One Sense per Tweeter ... and Other Lexical Semantic Tales of Twitter

Spandana Gella, Paul Cook and **Timothy Baldwin**
Department of Computing and Information Systems
The University of Melbourne
sgella@student.unimelb.edu.au, paulcook@unimelb.edu.au, tb@ldwin.net

Abstract

In recent years, microblogs such as Twitter have emerged as a new communication channel. Twitter in particular has become the target of a myriad of content-based applications including trend analysis and event detection, but there has been little fundamental work on the analysis of word usage patterns in this text type. In this paper — inspired by the one-sense-per-discourse heuristic of Gale et al. (1992) — we investigate user-level sense distributions, and detect strong support for "one sense per tweeter". As part of this, we construct a novel sense-tagged lexical sample dataset based on Twitter and a web corpus.

1 Introduction

Social media applications such as Twitter enable users from all over the world to create and share web content spontaneously. The resulting user-generated content has been identified as having potential in a myriad of applications including real-time event detection (Petrović et al., 2010), trend analysis (Lau et al., 2012) and natural disaster response co-ordination (Earle et al., 2010). However, the dynamism and conversational nature of the text contained in social media can cause problems for traditional NLP approaches such as parsing (Baldwin et al., 2013), meaning that most content-based approaches use simple keyword search or a bag-of-words representation of the text. This paper is a first step towards full lexical semantic analysis of social media text, in investigating the sense distribution of a range of polysemous words in Twitter and a general-purpose web corpus.

The primary finding of this paper is that there are strong user-level lexical semantic priors in Twitter, equivalent in strength to document-level lexical semantic priors, popularly termed the "one sense per discourse" heuristic (Gale et al., 1992). This has potential implications for future applications over Twitter which attempt to move beyond a simple string-based meaning representation to explicit lexical semantic analysis.

2 Related Work

The traditional approach to the analysis of word-level lexical semantics is via word sense disambiguation (WSD), where usages of a given word are mapped onto discrete "senses" in a pre-existing sense inventory (Navigli, 2009). The most popular sense inventory used in WSD research has been WordNet (Fellbaum, 1998), although its fine-grained sense distinctions have proven to be difficult to make for human annotators and WSD systems alike. This has resulted in a move towards more coarse-grained sense inventories (Palmer et al., 2004; Hovy et al., 2006; Navigli et al., 2007), or alternatively away from pre-existing sense inventories altogether, towards joint word sense induction (WSI) and disambiguation (Navigli and Vannella, 2013; Jurgens and Klapaftis, 2013).

Two heuristics that have proven highly powerful in WSD and WSI research are: (1) first sense tagging, and (2) one sense per discourse. First sense tagging is based on the observation that sense distributions tend to be Zipfian, such that if the predominant or "first" sense can be identified, simply tagging all occurrences of a given word with this sense can achieve high WSD accuracy (McCarthy et al., 2007). Unsurprisingly, there are significant differences in sense distributions across domains (cf. *cloud* in the COMPUTING and METEOROLOGICAL domains), motivating the need for unsupervised first sense learning over domain-specific corpora (Koeling et al., 2005).

One sense per discourse is the observation that a given word will often occur with a single sense across multiple usages in a single document (Gale

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 215–220,
Gothenburg, Sweden, April 26-30 2014.

et al., 1992). Gale et al. established the heuristic on the basis of 9 ambiguous words using a coarse-grained sense inventory, finding that the probability of a given pair of usages of a word taken from a given document having the same sense was 94%. However, Krovetz (1998) found that for a fine-grained sense inventory, only 67% of words exhibited the single-sense-per-discourse property for all documents in a corpus.

A radically different view on WSD is word usage similarity, whereby two usages of a given word are rated on a continuous scale for similarity, in isolation of any sense inventory (Erk et al., 2009). Gella et al. (2013) constructed a word usage similarity dataset for Twitter messages, and developed a topic modelling approach to the task, building on the work of Lui et al. (2012). To the best of our knowledge, this has been the only attempt to carry out explicit word-level lexical semantic analysis of Twitter text.

3 Dataset Construction

In order to study sense distributions of words in Twitter, we need a sense inventory to annotate against, and also a set of Twitter messages to annotate. Further, as a point of comparison for the sense distributions in Twitter, we require a second corpus; here we use the ukWaC (Ferraresi et al., 2008), a corpus built from web documents.

For the sense inventory, we chose the Macmillan English Dictionary Online[1] (MACMILLAN, hereafter), on the basis of: (1) its coarse-grained general-purpose sense distinctions, and (2) its regular update cycle (i.e. it contains many recently-emerged senses). These criteria are important in terms of inter-annotator agreement (especially as we crowdsourced the sense annotation, as described below) and also sense coverage. The other obvious candidate sense inventory which potentially satisfied these criteria was ONTONOTES (Hovy et al., 2006), but a preliminary sense-tagging exercise indicated that MACMILLAN better captured Twitter-specific usages.

Rather than annotating all words, we opted for a lexical sample of 20 polysemous nouns, as listed in Table 1. Our target nouns were selected to span the high- to mid-frequency range in both Twitter and the web corpus, and have at least 3 MACMILLAN senses. The average sense ambiguity is 5.5.

[1] `http://www.macmillandictionary.com`

band	*bar*	*case*	*charge*	*deal*
degree	*field*	*form*	*function*	*issue*
job	*light*	*match*	*panel*	*paper*
position	*post*	*rule*	*sign*	*track*

Table 1: The 20 target nouns used in this research

3.1 Data Sampling

We sampled tweets from a crawl made using the Twitter Streaming API from January 3, 2012 to February 29, 2012. The web corpus was built from ukWaC (Ferraresi et al., 2008), which was based on a crawl of the `.uk` domain from 2007. In contrast to ukWaC, the tweets are not restricted to documents from any particular country.

For both corpora, we first selected only the English documents using `langid.py`, an off-the-shelf language identification tool (Lui and Baldwin, 2012). We next identified documents which contained nominal usages of the target words, based on the POS tags supplied with the corpus in the case of ukWaC, and the output of the CMU ARK Twitter POS tagger v2.0 (Owoputi et al., 2012) in the case of Twitter.

For Twitter, we are interested in not just the overall lexical distribution of each target noun, but also per-user lexical distributions. As such, we construct two Twitter-based datasets: (1) $\text{TWITTER}_{\text{RAND}}$, a random sample of 100 usages of each target noun; and (2) $\text{TWITTER}_{\text{USER}}$, 5 usages of each target noun from each member of a random sample of 20 Twitter users. Naively selecting users for $\text{TWITTER}_{\text{USER}}$ without filtering resulted in a preponderance of messages from accounts that were clearly bots, e.g. from commercial sites with a single post per item advertised for sale, with artificially-skewed sense distributions. In order to obtain a more natural set of messages from "real" people, we introduced a number of user-level filters, including removing users who posted the same message with different user mentions or hashtags, and users who used the target nouns more than 50 times over a 2-week period. From the remaining users, we randomly selected 20 users per target noun, resulting in 20 nouns $\times$ 20 users $\times$ 5 messages = 2000 messages.

For ukWaC, we similarly constructed two datasets: (1) $\text{UKWAC}_{\text{RAND}}$, a random sample of 100 usages of each target noun; and (2) $\text{UKWAC}_{\text{DOC}}$, 5 usages of each target noun from 20 documents which contained that noun in at least

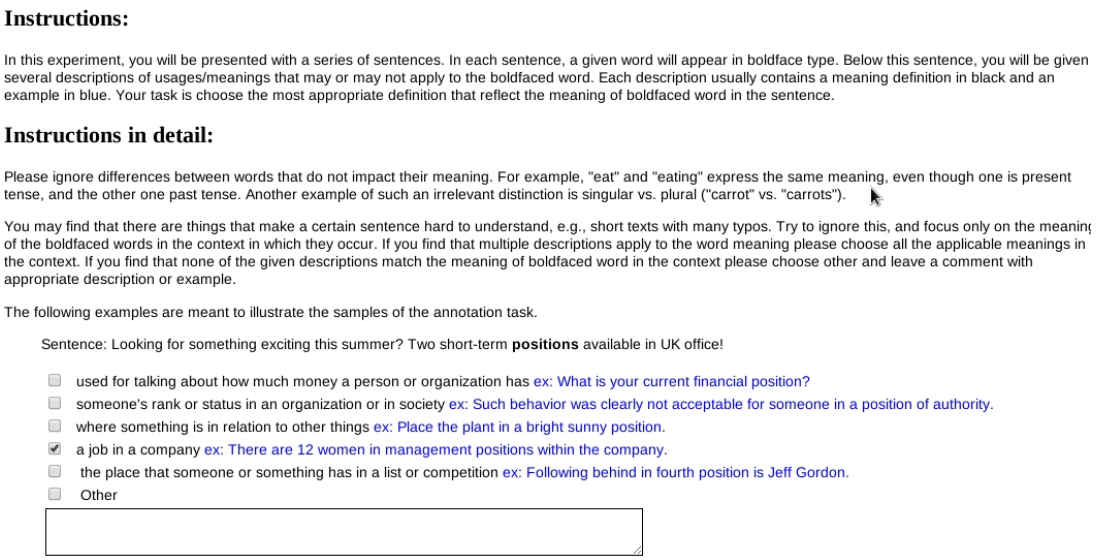

Figure 1: Screenshot of a sense annotation HIT for *position*

5 sentences. 5 such sentences were selected for annotation, resulting in a total of 20 nouns $\times$ 20 documents $\times$ 5 sentences = 2000 sentences.

3.2 Annotation Settings

We sense-tagged each of the four datasets using Amazon Mechanical Turk (AMT). Each Human Intelligence Task (HIT) comprised 5 occurrences of a given target noun, with the target noun highlighted in each. Sense definitions and an example sentence (where available) were provided from MACMILLAN. Turkers were free to select multiple sense labels where applicable, in line with best practice in sense labelling (Mihalcea et al., 2004). We also provided an "Other" sense option, in cases where none of the MACMILLAN senses were applicable to the current usage of the target noun. A screenshot of the annotation interface for a single usage is provided in Figure 1.

Of the five sentences in each HIT, one was a heldout example sentence for one of the senses of the target noun, taken from MACMILLAN. This gold-standard example was used exclusively for quality assurance purposes, and used to filter the annotations as follows:

1. Accept all HITs from Turkers whose gold-standard tagging accuracy was $\geq 80\%$;
2. Reject all HITs from Turkers whose gold-standard tagging accuracy was $\leq 20\%$;
3. Otherwise, accept single HITs with correct gold-standard sense tags, or at least 2/4 (non-gold-standard) annotations in common with Turkers who correctly annotated the gold-standard usage; reject any other HITs.

This style of quality assurance has been shown to be successful for sense tagging tasks on AMT (Bentivogli et al., 2011; Vuurens et al., 2011), and resulted in us accepting around 95% of HITs.

In total, the annotation was made up of 500 HITs (= 2000/4 usages per HIT) for each of the four datasets, each of which was annotated by 5 Turkers. Our analysis of sense distribution is based on only those HITs which were accepted in accordance with the above methodology, excluding the gold-standard items. We arrive at a single sense label per usage by unweighted voting across the annotations, allowing multiple votes from a single Turker in the case of multiple sense annotations. In this, the "Other" sense label is considered as a discrete sense label.

Relative to the majority sense, inter-annotator agreement post-filtering was respectably high in terms of Fleiss' kappa at $\kappa = 0.64$ for both UKWAC$_{\text{RAND}}$ and UKWAC$_{\text{DOC}}$. For TWITTER$_{\text{USER}}$, the agreement was actually higher at $\kappa = 0.71$, but for TWITTER$_{\text{RAND}}$ it was much weaker, $\kappa = 0.47$.

All four datasets have been released for public use: `http://www.csse.unimelb.edu.au/~tim/etc/twitter_sense.tgz`.

4 Analysis

In TWITTER$_{\text{USER}}$, the proportion of users who used a target noun with one sense across all 5 usages ranged from 7/20 for *form* to 20/20 for *degree*, at an average of 65%. That is, for 65% of users, a given noun (with average polysemy = 5.5 senses) is used with the same sense across 5 separate messages. For UKWAC$_{\text{DOC}}$ the proportion of documents with a single sense of a given target noun

	Partition	Agreement (%)
Gale et al. (1992)	document	94.4
TWITTER$_{\text{USER}}$	user	95.4
TWITTER$_{\text{USER}}$	—	62.9
TWITTER$_{\text{RAND}}$	—	55.1
UKWAC$_{\text{DOC}}$	document	94.2
UKWAC$_{\text{DOC}}$	—	65.9
UKWAC$_{\text{RAND}}$	—	60.2

Table 2: Pairwise agreement for each dataset, based on different partitions of the data ("—" indicates no partitioning, and exhaustive comparison)

across all usages ranged from 1/20 for *case* to 20/20 for *band*, at an average of 63%. As such, the one sense per tweeter heuristic is at least as strong as the one sense per discourse heuristic in UKWAC$_{\text{DOC}}$.

Looking back to the original work of Gale et al. (1992), it is important to realise that their reported agreement of 94% was calculated *pairwise* between usages in a given document. When we recalculate the agreement in TWITTER$_{\text{USER}}$ and UKWAC$_{\text{DOC}}$ using this methodology, as detailed in Table 2 (calculating pairwise agreement within partitions of the data based on "user" and "document", respectively), we see that the numbers for our datasets are very close to those of Gale et al. on the basis of more than twice as many nouns, and many more instances per noun. Moreover, the one sense per tweeter trend again appears to be slightly stronger than the one sense per discourse heuristic in UKWAC$_{\text{DOC}}$.

One possible interpretation of these results is that they are due to a single predominant sense, common to all users/documents rather than user-specific predominant senses. To test this hypothesis, we calculate the pairwise agreement for TWITTER$_{\text{USER}}$ and UKWAC$_{\text{DOC}}$ across all annotations (without partitioning on user/document), and also for TWITTER$_{\text{RAND}}$ and UKWAC$_{\text{RAND}}$. The results are, once again, presented in Table 2 (with partition indicated as "—" for the respective datasets), and are substantially lower in all cases ($< 66\%$). This indicates that the first sense preference varies considerably between users/documents. Note that the agreement is slightly lower for TWITTER$_{\text{RAND}}$ and UKWAC$_{\text{RAND}}$ simply because of the absence of the biasing effect for users/documents.

Comparing TWITTER$_{\text{RAND}}$ and UKWAC$_{\text{RAND}}$, there were marked differences in first sense preferences, with 8/20 of the target nouns having a different first sense across the two corpora. One surprising observation was that the sense distributions in UKWAC$_{\text{RAND}}$ were in general more skewed than in TWITTER$_{\text{RAND}}$, with the entropy of the sense distribution being lower (= more biased) in UKWAC$_{\text{RAND}}$ for 15/20 of the target nouns.

All datasets included instances of "Other" senses (i.e. usages which didn't conform to any of the MACMILLAN senses), with the highest relative such occurrence being in TWITTER$_{\text{RAND}}$ at 12.3%, as compared to 6.6% for UKWAC$_{\text{RAND}}$. Interestingly, the number of such usages in the user/document-biased datasets was around half these numbers, at 7.4% and 3.6% for TWITTER$_{\text{USER}}$ and UKWAC$_{\text{DOC}}$, respectively.

5 Discussion

It is worthwhile speculating why Twitter users would have such a strong tendency to use a given word with only one sense. This could arise in part due to patterns of user behaviour, in a given Twitter account being used predominantly to comment on a favourite sports team or political events, and as such is domain-driven. Alternatively, it can perhaps be explained by the "reactive" nature of Twitter, in that posts are often emotive responses to happenings in a user's life, and while different things excite different individuals, a given individual will tend to be excited by events of similar kinds. Clearly more research is required to test these hypotheses.

One highly promising direction for this research would be to overlay analysis of sense distributions with analysis of user profiles (e.g. Bergsma et al. (2013)), and test the impact of geospatial and sociolinguistic factors on sense preferences. We would also like to consider the impact of time on the one sense per tweeter heuristic, and consider whether "one sense per Twitter conversation" also holds.

To summarise, we have investigated sense distributions in Twitter and a general web corpus, over both a random sample of usages and a sample of usages from a single user/document. We found strong evidence for Twitter users to use a given word with a single sense, and also that individual first sense preferences differ between users, suggesting that methods for determining first senses on a per user basis could be valuable for lexical semantic analysis of tweets. Furthermore, we found that sense distributions in Twitter are overall less skewed than in a web corpus.

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 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pages 356–364, Nagoya, Japan.

Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting expert quality from the crowd for machine translation evaluation. *Proceedings of the MT Summmit*, 13:521–528.

Shane Bergsma, Mark Dredze, Benjamin Van Durme, Theresa Wilson, and David Yarowsky. 2013. Broadly improving user classification via communication-based name and location clustering on Twitter. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 1010–1019, Atlanta, USA.

Paul Earle, Michelle Guy, Richard Buckmaster, Chris Ostrum, Scott Horvath, and Amy Vaughan. 2010. OMG earthquake! can Twitter improve earthquake response? *Seismological Research Letters*, 81(2):246–251.

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.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, USA.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop: Can we beat Google*, pages 47–54, Marrakech, Morocco.

William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*, pages 233–237.

Spandana Gella, Paul Cook, and Bo Han. 2013. Unsupervised word usage similarity in social media texts. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, pages 248–253, Atlanta, USA.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Main Conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 57–60, New York City, USA.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299, Atlanta, USA.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 419–426, Vancouver, Canada.

Robert Krovetz. 1998. More than one sense per discourse. *NEC Princeton NJ Labs., Research Memorandum*.

Jey Han Lau, Nigel Collier, and Timothy Baldwin. 2012. 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.

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.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 4(33):553–590.

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain.

Roberto Navigli and Daniele Vannella. 2013. SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)*, pages 193–201, Atlanta, USA.

Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. 2007. SemEval-2007 task 07: Coarse-grained English all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 30–35, Prague, Czech Republic.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

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, Machine Learning Department, Carnegie Mellon University.

Martha Palmer, Olga Babko-Malaya, and Hoa Trang Dang. 2004. Different sense granularities for different applications. In *Proceedings of the HLT-NAACL 2004 Workshop: 2nd Workshop on Scalable Natural Language Understanding*, pages 49–56, Boston, USA.

Sasa Petrović, Miles Osborne, and Victor Lavrenko. 2010. Streaming first story detection with application to twitter. In *Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL HLT 2010)*, pages 181–189, Los Angeles, USA.

Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. 2011. How much spam can you take? an analysis of crowdsourcing results to increase accuracy. In *Proc. ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR 2011)*, pages 21–26.