TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text

Kalina Bontcheva, Leon Derczynski, Adam Funk,
Mark A. Greenwood, Diana Maynard, Niraj Aswani
University of Sheffield
Initial.Surname@dcs.shef.ac.uk

Abstract

Twitter is the largest source of microblog text, responsible for gigabytes of human discourse every day. Processing microblog text is difficult: the genre is noisy, documents have little context, and utterances are very short. As such, conventional NLP tools fail when faced with tweets and other microblog text. We present TwitIE, an open-source NLP pipeline customised to microblog text at every stage. Additionally, it includes Twitter-specific data import and metadata handling. This paper introduces each stage of the TwitIE pipeline, which is a modification of the GATE ANNIE open-source pipeline for news text. An evaluation against some state-of-the-art systems is also presented.

1 Introduction

Researchers have started recently to study the problem of mining social media content automatically (e.g. (Rowe et al., 2013; Nagarajan and Gamon, 2011; Farzindar and Inkpen, 2012; Bontcheva and Rout, 2013)). The focus of this paper is on information extraction, but other active topics include opinion mining (Maynard et al., 2012; Pak and Paroubek, 2010), summarisation (e.g. (Chakrabarti and Punera, 2011)), and visual analytics and user and community modelling (Bontcheva and Rout, 2013). Social media mining is relevant in many application contexts, including knowledge management, competitor intelligence, customer relation management, eHealth, and eGovernment.

Information extraction from social media content 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 (Derczynski et al., 2013a). Simple domain adaptation techniques (e.g. (Daumé and Marcu, 2007) are not so useful on this genre, in part due to its unusual structure and representation of discourse, which can switch between one-to-one conversation, multiparty conversation and broadcast messages. For instance, named entity recognition methods typically have 85-90% accuracy on longer texts, but 30-50% on tweets (Ritter et al., 2011; Liu et al., 2012).

This paper introduces the TwitIE information extraction system, which has been specifically adapted to microblog content. It is based on the most recent GATE (Cunningham et al., 2013) algorithms and is available as a GATE plugin available to download from `https://gate.ac.uk/wiki/twitie.html`, usable both via the GATE Developer user interface and via the GATE API. Comparisons against other state-of-the-art research on this topic are also made.

2 Related Work

In terms of Named Entity Recognition (NER), and Information Extraction (IE) in general, microblogs are possibly the hardest kind of content to process. First, their shortness (maximum 140 characters for tweets) makes them hard to interpret. Consequently, ambiguity is a major problem since IE 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 information extraction algorithms (e.g. named entity recognition for Twitter using CRFs (Ritter et al., 2011), Wikipedia-based topic and entity disambiguation (van Erp et al., 2013)). 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 (Derczynski et al., 2013a; Han and Baldwin, 2011; Han et al., 2012).

Named entity recognition of longer texts, such as news, is a very well studied problem (cf. (Nadeau and Sekine, 2007; Roberts et al., 2008; Marrero et al., 2009)).

For Twitter, some approaches have been proposed but often they are not freely available. Ritter et al. (Ritter et al., 2011) take a pipeline approach performing first tokenisation and POS tagging before using topic models to find named entities. Liu (Liu et al., 2012)

Proceedings of Recent Advances in Natural Language Processing, pages 83–90,
Hissar, Bulgaria, 7-13 September 2013.

propose a gradient-descent graph-based method for doing joint text normalisation and recognition, reaching 83.6% F1 measure.

We have also included in our evaluation of TwitIE, a Twitter-adapted version of the state-of-the-art Stanford NER (Finkel et al., 2005), which we trained using both tweets and newswire. It uses a machine learning-based method to detect named entities, and is distributed with CRF models for English newswire text.

NER apart, other actively researched IE topics are entity disambiguation (e.g. (Davis et al., 2012; van Erp et al., 2013)), event extraction and summarisation (e.g. (Becker et al., 2011b; Becker et al., 2011a; Chakrabarti and Punera, 2011)), and opinion mining (e.g. (Maynard et al., 2012; Pak and Paroubek, 2010)) to name just a few. Since at present, TwitIE's focus is currently on named entity recognition, we will not compare against these methods. In future work, TwitIE will be extended towards entity disambiguation and relation extraction.

3 The TwitIE IE Pipeline

The open-source GATE NLP framework (Cunningham et al., 2013) comes pre-packaged with the ANNIE general purpose IE pipeline (Cunningham et al., 2002). ANNIE consists of the following main processing resources: tokeniser, sentence splitter, POS tagger, gazetteer lists, finite state transducer (based on GATE's built-in regular expressions over annotations language), orthomatcher and coreference resolver. The resources communicate via GATE's annotation API, which is a directed graph of arcs bearing arbitrary feature/value data, and nodes rooting this data into document content.

The ANNIE components can be used individually or coupled together with new modules in order to create new applications. TwitIE re-uses the sentence splitter and name gazetteer components unmodified, though we re-trained and adapted all other components to the specifics of this genre.

The rationale behind adopting the sentence splitter unmodified, is that in most cases it tends to consider the text of the entire tweet as one sentence. Due to the limited local context, this did not present problems for the later components. Nevertheless, a more in-depth evaluation of the sentence splitter errors is necessary and envisaged as part of future work.

Similarly, the reuse of the ANNIE gazetteer lists was sufficient for the time being, due to their very generic nature (e.g. country names, days of the week, months, first names). However, the TwitIE POS tagger does come with customised in-built gazetteer lists, used for tagging unambiguous named entities, e.g. YouTube, Twitter, Yandex (see (Derczynski et al., 2013b) for details on the lists and how they were created and used).

For the rest of the TwitIE components, adaptation to the specifics of the microblog genre is required, in order to address the genre-specific challenges of noisiness, brevity, idiosyncratic language, and social context. General-purpose tools (e.g. POS taggers and entity recognisers) do particularly badly on such texts (see Sections 3.5 and 3.6).

Therefore, we have developed TwitIE – a customisation of ANNIE, specific to social media content, which has been tested most extensively on microblog messages.

Figure 1 shows the TwitIE pipeline and its components. TwitIE is distributed as a plugin in GATE, which needs to be loaded for these processing resources to appear in GATE Developer. Re-used ANNIE components are shown in dashed boxes, whereas the ones in dotted boxes are new and specific to the microblog genre.

The first step is language identification, which is discussed next (Section 3.2), followed by the TwitIE tokeniser (Section 3.3).

The **gazetteer** consists of lists such as cities, organisations, days of the week, etc. It not only consists of entities, but also of names of useful *indicators*, such as typical company designators (e.g. 'Ltd.'), titles, etc. The gazetteer lists are compiled into finite state machines, which can match text tokens. TwitIE reuses the ANNIE gazetteer lists, at present, without any modification.

The **sentence splitter** is a cascade of finite-state transducers which segments text into sentences. This module is required for the POS tagger. The ANNIE sentence splitter is reused without modification, although when processing tweets, it is also possible to just use the text of the tweet as one sentence, without further analysis.

The normaliser, the adapted POS tagger, and named entity recognition are discussed in detail in Sections 3.4, 3.5, and 3.6 respectively.

3.1 Tweet Import

The ability to collect corpora is particularly important with social media. Twitter, for example, currently forbids distribution of whole tweets, and so instead tweet corpora are distributed via tweet ID. Data is delivered from the Twitter API in JSON format. This is currently a process external to GATE, although we plan to address this in future work.

In the most recent GATE codebase, we added a new `Format_Twitter` plugin, which coverts automatically tweets in JSON, into fully-annotated GATE documents.

The JSON format ceonvertor is automatically associated with les whose names end in .json; otherwise the user needs to specify `text/x-json-twitter` as the document mime type. The JSON import works both when creating a single new GATE document and when populating a corpus.

Each tweet objects text value is converted into the document content, which is covered with a Tweet annotation whose features represent (recursively when appropriate, using HashMap and List) all the other key-value pairs in the tweet JSON object.

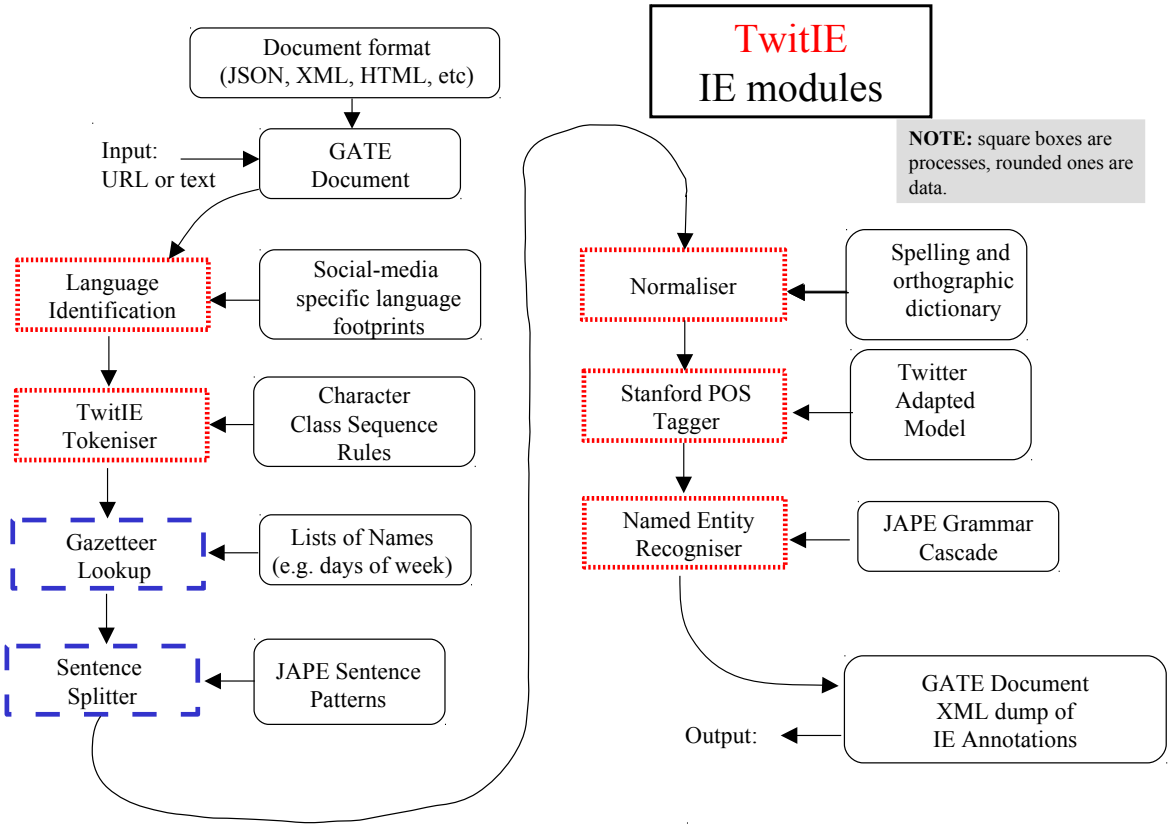

Figure 1: The TwitIE Information Extraction Pipeline

Multiple tweet objects in the same JSON le are separated by blank lines (which are not covered by Tweet annotations.

3.2 Language Identification

The TwitIE system uses the TextCat (Cavnar and Trenkle, 1994) language identification algorithm, which relies on n-gram frequency models to discriminate between languages. More specifically, we have integrated the TextCat adaptation to Twitter (Carter et al., 2013) which works currently on five languages. It is 97.4% accurate overall, with per language accuracy ranging between 95.2% for French and 99.4% for English (Derczynski et al., 2013a). These results demonstrate that language identification is hard on tweets, but nevertheless, can be achieved with reasonable accuracy.

Due to the shortness of tweets, TwitIE makes the assumption that each tweet is written in only one language. The choice of languages used for categorisation is specified through a configuration file, supplied as an initialisation parameter.

Figure 2 shows three tweets – one English, one German, and one French. TwitIE TextCat was used to assign automatically the lang feature to the tweet text (denoted by the Tweet annotation).

Given a collection of tweets in a new language, it is possible to train TwitIE TextCat to support that new language as well. This is done by using the Fingerprint Generation PR, included in the `Language_Identification` plugin. It builds a new ngerprint from a corpus of documents.

Reliable tweet language identification allows us to only process those tweets written in English with the TwitIE English POS tagger and named entity recogniser. This is achieved by making the execution of these components conditional on the respective tweet being in English, by using a Conditional Corpus Pipeline. GATE also provides POS tagging and named entity recognition in French and German, so it is possible to extend TwitIE towards these languages with some training and adaptation effort.

3.3 Tokenisation

Commonly distinguished types of tokens are numbers, symbols (e.g., $, %), punctuation and words of different kinds, e.g., uppercase, lowercase, mixed case. Tokenising well-written text is generally reliable and reusable, since it tends to be domain-independent, e.g. the Unicode tokeniser bundled with the ANNIE system in GATE.

However, such general purpose tokenisers need to be adapted to work correctly on social media, in order to

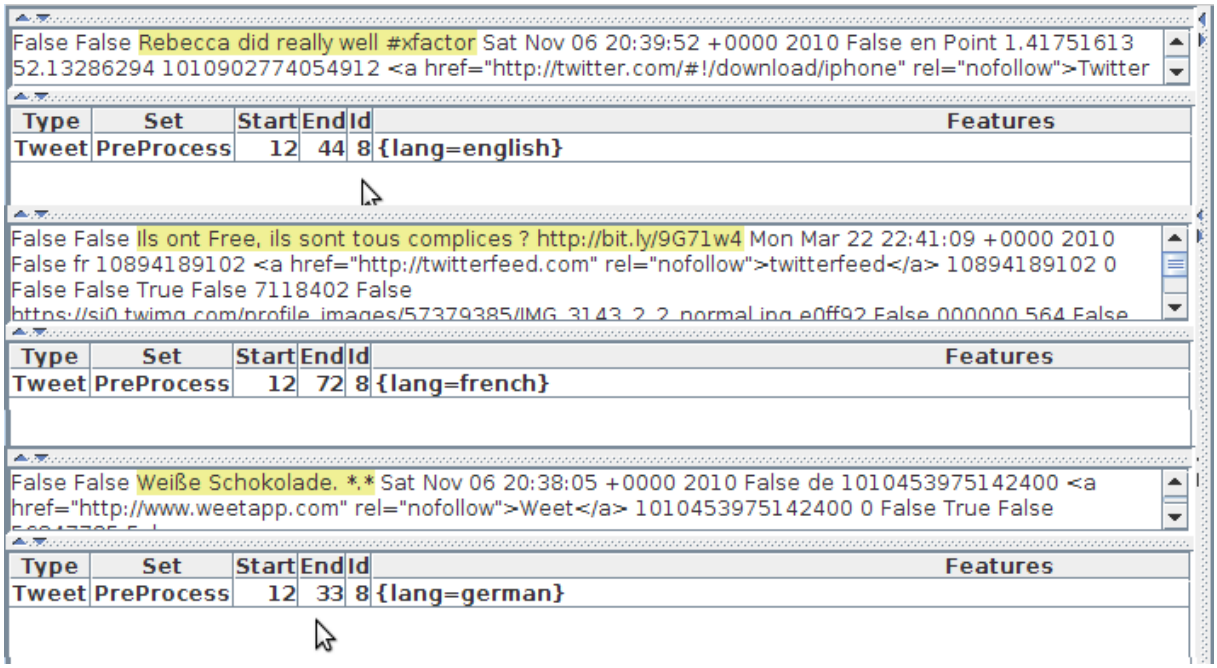

Figure 2: Example Tweets Annotated for Language

handle specific tokens like URLs, hashtags (e.g. #nlproc), user mentions in microblogs (e.g. @GateAcUk), special abbreviations (e.g. RT, ROFL), and emoticons. A study of 1.1 million tweets established that 26% of English tweets have a URL, 16.6% – a hashtag, and 54.8% – a user name mention (Carter et al., 2013). These elements prove particularly disruptive to conventional NLP tools (Derczynski et al., 2013a). Therefore, tokenising these accurately is important.

To take part of a tweet as an example:

```
#WiredBizCon #nike vp said when @Apple
saw what http://nikeplus.com did,
#SteveJobs was like wow I didn't...
```

One option is to tokenise on white space alone, but this does not work that well for hashtags and username mentions. In our example, if we have #nike and @Apple as one token each, this will make their recognition as company names harder, since the named entity recognition algorithm will need to look at sub-token level. Similarly, tokenising on white space and punctuation does not work well since URLs become split into many tokens (e.g. http, nikeplus), as do emoticons and email addresses.

The TwitIE tokeniser is an adaptation of ANNIE's English tokeniser. It follows Ritter's tokenisation scheme (Ritter et al., 2011). More specifically, it treats abbreviations (e.g. RT, ROFL) and URLs as one token each. Hashtags and user mentions are two tokens (i.e., `\#` and `nike` in the above example) with a separate annotation HashTag covering both. Capitalisation is preserved and an orthography feature added. Normalisation and emoticons are handled in optional separate modules, since information about them is not always needed. Consequently, tokenisation is fast and generic, while tailored to the needs of named entity recognition.

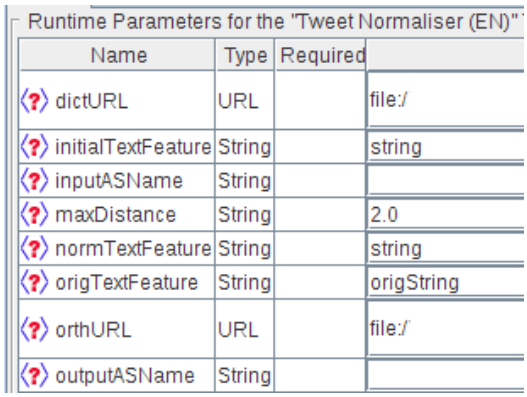

Figure 3: Configuration options for the TwitIE normaliser.

3.4 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 (Sproat et al., 2001). 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.

The TwitIE Normaliser is a combination of a generic spelling-correction dictionary and a spelling correction dictionary, specific to social media. The latter contains entries such as "2moro" and "brb", similar to Han et al. (2012). Figure 4 shows an example tweet, where the abbreviation "Govt" has been normalised to government.

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 misspelled. (Han and Baldwin, 2011) 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. We also experimented with this normalisation approach in TwitIE, and provide a toy corpus of various utterances that require normalisation. This method has higher recall (more wrong words can be corrected by the resource) but lower precision (some corrections are wrong).

3.5 Part-of-speech Tagging

Accuracy of the general-purpose English POS taggers is typically excellent (97-98%) on texts similar to those on which the taggers have been trained (mostly news articles). However, they are not suitable for microblogs and other short, noisy social media content, where their accuracy declines to 70-75% (Derczynski et al., 2013a).

TwitIE contains an adapted Stanford tagger (Toutanova et al., 2003), trained on tweets tagged with the Penn TreeBank (PTB) tagset. Extra tag labels have been added for retweets, URLs, hashtags and user mentions. We trained this tagger using hand-annotated tweets (Ritter et al., 2011), the NPS IRC corpus (Forsyth and Martell, 2007), and news text from PTB (Marcus et al., 1993). The resulting model achieves 83.14% token accuracy, which is still below that achieved on news content.

The most common mistakes (just over 27%) arise from words which are common in general, but do not occur in the training data, indicating a need for a larger training POS-tagged corpus of social media content. Another 27% of errors arise from slang words, which are ubiquitous in social media content and are also often misspelled (e.g. *LUVZ*, *HELLA* and *2night*) and another 8% from typos. Many of these can be addressed using normalisation (see Section 3.4). Close to 9% of errors arise from tokenisation mistakes (e.g. joined words). Lastly, 9% of errors are words, to which a label may be reliably assigned automatically, including URLs, hash tags, re-tweets and smileys, which we now pre-tag automatically with regular expressions and lookup lists.

Another frequently made mistake is tagging proper noun (`NN`/`NNP`) – an observation also made by (Ritter et al., 2011). Therefore, we use ANNIE's gazetteer lists of personal first-names and cities and, in addition, a list of unambiguous corporation and website names frequently-mentioned in the training data (e.g. *YouTube*, *Toyota*).

By combining normalisation, gazetteer name lookup, and regular expression-based tagging of Twitter-specific POS tags, we increase performance from 83.14% accuracy to 86.93%. By generating additional 1.5M training tokens from tweets annotated automatically using two existing POS taggers (namely (Ritter et al., 2011) and (Gimpel et al., 2011)), we further improve the performance of our Twitter-adapted tagger to 90.54% token accuracy using the PTB tagset (better than state-of-the-art).

Figure 4 shows an example tweet, which has been tagged both without normalisation (upper row of POS tags) and with tweet normalisation (the lower row of POS tags). The word *"Govt"* is normalised to government, which is then tagged correctly as `NN`, instead of `NNP`.

3.6 Named Entity Recognition

Named entity recognition (**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 how the default ANNIE named entity recognition pipelines performs in comparison to a Twitter-specific approach, on a corpus of 2 400 tweets comprising 34 000 tokens (Ritter et al., 2011).

We did not consider Percent-type entity annotations in these evaluations because there were so few (3 in the whole corpus) and they were all annotated correctly. Note also that twitter-specific UserID annotation as a Person annotation is *not* included in these results, as they can be matched using a simple, public regular expression provided by Twitter, and as a result were all 100% correct.

As we can see in Table 1, the performance of ANNIE and the Stanford NER tagger degrades significantly on microblog content, in comparison to newswire, which motivates the need for microblog domain adaptation. Thanks to adaptation in the earlier components in TwitIE (especially the POS tagger (Derczynski et al., 2013b)), we demonstrate a +30% absolute precision and +20% absolute F1 performance increase, as compared to ANNIE, mainly with respect to Date, Organization and in particular Person. TwitIE also outperforms Ritter's Twitter NER algorithm (Ritter et al., 2011) and our adaptation of the Stanford NER, which we trained using both tweets and newswire (see (Derczynski et al., 2013a) for details).

However, as shown in Table 1, when compared against state-of-the-art NER performance on longer news content, an overall F1 score of 80% leaves notable amounts of missed annotations and false positives.

Labelling Organizations in tweets proved particularly hard, where errors were often caused by miscategorisations. 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, due to insufficient surrounding disambiguating context (typical in microblogs). For example, *Amazon* was labelled as a Location when it should have been an Organization.

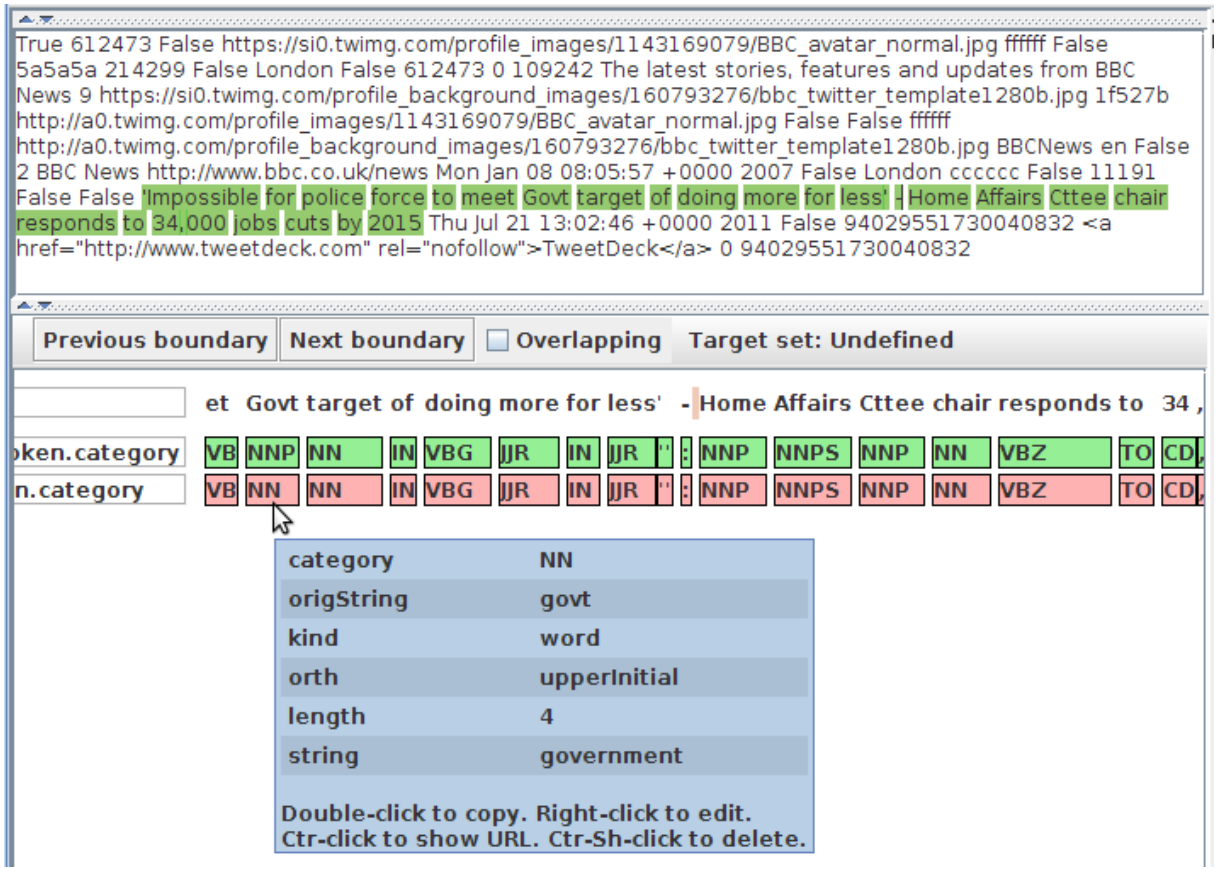

Figure 4: Comparing POS Tagger Output: A Normalisation Example

System	Precision	Recall	F1
	Newswire		
ANNIE	78%	74%	77%
Stanford	-	-	**89%**
	Microblog		
ANNIE	47%	**83%**	60%
TwitIE	**77%**	**83%**	**80%**
Stanford	59%	32%	41%
Stanford-twitter	54%	45%	49%
Ritter	73%	49%	59%

Table 1: 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.

NEs represented in lowercase (e.g. *skype*) were frequently ignored. However, handling capitalisation is hard from trivial (Derczynski et al., 2013a) and this is an area where we plan more future work, combined with the creation of a larger, human-annotated corpus of NER-annotated tweets.

4 Conclusion

This paper presented the TwitIE open-source NER pipeline, specifically developed to handle microblogs. Issues related to microblog NER were discussed, and the requirement for domain adaptation demonstrated. As can be seen from the evaluation results reported here, significant inroads have been made into this challenging problem. By releasing TwitIE as open source, we hope to give researchers also an easily repeatable, baseline system against which they can compare new Twitter NER algorithms.

As already discussed, there is still a significant gap in NER performance on microblogs, as compared against news content. This gap is due to some degree to insufficient linguistic context and the noisiness of tweets. However, there is also a severe lack of labeled training data, which hinders the adaptation of state-of-the-art NER algorithms, such as the Stanford CRF tagger. These are all areas of ongoing and future work, as well as the adaptation of the entire TwitIE pipeline to languages other than English.

Acknowledgments

This work was supported by UK EPSRC grants Nos. EP/I004327/1 and EP/K017896/1 uComp,[1] and by the European Union under grant agreement No. 270239 Arcomem.[2]

References

H. Becker, M. Naaman, and L. Gravano. 2011a. Selecting Quality Twitter Content for Events. In *Pro-*

[1] http://www.ucomp.eu
[2] http://www.arcomem.eu

ceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM).

H. Becker, M. Naaman, and L. Gravano. 2011b. Beyond Trending Topics: Real-World Event Identification on Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*.

K. Bontcheva and D. Rout. 2013. Making sense of social media through semantics: A survey. *Semantic Web - Interoperability, Usability, Applicability*.

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

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

D. Chakrabarti and K. Punera. 2011. Event Summarization Using Tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM)*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. Gate: an architecture for development of robust hlt applications. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 168–175.

H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva. 2013. Getting more out of biomedical documents with gate's full lifecycle open source text analytics. *PLoS Computational Biology*, 9(2):e1002854, 02.

H. Daumé and D. Marcu. 2007. Frustratingly easy domain adaptation. In *Proceedings of the Annual meeting of the Association for Computational Linguistics*.

A. Davis, A. Veloso, A. Soares, A. Laender, and W. Meira Jr. 2012. Named entity disambiguation in streaming data. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, Jeju Island, Korea, July. Association for Computational Linguistics.

L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. 2013a. Microblog-Genre Noise and Impact on Semantic Annotation Accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM.

L. Derczynski, A. Ritter, S. Clark, and K. Bontcheva. 2013b. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*. Association for Computational Linguistics.

A. Farzindar and D. Inkpen, editors. 2012. *Proceedings of the Workshop on Semantic Analysis in Social Media*. Association for Computational Linguistics, Avignon, France, April.

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

E. Forsyth and C. Martell. 2007. Lexical and discourse analysis of online chat dialog. In *International Conference on Semantic Computing*, pages 19–26. IEEE.

K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. Smith. 2011. Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47. ACL.

B. Han and T. Baldwin. 2011. Lexical normalisation of short text messages: makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, HLT '11, pages 368–378.

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

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

M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

M. Marrero, S. Sanchez-Cuadrado, J. Lara, and G. Andreadakis. 2009. Evaluation of Named Entity Extraction Systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

D. Maynard, K. Bontcheva, and D. Rout. 2012. Challenges in developing opinion mining tools for social media. In *Proceedings of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012*, Turkey.

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

M. Nagarajan and M. Gamon, editors. 2011. *LSM '11: Proceedings of the Workshop on Languages in Social Media*, Stroudsburg, PA, USA. Association for Computational Linguistics.

A. Pak and P. Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*.

A. Ritter, S. Clark, Mausam, and O. Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, Edinburgh, UK.

A. Roberts, R. Gaizauskas, M. Hepple, and Y. Guo. 2008. Combining terminology resources and statistical methods for entity recognition: an evaluation. *Proceedings of the Conference on Language Resources and Evaluation (LRE'08)*.

M. Rowe, M. Stankovic, A. Dadzie, B. Nunes, and A. Cano. 2013. Making sense of microposts (#msm2013): Big things come in small packages. In *Proceedings of the WWW Conference - Workshops*.

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

K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, NAACL '03, pages 173–180.

M. van Erp, G. Rizzo, and R. Troncy. 2013. Learning with the Web: Spotting Named Entities on the intersection of NERD and Machine Learning. In *Proceedings of the 3^{rd} Workshop on Making Sense of Microposts (#MSM2013)*.