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

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task*, pages 248–253, Atlanta, Georgia, June 13-14, 2013.

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 usage of the word (Manandhar et al., 2010; Lau et al., 2012b). WSD and WSI have been the predominant paradigms for capturing and evaluating lexical semantics, and both assume that each usage corresponds 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 inventory, 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 usage 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 modelling approach over. Based on extensive experimentation, 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 latent semantic approaches can be boosted by expanding the text to include "missing" words. Based on this, they proposed Weighted Textual Matrix Factorization (WTMF), based on weighted matrix factorization (Srebro and Jaakkola, 2003). Here we experiment 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 background 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 convert 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*, *investigator*, *match*, *paper*, *post*), and identified tweets containing 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 similar annotation guidelines to Erk et al. We randomly 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 adverbs) and at least 70+% of its post-normalized tokens in the Aspell dictionary (v6.06)[2]; these restrictions were included in an effort to ensure the tweets would contain sufficient linguistic content to be interpretable.[3] We created 110 Mechanical Turk jobs (referred to as HITs), with each HIT containing 5 randomly-selected message pairs. For this annotation 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 restricted to turkers based in the United States having had at least 95% of their previous HITs accepted. In total, the annotation was carried out by 68 turkers, each completing between 1 and 100 HITs.

To detect outlier annotators, we calculated the average Spearman correlation score (ρ) of every annotator by correlating their annotation values with every other annotator and taking the average. We

[1] `http://trec.nist.gov/data/tweets/`

[2] `http://aspell.net/`

[3] In future analyses we intend to explore the potential impact of these restrictions on the resulting dataset.

Word	Orig	Exp	Word	Orig	Exp
bar	180k	186k	function	26k	27k
charge	41k	43k	investigator	17k	19k
execution	28k	30k	field	72k	75k
figure	28k	29k	match	126k	133k
paper	210k	218k	post	299k	310k

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).

accepted all the annotations of annotators whose average ρ is greater than 0.6; this corresponded to 95% of the annotators. Two annotators had a negative average ρ 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 ± 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

Model	ORIGINAL	EXPANDED	RANDEXPANDED
Baseline	0.09	0.08	0.09
WTMF	0.02	0.09	0.06
LDA	**0.20**	**0.29**	0.18

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

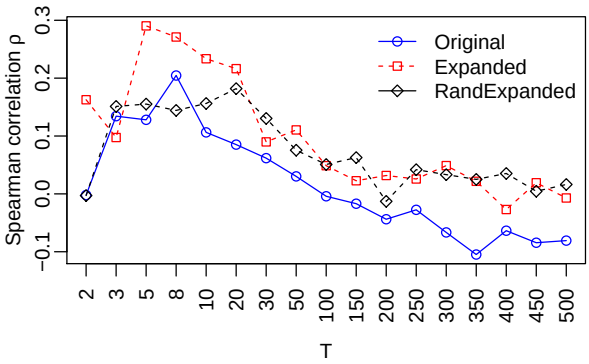

Figure 1: Spearman rank correlation (ρ) for LDA for varying numbers of topics (T) using different background corpora.

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-

Lemma	ORIGINAL		EXPANDED	
	Per lemma	Global	Per lemma	Global
	ρ (T)	ρ (T=8)	ρ (T)	ρ (T=5)
bar	0.39 (10)	**0.28**	**0.35 (50)**	0.1
charge	**0.27 (30)**	0.04	**0.33 (20)**	−0.08
execution	**0.43 (8)**	**0.43**	**0.58 (5)**	**0.58**
field	**0.46 (5)**	**0.33**	**0.53 (10)**	0.32
figure	0.24 (150)	0.06	0.24 (250)	0.14
function	**0.44 (8)**	**0.44**	**0.40 (10)**	**0.27**
investigator	**0.3 (30)**	0.05	**0.50 (5)**	**0.50**
match	**0.28 (5)**	0.26	**0.45 (5)**	**0.45**
paper	**0.29 (30)**	0.20	0.32 (30)	0.22
post	0.1 (3)	−0.13	0.2 (30)	−0.01

Table 3: Spearman's ρ using LDA for the optimal T for each lemma (Per lemma) and the best T over all lemmas (Global) using ORIGINAL and EXPANDED. ρ values that are significant at the 0.05 level are shown in **bold**.

ing 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

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393, Montreal, Canada.

Shea Bennett. 2012. Twitter on track for 500 million total users by March, 250 million active users by end of 2012. `http://www.mediabistro.com/alltwitter/twitter-active-total-users_b17655`.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2009)*, pages 10–18, Singapore.

Weiwei Guo and Mona Diab. 2012a. Modeling sentences in the latent space. In *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 864–872, Jeju, Republic of Korea.

Weiwei Guo and Mona Diab. 2012b. Weiwei: A simple unsupervised latent semantics based approach for sentence similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 586–590, Montreal, Canada.

Bo Han, Paul Cook, and Timothy Baldwin. 2012. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012*, pages 421–432, Jeju, Republic of Korea.

Liangjie Hong and Brian D Davison. 2010. Empirical study of topic modeling in twitter. In *Proc. of the First Workshop on Social Media Analytics*, pages 80–88.

Lichan Hong, Gregoria Convertino, and Ed H. Chi. 2011. Language matters in Twitter: A large scale study. In *Proceedings of the 5th International Conference on Weblogs and Social Media (ICWSM 2011)*, pages 518–521, Barcelona, Spain.

Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. 2011. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proc. of the 20th ACM International Conference on Information and Knowledge Management*, pages 775–784.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012a. On-line trend analysis with topic models: #twitter trends detection topic model online. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1519–1534, Mumbai, India.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012b. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 591–601, Avignon, France.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012) Demo Session*, pages 25–30, Jeju, Republic of Korea.

Marco Lui, Timothy Baldwin, and Diana McCarthy. 2012. Unsupervised estimation of word usage similarity. In *Proceedings of the Australasian Language Technology Workshop 2012 (ALTW 2012)*, pages 33–41, Dunedin, New Zealand.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. SemEval-2010 Task 14: Word sense induction & disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 63–68, Uppsala, Sweden.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2).

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International Conference on Weblogs and Social Media*, pages 122–129, Washington, USA.

Miles Osborne, Sasa Petrović, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2012. Bieber no more: First story detection using Twitter and Wikipedia. In *Proceedings of the SIGIR 2012 Workshop on Time-aware Information Access*, Portland, USA.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, and Nathan Schneider. 2012. Part-of-speech tagging for Twitter: Word clusters and other advances. Technical Report CMU-ML-12-107, Carnegie Mellon University.

Daniel Ramage, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*, volume 5, pages 130–137.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534, Edinburgh, UK.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Nathan Srebro and Tommi Jaakkola. 2003. Weighted low-rank approximations. In *Proceedings of the 20th International Conference on Machine Learning*, Washington, USA.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101:1566–1581.