Unsupervised Word Usage Similarity in Social Media Texts

Spandana Gella,♣ Paul Cook,♣ and Bo Han♠♣
♠♣ NICTA Victoria Research Laboratory
♣ Department of Computing and Information Systems, The University of Melbourne
sgella@student.unimelb.edu.au, paulcook@unimelb.edu.au, hanb@student.unimelb.edu.au

Abstract

We propose an unsupervised method for automatically calculating word usage similarity in social media data based on topic modelling, which we contrast with a baseline distributional method and Weighted Textual Matrix Factorization. We evaluate these methods against a novel dataset made up of human ratings over 550 Twitter message pairs annotated for usage similarity for a set of 10 nouns. The results show that our topic modelling approach outperforms the other two methods.

1 Introduction

In recent years, with the growing popularity of social media applications, there has been a steep rise in the amount of “post”-based user-generated text (including microblog posts, status updates and comments) (Bennett, 2012). This data has been identified as having potential for applications ranging from trend analysis (Lau et al., 2012a) and event detection (Osborne et al., 2012) to election outcome prediction (O’Connor et al., 2010). However, given that posts are generally very short, noisy and lacking in context, traditional NLP approaches tend to perform poorly over social media data (Hong and Davison, 2010; Ritter et al., 2011; Han et al., 2012).

This is the first paper to address the task of lexical semantic interpretation in microblog data based on word usage similarity. Word usage similarity (USIM: Erk et al. (2009)) is a relatively new paradigm for capturing similarity in the usages of a given word independently of any lexicon or sense inventory. The task is to rate on an ordinal scale the similarity in usage between two different usages of the same word. In doing so, it avoids common issues in conventional word sense disambiguation, relating to sense underspecification, the appropriateness of a static sense inventory to a given domain, and the inability to capture similarities/overlaps between word senses. As an example of USIM, consider the following pairing of Twitter posts containing the target word paper:

1. Deportation of Afghan Asylum Seekers from Australia: This paper aims to critically evaluate a newly signed agree.

2. @USER has his number on a piece of paper and I walkd off!

The task is to predict a real-valued number in the range [1, 5] for the similarity in the respective usages of paper, where 1 indicates the usages are completely different and 5 indicates they are identical.

In this paper we develop a new USIM dataset based on Twitter data. In experiments on this dataset we demonstrate that an LDA-based topic modelling approach outperforms a baseline distributional semantic approach and Weighted Textual Matrix Factorization (WTMF: Guo and Diab (2012a)). We further show that context expansion using a novel hashtag-based strategy improves both the LDA-based method and WTMF.

2 Related Work

Word sense disambiguation (WSD) is the task of determining the particular sense of a word from a given set of pre-defined senses (Navigli, 2009). It
contrasts with word sense induction (WSI), where
the senses of a given target word are induced from
an unannotated corpus of usages, and the induced
senses are then used to disambiguate each token us-
age of the word (Manandhar et al., 2010; Lau et
al., 2012b). WSD and WSI have been the predomi-
nant paradigms for capturing and evaluating lexical
semantics, and both assume that each usage corre-
sponds to exactly one of a set of discrete senses of
the target word, and that any prediction other than
the “correct” sense is equally wrong.

Erk et al. (2009) showed that, given a sense in-
vventory, there is a high likelihood of multiple senses
being compatible with a given usage, and proposed
USIM as a means of capturing the similarity in us-
age between a pairing of usages of a given word.
As part of their work, they released a dataset, which
Lui et al. (2012) recently developed a topic mod-
elling approach over. Based on extensive experi-
mentation, they demonstrated the best results with
a single topic model for all target words based on
full document context. Our topic modelling-based
approach to USIM builds off the approach of Lui
et al. (2012). Guo and Diab (2012a) observed that,
when applied to short texts, the effectiveness of la-
tent semantic approaches can be boosted by expand-
ing the text to include “missing” words. Based on
this, they proposed Weighted Textual Matrix Factor-
ization (WTMF), based on weighted matrix factor-
ization (Srebro and Jaakkola, 2003). Here we ex-
periment with both LDA based topic modeling and
WTMF to estimate word similarities in twitter data.
LDA based topic modeling has been earlier studied
on Twitter data for tweet classification (Ramage et
al., 2010) and tweet clustering (Jin et al., 2011).

3 Data Preparation

This section describes the construction of the USIM-
tweet dataset based on microblog posts (“tweets”) from Twitter. We describe the pre-processing steps
taken to sample the tweets in our datasets, outline
the annotation process, and then describe the back-
ground corpora used in our experiments.

3.1 Data preprocessing

Around half of Twitter is non-English (Hong et al.,
2011), so our first step was to automatically identify
English tweets using langid.py (Lui and Baldwin,
2012). We next performed lexical normalization using
the dictionary of Han et al. (2012) to con-
vert lexical variants (e.g., tmrw) to their standard
forms (e.g., tomorrow) and reduce data sparseness.
As our target words, we chose the 10 nouns from
the original USIM dataset of Erk et al. (2009) (bar,
charge, execution, field, figure, function, investiga-
tor, match, paper, post), and identified tweets con-
taining the target words as nouns using the CMU
Twitter POS tagger (Owoputi et al., 2012).

3.2 Annotation Settings and Data

To collect word usage similarity scores for Twitter
message pairs, we used a setup similar to that of
Erk et al. (2009) using Amazon Mechanical Turk: we asked the annotators to rate each sentence pair
with an integer score in the range [1, 5] using sim-
ilar annotation guidelines to Erk et al. We ran-
domly sampled twitter messages from the TREC
2011 microblog dataset,1 and for each of our 10
nouns, we collected 55 pairs of messages satisfying
the preprocessing described in Section 3.1. These
55 pairs are chosen such that each tweet has at least
4 content words (nouns, verbs, adjectives and ad-
verbs) and at least 70+% of its post-normalized to-
kens in the Aspell dictionary (v6.06); these restric-
tions were included in an effort to ensure the tweets
would contain sufficient linguistic content to be in-
terpretable.3 We created 110 Mechanical Turk jobs
(referred to as HITs), with each HIT containing 5
randomly-selected message pairs. For this annota-
tion the tweets were presented in their original form,
i.e., without lexical normalisation applied. Each HIT
was completed by 10 “turkers”, resulting in a total
of 5500 annotations. The annotation was carried out
by 68 turkers, each completing between 1 and 100 HITs.

To detect outlier annotators, we calculated the av-
average Spearman correlation score ($\rho$) of every an-
notator by correlating their annotation values with
every other annotator and taking the average. We

1http://trec.nist.gov/data/tweets/
2http://aspell.net/
3In future analyses we intend to explore the potential impact
of these restrictions on the resulting dataset.
Table 1: The number of tweets for each word in each background corpus (“Orig” = ORIGINAL; “Exp” = EXPANDED; RANDEXPANDED, not shown, contains the same number of tweets as EXPANDED).

<table>
<thead>
<tr>
<th>Word</th>
<th>Orig</th>
<th>Exp</th>
<th>Word</th>
<th>Orig</th>
<th>Exp</th>
</tr>
</thead>
<tbody>
<tr>
<td>bar</td>
<td>180k</td>
<td>186k</td>
<td>function</td>
<td>26k</td>
<td>27k</td>
</tr>
<tr>
<td>charge</td>
<td>41k</td>
<td>43k</td>
<td>investigator</td>
<td>17k</td>
<td>19k</td>
</tr>
<tr>
<td>execution</td>
<td>28k</td>
<td>30k</td>
<td>field</td>
<td>72k</td>
<td>75k</td>
</tr>
<tr>
<td>figure</td>
<td>28k</td>
<td>29k</td>
<td>match</td>
<td>126k</td>
<td>133k</td>
</tr>
<tr>
<td>paper</td>
<td>210k</td>
<td>218k</td>
<td>post</td>
<td>299k</td>
<td>310k</td>
</tr>
</tbody>
</table>

accepted all the annotations of annotators whose average \( \rho \) is greater than 0.6; this corresponded to 95% of the annotators. Two annotators had a negative average \( \rho \) and their annotations (only 4 HITs total) were discarded. For the other annotators (i.e., \( 0 \leq \rho \leq 0.6 \)), we accepted each of their HITs on a case by case basis; a HIT was accepted only if at least 2 out of 5 of the annotations for that HIT were within \( \pm 2.0 \) of the mean for that annotation based on the judgments of the other turkers. (21 HITs were discarded using this heuristic.) We further eliminated 7 HITs which have incomplete judgments. In total only 32 HITs (of the 1100 HITs completed) were discarded through these heuristics. The weighted average Spearman correlation over all annotators after this filtering is 0.681, which is somewhat higher than the inter-annotator agreement of 0.548 reported by Erk et al. (2009). This dataset is available for download.

3.3 Background Corpus

We created three background corpora based on data from the Twitter Streaming API in February 2012 (only tweets satisfying the preprocessing steps in Section 3.1 were chosen).

**ORIGINAL:** 1 million tweets which contain at least one of the 10 target nouns;

**EXPANDED:** ORIGINAL plus an additional 40k tweets containing at least 1 hashtag attested in ORIGINAL with an average frequency of use of 10–35 times/hour (medium frequency);

**RANDEXPANDED:** ORIGINAL plus 40k randomly sampled tweets containing the same target nouns.

We select medium-frequency hashtags because low-frequency hashtags tend to be ad hoc and non-thematic in nature, while high-frequency hashtags are potentially too general to capture usage similarity. Statistics for ORIGINAL and EXPANDED/RANDEXPANDED are shown in Table 1. RANDEXPANDED is sampled such that it has the same number of tweets as EXPANDED.

4 Methodology

We propose an LDA topic modelling-based approach to the USIM task, which we contrast with a baseline distributional model and WTMF. In all these methods, the similarity between two word usages is measured using cosine similarity between the vector representation of each word usage.

4.1 Baseline

We represent each target word usage in a tweet as a second-order co-occurrence vector (Schütze, 1998). A second-order co-occurrence vector is built from the centroid (summation) of all the first-order co-occurrence vectors of the context words in the same tweet as the target word.

The first-order co-occurrence vector for a given target word represents the frequency with which that word co-occurs in a tweet with other context words. Each first-order vector is built from all tweets which contain a context word and the target word categorized as noun in the background corpus, thus sensitizing the first-order vector to the target word. We use the most frequent 10000 words (excluding stopwords) in the background corpus as our first-order vector dimensions/context words. Context words (dimensions) in the first-order vectors are weighted by mutual information.

Second-order co-occurrence is used as the context representation to reduce the effects of data sparseness in the tweets (which cannot be more than 140 codepoints in length).

4.2 Weighted Textual Matrix Factorization

WTMF (Guo and Diab, 2012b) addresses the data sparsity problem suffered by many latent variable
models by predicting “missing” words on the basis of the message content, and including them in the vector representation. Guo and Diab showed WTMF to outperform LDA on the SemEval-2012 semantic textual similarity task (STS) (Agirre et al., 2012). The semantic space required for this model as applied here is built from the background tweets corresponding to the target word. We experimented with the missing weight parameter \(w_m\) of WTMF in the range \([0.05, 0.01, 0.005, 0.0005]\) and with dimensions \(K=100\) and report the best results \((w_m = 0.0005)\).

### 4.3 Topic Modelling

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a generative model in which a document is modeled as a finite mixture of topics, where each topic is represented as a multinomial distribution of words. We treat each tweet as a document. Topics sensitive to each target word are generated from its corresponding background tweets. We topic model each target word individually,\(^4\) and create a topic vector for each word usage based on the topic allocations of the context words in that usage. We use Gibbs sampling in Mallet (McCallum, 2002) for training and inference of the LDA model. We experimented with the number of topics \(T\) for each target word ranging from 2 to 500. We optimized the hyper parameters by choosing those which best fit the data every 20 iterations over a total of 800 iterations, following 200 burn-in iterations.

\(^4\)Unlike Lui et al. (2012) we found a single topic model for all target words to perform very poorly.

### 5 Results

We evaluate the above methods for word usage similarity on the dataset constructed in Section 3.2. We evaluate our models against the mean human ratings using Spearman’s rank correlation. Table 2 presents results for each method using each background corpus. The results for LDA are for the optimal setting for \(T\) (8, 5, and 20 for ORIGINAL, EXPANDED, and RANDEXPANDED, respectively). LDA is superior to both the baseline and WTMF using each background corpus. The performance of LDA improves for EXPANDED but not RANDEXPANDED, over ORIGINAL, demonstrating the effectiveness of our hashtag based corpus expansion strategy.

In Figure 1 we plot the rank correlation of LDA across all words against the number of topics \((T)\). As the number of topics increases beyond a certain number, the rank correlation decreases. LDA trained on EXPANDED consistently outperforms ORIGINAL and RANDEXPANDED for lower values of \(T\) (i.e., \(T <= 20\)).

In Table 3, we show results for LDA over each target word, for ORIGINAL and EXPANDED. (Results for RANDEXPANDED are not shown but are similar to ORIGINAL.) Results are shown for the optimal \(T\) for each lemma, and the optimal \(T\) over all lemmas. Optimizing \(T\) for each lemma gives an indication of the upperbound of the performance of LDA, and unsurprisingly gives better performance than us-

### Table 2: Spearman rank correlation \((\rho)\) for each method based on each background corpus. The best result for each corpus is shown in **bold**.

<table>
<thead>
<tr>
<th>Model</th>
<th>ORIGINAL</th>
<th>EXPANDED</th>
<th>RANDEXPANDED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.09</td>
<td>0.08</td>
<td>0.09</td>
</tr>
<tr>
<td>WTMF</td>
<td>0.02</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>LDA</td>
<td><strong>0.20</strong></td>
<td><strong>0.29</strong></td>
<td>0.18</td>
</tr>
</tbody>
</table>

![Figure 1: Spearman rank correlation \((\rho)\) for LDA for varying numbers of topics \((T)\) using different background corpora.](image)

---

\(^4\)Unless otherwise stated, all results are shown for the optimal \(T\) for each lemma, and the optimal \(T\) over all lemmas. Optimizing \(T\) for each lemma gives an indication of the upperbound of the performance of LDA, and unsurprisingly gives better performance than usual.
Table 3: Spearman’s $\rho$ using LDA for the optimal $T$ for each lemma (Per lemma) and the best $T$ over all lemmas (Global) using ORGINAL and EXPANDED. $\rho$ values that are significant at the 0.05 level are shown in bold.

<table>
<thead>
<tr>
<th>Lemma</th>
<th>ORIGINAL</th>
<th>EXPANDED</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Per lemma</td>
<td>Global</td>
</tr>
<tr>
<td></td>
<td>$\rho$ ($T$)</td>
<td>$\rho$ ($T=5$)</td>
</tr>
<tr>
<td>bar</td>
<td>0.39 (10)</td>
<td>0.28</td>
</tr>
<tr>
<td>charge</td>
<td>0.27 (30)</td>
<td>0.04</td>
</tr>
<tr>
<td>execution</td>
<td>0.43 (8)</td>
<td>0.43</td>
</tr>
<tr>
<td>field</td>
<td>0.46 (5)</td>
<td>0.33</td>
</tr>
<tr>
<td>figure</td>
<td>0.24 (150)</td>
<td>0.06</td>
</tr>
<tr>
<td>function</td>
<td>0.44 (8)</td>
<td>0.44</td>
</tr>
<tr>
<td>investigator</td>
<td>0.3 (30)</td>
<td>0.05</td>
</tr>
<tr>
<td>match</td>
<td>0.28 (5)</td>
<td>0.26</td>
</tr>
<tr>
<td>paper</td>
<td>0.29 (30)</td>
<td>0.20</td>
</tr>
<tr>
<td>post</td>
<td>0.1 (3)</td>
<td>−0.13</td>
</tr>
</tbody>
</table>

Regarding a fixed $T$ for all lemmas. This suggests that approaches that learn an appropriate number of topics (e.g., HDP, (Teh et al., 2006)) could give further improvements; however, given the size of the dataset, the computational cost of HDP could be a limitation.

Contrasting our results with a fixed number of topics to those of Lui et al. (2012), our highest rank correlation of 0.29 ($T = 5$ using EXPANDED) is higher than the 0.11 they achieved over the original USIM dataset (where the documents offer an order of magnitude more context). The higher inter-annotator agreement for USIM-tweet compared to the original USIM dataset (Section 3.2), combined with this finding, demonstrates that USIM over microblog data is indeed a viable task.

Returning to the performance of LDA relative to WTMF in Table 2, the poor performance of WTMF is somewhat surprising here given WTMF’s encouraging performance on the somewhat similar SemEval-2012 STS task. This difference is possibly due to the differences in the tasks: usage similarity measures the similarity of the usage of a target word while STS measures the similarity of two texts. Differences in domain — i.e., Twitter here and more standard text for STS — could also be a factor. WTMF attempts to alleviate the data sparsity problem by adding information from “missing” words in a text by assigning a small weight to these missing words. Because of the prevalence of lexical variation on Twitter, some missing words might be counted multiple times (e.g., coool, kool, and kewl all meaning roughly cool) thus indirectly assigning higher weights to the missing words leading to the lower performance of WTMF compared to LDA.

6 Summary

We have analysed word usage similarity in microblog data. We developed a new dataset (USIM-tweet) for usage similarity of nouns over Twitter. We applied a topic modelling approach to this task, and contrasted it with baseline and benchmark methods. Our results show that the LDA-based approach outperforms the other methods over microblog data. Moreover, our novel hashtag-based corpus expansion strategy substantially improves the results.

In future work, we plan to expand our annotated dataset, experiment with larger background corpora, and explore alternative corpus expansion strategies. We also intend to further analyse the difference in performance LDA and WTMF on similar data.

Acknowledgements

We are very grateful to Timothy Baldwin for his tremendous help with this work. We additionally thank Diana McCarthy for her insightful comments on this paper. We also acknowledge the European Erasmus Mundus Masters Program in Language and Communication Technologies from the European Commission.

NICTA is funded by the Australian government as represented by Department of Broadband, Communication and Digital Economy, and the Australian Research Council through the ICT Centre of Excellence programme.

References


