Identifying Twitter Location Mentions

Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay, Qiang Chen
IBM Research
Melbourne, VIC, Australia
bohan.ibm@au1.ibm.com, antonio.jimeno@au1.ibm.com,
admackin@au1.ibm.com, qiangchen@au1.ibm.com

Abstract

This paper describes our system in the ALTA shared task 2014. The task is to identify location mentions in Twitter messages, such as place names and point-of-interests (POIs). We formulated the task as a sequential labelling problem, and explored various features on top of a conditional random field (CRF) classifier. The system achieved 0.726 mean-F measure on the held-out evaluation data. We discuss our results and suggest ideas for future work on location mention recognition in social media.

1 Introduction

The ALTA shared task 2014 aims to identify location mentions in Twitter data. The input is plain text messages, and the expected output is location entities such as country names, city names and POIs for each message. For instance, *Auckland* and *#eqnz* are identified as location mentions in *@USER are you considering an Auckland visit after #eqnz today?*.[1] This shared task is very similar to a well-established NLP task — named entity recognition (NER) but with a focus on location entities in social media. Each token in a text message is categorised as either a location mention or not. The nearby tokens (i.e., context) may influence a token's labelling, hence we incorporate context information in our system. Following the literature on NER (Lingad et al., 2013), we formulate it as a sequential labelling task and use a conditional random field (CRF) as the classifier.

The main contributions of the paper are: (1) A sequential labeller for identifying location mentions in social media; (2) Feature analysis and comparison in NER between social media and other genres. (3) Discussion on errors and extensions to current sequential labeller.

[1] *#eqnz* is a short form for earthquake in New Zealand.

2 Challenges

Although CRF models for NER are widely used and are reported to achieve state-of-the-art results in literature (Finkel et al., 2005; Liu et al., 2011; Ritter et al., 2011), NER in social media still raises several non-trivial challenges.

First, Twitter text is noisy, with more non-standard words than polished text (Baldwin et al., 2013) including typos (e.g., *challanges* "challenges"), abbreviations (e.g., *ppl* "people") and phonetic substitutions (e.g., *4eva* "forever"). These non-standard words often cause generalisation issues (Han and Baldwin, 2011). For instance, lexical variants (e.g., *Melb*, *Mel*, *melbn*) will not be recognised in the test data when only standard forms (e.g., "Melbourne") are observed in the training data.

In addition to non-standard words, informal writing style further reduces NER accuracy. One example is that conventional features relying on capitalisation are less reliable. For instance, *LOL* is capitalised but it is not a location entity, while *brisbane* may be a valid location mention even though it is in lowercase.

Similarly, Twitter specific entities sometimes are sentence constituents, e.g., *#Melbourne* in *#Melbourne is my fav city.* However, they may be a topic tag that does not form part of the syntactic structure of the sentence, such as the hashtags in *I like travel to beautiful places, #travel #melbourne*, in which case syntactic features would be less effective.

For this reason, widely-used NER features may need to be re-engineered for use over social media text.

Bo Han, Antonio Jimeno Yepes, Andrew MacKinlay and Qiang Chen. 2014. Identifying Twitter Location Mentions. In *Proceedings of Australasian Language Technology Association Workshop*, pages 157−162.

3 Feature Engineering

3.1 Related work for NER

The starting point for our features comes from some other representative systems that are summarised in Table 1.

STANFORD NER (Finkel et al., 2005) combined Gibbs sampling and a widely used CRF model. The Gibbs sampling offers non-local constraints to the conventional CRF model that utilises a range of local features. The features in the CRF model are based on words, POS tags, character *n*-grams , word shapes and the presence of words in a predefined window. The word and POS tag features also include the surrounding tokens and tags to capture the local context information.

Liu et al. (2011) proposed a two-stage CRF-based tagger MSRA for Twitter NER. First, a *k*-NN classifier pre-categorises words, and then feeds results to a downstream CRF modeller. The features they adopted in *k*-NN are two word text windows including the target word (i.e., five words in total). The gazetted resources (from Wikipedia) are also utilised and shown to be effective in their experiments. As for the features for building the second stage CRF model, they followed Ratinov and Roth (2009) and made use of tokens, word types (e.g., whether the word is alphanumeric or capitalised), word morphological features (e.g., suffix and prefix of words), previous tagging labels, word context windows, and conjunction features that combine both tags and word context windows.

Recently, another WASHINGTON NER tool (Ritter et al., 2011) was developed by rebuilding a Twitter-specific NLP pipeline (from tokenisation and POS tagging to chunking and NER). They adopted rich information generated in the pipeline, such as POS tags, chunking and predicted capitalisation information, as well as clustering of lexical variants (Brown et al., 1992) and gazetted features from Freebase.

3.2 Proposed Features

Based on the previous representative NER work, we considered the following features:

- `Word`. Lowercased word types are included as a default feature as suggested by existing systems. Previous and next two words are also included to capture local context information. Larger context window size is not considered as Twitter data is fairly terse and ungrammatical (Baldwin et al., 2013), so incorporating long distance context may bring little context information and introduce more noise.
- `POS`. Based on the fact that location named entities are primarily nouns. A reliable POS tagger generates valuable clues for locations. Instead of re-building a NLP pipeline, we adopt an off-the-shelf Twitter POS tagger CMU that generates coursed-grained POS tags with high accuracy ($\geq 90\%$) (Owoputi et al., 2013). Similar to `word`, the previous and next two `POS` tags are also included. We also consider `POS` bigrams.
- `Capitalisation`. Instead of predicting token case in Twitter (e.g., (Ritter et al., 2011)), four types of capitalisation information are retrieved based on the original surface form. Namely, they are all character uppercased (AU), all character lowercased (AL), first character uppercased and the rest are lowercased (UL) and mixed capitalisation (MC). We also consider `capitalisation` bigrams.
- `Domain`. Twitter specific entities such as user mentions, hashtags and URLs are considered as normal words. This is because many location mentions are embedded in these entities. For instance, *@Iran*, *#brisbane* and `http:www.abc.net.au/melbourne/`. Furthermore, we distinguish whether a word is in a stop word or not. Moreover, some location clues such as *street* are also categorised as task-specific features in this feature group.
- `Gazetteer`. Literature has shown that external gazetted resources are helpful in identifying named entities. Therefore, we incorporate features based on external place names, e.g., whether a current word is in a refined list of locations. Details are in Section 3.3.
- `Gazetteer Morphology`. As an extension of previous `gazetteer` features, we also observed that gazetted names may form part of a token and this is particularly common for Twitter specific entities, e.g., *#IranEQ* and *@zdnetaustralia*. As a result, we also perform partial string matching in Section 3.4.

Features	STANFORD	MSRA	WASHINGTON
Word	✓	✓	✓
Word Context	✓	✓	✓
Word Morphology	Character *n*-gram	Affix	Brown Cluster
POS	✓	✗	in-domain POS tagger
Chunking	✗	✗	in-domain chunker
Capitalisation	✗	✓	in-domain capitalisation restoration
Gazetteers	✗	Wikipedia	Freebase

Table 1: Features comparison of representative NER Systems

3.3 Gazetteers

We adopted GeoNames as our primary source of gazetted features. It is a geographical database with information about all countries with over eight million places, such as cities and points of interest.[2] However, as noted by Liu et al. (2011), some place names are also commonly used to denote something other than a location. Examples of these terms include people's names, natural disasters (e.g., *storm*), and names that usually do not denote a location (e.g., *Friday* or *Friend*). To alleviate the negative impact of these unreliable place names, we collected stopwords starting with a standard one and then added 5K most frequent English terms,[3] natural disaster names from Wikipedia and a list of popular personal names.[4]

After extracting and cleaning the terms from GeoNames, the list had over 9.8 million terms.[5] The dictionary was used to annotate the tweets using ConceptMapper (Tanenblatt et al., 2010) and the GeoNames annotation was used as a CRF feature.

On top of refined gazetteers, we also collected country names, state abbreviations, airport IATA codes and place abbreviations (e.g., *st* for street) in some English speaking countries from Wikipedia and Google.[6] The list is also filtered by stopword removal so that it represents a high quality place names and we can separately use them as gazetted features from GeoNames.

[2] http://www.geonames.org

[3] http://www.wordfrequency.info

[4] https://online.justice.vic.gov.au/bdm/popular-names

[5] Locations might have more than one name to include variants.

[6] The data is available at https://github.com/tq010or/alta-shared-task-2014-ibm

3.4 Gazetteer Morphology

The unique genre in Twitter generates many composite and non-standard location mentions. For instance, *chch* represents Christchurch in New Zealand in *Thoughts with everyone in chch #eqnz - not again!*; A standard place name may be concatenated with other tokens, e.g., *#pakistanflood*. A naive string match will miss these gazetted terms, therefore, we also match the prefix and suffix of all refined gazetted terms in Section 3.3 for each token in the tweet. The side effect of this approach is it also produces some false positives, e.g., *sa* (as South Australia or South Africa) also matches *samsung*. To avoid matching false positives, we further restrict the other part (e.g., *msung* in the *samsung* example) must be a valid word from a 81K English lexicon.[7].

Additionally, we also stripped spaces for higher order *n*-gram for this gazetteer morphology matching so that *newzealand* and *losangeles* would be recognised as well.

4 Experiments and Discussion

The training/dev/test data is offered by ALTA-shared task organisers. The collected and filtered tweets correspond to short time period when several disastrous events happened. 2K tweets are used to training and 1K tweets are randomly selected and equally split for dev and test purpose. The data set, however, is skewed to a number disastrous events such as New Zealand earthquake and Queensland floods.

The evaluation metric is mean-F score which averages the F1 number for each tweet. In addition to this official evaluation metric, we also present precision, recall and F1 numbers.

We adopted a state-of-the-art CRF implemen-

[7] 2of12inf.txt in http://goo.gl/4c49gv

tation named CRF-SUITE (Okazaki, 2007) with default parameters. We used built-in tokenisation and POS tags in CMU and all our string matching is in lowercase. Furthermore, the features are represented in BIO notation, in which B and I represent the beginning and continuity of a location entity, respectively. O denotes non-entity words. Using lowercased features and BIO (instead of BIOLU (Ratinov and Roth, 2009)) notations are to avoid potential data sparsity issues in generalisation, as we only have 2K training tweets and many labels such as *#eqnz* are repeated fairly frequently.

To correct CRF tagging errors and misses, we further imposed some post-processing rules. Words satisfying the following rules are also added as location mentions:

1. A word is in the refined `GeoNames` dictionary or a Twitter-specific entity is a gazetted term combined with an English word;
2. A word is in a closed set of direction names (e.g., *north*, *nth* or *north-east*) or location clues (e.g., *street*, *st*);
3. An URL contains an entry in the refined `GeoNames` dictionary;
4. If the tokens preceding and following *of* are labelled as locations, then the middle *of* is counted as part of a location mention, e.g., *north **of** Brunswick Heads*;
5. *CFA* and *CFS* with the following two words are labelled as location mentions, e.g., *CFA district 123*.

The evaluation numbers of our system for overall and feature ablations are presented in Table 2. Kaggle's site shows the results on the test set of our system compared to other systems.[8] Our system seems to perform below other participating systems, which is cannot be discussed since we are not aware of the implementation of the other systems. Overall, our best tagger achieved 0.758 and 0.726 mean-F1 on dev and test data, respectively. The noticeable disagreements between the results on the dev and test data indicates that there is a large difference between the two sets and that larger training sets are required to avoid overfitting or that additional sets of features might be considered.

[8]Challenge results on the dev set: `https://inclass.kaggle.com/c/alta-2014-challenge/leaderboard`, and public and private split of the test set: `https://inclass.kaggle.com/c/alta-2014-challenge/forums/t/10702/and-the-winner-is/57341#post57341`

Among all features, we saw that `word` and `post-processing` are the most important features to NER. By contrast, `domain`, `gazetteer` and `gazetteer morphology` contribute little to the overall performance. It makes sense that `word` are effective features, because many specific tokens (e.g., *eqnz*) are strong signals showing the token is a location mention. However, it is counter-intuitive that `gazetteer` and `gazetteer Morphology` failed to boost the performance (Ratinov and Roth, 2009; Liu et al., 2011). We hypothesise this may be because our CRF model down-weighted `gazetteer` features, when some location mentions (such as non-gazetted POIs) are not in the refined `GeoNames` and there might be common words that share the same surface forms with entries in `GeoNames`. Nonetheless, this doesn't indicate the gazetted data is not useful, but rather it should be integrated appropriately. Because when we added gazetted data in the `post-processing`, a considerable boost in performance is observed.

Notably, `capitalisation` and `POS` are useful in identifying location mentions. This suggests developing reliable capitalisation restoration and NLP pipeline will be beneficial to downstream NER.

5 Error Analysis

Our system incorrectly identified some tokens as locations. Most of the false positives were due to CRF mistakes. Examples of these mistakes are annotation of tokens like *bushfires*, Probably a larger data set would allow the CRF model to avoid these mistakes. On the other hand, many false positives produced by our system look as genuine locations. For instance, *bakery* was not annotated in *Chinatown bakery* but was annotated in *Manchester Wong Wong's bakery* a few tweets below. Some locations such as *Kumbarilla State Forest* seem to be false positives as well. Possibly the noise in the data set is also responsible for errors produced by our CRF tagger.

Even with our best efforts to remove location names that would not typically denote a location, there are some `GeoNames` locations in our dictionary that typically do not denote a location, e.g., *The End of the World*.

Our system missed some Twitter user names or hashtags with location information, e.g., *@FireRescueNSW*, *@abcsouthqld*. Although these lo-

Data	Dev				Test			
Evaluation metrics	mean-F1	F1	P	R	mean-F1	F1	P	R
Overall	0.758	0.774	0.784	0.764	0.726	0.756	0.770	0.742
-Word	0.716	0.738	0.717	0.760	0.683	0.715	0.702	0.729
-POS	0.744	0.767	0.780	0.755	0.713	0.742	0.772	0.715
-Capitalisation	0.748	0.761	0.769	0.752	0.723	0.753	0.769	0.737
-Domain	0.758	0.772	0.781	0.763	0.715	0.749	0.768	0.732
-Gazetteer	0.751	0.770	0.776	0.763	0.725	0.749	0.758	0.741
-Gazetteer Morph.	0.754	0.772	0.780	0.763	0.727	0.756	0.770	0.742
-Post-processing	0.714	0.743	0.814	0.684	0.700	0.736	0.814	0.672

Table 2: Overall experiment results and feature ablations

cations contain an acronym or abbreviation denoting a location as prefix or suffix, the rest part is not a valid single word in our English lexicon. For some locations, their variants were not in our location dictionary, e.g., *Melb* for *Melbourne*.

Some location names were not in our GeoNames dictionary and nor were identified by the CRF. Examples of these location names include *Coal Quay* or *Massabielle grotto*. Some two letter US state abbreviations were not recognised by our system, e.g., *OR* or *ID*; this could possibly be alleviated by less aggressively filtering the stopwords such as *OR* from the gazetteer but this would in many cases result in many false positives.

In a few cases, our system missed part of the location name when it was a generic location token attached to a specific named location. For instance, *markets* was not annotated in *Kelvin Grove markets* and *grounds* was not annotated in *UTS grounds*.

6 Discussion

In addition to standard CRF experiments and feature ablation analysis, we also tried to improve the accuracy through two extensions. First, we leveraged embedded topics to represent features, i.e., a feature is represented by the distribution of a limited number of related topics. The feature to topic distribution map is generated on a larger number of English tweets using WORD2VEC.[9] The results, compared with the CRF experiments, turn to be negative or minor positive in various settings. We infer this may be due to the sub-domain difference in the representation. We used general English tweets from Twitter Streaming API to obtain the embedded topics, which is different from disaster-related tweets with location mentions. Alternatively, this may be due to the high noise/signal ratio, i.e., expanding original feature to embedded topics brings more noise than the useful information.

Additionally, we also tried semi-supervised learning by first training a CRF model to annotate locations in a large amount of unseen new tweets, then feeding all locations and tweets into a CRF learner to train a new model for future tagging. This approach didn't show improvement either. We hypothesise that this is due to the data set being skewed towards disaster-related location mentions, adding more training data from general tweets does not improve the results.

7 Conclusion and Future Work

In this paper, we described our system in participating ALTA-shared task — identifying location mentions in Twitter. We formulated the problem as a location entity recognition task to scope our efforts in NER literature. Having examined and compared NER feature of existing systems, we proposed our own feature set with justifications. We further built a CRF-based location mention tagger and analysed the feature contributions. Overall, our tagger achieved 0.726 mean-F1 in the shared task. Although our extension experiments both show negative results, there is certainly room for further improvements. Our discussion and error analysis shed light on the future work in this research topic.

[9] https://code.google.com/p/word2vec/

References

Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources? In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 356–364, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479.

Jenny Rose Finkel, Trond Grenager, and Christopher 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 (ACL 2005)*, pages 363–370, Ann Arbor, USA.

Bo Han and Timothy 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 (ACL HLT 2011)*, pages 368–378, Portland, USA.

John Lingad, Sarvnaz Karimi, and Jie Yin. 2013. Location extraction from disaster-related microblogs. In *WWW 2013 Companion*, pages 1017–1020, Rio de Janeiro, Brazil.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 359–367, Portland, USA.

Naoaki Okazaki. 2007. CRFsuite: a fast implementation of conditional random fields (CRFs).

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 380–390, Atlanta, Georgia.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the 13th Conference on Natural Language Learning (CoNLL-2009)*, pages 147–155, Boulder, USA.

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

Michael A Tanenblatt, Anni Coden, and Igor L Sominsky. 2010. The conceptmapper approach to named entity recognition. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, Malta.