Lexical Comparison Between Wikipedia and Twitter Corpora by Using Word Embeddings

Luchen Tan[1], **Haotian Zhang**[1*], **Charles L.A. Clarke**[1], **and Mark D. Smucker**[2]
[1]David R. Cheriton School of Computer Science, University of Waterloo, Canada
{luchen.tan, haotian.zhang, claclark}@uwaterloo.ca
[2]Department of Management Sciences, University of Waterloo, Canada
mark.smucker@uwaterloo.ca

Abstract

Compared with carefully edited prose, the language of social media is informal in the extreme. The application of NLP techniques in this context may require a better understanding of word usage within social media. In this paper, we compute a word embedding for a corpus of tweets, comparing it to a word embedding for Wikipedia. After learning a transformation of one vector space to the other, and adjusting similarity values according to term frequency, we identify words whose usage differs greatly between the two corpora. For any given word, the set of words closest to it in a particular embedding provides a characterization for that word's usage within the corresponding corpora.

1 Introduction

Users of social media typically employ highly informal language, including slang, acronyms, typos, deliberate misspellings, and interjections (Han and Baldwin, 2011). This heavy use of nonstandard language, as well as the overall level of noise on social media, creates substantial problems when applying standard NLP tools and techniques (Eisenstein, 2013). For example, Kaufmann and Kalita (2010) apply machine translation methods to convert tweets to standard English in an attempt to ameliorate this problem. Similarly, Baldwin et al. (2013) and Han et al. (2012) address this problem by generating corrections for irregularly spelled words in social media.

In this short paper, we continue this line of research, applying word embedding to the problem of translating between the informal English of social media, specifically Twitter, and the formal English of carefully edited texts, such as those found in Wikipedia. Starting with a large collection of tweets and a copy of Wikipedia, we construct word embeddings for both corpora. We then generate a transformation matrix, mapping one vector space into another. After applying a normalization based on term frequency, we use distances in the transformed space as an indicator of differences in word usage between the two corpora. The method identifies differences in usage due to jargon, contractions, abbreviations, hashtags, and the influence of popular culture, as well as other factors. As a method of validation, we examine the overlap in closely related words, showing that distance after transformation and normalization correlates with the degree of overlap.

2 Related Work

Mikolov et al. (2013b) proposed a novel neural network model to train continuous vector representation for words. The high-quality word vectors obtained from large data sets achieve high accuracy in both semantic and syntactic relationships (Goldberg and Levy, 2014).

Some probabilistic similarity measures, based on Kullback-Leibler (KL) divergence (or relative entropy), give an inspection of relative divergence between two probability distributions of corpus (Kullback and Leibler, 1951; Tan and Clarke, 2014). For a given token, KL divergence measures the distribution divergence of this word in different corpora according to its corresponding probability. Intuitively, the value for KL divergence increases as two distributions become more different. Verspoor et al. (2009) found that KL divergence could be applied to analyze text in terms of two characteristics: the magnitude of the differences, and the semantic nature of the characteristic words.

Subašić and Berendt (2011) applied a symmetrical variant of KL divergence, the Jensen-Shannon (JS) divergence (Lin, 1991), to compare various aspects of the corpora such as language

[*] Luchen Tan and Haotian Zhang contributed equally to this work.

Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers), pages 657–661,
Beijing, China, July 26-31, 2015. ©2015 Association for Computational Linguistics

divergence, headline divergence, named-entity divergence and sentiment divergence. As for the applications derived from above methods, Tang et al. (2011) studied the lexical semantics and sentiment tendency of high frequency terms in each corpus by comparing microblog texts with general articles. Baldwin et al. (2013) analyzed non-standard language on social media in the aspects of lexical variants, acronyms, grammaticality and corpus similarity. Their results revealed that social media text is less grammatical than edited text.

3 Methods of Lexical Comparison

Mikolov et al. (2013a) construct vector spaces for various languages, including English and Spanish, finding that the relative positions of semantically related words are preserved across languages. We adapt this result to explore differences between corpora written in a single language, specifically to explore the contrast between the highly informal language used in English-language social media with the more formal language used in Wikipedia. We assume that there exists a *linear* transformation relationship between the vectors for the most frequent words from each corpus. Working with these frequent terms, we learn a linear projection matrix that maps source to target spaces. We hypothesize that usage of those words appearing far apart after this transformation differs substantially between the two corpora.

Let $a \in \mathbb{R}^{1 \times d}$ and $b \in \mathbb{R}^{1 \times d}$ be the corresponding source and target word vector representation with dimension d. We construct a source matrix $A = [a_1^T, a_2^T, ..., a_c^T]^T$ and a target matrix $B = [b_1^T, b_2^T, ..., b_c^T]^T$, composed of vector pairs $\{a_i, b_i\}_{i=1}^c$, where c is the size of the vocabulary common between the source and target corpora. We order these vectors according to frequency in the target corpus, so that a_i and b_i correspond to the i-th most common word in the target corpus.

These vectors are used to learn a linear transformation matrix $M \in \mathbb{R}^{d \times d}$. Once this transformation matrix M is obtained, we can transform any a_i to $a_i' = a_i M$ in order to approximate b_i. The linear transformation can be depicted as:

$$AM = B \tag{1}$$

Following the solution provided by (Mikolov et al., 2013a), M can be approximately computed by using stochastic gradient descent:

$$\min_M \sum_{i=1}^{n} \parallel a_i M - b_i \parallel^2 \tag{2}$$

where we limit the training process to the top n terms.

After the generation of M, we calculate $a_i' = a_i M$ for each word. For each a_i where $i > n$, we determine the distance between a_i' and b_i:

$$Sim(a_i', b_i), n \leq i \leq c. \tag{3}$$

Let Z be the set of these words ordered by distance, so that z_j is the word with the j-th greatest distance between the corresponding a' and b vectors. For the experiments reported in this paper, we used cosine distance to calculate this Sim metric.

4 Experiments

In this section, we describe the results of applying our method to Twitter and Wikipedia.

4.1 Experimental Settings

The Wikipedia dataset for our experiments consists of all English Wikipedia articles downloaded from MediaWiki data dumps[1]. The Twitter dataset was collected through the Twitter Streaming API from November 2013 to March 2015. We restricted the dataset to English-language tweets on the basis of the language field contained in each tweet. To obtain distributed word representation for both corpora, we trained word vectors separately by applying the *word2vec*[2] tool, a well-known implementation of word embedding.

Before applying the tool, we cleaned Wikipedia and Twitter corpora. The clean version of Wikipedia retains only normally visible article text on Wikipedia web pages. The Twitter clean version removes HTML code, URLs, user mentions(@), the # symbol of hashtags, and all the retweeted tweets. The sizes of document and vocabulary in both corpora are listed in Table 1.

Corpora	# Documents	# Vocabulary
Wikipedia	3,776,418	7,267,802
Twitter	263,572,856	13,622,411

Table 1: Corpora sizes

[1] https://dumps.wikimedia.org/enwiki/20150304/

[2] https://code.google.com/p/word2vec/

There are two major parameters that affect *word2vec* training quality: the dimensionality of word vectors, and the size of the surrounding words window. We choose 300 for our word vector dimensionality, which is typical for training large dataset with *word2vec*. We choose 10 words for the window, since tweet sentence length is 9.2 ± 6.4 (Baldwin et al., 2013).

4.2 Visualization

In Figure 1, we visualize the vectors of some most common English words by applying principal component analysis (PCA) to the vector spaces. The words “and”, “is”, “was” and “by” have similar geometric arrangements in Wikipedia and in Twitter, since these common words are not key differentiators for these corpora. On the other hand, the pronouns “I” and “you”, are heavily used in Twitter but rarely used in Wikipedia. Despite this difference in term frequency, after transformation, the vectors for these terms appear close together.

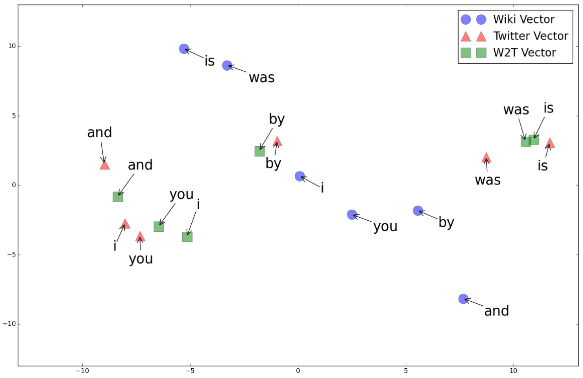

Figure 1: Word representations in Wikipedia, Twitter and transformed vectors after mapping from Wikipedia to Twitter.

4.3 Results

As our primary goal, we hope to demonstrate that our transformation method reflects meaningful lexical usage differences between Wikipedia and Twitter. To train our space transformation matrix, we used the top $n = 1,000$ most frequent words from the 505,121 words that appear in both corpora. The transformation can be either from Twitter to Wikipedia (*T2W*) or the opposite direction *W2T*. We observed that the two transformation matrices are not exactly the same, but they produce similar results. Mikolov et al. (2013c) suggest that a simple vector offset method based on cosine distance was remarkably effective to search both syntactic and semantic similar words. They also report that cosine similarity preformed well, given that the embedding vectors are all normalized to unit norm.

Figure 2 illustrates how *T2W* word vectors are similar to their original word vectors. For the purpose of explaining Figure 2, we define new notation as follows: Let $\mathcal{T}$ and $\mathcal{W}$ be the word sets of Twitter and Wikipedia respectively, and let $\mathcal{C} = \mathcal{T} \cap \mathcal{W}$. Denote the document frequency of a word t in the Twitter corpus as $df(t)$. Sorting the whole set $\mathcal{C}$ by $df(t)$ in an ascending order, we obtain a sequence $\bar{S} = \{c_0, \cdots, c_{m-1}\}$, where $c_i \in \mathcal{C}$; $m = 505,121$; and $df(c_i) \leq df(c_j)$, $\forall i < j$. We partition the sequence $\bar{S}$ into 506 buckets, with a bucket size $b = 1000$. $B_i = \{c_{i*b}, \cdots, c_{(i+1)*b-1}\}$ represents the i-th bucket. We number the curves in Figure 2 from the top to the bottom. The points on the i-th curve demonstrates the cosine similarity of the $(i-1)*100$-th word in each bucket. From this figure, it is apparent that words with higher frequencies have higher average cosine similarity than those words with lower frequencies. Since our goal is to find words with lower than average similar, we apply the median curve of Figure 2 to adjust word distances.

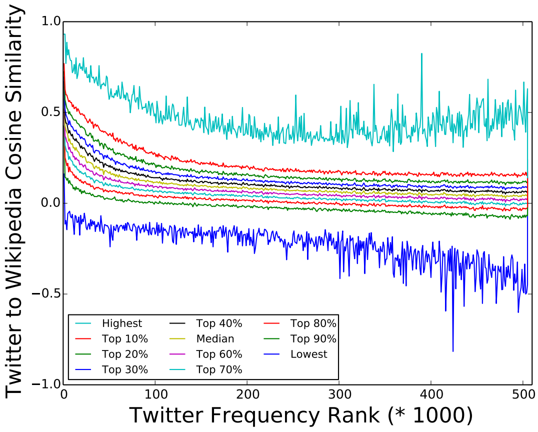

Figure 2: T2W transformated similarity curves.

Defining **adjusted distance** as $D_{adjusted}(t)$ of a given word t, we calculate the cosine distance between t and the median point c_{median} from its corresponding bucket B_i.

$$D_{adjusted}(t) = Sim(c_{median}) - Sim(t) \quad (4)$$

where the index of median point should be $i * b + b/2$. A negative adjusted distance value means the word is more similar than at least half of

Word	Twitter Most Similar	Wikipedia Most Similar
bc	because bcus bcuz cuz cos	bce macedon hellenistic euthydemus ptolemaic
ill	ll imma ima will youll	unwell sick frail fated bedridden
cameron	cam nash followmecam camerons callmecam	gillies duncan mckay mitchell bryce
mentions	unfollow reply respond strangerswelcomed offend	mentions mentioned mentioning reference attested
miss	misss love missss missssss imiss	pageant pageants titlehoder titlehoders pageantopolis
yup	yep yupp yeah yea yepp	chevak yupik gwaii tlingit nunivak
taurus	capricorn sagittarius pisces gemini scorpio	poniatovii scorpio subcompact sagittarius chevette

Table 2: Characteristic Words in Twitter Corpora

words in its bucket. On the other hand, the words that are less similar than at least half of words in their buckets have positive adjusted distance values. The larger an adjusted distance, the less similar the word is between the corpora.

4.4 Examples

Table 2 provides some examples of common words with large adjusted distance, suggesting that their usage in the two corpora are quite different. For each of these words, the example shows the closest terms to that word in the two corpora. In Twitter, "bc" is frequently an abbreviation for "because", while in Wikipedia "bc" is more commonly used as part of dates, e.g. 900 BC. Similarly, in Twitter "ill" is often a misspelling of the contraction "I'll", rather than a synonym for sickness, as in Wikipedia. In Twitter, the most similar words to "cameron" relate to a YouTube personality, whereas in Wikipedia they relate to notable Scottish persons. In Wikipedia, "miss" is related to beauty pageants, while in Twitter it is related to expressions of affection ("I missssss you"). The other examples also have explanations related to popular culture, jargon, slang, and other factors.

5 Validation

To validate our method of comparing lexical distinctions in the two corpora, we employ a ranking similarity measurement. Within a single corpus, the most similar words to a word t can be generated by ranking cosine distance to t. We then determine the overlap between the most similar words to t from Twitter and Wikipedia. The more the two lists overlap, the greater the similarity between the words in the two corpora. Our hypothesis is that larger rank similarity correlates with smaller adjusted distance.

Rank biased overlap (RBO) provides a rank similarity measure designed for comparisons between top-weighted, incomplete and indefinite rankings. Given two ranked lists, A and B, let $A_{1:k}$ and $B_{1:k}$ denote the top k items in A and B (Webber et al., 2010). RBO defines the *overlap* between A and B at depth k as the size of the intersection between these lists at depth k and defines the agreement between A and B at depth k as the overlap divided by the depth. Webber et al. (2010) define RBO as a weighted average of agreement across depths, where the weights decay geometrically with depth, reflecting the requirement for top weighting:

$$RBO = (1-\varphi)\sum_{k=1}^{\infty}\varphi^{k-1}\frac{|A_{1:k}\cap B_{1:k}|}{k} \qquad (5)$$

Here, φ is a persistence parameter. As suggested by Webber et al., we set $\varphi = 0.9$. In practice, RBO is computed down to some fixed depth K. We select $K = 50$ for our experiments. For a word t, we compute RBO value between its top 50 similar words in Wikipedia and top 50 similar words in Twitter.

In Figure 3, we validate consistency between results of our space transformation method and RBO. For the top 5,000 terms in the Twitter corpus, we sort them by their adjusted distance value. Due to properties of RBO, there are many zero RBO values. To illustrate the density of these zero overlaps, we smooth our plot by sliding a 100-word window with a step of 10 words. As shown sharply in the figure, RBO and adjusted distance is negatively correlated.

6 Conclusion

This paper analyzed the lexical usage difference between Twitter microblog corpus and Wikipedia corpus. A word-level comparison method based on word embedding is employed to find the characterisic words that particularly discriminating corpora. In future work, we plan to introduce this method to normalize the nonstandard language used in Twitter, applying the methods to problems in search and other areas.

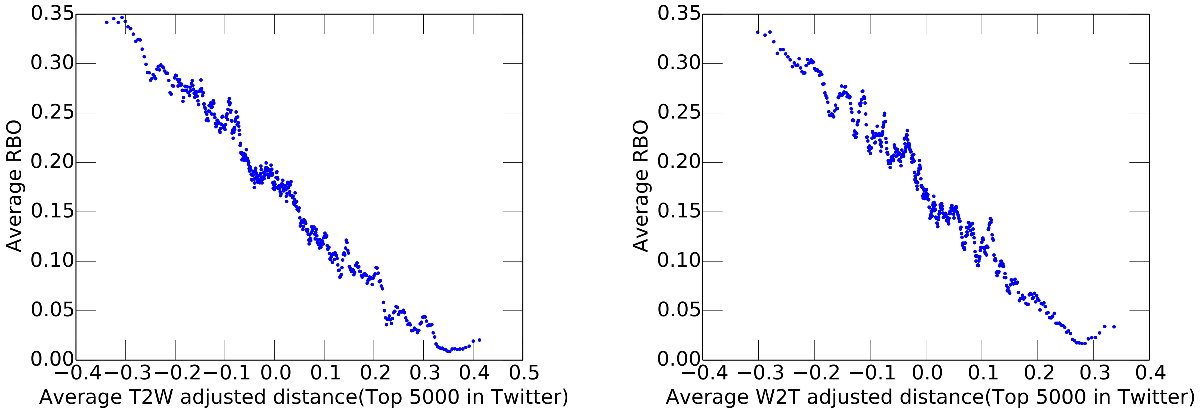

Figure 3: T2W and W2T negative correlation between adjusted distance and RBO.

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Bo Han and Timothy Baldwin. 2011. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 421–432. Association for Computational Linguistics.

Max Kaufmann and Jugal Kalita. 2010. Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, pages 79–86.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *HLT-NAACL*, pages 746–751.

Ilija Subašić and Bettina Berendt. 2011. Peddling or creating? investigating the role of twitter in news reporting. In *Advances in Information Retrieval*, pages 207–213. Springer.

Luchen Tan and Charles L.A. Clarke. 2014. Succinct queries for linking and tracking news in social media. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1883–1886, New York, NY, USA. ACM.

Yi-jie Tang, Chang-Ye Li, and Hsin-Hsi Chen. 2011. A comparison between microblog corpus and balanced corpus from linguistic and sentimental perspectives. In *Analyzing Microtext*.

Karin Verspoor, K Bretonnel Cohen, and Lawrence Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC bioinformatics*, 10(1):183.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20.