Microblog-Genre Noise and Impact on Semantic Annotation Accuracy

Leon Derczynski, Diana Maynard, Niraj Aswani and Kalina Bontcheva
University of Sheffield
Department of Computer Science
Sheffield, S1 4DP, UK
{leon,diana,niraj,kalina}@dcs.shef.ac.uk

ABSTRACT

Using semantic technologies for mining and intelligent information access to microblogs is a challenging, emerging research area. Unlike carefully authored news text and other longer content, tweets pose a number of new challenges, due to their short, noisy, context-dependent, and dynamic nature. Semantic annotation of tweets is typically performed in a pipeline, comprising successive stages of language identification, tokenisation, part-of-speech tagging, named entity recognition and entity disambiguation (e.g. with respect to DBpedia). Consequently, errors are cumulative, and earlier-stage problems can severely reduce the performance of final stages. This paper presents a characterisation of genre-specific problems at each semantic annotation stage and the impact on subsequent stages. Critically, we evaluate impact on two high-level semantic annotation tasks: named entity detection and disambiguation. Our results demonstrate the importance of making approaches specific to the genre, and indicate a diminishing returns effect that reduces the effectiveness of complex text normalisation.

Categories and Subject Descriptors

I.7.2 [**Document and Text Processing**]: Document Preparation—*Hypertext*; I.2 [**Computing Methodologies**]: Artificial Intelligence

Keywords

semantic annotation, entity disambiguation, microblog, twitter, entity recognition, text normalisation

1 Introduction

Semantic annotation is the process of tying machine tractable semantic models to natural language text. From a technological perspective, semantic annotation is about annotating in texts all mentions of concepts from the ontology (i.e. classes, instances, properties, and relations), through metadata referring to their URIs. Entity recognition is a kind of semantic annotation, typically broken down into two main phases: *entity annotation* and *entity linking* (also called reference disambiguation or entity resolution) [34]. State-of-the-art automatic semantic annotation, browsing, and search algorithms are typically developed and evaluated on news articles and other carefully written, long web content [3].

However, in recent years, social media – and microblogging in particular – have established themselves as high-value, high-volume content, which organisations increasingly wish to analyse automatically. The leading microblogging platform is currently Twitter [21], which has around 288 million active users, posting over 500 million tweets a day,[1] and has the fastest growing network in terms of active usage.[2]

Reliable semantic annotation of user-generated content is an enabler for other semantic technologies [4], including opinion mining [28], summarisation [38], semantic-based search, recommendation, visual analytics, and user and community modelling [41]. It is relevant in many application contexts [12], including knowledge management, competitor intelligence, customer relation management, eBusiness, eScience, eHealth, and eGovernment.

Semantic annotation of microblogs has only recently become an active research topic, following early experiments which showed this genre to be extremely challenging for state-of-the-art algorithms. For instance, named entity recognition methods typically have 85-90% accuracy on longer texts, but 30-50% on tweets [35, 22]. First, the shortness of microblogs (maximum 140 characters) makes them hard to interpret. Consequently, ambiguity is a major problem since semantic annotation methods cannot easily make use of coreference information. Unlike longer news articles, there is a low amount of discourse information per microblog document, and threaded structure is fragmented across multiple documents, flowing in multiple directions. Second, microtexts also exhibit much more language variation, tend to be less grammatical than longer posts, contain unorthodox capitalisation, and make frequent use of emoticons, abbreviations and hashtags, which can form an important part of the meaning.

To combat these problems, research has focused on microblog-specific semantic annotation algorithms (e.g. named entity recognition for Twitter [35], Wikipedia-based topic and entity disambiguation [30]). Particular attention is given to microtext normalisation, as a way of removing some of the linguistic noise prior to part-of-speech tagging and entity recognition.

In light of the above, this paper aims to answer the following research questions:

- How robust are entity recognition and disambiguation methods on shorter and noisier microblog texts, in comparison

[1] See http://news.cnet.com/8301-1023_3-57541566-93/report-twitter-hits-half-a-billion-tweets-a-day/.
[2] See http://globalwebindex.net/thinking/social-platforms-gwi-8-update-decline-of-local-social-media-platforms/.

24th ACM Conference on Hypertext and Social Media
1–3 May 2013, Paris, France

System	Overall accuracy	English	Dutch	French	German	Spanish
TextCat	89.5%	88.4%	90.2%	86.2%	94.6%	88.0%
langid	89.5%	92.5%	89.1%	89.4%	94.3%	83.0%
TextCat (twitter)	**97.4%**	**99.4%**	**97.6%**	**95.2%**	**98.6%**	**96.2%**
langid (twitter)	87.7%	88.7%	88.8%	88.0%	92.5%	81.6%

Table 1: Language classification accuracy on the ILPS dataset for systems before and after adaptation to the microblog genre.

with the longer, cleaner news texts on which the algorithms have been trained and evaluated?

- What problem areas are there in the semantic annotation of microblog posts, and what is the cause of the majority of errors made? Can normalisation help reduce the detrimental effect of microblog noise?
- Which directions for future improvement are the most promising, given the state-of-the-art?

The rest of the paper is structured as follows. Section 2 compares the performance of state-of-the-art algorithms on the various components of a semantic annotation pipeline, trained specifically on microblog data versus those trained on longer text. Section 3 introduces the microblog normalisation task, compares different methods, and measures the impact normalisation has on the accuracy of semantic annotation of tweets. The third research question is answered in Section 4. In this paper, we focus only on microposts in English, since few Twitter-based linguistic tools have currently been developed for other languages. We do not investigate how successful such techniques might be on other languages, or indeed, whether all problems are similarly prevalent in other languages.

2 Method Comparison

Semantic annotation with linked data comprises of a sequence of tasks, with a different component for each one. Each component relies on the information provided by one or more of the previous algorithms, typically ordered in a pipeline: language detection; tokenisation; part-of-speech tagging; named entity recognition; and entity disambiguation, as exemplified in GATE's ANNIE system [9]. As the performance of each component relies on that of its predecessors, errors can cascade through the pipeline. We present approaches to each stage of the semantic annotation pipeline, contrasting a variety of methods, and comparing performance at each stage on microblog and standard-format (e.g. newswire) text. Since the overall goal is error reduction, attention is given to the specific kinds of errors made by each method.

2.1 Language Identification

It is critical to determine the language in which a document is written, in order to know which tools to apply. The *language identification* task is thus typically performed before other linguistic processing, as its goal is to output a language suggestion given some unprocessed text. We consider two types of approach: n-gram frequency-based and n-gram information gain-based.

Both approaches include an implicit tokenisation step, though this is speculative and does not inform later processes in the entity-linking pipeline. TextCat [6] relies on n-gram frequency models to discriminate between languages, relying on token sequences that are strong language differentiators. Information gain-based langid.py [24] uses n-grams to learn a multinomial event model, with a feature selection process designed to cope with variations of expression between text domains. They have both been adapted for microblog text, using human-annotated data from Twitter. The TextCat adaptation [5] works on a limited set of languages; the langid.py adaptation [33] on 97 languages.

Approach	Precision	Recall	F1
PTB Regexp	90%	72%	80%
PTB Regexp (twitter)	**98%**	**94%**	**96%**

Table 2: Tokeniser performance on sample microblog text

We evaluated four system runs on the ILPS TextCat microblog evaluation dataset.[3] Results are given in Table 1, with the Twitter-specific versions marked "twitter". Note that the adapted version of TextCat has a slightly easier task than that for langid.py because it expects only five language choices, whereas the adapted langid.py is choosing labels from a set of 97 languages. The latter assigned a language outside the five available to 6.3% of tweets in the evaluation set. Why the adapted langid.py performed worse than the generic version is not clear; the results are quite close for some languages, and so if an approximate 6% improvement could be made in these cases, the Twitter-adapted version would be better. Although language identification is harder on tweets than longer texts, it is doable and has been achieved to a reasonably high level – enough to inform choices of later tools.

2.2 Tokenisation

Tokenisation is the process of dividing up a string of characters and punctuation into constituent words. Strings must be broken into tokens to enable subsequent processing, by determining the atomic units that make up the text. Different languages require different tokenisers, with some easier than others [29]. Even punctuation use can differ between languages for the microblog genre, in which "smileys" (comprised of extended sequences of punctuation symbols) are prevalent.

There is some disagreement regarding tokenisation schemes for microblogs, which contain some unusual word types. While conventionally the Penn Treebank (PTB) rules[4] are suitable for English, variations of tokenisation for microblog text have been proposed, particularly accounting for user mentions and hashtags in Twitter, which both include punctuation. The convention for this paper is to use Ritter's tokenisation scheme [35].

For our evaluation, we took the ANNIE tokenizer from GATE, which uses the PTB tagset, and measured its performance on Ritter's tweet dataset [35]. To gauge the impact of making Twitter-specific considerations, we compared this with a Twitter-adapted version of the ANNIE tokeniser. Performance was measured in terms of precision and recall, as well as F1 measure, and is given in Table 2.

The original tokeniser performed poorly, reaching an F1 of only 80% (near 100% is typical), and with many errors around punctuation and Twitter-specific entities. This is too weak to accurately inform later tasks. Smileys cause some trouble for tokenisers, many of which do not occur in the training data. Orthographic errors are also rife in this genre, an analysis of which can be found in [15]. Aside from smileys and typos, the low performance of a conventional tokeniser such as this is mostly due to differing tokenisation rules regarding usernames and Twitter-specific content.

[3] See http://ilps.science.uva.nl/resources/twitterlid.

[4] See http://www.cis.upenn.edu/~treebank/tokenization.html.

Approach	Accuracy	Sentence	On unknowns
Newswire			
Brill	93.9%	28.4%	-
Stanford Tagger	**97.3%**	**56.8%**	**90.5%**
Microblog			
Brill	70.5%	2.54%	13.1%
Stanford Tagger	73.6%	4.24%	26.6%
Brill (twitter)	78.6%	8.47%	28.5%
Stanford (twitter)	**88.4%**	**25.4%**	**72.1%**
Ritter (twitter)	88.3%	-	-
Incorrect for all	9.0%	-	16.1%

Table 3: Part-of-speech tagging performance on two corpora: extracts from the Wall Street journal (newswire), and a 25% part of Ritter's corpus. The latter is tagged with and without in-genre training data. Accuracy measured at both token and sentence level.

2.3 Part-of-speech Tagging

Part-of-speech (**POS**) tagging is generally the next stage in a semantic annotation pipeline following tokenisaton, and is necessary for many tasks such as named entity recognition and disambiguation. Early high-performance tagging approaches include the Brill tagger, which uses transformation-based learning, and has the benefit of being fast [20]. Later, Toutanova et al. [44] introduced the Stanford tagger, trained on newswire texts, and which has sophisticated feature generation, especially for unknown words, and a highly configurable re-trainable engine. This is generally thought to represent the current state of the art for such texts. Models using the PTB set are available for both these taggers.

However, these are not necessarily suitable for microblogs, and thus specific taggers have been developed to handle these. We concentrate on those using the PTB tagset, as many later components in our pipeline rely on this labelling schema, and changes in tagset hamper tagger comparison. Ritter et al. [35] trained on PTB-tagged tweets, adding extra tag labels for retweets, URLs, hashtags and user mentions. Their work also included the distribution of a ground truth annotated tweet dataset.

2.3.1 POS Tagging Comparison

We compare three taggers on a subset of Ritter's corpus: Brill's transformation-based tagger, the Stanford tagger and Ritter's tagger. We use the experimental setup detailed in Ritter's paper, apart from using a fixed train/test split in the Twitter data, where the evaluation portion had 2,242 tokens. Results are given in Table 3, including comparison against sections 22-24 of the Wall Street Journal part of the Penn Treebank [26].

The results for the Stanford and Brill taggers trained on newswire text show poor success rates; it was not possible to re-train the Ritter tagger on newswire data only. Results are also given for the taggers trained on Twitter data – a re-trained, re-tuned Stanford tagger with unknown word parameters scaled to the size of the corpus[5] and Ritter's stock tagger.[6] These figures are far below the state-of-the-art in POS tagging for some other well-formed genres (e.g. newswire, where accuracy is around 98%), which warrants some analysis of the problem and of error cases. However, even the performance of the worst tagger trained with microblog text was better than the best tagger not using microblog data. Ritter's system did not provide information about unknown word tagging accuracy.

The "incorrect for all" row shows the proportion of tokens that none of the taggers could label correctly. This gives an initial idea of whether a voting approach would be successful for improving performance: higher scores indicate a greater proportion of particularly tricky tokens. Not included in the table are whole-sentence accuracy figures – that is, the proportion of sentences in which every token was correctly tagged. These vary from 2% for a standard Brill tagger to 22% for the best-performing Stanford tagger's in-genre approach.

While ground truth data is available for the microblog genre, there is not enough of it (Figure 1). As the curve in the figure does not appear to be levelling off (or dipping) when more data is added, its shape suggests that increasing the amount of tagged ground truth data will improve in-genre POS tagging performance. Aside from providing an incomplete model of the language used in microblogs, the smallness of the volume of tagged text increases the chance of encountering words not found in the training data when tagging new text.

2.3.2 Difficult Token Failure Analysis

Tagging previously unseen words forces the tagger to rely on contextual clues. These unknown words make up a large part of the mis-tagged tokens. One can see the effect that improving accuracy on unknown words has on overall performance by comparing the Stanford tagger when trained on off-genre vs. in-genre data in Table 3. We identified the unknown words that were tagged incorrectly (hereafter referred to as incorrectly tagged unknowns or the **ITU set**), and categorised them into eight groups.

Gold standard error – Where the ground truth data is wrong; 6.7% of the ITU set. For example, the Dutch *dank je* should in an English corpus be tagged as foreign words (FW), but in our dataset is marked *dank/URL je/IN*. These are not tagger errors but rather evaluation errors, and may be repaired by amending the dataset.

In-vocabulary – Tokens that are common in general, but do not occur in the training data; 27% of the ITU set. For example, *Internet* and *bake* are unknown words and mis-tagged in the evaluation corpus. This kind of error may be fixed by a larger training set or the use of a lexicon of single-tag words.

Pre-taggable – Words to which a label may be reliably assigned automatically; 9% of the ITU set. This includes well-formed URLs, hashtags, cardinals (e.g. *3rd*) and smileys. Regular expression based labellings added pre- or post-tagging should improve performance here.

Proper noun – Proper nouns not in the training data; 11.2% of the ITU set. Most of these should be tagged NNP, and are often useful for later named entity recognition. Proper nouns in the ITU set often had incorrect capitalisation; for example, *derek* and *birmingham*. Gazetteer approaches may help annotate these, in cases of words that can only occur as proper nouns.

Slang – An abundance of slang is a characterising feature of microblog text, and these words are often incorrectly tagged, as well as being rarely seen due to a proliferation of spelling variations (all incorrect). They comprise 27% of the ITU set. Examples include *LUVZ*, *HELLA* and *2night*. Some kind of automatic correction or expanded lexicon could be employed to either map these back to dictionary words or to include previously-seen spelling variations.

Tokenisation error – Occasionally the tokeniser or original author makes tokenisation errors; these are 9% of the ITU set. Examples include *ass**sneezes*, which should have been split into more than one token as indicated by special/punctuation characters, and *eventhough*, where the author has missed a space. These are hard to correct. Specific subtypes of error, such as the joined words in the

[5] Props changes from `l3w-gen`: `arch` include `naacl2003unknowns`, `lnaacl2003unknowns`, and `veryCommonWordThresh = 100`.
[6] See https://github.com/aritter/twitter_nlp.

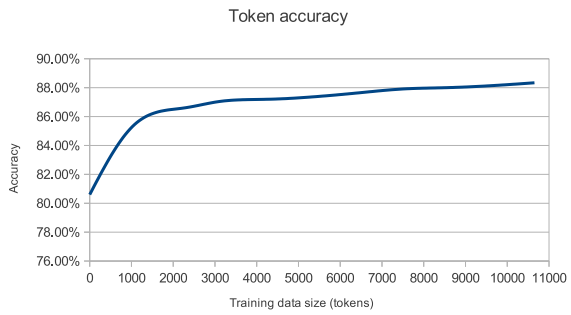

Figure 1: Stanford tagger token-level performance with increasing amounts of in-genre text in the training set, using Ritter's experimental setup and segregated evaluation/training sets.

example, could be checked for and forcibly fixed, though this introduces the problem of distinguishing intentional from unintentional word usage.

Genre-specific – Genre- and site-specific words are 2.2% of the set. These are words that are unique to specific sites, often created for microblog usage, such as *unfollowing*. Extra genre-specific training data may to reduce genre-specific word errors.

Orthographic error – Finally, although it is difficult to detect the intent of the user, some content seems likely to have been accidentally mis-spelled; this is 7.9% of the ITU set. Examples include *Handle]* and *suprising*. Automatic spelling correction may improve performance in these cases.

2.3.3 POS Tagging Summary

Data suggests that genre-specific changes must be made to attain any kind of reasonable POS tagging accuracy in this genre. Enhancements including improved tokenisation, pre-tagging labelling of unambiguous tokens and also using a small proportion of annotated microblog text in the training data led to +15% absolute accuracy improvement – an error reduction of over half. However, current genre-adapted performance is still not high enough, given the early position of this process in the pipeline.

2.4 Named Entity Recognition

Named entity recognition (**NER**) is critical to semantic annotation, in this case involving finding which utterances in a discourse may be linked to ontological concepts via a URI. NER is difficult on user-generated content in general, and in the microblog genre specifically, because of the reduced amount of contextual information in short messages and a lack of curation of content by third parties (e.g. that done by editors for newswire). In this section, we examine some state-of-the-art NER methods for semantic annotation. We then compare their performance on microblog data and analyse the task of entity recognition in this genre.

2.4.1 Existing NER systems

A plethora of named entity recognition techniques and systems is available in general [32, 37, 27]. For Twitter, a selection have been proposed but most are not publicly accessible.

General purpose Of the many NER systems available, we chose those that take different approaches and are very readily available. ANNIE uses gazetteer-based lookups and finite state machines to identify and type named entities in newswire text. The Stanford NER system [14] uses a machine learning-based method to detect named entities, and is distributed with CRF models for English newswire text.

System	Precision	Recall	F1
Newswire			
ANNIE	78%	74%	77%
ARCOMEM	-	-	-
Stanford	-	-	**89%**
Microblog			
ANNIE	47%	**83%**	60%
ANNIE-twitter	77%	**83%**	**80%**
ARCOMEM	55%	74%	63%
ARCOMEM-twitter	**83%**	64%	72%
ARCOMEM-lang	48%	82%	61%
ARCOMEM-lang-twitter	79%	82%	**80%**
Stanford	59%	32%	41%
Stanford-twitter	54%	45%	49%
Ritter (twitter only)	73%	49%	59%

Table 4: Whole-pipeline named entity recognition performance, before and after genre adaptation. Newswire performance is over the CoNLL 2003 English dataset; microblog performance is over the development part of the Ritter dataset

Twitter-specific The general picture is that Twitter NER is difficult. There are few datasets, which are small and each custom to a single approach.

Ritter et al. [35] take a pipeline approach, performing first tokenisation and POS tagging before using topic models to find named entities. Liu [22] propose a gradient-descent graph-based method for doing joint text normalisation (see Section 3 below) and recognition, reaching 83.6% F1 measure. Finally, Freire [16] use a CRF classifier to define entity bounds while minimising reliance upon having well-structured text; this reaches F1 of 79%.

2.4.2 NER Comparison

We chose four high-performance and easy-to-evaluate named entity recognition tools, three of which are widely available, and evaluated them on the corpus of tweets developed by Ritter (2,400 tweets comprising 34,000 tokens). The first was ANNIE. The second used the ARCOMEM NER system,[7] [28] which is based on ANNIE but designed to deal with multi-genre entity detection. The third was the Stanford general-purpose NER system, and the fourth Ritter's NER approach. Results are given in Table 4, including comparison with newswire NER using the CoNLL 2003 dataset [43]. We did not consider Percent-type entity annotations in these evaluations because there were so few (three in the whole corpus) and they were all annotated correctly in any case.

In the first experiment, we compared the default ANNIE with the Twitter-specific version (ANNIE-twitter), and found an absolute precision increase of 30% – mainly with respect to Date, Organization and in particular Person. Recall remained identical after microblog adaptation, which led to an increase of 20% in F1. Note that we did not consider the Twitter-specific UserID annotation as a Person annotation, since these were all 100% correct. Because the regular ANNIE does not consider these, there were many false positive Person annotations as part of UserIDs.

In the second experiment, we compared the default ARCOMEM application with the Twitter-specific version, and found a precision increase of 18% but a recall decrease of 10%. Nevertheless, this had the effect of increasing F1 by 9% – mainly with respect to Date, Organization and Person entities. However, the dramatic increase in Person recognition seen in the first experiment was not

[7] See http://www.arcomem.eu.

Ritter	Stanford
company	Organisation
facility	Location
geo-loc	Location
movie	Misc
musicartist	Misc
other	Misc
person	Person
product	Misc
sportsteam	Organisation
tvshow	Misc

Table 5: Mappings from Ritter entity types to the Stanford NER categories. Most "musicartist" annotations referred to either groups or were of indeterminate size, so Misc was chosen for this category instead of Person or Organisation.

seen here – largely we suspect because already, the grammars used in ARCOMEM take account of some specific issues (ambiguity with common words used on Twitter, for example). Note also that the ARCOMEM application includes a language ID processing resource (using TextCat – see Section 2.1) which separates English sentences from those deemed to be in other languages, and only processes the English ones. Since this component is not 100% accurate, it has the effect that some English sentences do not get processed and thus recall is lower.

In the third experiment, we compared a version ARCOMEM without the language ID component (so every sentence gets processed), with a Twitter-specific version. Interestingly, we got lower precision on each of these compared with the applications in experiment 2 – presumably because it was now over-generating some named entities which were part of some of the plain sentences in the previous application. Comparing the normal and Twitter-specific versions, we see a 31% absolute increase in precision, though no change in recall, leading to a 19% increase in F1. Slightly surprising is the fact that precision of Dates actually dropped with the Twitter-specific version, although as with Experiment 1, precision on Person entities rose dramatically, and precision on the other types all rose slightly.

In the fourth experiment, we compared the standard Stanford NER system with a version trained including some Twitter data (with the same Ritter train/test splits used earlier). It was necessary to translate Ritter's NE categories to those native to this system, using the mappings given in Table 5. While precision was high for some categories (Person 65% and Location 74%), Organizations were recognised poorly, having both low recall and low precision for a category F1 of 21%. Errors in labelling Organizations were more often caused by mis-categorisations than by missing the entity entirely. For example, *Vista del Lago* and *Clemson Auburn* were both labelled as Organizations when they should have been Locations. Polysemous named entities were also handled poorly, possibly due to a lack of surrounding disambiguating context (typical in this genre). For example, *UT* and *Amazon* were labelled as Locations when they occurred as Organizations (University of Texas and the international web retailer) – though both these could also occur as Locations (the state of Utah and the large river, respectively). Adding in-genre training material helped somewhat with recall, giving a 13% absolute increase, though performance was still markedly weaker than the other gazetteer and FSM based approaches. Organization recognition accuracy improved to an F1 of 29%, though Person and Location recognition was weaker. NEs represented in lowercase (e.g. *skype*, *pete*) were frequently ignored.

In summary, conventional tools perform poorly in this genre, and we can see that microblog domain adaptation is critical to good NER in a semantic annotation pipeline. We demonstrate a +30% absolute precision and +20% absolute F1 performance increase using off-the-shelf tools adapted using publicly available resources. However, even the state-of-the-art is still lacking, leaving significant amounts of missed annotations and generating false positives. Common themes in the mis-annotated set suggest that better tokenisation and better POS tagging can help improve performance, but better NER techniques and models must be developed as well.

2.4.3 Facebook NER

We also investigated the effect of the processing resources on other kinds of social media, namely Facebook posts. Facebook messages are a little more diverse than tweets, and overlap with the microblog genre (rather than all being in microblog format). Specifically, the message length limit is much greater, and messages do not necessarily suffer the compression that is present in other sites. However, the site is still un-curated and permits bi-directional communication, as opposed to newswire which exhibits neither of these properties. Further, candid observations suggest that many of the linguistic difficulties of microblog text are present in Facebook messages, and that they too can benefit from semantic annotation for the same reasons that microblog text can.

For the evaluation, we used a small corpus of Facebook posts about the financial crisis, collected as part of ARCOMEM. Due to its small size, we do not provide quantitative results, but do make qualitative observations.

Because these posts were written in a fairly formal style (not typical of microblog text), there was no difference between using the standard ARCOMEM application and the Twitter one. In this case author style may be biased by the type of reader that posts on this topic. However, adapting the pipeline still offered recall improvements over the baseline of standard newswire-oriented ANNIE, being able to detect a wider range of entities at the cost of precision, with a net positive effect. We identified three types of error behind the precision drop.

Multiword expressions Some spurious NE annotations actually formed part of multiword expressions. For example, in the headline *Green Europe Imperiled as Debt Crises Triggers Carbon Market Drop*), the word *Europe* does not refer to a Location, but was annotated as such (the phrase *Green Europe* does not refer to either a Location or Organization, but a project).

Complex punctuation Some false positives were due to annotations that were non-initial substrings of complex URLs which had slipped past earlier URL tokenisation.

Adjectival use Finally, some NEs were spuriously annotated over words acting as adjectives. For example, in *the Democratic chairman of the committee*, the capitalised word *Democratic* is not a mention of an Organisation (although the whole might be considered a Person).

Overall, named entity recognition on Twitter remains harder than on Facebook data, but learnings and adaptations from dealing with microblog text can be used for performance gains when processing Facebook messages. There is some support of an association between genre and formality of context with ease-of-processing.

2.5 Entity Linking

Having determined which expressions in text refer to entities, the next semantic annotation task is entity linking (or entity disambiguation). It typically requires annotating a potentially ambiguous entity mention with a link to a canonical URI describing a unique entity. Approaches have used different entity databases as a disambiguation target, including Wikipedia (e.g. Cucerzan et al. [8])

Name	Precision	Recall	F1
Zemanta	89%	65%	75%
LODIE	68%	69%	68%
LODIE + hashtag expansion	70%	69%	69%

Table 6: Tweet-level entity linking performance. LODIE results are for a whole-pipeline approach

and Linked Open Data resources (e.g. DBpedia [31], YAGO [39]). Many disambiguation targets share common ground, and it is often possible to map between them [36].

Microblog entity disambiguation is a relatively new, underexplored task. Recent Twitter-specific disambiguators discovered problems with using state-of-the-art entity linking in tweets [1, 30, 17]. Another approach using Twitter concentrated on hashtags, automatically annotating them with DBpedia entries to assist semantic search on Twitter [23]. We compare approaches to linking entities within microtext, and assigning entities to whole messages.

2.5.1 Word-level Entity Linking

Evaluating entity linking on microblogs is currently hampered by the lack of a shared corpus of manually disambiguated entities. The only currently available relevant dataset consists of 1,800 tweets from the RepLab 2012 [2] evaluation of reputation analysis methods. One of its subtasks was to determine whether a user-generated message pertains to a given organisation (e.g. Apple, Lufthansa). For each of the targets, we manually found their URIs in DBpedia and used these in the gold-standard instead. Effectively, this transformed the entity disambiguation task into a binary classification task, i.e. whether the given tweet is about the target URI. Using this dataset, we evaluated two systems for their accuracy in entity disambiguation, as well as the impact of additional information from hashtag expansion (for LODIE) on entity linking accuracy.

We evaluated the LODIE pipeline system [10], which links entities to DBpedia URIs chosen using a combination of four metrics: string similarity; semantic similarity between nearby entities; word level contextual similarity; and URI frequency in Wikipedia articles. This system contains the Twitter tokeniser, POS tagger, and ANNIE NER system that we described in earlier sections and is thus an end to end evaluation. The second system was Zemanta,[8] a system designed for semantic annotation of blog posts which are similar in genre to microblog. Zemanta was found to perform best in a recent comparative evaluation on longer texts [36], and relies on a combination of machine learning and user feedback to disambiguate entities. Preliminary results are given in Table 6. Extra matches outside the bounds of recognised entities are discarded, and so this reports on the joint recognition/linking task.

Matches only count if an annotation mentions one of the six target organisations, while disambiguation of other entities is disregarded. This makes the RepLab corpus of somewhat limited utility for evaluation of entity linking, and emphasises the need for a more comprehensive microblog gold standard for evaluation of entity linking.

It should also be noted that Zemanta's precision on this dataset is a maximum, due in part to annotation choices (producing overlapping annotations) and in part to the specifics of the RepLab dataset. For example, in *Lufthansa Cargo*, both *Lufthansa* and *Lufthansa Cargo* are annotated as Organization. In contrast, the LODIE only annotates *Lufthansa Cargo*, since it is the more specific entity. This results in lower precision – the manual RepLab annotation is for Lufthansa only, since it indicates the tweet is relevant to the company. On other entity linking datasets (e.g. TAC-KBP), producing overlapping entities would result in Zemanta having higher recall at the cost of lower precision. However, due to the specifics of the RepLab task, for Zemanta we discard entities that do not match those recognised *after* linking, instead of before. This acts as applying an extra layer of filtering.

Further, LODIE does not preserve named entities recognised on lowercase strings, unless the part-of-speech tag suggests a proper noun. This was originally intended to reduce the number of incorrectly recognised entities and worked well with e.g. library texts and Wikipedia articles, but since entity names in tweets are often specified in lower case, this strategy backfires. Similar to named entity recognition, other capitalisation issues also cause problems. One difficult tweet is shown in Example 1.

(1) *Apple trees in the home garden bit.ly/yOztKs*

Here, the ground truth indicates that this tweet is not relevant to Apple Inc., the company (a correct annotation). However, some methods incorrectly linked the first word to the Apple company, possibly because it is unusual to see the word used to refer to the fruit in a sentence-initial position. Further, in cases where the word refers to the company and is capitalised (perhaps to relay a strength of sentiment regarding one's iPhone), *APPLE* may be mis-annotated as the APPLE related to the Ariane rocket program, which is not the correct choice.

Lastly, our experiments showed a significantly impaired recall and overall F1-score, due to many tweets not actually containing directly the company names, but only names of their products (e.g. iPod and iPad in tweets relevant to Apple). In order to overcome microblog terseness and introduce additional context, we experimented with expanding hashtags in tweets with their textual definions, using an online lookup service.[9] Hashtags are used in some tweet messages to indicate a topic. As can be seen from Table 6, this additional text does result in improved performance for LODIE, but unfortunately only a relatively small number of tweets in the dataset contain hashtags.

2.5.2 "Whole tweet" Entity Linking

The "aboutness" task determines which topics and entities describe the main themes of a text. For Twitter, the topic-disambiguated dataset detailed in [30] contains both tweets and the corresponding Wikipedia articles, which are the topic of these tweets. Since each tweet is topic-annotated, there is no guarantee that the corresponding Wikipedia article is an entity (e.g. some topics are website, usability, and audience). Even when the assigned topics are entities, the tweets often contain other entities as well, which are not annotated. These characteristics make this dataset only a rough indicator of entity linking performance. We evaluated the performance of DBpedia Spotlight [31], Zemanta and TextRazor[10] (an open NLP tagging service offering a variety of pipeline stages) on this corpus (see Table 7). Only the top-ranked entities were considered for matching, and we discarded recognised but unlinked entities. Zemanta and DBpedia Spotlight performed quite poorly, probably due to the nature of the dataset. Of interest is the performance of TextRazor at entity extraction and linking, being connected to a variety of linking resources.

In general, where entities are annotated at tweet level, it is unclear whether these are based on mentions of an entity in the text, or are meant to be mentions of a topic describing the text. Examination suggests that this is a primary source of problems when using such corpora for entity linking evaluation.

[8] See http://developer.zemanta.com/.

[9] See http://www.tagdef.org.

[10] See http://www.textrazor.com/technology.

System	Precision	Recall	F1
DBpedia Spotlight	20.1%	**47.5%**	28.3%
TextRazor	**64.6%**	26.9%	38.0%
Zemanta	57.7%	31.8%	**41.0%**

Table 7: Performance of entity linking over the WSDM2012 tweet dataset; note low performance due to varying levels of specificity in tweet-level annotations.

2.5.3 Entity Linking Summary

To summarise, entity linking in microtexts is a challenging task, due to the limited surrounding context, as well as the implicit semantics of hashtags and user name mentions. Furthermore, comparing methods reliably is currently very difficult, due to the limited datasets available in this genre.

Variation in capitalisation is often an indicator of a pronominal use of a word (suggesting a named entity); the genre sometimes disregards capitalisation and performance suffers as a result. This can be attributed to prior named entity recognition problems.

3 Normalisation

Noisy environments such as microblog text pose challenges to existing tools, being rich in previously unseen tokens, elision of words, and unusual grammar. Normalisation is commonly proposed as a solution for overcoming or reducing linguistic noise [40]. The task is generally approached in two stages: first, the identification of orthographic errors in an input discourse, and second, the correction of these errors.

Example 2 shows an original microblog message, including a variety of errors, and the post-normalisation repaired version.

(2) Source text: *@DORSEY33 lol aw . i thought u was talkin bout another time . nd i dnt see u either !*

Normalised text: *@DORSEY33 lol aww . I thought you was talking about another time . And I didn't see you either !*

As can be seen, not all the errors can be corrected (*was* ought to be *were*, for example) and some genre-specific slang remains – thought not in a particularly ambiguous sense or grammatically crucial place. Note the frequent shortening of words in messages entered by users, possibly attributable to both their effort to minimise the energy cost of communication and also out of habit of fitting within the tight message length limits typical of the genre.

Custom approaches to normalisation have been developed for Twitter. In this section, we present state-of-the-art normalisers; following this, we evaluate the effectiveness of two approaches to improving entity extraction and linking.

3.1 Normalisation Approaches

Normalisation approaches are typically based on a correction list, edit-distance based, cluster-based, or a mixture of these, with hybrid approaches common. We present a brief survey of the most recent techniques for normalising, in general and for Twitter.

One common source of errors across many genres is variation of spelling of proper names. Correct detection and resolution of a name may be partially dependent on it being spelled correctly. Therefore, resources have been developed to capture the many variations on spellings seen for given entities. For example, JRC-Names [42] is a list-based collection of name variations for many entities, coupled with an algorithm for matching target words to a given entity.

Gazetteer-based approaches can be used to repair errors on all kinds of words, not just named entities; for example, Han et al. [19] construct a general-purpose spelling correction dictionary for microblogs. This achieves state-of-the-art performance on both the detection of mis-spelled words and also applying the right correction.

Instead of a fixed list of variations, it is also possible to use a heuristic to suggest correct spellings. Both text edit distance and phonetic distance can be used to find candidate matches for words identified as mis-spelled. Han and Baldwin [18] achieved good corrections in many cases by using a combination of Levenshtein distance and double-metaphone distance between known words and words identified as incorrectly entered.

Some choose to view normalisation as recovery from a noisy channel, where the well-formed utterance is the original signal and the source of noise is the user's interpretation of that utterance as a microblog message. This model has been used with some success in, for example, interpreting SMS text [7].

In the case of microblogs, our analysis of difficult unknown words in POS tagging (in Section 2.3) showed that slang made up for three times as many mis-taggings as did orthographic errors. To this end, as well as investigating a heuristic normaliser, we investigate the performance impact that a custom gazetteer-based general-purpose pre-recognition dictionary has on NER.

3.2 Normalisation Evaluation Setup

Linguistic noise is an intrinsic quality of user generated content, especially in the microblog genre. As discussed in Section 2.3, POS tagging and tokenisation are both prone to noise. We are primarily interested in the impact of noise and the noise correction afforded by normalisation when it comes to the named entity recognition task. This is because not only is this task critical to entity-level linking (e.g. semantic enrichments), but also because most of the entity linking approaches discussed above are capable of overcoming a degree of noise. In fact, the biggest problem with entity linking over microblog text is the terseness of the documents, which cannot be overcome directly by cleaning up existing data.

In light of our findings so far, we also investigate various approaches to improving performance of automatic approaches in this genre – the so-called "normalisation" task. This is typically broken into two parts – determining the wrongly-spelled words and then correcting them. Therefore, we evaluate approaches to normalisation and their impact on two different styles of NER. This constitutes a qualitative and quantitative evaluation of normalisation systems and their ability to enable high-quality semantic annotation of user-generated microblog content.

3.2.1 Basic and Strong Normalisation

Normalisation approaches typically include a dictionary of known correctly-spelled terms, and refer to in-vocabulary (IV) and out-of-vocabulary (OOV) terms with respect to this dictionary. We contrast "basic" and "strong" normalisation. Basic normalisation is designed to deal with the majority of errors detected at the POS tagging stage (which is performed at token-level and sensitive to unknown/mis-used/mis-spelled words). Strong normalisation is more flexible, taking a lightly-supervised automatic approach trained on an external dataset.

Our basic normalisation approach is roughly equivalent to a rule-engineered approach: find resources with errors, and compile a lookup list for translating erroneous tokens to their correct equivalent. This is computationally fast, and can cover the majority of instances of slang words and even common mis-spellings. However, creation of the translation set is labour-intensive, and the approach is not flexible when taken across domain, across languages or presented with new kinds of slang or nicknames.

Entity type	No norm	Basic norm	Strong norm
ANNIE-twitter pipeline			
Organisation	64%	64%	64%
Location	77%	77%	77%
Person	67%	**70%**	**70%**
Date	90%	**91%**	**91%**
Overall	80%	**81%**	**81%**
Stanford NER (twitter)			
Organisation	28.6%	28.6%	**29.6%**
Location	60.5%	60.5%	**62.6%**
Person	63.2%	**63.6%**	62.8%
Overall	49.1%	47.6%	**49.3%**

Table 8: Impact of various normalisation techniques on F1 measure when extracting named entities, using the ANNIE-twitter pipeline compared with Stanford NER with a Twitter-trained model.

The strong normalisation approach is taken from Han and Baldwin [18]. It uses a combination of Brown clustering, edit distance, phonetic distance (double metaphone), contextual features and a small fixed list of slang words. For our experiment, we use a representation of this normalisation approach – an exhaustive autogenerated lookup list [19] based on English tweets.

These basic and strong normalisation appraoches had different error coverage. Effectively, strong has higher recall – more wrong words can be corrected by the resource – but lower precision, in that some corrections are wrong. These precision differences mostly caused one of three types of error:

Sentiment change Some words can be corrected to another word with differing sentiment, that is orthographically close. Tolerating mistakes within a Levenshtein edit distance of 2 – a common limit in this task – allows *in-*, *im-*, *un-* and *an-* prefixes to be stripped, thus reversing the meaning of a word in some cases. For example, *unpossible* → *possible* and *untalented* → *talented*.

IV errors The core vocabulary for discriminating between IV/OOV can be too small, or have too low a weight over other factors. Corrections are accidentally made to correctly-spelled words as a result, e.g., *roleplay* → *replay* and *armwarmers* → *armorers*.

Proper name variations Corrections to the spelling of variations of family names were sometimes made incorrectly. These seem minor but alter meaning. For example, there are many variations of spellings of English surnames; *Walmesley* may be alternatively represented as *Walmsley* or *Walmesly*. This can create critical changes in meaning. From the data, *She has Huntington's* can be misnormalised to *She has Huntingdon's*, reducing the chances of this already minimal-context mention of a disease being correctly recognised and linked, especially given the reliance of many entity linking techniques upon textual context.

3.2.2 Normalisation Analysis

We measured the performance impact of basic and strong normalisation over two different types of NE approach – gazetteer/FSM based (ANNIE) and machine-learning based (Stanford CRF NER). Results are given in Table 8. It can be seen that normalisation has a much lower impact on gazetteer-based methods than on machine learning ones, which are somewhat weaker to begin with, as described in Section 2.4. This can be attributed to the way that ANNIE pays less attention to POS tags and surrounding context than many systems, and is thus somewhat robust in the face of orthographic errors and grammar mistakes. It is also interesting to see that both basic and strong normalisation improve ANNIE performance to the same extent. This suggests that the extra effort involved in complex normalisation techniques does not offer performance increases near the upper end of the performance scale.

The Stanford NER system places more reliance on context, and so might be more affected by normalisation, which is reflected in the figures given. In fact, it requires strong normalisation in order to realise performance benefits from any normalisation for this system. While per-category F1 scores are roughly the same (or even slightly improved) with basic normalisation, entities of other types are missed more frequently than without normalisation, leading to an overall decrease in performance. Use of full normalisation did not cause any more entities to be missed than the no-normalisation baseline, but did slightly reduce the false positive rate. Basic normalisation in fact decreases performance overall. Nevertheless, even given strong normalisation, performance increases remain modest.

4 Directions for Future Work

This paper demonstrated experimentally that semantic annotation of microblog text is a challenging task. Problems lie at each stage of the semantic annotation pipeline, having negative impact on each subsequent task. Some genre-adaptation has been attempted, often successfully, but far from the performance levels found on established text genres.

In this section, we identify common trends in the above findings. These lead to evidence-based suggestions for future research directions towards improved semantic annotation of microblog text.

4.1 Prevalent Errors

Common themes lie in the difficult cases for many stages, specific to this genre. Capitalisation causes problems for POS tagging, NER and entity linking. In each case, where capitalisation is used in well-formed text to differentiate between normal nouns and proper nouns, altering this information (e.g. through use of lower caps for convenience or all caps for emphasis) has led to incorrect decisions being made. Correcting capitalisation is difficult, especially in the cases of polysemous nouns that have named entity senses (e.g. *eat an apple* vs. *Apple Inc.*, or *the town Little Rock* vs. *throw a little rock*). It seems that contextual clues are the only disambiguator, but most POS tagging approaches already take these into account; there is perhaps just not enough combined capitalisation information and context to resolve this class of typing variation.

Typographic errors confuse many pre-linking stages, specifically tokenisation and POS tagging. Added, skipped and swapped letters have all been found to cause problems. Although normalisation offers some help, the state-of-the-art does not yet give a good enough balance between ignoring correct OOV words and correctly repairing mistyped IV words, demonstrated in the low impact that normalisation has on NER.

One defining aspect of microtext is short messages, often enforced by a character limit. This limit demands that users reduce anything but the shortest utterance to a shorter form. The constraint on the amount of information content per message encourages using a compressed form of language. This leads to rare, uncommon or incomplete grammatical structures being used, as well as abbreviations, heavy pronoun use and other shortenings. Typically, one would use linguistic context to unpack these information-dense, peculiarly-structured utterances; however, that is also at a premium. This creates many problems, and perhaps the best way to overcome it is to create large and more-richly annotated resources.

The lack of context is a particular problem in the entity linking stage. Even the best performing systems reach scores much lower than they would on well-formed text. As with other linguistic disambiguation tasks, context is critical to resolving polysemous

words. Methods for increasing the amount of available context are therefore of interest.

Finally, there is a need for larger linguistic resources. Annotated microblog text is particularly rare, and in this difficult genre, very much needed. The performance of ML-based approaches relies on more annotated tweets.

4.2 Areas for Improvement

Named entity recognition is a bottleneck for semantic annotation pipelines. If they perform poorly at this task, entity linking accuracy (low already due to the lack of context), becomes even lower. Some Twitter-specific methods reach F1 measures of over 80%, but are still far from state-of-the-art. Next we discuss ways for improving the NER and entity linking stages.

Alternatives to normalisation As demonstrated, current normalisation approaches do not help with noise reduction. At best, they offer only +1% F1 improvements, while simple methods can even be harmful, due to mis-normalisation of correct words. Fuzzy list-based resources like JRC-Names may be more effective at detecting and resolving variations of spellings of named entities. The precision/recall balance still remains to be explored and optimised.

Better data Even though inroads have been made, current methods for semantic annotation of microtexts have many limitations. Firstly, entity recognition and linking does not reach the significantly higher precision and recall results obtained on longer text documents. One way to improve the currently poor automatic performance is through crowdsourcing. The ZenCrowd system [11], for instance, combines algorithms for large-scale entity linking with human input through micro-tasks on Amazon Mechanical Turk. In this way, NE mentions that can be linked automatically and with high confidence to instances in the Linked Open Data cloud, are not shown to the human annotators. The latter are only consulted on hard-to-solve cases, which not only significantly improves the quality of the results, but also limits the amount of manual intervention required.

Secondly, it is axiomatic that semantic annotation methods are only as good as their training and evaluation data. Algorithm training on microblog gold standard datasets is currently very limited. For example, there are currently fewer than 10,000 tweets annotated with named entity types. Bigger, shared evaluation corpora are therefore badly needed. Creating these through traditional manual text annotation methodologies is unaffordable, if a significant mass is to be reached. Research on crowdsourcing evaluation gold standards has been limited, primarily with focus on using Amazon Mechanical Turk to acquire small datasets (e.g. tweets with named entity types) [13].

The few ground truth errors in the POS dataset are comparable to error rates in POS datasets in general. However, recognising that they have broken new ground and made this possible at all, evaluating and comparing microtext entity linking approaches on currently-available entity linking sets is still a delicate process.

Unexploited information Apart from noise, shortage of contextual information is the biggest problem. Luckily, the metadata of microblog messages (and the typical structure of microblogging sites) gives access to text outside of a given message. This abundant linguistic context is available but currently unexplored.

To start with, for extra context, it is possible to examine other messages belonging to the creator. One may access more content from the user, giving a better sample of their writing style and perhaps providing instances of easier-to-disambiguate entities when faced with a difficult text. Links posted by the user, and the content of their user profile, also provide additional linguistic context.

Outside of the creator's content, methods may examine their friends' content for use as additional context. One might also link to other sites, using a multi-layered social network model [25], to extract a user's and their friends' messages on other sites.

Lastly, as temporal information is available, it is possible to mine contemporary events to check for co-mentions of NEs. Also, one can use spatial metadata to check for special word senses near a user, informing for example resolution of the *little rock (stone)*/*Little Rock (town)* confusion mentioned earlier. Using spatial and temporal information, one can search for stories and content outside the author's social network to find descriptions of nearby occurrences.

5 Conclusion

This paper has explored a number of research questions arising from applications of semantic annotation to microblogs. Experiments demonstrate that current methods are not sufficiently robust on this ill-formed, terse, and linguistically compressed content.

To answer our second research question, error analysis was carried out at the semantic annotation stage and specific difficulties with the task were identified. In summary, all machine learning methods suffer from the lack of annotated training data, while entity linking methods also suffer from the lack of sufficient context. Normalisation, as a noise reduction method, is not a silver bullet, although it is helpful and could be improved further by precision/recall tuning.

To overcome these problems, we argue that microblog-specific metadata and content needs to be brought into the algorithms, as a way to combat the shortness of the microtext itself. More specifically, we propose including other posts from the same author, information from the author's network and spatio-temporal information.

Specific to linking, state-of-the-art algorithms rely substantially on the surrounding textual context for disambiguation. However, such context is extremely limited in microblog documents, and cannot be improved significantly through additional information from e.g. hashtag definitions. This makes entity linking in microblogs a particularly challenging, open research problem. Other methods of combating this problem involve leveraging structured knowledge bases such as Wikipedia to improve the semantic annotation of microposts with the hope that injecting more semantic information will counterbalance the lack of context.

6 Acknowledgments

This work was partially supported by the UK EPSRC grants Nos. EP/I004327/1 and EP/K017896/1 uComp,[11] and by the European Union under grant agreements No. 287863 TrendMiner,[12] and No. 270239 Arcomem.[13]

7 References

[1] F. Abel, Q. Gao, G. Houben, and K. Tao. Semantic enrichment of twitter posts for user profile construction on the social web. *The Semanic Web: Research and Applications*, pages 375–389, 2011.

[2] E. Amigó, A. Corujo, J. Gonzalo, E. Meij, and M. Rijke. Overview of Replab 2012: Evaluating online reputation management systems. In *CLEF 2012 Labs and Workshop Notebook Papers*, 2012.

[3] K. Bontcheva and H. Cunningham. Semantic annotation and retrieval: Manual, semi-automatic and automatic generation. In J. Domingue, D. Fensel, and J. A. Hendler, editors, *Handbook of Semantic Web Technologies*. Springer, 2011.

[4] K. Bontcheva and D. Rout. Making sense of social media streams through semantics: a survey. *Semantic Web Journal*, 2012.

[11] See http://www.chistera.eu/projects/ucomp.

[12] See http://www.trendminer-project.eu/.

[13] See http://www.arcomem.eu.

[5] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, 2013.

[6] W. Cavnar and J. Trenkle. N-gram-based text categorization. In *Proceedings of the Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[7] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10(3):157–174, 2007.

[8] S. Cucerzan. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of EMNLP-CoNLL*, volume 6, pages 708–716, 2007.

[9] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the Meeting of the Association for Computational Linguistics*, 2002.

[10] D. Damljanovic and K. Bontcheva. Named Entity Disambiguation using Linked Data. In *Proceedings of the 9th Extended Semantic Web Conference (ESWC)*, 2012.

[11] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proceedings of the 21st Conference on World Wide Web*, pages 469–478, 2012.

[12] L. Derczynski, B. Yang, and C. Jensen. Towards Context-Aware Search and Analysis on Social Media Data. In *Proceedings of the 16th Conference on Extending Database Technology*. ACM, 2013.

[13] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88, 2010.

[14] J. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.

[15] J. Foster, Ö. Çetinoglu, J. Wagner, J. Le Roux, S. Hogan, J. Nivre, D. Hogan, J. Van Genabith, et al. #hardtoparse: POS Tagging and Parsing the Twitterverse. In *Proceedings of the AAAI Workshop On Analyzing Microtext*, pages 20–25, 2011.

[16] N. Freire, J. Borbinha, and P. Calado. An approach for named entity recognition in poorly structured data. *The Semantic Web: Research and Applications*, pages 718–732, 2012.

[17] M. Greenwood, N. Aswani, and K. Bontcheva. Reputation Profiling with GATE. In *CLEF 2012 Labs and Workshop Notebook Papers*, 2012.

[18] B. Han and T. Baldwin. 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, 2011.

[19] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the conference on Empirical Methods in Natural Language Processing*, pages 421–432. ACL, 2012.

[20] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based POS taggers. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 278–277. ACL, 2000.

[21] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[22] X. Liu, M. Zhou, F. Wei, Z. Fu, and X. Zhou. Joint inference of named entity recognition and normalization for tweets. In *Proceedings of the Association for Computational Linguistics*, pages 526–535, 2012.

[23] U. Lösch and D. Müller. Mapping microblog posts to encyclopedia articles. *Lecture Notes in Informatics*, 192(150), 2011.

[24] M. Lui and T. Baldwin. langid. py: An off-the-shelf language identification tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[25] M. Magnani and L. Rossi. The ML-model for multi-layer social networks. In *Proceedings of the conference on Advances in Social Networks Analysis and Mining*, pages 5–12. IEEE, 2011.

[26] M. Marcus, M. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993.

[27] M. Marrero, S. Sanchez-Cuadrado, J. Lara, and G. Andreadakis. Evaluation of named entity extraction systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58, 2009.

[28] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In *Proceedings of the @NLP can u tag #usergeneratedcontent?! workshop, LREC*, pages 15–22, 2012.

[29] P. Mcnamee and J. Mayfield. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1):73–97, 2004.

[30] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In *Proc. of the Fifth Int. Conf. on Web Search and Data Mining (WSDM)*, 2012.

[31] P. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.

[32] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[33] D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *Proceedings of the workshop on Real-Time Analysis and Mining of Social Streams*, 2012.

[34] D. Rao, P. McNamee, and M. Dredze. Entity linking: Finding extracted entities in a knowledge base. In *Multi-source, Multi-lingual Information Extraction and Summarization*. 2011.

[35] A. Ritter, S. Clark, O. Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. ACL, 2011.

[36] G. Rizzo and R. Troncy. Nerd: evaluating named entity recognition tools in the web of data. In *Workshop on Web Scale Knowledge Extraction (WEKEX)*, pages 1–16, 2011.

[37] A. Roberts, R. Gaizauskas, M. Hepple, and Y. Guo. Combining terminology resources and statistical methods for entity recognition: an evaluation. *Proceedings of the conference on Language Resources and Evaluation*, 2008.

[38] D. Rout, K. Bontcheva, and M. Hepple. Reliably evaluating summaries of twitter timelines. In *Proceedings of the AAAI Workshop on Analyzing Microtext*, 2013.

[39] W. Shen, J. Wang, P. Luo, and M. Wang. LINDEN: Linking named entities with knowledge base via semantic knowledge. In *Proceedings of the 21st Conference on World Wide Web*, pages 449–458, 2012.

[40] R. Sproat, A. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards. Normalization of non-standard words. *Computer Speech & Language*, 15(3):287–333, 2001.

[41] M. Stankovic, M. Rowe, and P. Laublet. Finding co-solvers on twitter, with a little help from linked data. *The Semantic Web: Research and Applications*, pages 39–55, 2012.

[42] R. Steinberger, B. Pouliquen, M. Kabadjov, J. Belyaeva, and E. van der Goot. JRC-Names: A freely available, highly multilingual named entity resource. In *Proceedings of the 8th International Conference in Recent Advances in Natural Language Processing*, pages 104–110, 2011.

[43] E. Tjong Kim Sang and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh Conference on Natural Language Learning*, pages 142–147. ACL, 2003.

[44] K. Toutanova, D. Klein, C. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the conference of the N. American Chapter of the Association for Computational Linguistics*, pages 173–180, 2003.