equation (2) is used to measure the similarity of word w_i and w_j based on co-occurrence statistics obtained from a background corpus (e.g. Wikipedia or a large sample of tweets - detailed in Section 5) along with Equation (1). Note that the PMIs of word pairs need to be pre-calculated from these external datasets.

$$Coherence(t) = \frac{1}{\sum_{m=1}^{n-1} m} \sum_{i=1}^{n} \sum_{j=i+1}^{n} f_{ss}(w_i, w_j) \quad (1)$$

$$f_{ss}(w_i, w_j) = PMI(w_i, w_j) = log\frac{p(w_i, w_j)}{p(w_i) \times p(w_j)} \quad (2)$$

LSA can also be used to capture the semantic similarity of word pairs [13]. In applying LSA, each word is represented by a dense vector in the reduced LSA space, V_{m_i}, obtained by applying Singular Value Decomposition on a background corpus. Therefore, the LSA metric determines the similarity of two words by measuring the distance between the vectors of the words using a cosine function, by replacing Equation (2) in Equation (1) with:

$$f_{ss}(w_i, w_j) = cosine(V_{w_i}, V_{w_j}) \quad (3)$$

WE metric. Recently, as highlighted above, Word Embeddings have been shown to produce more effective word representations than LSA. Hence, we propose the use of WE vectors V_{w_i}, obtained from a pre-trained Word Embedding model on a large text dataset. If two words are semantically similar, the *cosine* similarity – as per Equation (3) – of their word vectors is higher. We describe how we train the Word Embedding models in Section 5.

We use the methodology explained in Section 4 to examine whether the WE-based metric can capture the coherence of topics from tweets, and how well WE, PMI, and LSA metrics compare with human judgements.

4. EVALUATION METHODOLOGY

We follow Fang et al. [3], and adopt a pairwise user study to gather human preferences to evaluate the effectiveness of the aforementioned metrics. Since it is difficult for humans to generate graded coherence scores of topics, we select a pairwise study, where a human is asked to choose which of two topics is more coherent. For our study, we first generate topic pairs using the three topic modelling approaches: LDA, PAM, and TLDA. We then determine whether the metrics can accurately identify the more coherent topics compared with human coherence assessments.

Ground Truth Generation. We use pairwise comparisons for the three topic modelling approaches, specifically: LDA vs. TLDA, LDA vs. PAM and TLDA vs. PAM. Each *comparison unit* consists of a certain number of topic pairs, where each pair contains a topic from topic models T_1 and T_2, respectively. Note that T_1 and T_2 are generated using any two topic modelling approaches among the three. In the pairwise user study, a human is asked to choose the more coherent topic among two topics presented in a given topic pair. For ease of assessment, we present the human with two similar topics in a topic pair. We first randomly select a number of topics from the topic model T_1. For each selected topic, we use Equation (4) to select its closest topic in T_2, where V_t is a vector representation using the term distribution of topic t. The selected topic pair is denoted as Pair($T_1 \rightarrow T_2$). Similarly, we also generate the same number of Pair($T_2 \rightarrow T_1$) for the comparison unit (T_1, T_2). Hence, for each comparison unit, we obtain a set of topic pairs. For example, we generate Pairs(LDA→TLDA & TLDA→LDA) for the comparison unit (LDA, TLDA).

$$closest_{t_j^{T_1}} = argmax_{i<K}\ (cosine(V_{t_j^{T_1}}, V_{t_i^{T_2}})) \quad (4)$$

A given coherence metric generates a coherence score for each topic in a topic pair. Thus for each comparison unit, we have a group of data pairs. We then apply the Wilcoxon signed-rank test to compute the statistical significance level of the difference between the two sets of data sampled in order to determine the better topic model between the two approaches utilised (e.g. TLDA > LDA). Therefore the outcomes of three comparison units gives the performance ranking order of the three topic modelling approaches. For instance, we obtain the ranking order LDA(1^{st})>TLDA(2^{nd}) >PAM(3^{rd}) from LDA>TLDA, LDA>PAM & TLDA>PAM. Similar to Fang et al., we do not observe a Condorcet paradox (e.g. TLDA>LDA, LDA>PAM & PAM>TLDA) in our experiments. Turning to our user study, a topic receives a vote if it is preferred by a human. Using the vote fraction of topics, we can also obtain an ordering of the three topic modelling approaches, i.e. the human ground-truth ranking.

Comparison of Coherence Metrics. A good metric should rank the three topic modelling approaches in a high

Table 1: Two used Twitter datasets.

Name	Time Period	Users#	Tweets#
(1) NYJ	20/05/2015-19/08/2015	2,853	946,006
(2) TVD	8pm-10pm 02/04/2015	121,594	343,511

agreement with humans. First, the three topic modelling approaches are ranked using each of the deployed coherence metrics. Second, the rankings are compared to the rankings from the generated ground-truth to identify which of the metrics agree most with the human assessments.

5. DATASETS & EXPERIMENTAL SETUP

Datasets. In this paper, we use the same two Twitter datasets from [3]. The first dataset[1] is comprised of the tweets of 2,852 newspaper journalists in the US state of New York posted from 20 May 2015 to 19 August 2015, denoted here as NYJ. The second dataset consists of tweets related to the first TV political leaders debate during the UK General Election held in April 2015, denoted as TVD[2]. Details of these two datasets are shown in Table 1.

Metrics Setup. We deploy 12 coherence metrics in total, which are implemented by two external sources: Wikipedia and a separate background Twitter dataset. The background Twitter dataset, which is also identical to the one used in [3], represents 1%-5% random tweets crawled from 01 Jan 2015 to 30 June 2015. Following [3], we remove stopwords, terms occurring in less than 20 tweets, and the retweets. The remaining tweets (30,151,847) are used to pre-calculate the PMI data, LSA word representation, and the WE models. The setup of the 12 metrics is described below:
Existing PMI/LSA metrics (4). We use the LSA word representation (1M tokens) and the PMI data (179M word pairs) from the *SEMILAR*[3] platform to implement the Wikipedia PMI/LSA-based metrics (W-PMI/W-LSA). For the Twitter PMI/LSA-based metrics (T-PMI/T-LSA), the Twitter background data contains 609k tokens and 354M word pairs.
WE metrics using GloVe (4). These metrics are instantiated using Word Embedding models from Wikipedia[4] and Twitter, pre-trained using the *GloVe* [12] tool. The metrics are denoted as G-W-WE$_{d=\{200,300\}}$ and G-T-WE$_{d=\{100,200\}}$, respectively, where $d = 100$ means that the size of word vectors in the WE model is 100.
WE metrics using word2vec (4). These metrics use Word Embedding models newly trained using the separate Twitter background dataset, but making use of the *word2vec*[5] tool. We denote the coherence metrics using our newly trained WE models (504K tokens) as T-WE$^{w=\{1,3\}}_{d=\{200,500\}}$, where w is the size of the context window size in the trained models.

We noticed that the unstemmed WE word representation performs poorly in our experiments. Hence, we stem the words in our 4 newly trained (*word2vec*) WE models. Note that the WE models of *GloVe* are not stemmed. We chose to use WE models with different pre-set parameters (e.g. context window and vector size) as we wish to examine whether these parameters affect the coherence evaluation task.

Topic Pairs Setup. Mallet[6] and Twitter LDA[7] are used to implement the three topic modelling approaches for the two Twitter datasets. The LDA parameters α and β are set to $50/K$ and 0.01 according to [14], and for TLDA $\gamma = 20$ according to [15]. Since the NYJ dataset contains many topics given the length of time and the fact that journalists discuss many issues, we use a high number of topics, $K = 100$. On the other hand, because the TVD dataset covers only one debate and the accompanying 2 hours' tweets, we use a smaller number of topic, $K = 30$. Each topic modelling approach is repeated 5 times. Therefore, for each topic modelling approach, we obtain 500 (150) topics in the NYJ (TVD) dataset, respectively. We use the methodology described in Section 4 to generate 100 topic pairs for each comparison unit. In total, we obtain 600 topic pairs from the two used Twitter datasets. In Section 6, we explain how we perform the pairwise user study using these 600 topic pairs.

6. PAIRWISE USER STUDY

We now describe how we use CrowdFlower[8] workers to perform the topic preference task while ensuring job quality.

Job Description. We show a CrowdFlower worker the top 10 words (ranked by their probabilities in a topic) of 2 topics in a topic pair, and the 3 most retweeted tweets in these 2 topics. We ask the workers to select the more coherent topic among the two presented using these 10 words. We describe a more coherent topic as one that is less mixed and that can be easily interpreted. The workers are instructed to take into account: 1) the number of semantically similar words (e.g. Knicks & basketball) among the 10 shown words, 2) whether the presented words suggest a mixed topic (i.e. more than one discussion) and 3) whether the displayed words gives more information about a discussion. To reach a decision, a worker can also use three associated tweets for the two topics. We give two rules for using these tweets: 1) whether the 10 shown words are reflected by their tweets and 2) whether these tweets are related with the two topics. We collect 5 judgements from 5 different workers for each topic pair. For each judgement, we paid a worker $0.05.

Quality Control. To ensure quality control, only those workers who passed a test were allowed to enter the topic preference task. For the test, we choose a number of topic pairs. The topic preference of the selected topic pairs were verified in advance, and they were used to set the test questions for the quality control. The worker must have maintained more than 70% accuracy on the test questions through the whole task, otherwise their judgements were nullified. Overall, we used 168 trusted workers for this user study.

7. RESULTS

We first demonstrate whether the 12 used coherence metrics can differentiate the three topic modelling approaches in comparison with human assessments. We also show whether they can distinguish the more coherent topic from a topic pair in a manner similar to that of humans.

The column "Ranking Order Matching" in Table 2 shows the extent to which the ranking order of the 12 metrics exactly or partially matches that of human judgements for our two Twitter datasets. The human ground-truth ranking order is LDA^{1st}>TLDA^{2nd} >PAM^{3rd} for the NYJ dataset, and TLDA^{1st}>LDA$^{2nd/3rd}$>PAM$^{2nd/3rd}$ for the TVD one. If there are no significant differences between two topic modelling approaches, they share the same rank in the table. If a metric receives a ranking such as LDA^{1st}>TLDA$^{2nd/3rd}$ >PAM$^{2nd/3rd}$ in the NYJ dataset, we say that the ranking order partially matches the human ground-truth one.

We find, first, that the PMI-based metrics differentiate the three topic modelling approaches very well, with the ex-

[1] This dataset was collected by tracking the journalists' Twitter handles using Twitter Streaming API. [2] Collected by searching for debate-related hashtags using the Twitter Streaming API [3] semanticsimilarity.org [4] It also contains *English Gigaword V5.* [5] deeplearning4j.org [6] mallet.cs.umass.edu [7] github.com/minghui/Twitter-LDA

[8] crowdflower.com

Table 2: The matching of the ranking order of the three topic modelling approaches from each metric and from humans, and the agreement of the topic preferences between each metric and humans.

Metrics	Ranking Order Matching		Preferences Agreement	
	NYJ	TVD	NYJ	TVD
(1) T-PMI	✔	✔	75.7%▲	51.0%
(2) W-PMI	✔	†	65.5%▲	56.0%
(3) T-LSA	✖	✖	47.6%	46.6%
(4) W-LSA	†	†	61.3%▲	47.3%
(5) G-T-WE$_{d=100}$	✖	✖	44.6%	39.3%
(6) G-T-WE$_{d=200}$	✖	✖	48.0%	39.3%
(7) G-W-WE$^{w=10}_{d=200}$	✖	✖	48.0%	43.3%
(8) G-W-WE$^{w=10}_{d=300}$	†	✖	50.6%	46.0%
(9) T-WE$^{w=1}_{d=200}$	†	†	65.3%▲	60.3%▲
(10) T-WE$^{w=1}_{d=500}$	†	†	66.7%▲	60.0%▲
(11) T-WE$^{w=3}_{d=200}$	†	†	64.3%▲	61.0%▲
(12) T-WE$^{w=3}_{d=500}$	†	†	70.7%▲	61.0%▲

"✔"/"†" means that the ranking order from a metric **exactly/partly** matches that from human judgements. "✖" indicates that the ranking order from a metric does not match that from human judgements or the metric cannot give a significant ranking order. "▲" represents that a metric have a high agreement ($\geq 60\%$) with humans.

ception of the W-PMI metric in the TVD dataset, whose ranking order only partially matches the ground-truth order. Second, all of our WE-based metrics using our trained WE models partially match the ground-truth ranking order. This finding indicates that the WE-based metrics have the ability to differentiate the three topic modelling approaches comparably to humans. However, the WE-based metrics using the $GloVe$ pre-trained models do not differentiate well between the modelling approaches. One possible reason is that the words in those WE models are not stemmed. The coherence metric focuses on capturing the semantic similarity between words rather than their syntax. The other possible reason is that words on Twitter are different from the words in Wikipedia, as tweets contain abbreviations, misspellings and hashtags. Moreover, if the time period of the Twitter background dataset does not match that of the testing datasets form which we extract the topics, then the trained WE models may not adequately capture the semantic similarity of words, as is likely the case for $GloVe$.

The column "Preference Agreement" in Table 2 lists the agreement rate on choosing the preferred topics from 300 topic pairs in the two Twitter datasets between each metric and humans. As there are three options ("Topic 1", "Topic 2" and "No preference") in the topic preference task, the baseline agreement rate is 33.3%. We obverse that the WE-based metrics using our trained WE models have consistently high agreement rates across both datasets. In addition, while most of the metrics do not perform well on the TVD dataset, the WE-based metrics using our trained WE models have the hignest agreement with humans. We also find that a slightly higher dimension and context window are likely to get a better agreement rate, such as T-WE$^{w=3}_{500}$. It is interesting to note that unlike the newly trained WE models, the WE-based metric using WE models from $GloVe$ do not have a high agreement with humans.

In summary, the WE-based metrics can effectively capture the coherence of topics from tweets, with a high agreement with humans. The WE-based metrics perform more robustly across the two Twitter datasets. Besides, the WE based metrics have an additional important benefit. In contrast to the PMI-based coherence that leverages the PMI data of a few hundreds millions of word pairs, the WE-based metrics use only a few hundred thousand (the size of vocabulary) words' vectors. Moreover, since *skip gram* is highly parallelizable, it is much faster to train a WE model than an LSA one [6]. Hence, we conclude that the WE-based metrics are more robust, efficient, and practical to use than the PMI/LSA-based ones.

8. CONCLUSIONS

We proposed a new *Word Embedding*-based topic coherence metric, and instantiated it using 8 different WE models. To identify the usefulness of these WE-based metrics, we conducted a large-scale pairwise user study to gauge human preferences. We examined which of the 8 WE-based metrics and the 4 existing PMI/LSA-based metrics align best with human assessments. We found that the WE-based metrics can effectively capture the coherence of topics from tweets. In addition, they performed more robustly and more efficiently than the PMI/LSA-based metrics. In future work, we will explore how the *Word Embedding* training parameters affect the coherence evaluation task.

9. REFERENCES

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3, 2003.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[3] A. Fang, C. Macdonald, I. Ounis, and P. Habel. Topics in tweets: A user study of topic coherence metrics for Twitter data. In *Proc. of ECIR*, 2016.

[4] A. Fang, I. Ounis, P. Habel, C. Macdonald, and N. Limsopatham. Topic-centric classification of Twitter user's political orientation. In *Proc. of SIGIR*, 2015.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of NAS*, 2004.

[6] R. Lebret and R. Collobert. N-gram-based low-dimensional representation for document classification. In *Proc. of ICLP*, 2015.

[7] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proc. of ICML*, 2006.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. efficient estimation of word representations in vector space. In *Proc. of ICLR workshop*, 2013.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*, 2013.

[10] T. Mikolov, W. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proc. of HLT-NAACL*, 2013.

[11] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *Proc. of NAACL*, 2010.

[12] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Proc. of EMNLP*, 2014.

[13] G. Recchia and M. N. Jones. More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior research methods*, 41:647–656, 2009.

[14] M. Steyvers and T. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427:424–440, 2007.

[15] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing Twitter and traditional media using topic models. In *Proc. of ECIR*, 2011.

[16] W. Y. Zou, R. Socher, D. M. Cer, and C. D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proc. of EMNLP*, 2013.