Named Entity Recognition from Tweets*

Ayan Bandyopadhyay[1], Dwaipayan Roy[1]
Mandar Mitra[1], and Sanjoy Kumar Saha[2]

[1] Indian Statistical Institute, India
{bandyopadhyay.ayan, dwaipayan.roy, mandar.mitra}@gmail.com,
[2] Jadavpur University, India
sks_ju@yahoo.co.in

Abstract. Entries in microblogging sites are very short. For example, a 'tweet' (a post or status update on the popular microblogging site Twitter) can contain at most 140 characters. To comply with this restriction, users frequently use abbreviations to express their thoughts, thus producing sentences that are often poorly structured or ungrammatical. As a result, it becomes a challenge to come up with methods for automatically identifying named entities (names of persons, organizations, locations etc.). In this study, we use a four-step approach to automatic named entity recognition from microposts. First, we do some preprocessing of the micropost (e.g. replace abbreviations with actual words). Then we use an off-the-shelf part-of-speech tagger to tag the nouns. Next, we use the Google Search API to retrieve sentences containing the tagged nouns. Finally, we run a standard Named Entity Recognizer (NER) on the retrieved sentences. The tagged nouns are returned along with the tags assigned by the NER. This simple approach, using readily available components, yields promising results on standard benchmark data.

1 Introduction

Microblogging emerged as a form of communication about ten years ago. Over the last decade, microblogging has evolved into an enormously popular platform for communicating via "microposts" (short text messages). According to a study, most tweets are either personal or conversational, but a large number do carry information in the form of Web links, music recommendations, and news [11]. In particular, microblogging has been demonstrated to be a particularly effective communication medium during disasters [10, 16]. Given the growing amount of information available through microblogging sites, techniques for efficiently and effectively processing this information are becoming increasingly important. One such information processing task that has attracted attention within the research community in recent times is Named Entity Recognition

 In: T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, Aachen, Germany, 8-10 September 2014, published at http://ceur-ws.org

(NER), the task of locating and classifying names in text [6]. NER from microblogs is challenging for the following reason. Entries in microblogging sites are required to be very short. For example, a 'tweet' (a post or status update on the popular microblogging site Twitter) can contain at most 140 characters. To comply with this restriction, users frequently use abbreviations to express their thoughts, thus producing text that is characterised by poor spelling, grammar or structure. Existing named entity recognition (NER) tools have generally been designed for (and tested on) full-text documents. It is quite likely that these tools will not perform well on microposts [14]. In this study, we try a simple approach to NER from microposts using existing, readily available Natural Language Processing (NLP) tools. In order to circumvent the problem mentioned above regarding the use of such tools, we first identify some candidate NEs, and then look for *full-text documents* containing these candidates. For this purpose, we use the Web as a source of pages that are likely to be full-text and properly structured in nature. These pages are expected to contain longer and more grammatical passages that provide better context for the standard NLP tools. We evaluated our method using benchmark data that was created as part of the MSM2013 Challenge (`http://oak.dcs.shef.ac.uk/msm2013/`). Our approach combines simplicity with effectiveness: it compares favourably with the methods that topped the MSM2013 Challenge Task.

2 Related Work

Named Entity Recognition is a well known problem in the field of NLP. Some named entity (NE) taggers like the Stanford Tagger [7] and the Illinois Named Entity Tagger [12] have been shown to work well for properly structured sentences. However, these NE taggers are unlikely to perform satisfactorily on the incomplete, fragmented and ungrammatical sentences typically found in microposts. As a result, NE tagging for microposts has emerged as a challenging research problem. Ritter et al. [14] were among the earliest to study NER from tweets. They show that "the performance of standard NLP tools is severely degraded on tweets." Their approach, based on Latent Dirichlet Allocation (LDA), utilises the Freebase dictionaries (`http://www.freebase.com`), and significantly outperforms the Stanford NER system. *Making Sense of Microposts* (#MSM) is a workshop series that started in 2011. It focuses on the problem of Information Extraction from microposts in general. A Concept Extraction Challenge (or contest) was organised as a part of #MSM2013. Contest participants were required to correctly identify entities belonging to one of four possible types: 'Person', 'Location' 'Organization' and 'Miscellaneous' (please see Section 3 for more details about these categories). The best challenge submission was by Habib et al. [9]. They used a hybrid approach that combines Conditional Random Fields (CRF) and Support Vector Machines (SVM) to tag named entities in microposts. The next best group [15] made use of the Wikipedia for the NER task. Dlugolinsky et al. [5] fused some well-known NER tools like GATE [4], Apache OpenNLP (`https://opennlp.apache.org/`), Illinois Named

Entity Tagger, Illinois Wikifier [13], LingPipe (http://alias-i.com/lingpipe) (with English News - MUC-6 model), OpenCalais (http://www.opencalais.com/about), Stanford Named Entity Recognizer (with 4 class caseless model), and WikiMiner (http://wikipedia-miner.cms.waikato.ac.nz) for named entity tagging in microposts.

3 Our Approach

As mentioned in the Introduction, our goal in this study is to recognise named entities (NEs) in microposts. Specifically, we try to identify and classify NEs belonging to the following four categories.

- **Person (PER)**: full or partial person names, e.g., Isaac Newton, Einstein.
- **Location (LOC)**: full or partial (geographical or physical) location names, including cities, provinces or states, countries, continents, e.g. Kolkata, Europe, Middle East.
- **Organization (ORG)**: full or partial organisation names, including academic, state, governmental, military and business or enterprise organizations, e.g., NASA, Reserve Bank of India.
- **Miscellaneous (MISC)**: any concept not covered by any of the categories above, but limited to one of the entity types: film/movie, entertainment award event, political event, programming language, sporting event and TV show, e.g. World Cup, Java.

Original string	**Replaced by**
AFAIK	as far as I know
B4	before
TTYL	talk to you later
!!!!!!	!
greeeeat	great

Table 1. Examples of changes made during preprocessing

1. **Preprocessing.** We replaced commonly used abbreviations with their expanded forms. For this step, we have used a simple lookup table consisting of 4704 commonly used abbreviations and their expansions. These were mostly collected from various Web sites (e.g., http://osakabentures.com/2011/06/twitter-acronyms-who-knows-them/). We also replaced strings of consecutive punctuation marks by a single punctuation mark. Finally, if a letter is repeated for emphasis, it is replaced by a single occurrence of that letter. This step is implemented via a simple lookup table of replacements. Table 1 gives some examples of changes made during preprocessing.

This preprocessing generally does not have a direct impact on NEs, but is likely to make the text more grammatical. The subsequent language processing tools that we apply (e.g., a Part of Speech tagger) are thus expected to give more accurate results. However, if a named entity coincidentally matches an abbreviation, it will also be replaced. For example, using Table 1, "B4" — the paper size — is replaced by "before", leading to a false negative.

2. **Part of speech tagging.** We use a readily available part-of-speech (POS) tagger for microposts [8] to tag each word in a micropost with its POS. Since named entities are proper nouns, we select only the proper nouns from the tagged tweet. Neighbouring proper nouns (words that are tagged as proper nouns and have only space(s) separating them) are taken together as a group. The list of nouns / noun-groups thus extracted constitute the list of candidate NEs.
3. **Google search.** Once the candidates have been identified above, we need to eliminate the candidates that are not actually NEs, and to classify the remainder into one of the four categories listed above. This step can be viewed as a five-class classification problem, with one of the classes being "**Not an NE**". If enough textual context were provided for each candidate, this classification task would be relatively easier. Unfortunately, because the tweets themselves are very short, they provide very little context.
   Since the Web can be regarded as a large natural language corpus, we turn to this obvious source in order to find longer texts containing a candidate NE. Each candidate NE is submitted as a query to the Google Search API (GSA) `http://code.google.com/apis/websearch/`. The webpages corresponding to the top 10 URLs (or fewer, if GSA returns fewer results) returned in the result list are fetched. If the original micropost is also returned among the top 10, it is neither counted nor fetched. Since Google may return slight variants of the submitted query term(s), we select only those pages that contain at least one exact match. In other words, if a page does not contain any exact match, it is discarded. If all pages are eliminated in the process, then we repeat the process once more with the next 10 results. The selected pages are likely to contain properly structured, grammatically correct sentences with the candidate NEs.
4. **NE tagging.** From the pages obtained in the above step, we extract sentences containing the candidate NEs and submit these to a standard NE tagger (the Stanford NE tagger [7]).

4 Evaluation

One standard measure used to evaluate (binary) classifiers is the F_β-score or F_β-measure. F_β is a weighted harmonic mean of the precision p and the recall r of the classifier. For the NER task, p and r are defined as follows.

Consider one of the four NE categories considered in the present study, say **PER**. Let N be the number of *actual* **PER**s present in the corpus; let n be the number of entities (words or phrases) that are tagged as **PER** by an NER

	PER	LOC	ORG	MISC	All
OurApproach	0.8402	0.3800	0.2836	0.0233	0.6359
StanfordNER	0.7932	0.3211	0.1395	0.0556	0.5112
openNLP	0.4968	0.2235	0.0483	0.0000	0.3889
LabelledLDA	0.7884	0.4227	0.4364	0.0954	0.5881
14 - 1	**0.9230**	0.6730	**0.8770**	0.6220	**0.7740**
21 - 3	0.8760	0.6030	0.8640	**0.7140**	0.7640
15 - 3	0.8790	**0.6860**	0.8440	0.5250	0.7340

Table 2. Overall and category-wise precision results

system; and let m be the number of actual **PER**s that are *correctly* identified by the NER system. Then p, r and F_β are given by:

$$p = \frac{m}{n} \qquad r = \frac{m}{N} \qquad F_\beta = \frac{(1+\beta^2) \ * \ p \ * \ r}{(\beta^2 * p) \ + \ r}$$

For this work, we adopt the common policy of setting β to 1 to allow precision and recall to be weighted equally. With $\beta = 1$, the F_β-measure reduces to the conventional harmonic mean of p and r, and is referred to as the F_1-measure. The F-measure is computed separately for each of the four NE categories mentioned in Section 3 and then averaged across the four categories to obtain a single overall measure of performance.

5 Results

For evaluation, we used the data set provided by "Making Sense of Microposts (#MSM2013)" [2]. The data consists of 1450 tweets contained in a single file, with one tweet per line. Each tweet has a unique tweet-id and the tweet text.

Tables 2–4 compare our approach with several readily available NER tools applied directly on the tweet text: openNLP tool, Stanford NER [7], and Labeled LDA method [14]. Since we used the MSM2013 data, we also compare our method with the three best submissions to the MSM2013 challenge (these are identified by their submission numbers in the tables). More details about the MSM2013 results can be found in the MSM2013 overview paper [3].

	PER	LOC	ORG	MISC	All
OurApproach	0.6922	0.5700	0.3305	0.0211	0.5884
StanfordNER	0.7269	0.6100	0.3263	0.0632	0.6180
openNLP	0.2794	0.1900	0.0424	0.0000	0.2206
LabelledLDA	0.7358	0.4100	0.1017	**0.3053**	0.5923
14 - 1	0.9080	0.6110	0.6200	0.2770	0.6040
21 - 3	0.9380	**0.6140**	0.6130	0.2870	**0.6130**
15 - 3	**0.9520**	0.4850	**0.7390**	0.2690	0.6110

Table 3. Overall and category-wise recall results

	PER	LOC	ORG	MISC	All
OurApproach	0.7583	0.4542	0.3041	0.0220	0.6123
StanfordNER	0.7586	0.4207	0.1954	0.0591	0.5474
openNLP	0.3576	0.2054	0.0451	0.0000	0.2815
LabelledLDA	0.7612	0.4162	0.1649	0.1454	0.5902
14 - 1	**0.9200**	**0.6400**	**0.7380**	0.3830	**0.6700**
21 - 3	0.9100	0.6090	0.7210	**0.4100**	0.6620
15 - 3	0.9180	0.5680	0.7900	0.3560	0.6580

Table 4. Overall and category-wise F_1 results

In general, we find that our method fails to identify NEs in the MISC category. Though the named entities are recognised, they are misclassified in most cases. One reason for misclassification is the occurrence of named entities like Annie Hall (tweet id 2904). Since this is the name of a fictional character, it is classified as PER. However, the tweet is about the movie by this name; thus, the entity actually belongs to the MISC category. This is one of the reasons affecting the precision of our method.

However, it is encouraging to note that the overall results obtained by our method are not statistically significantly different from the best results reported at MSM2013. We used the Welch Two Sample t-test [17] to determine the statistical significance of the differences between our approach and the top three submissions at MSM2013 (run IDs 14-1, 21-3 and 15-3). Table 5 shows the p-values for the three tests.

	14 - 1	21 - 3	15 - 3
Our Approach	0.1878	0.1916	0.2171

Table 5. p-values for Welch Two Sample t-test

Discussion Table 6 analyses the nature of false negatives for our method. Our method is based on the following assumption: while the text surrounding an NE may be of poor-quality, users are careful / accurate when mentioning names. This assumption turns out not be completely correct. For example, one tweet mentions 'britnay spers' (instead of 'Britney Spears'). Similarly, tweet ID 4261

Total # of NEs in dataset	1555
(Step 2) # of NEs not tagged as candidate by POS tagger	396
(Step 3) # of candidates for which no results found	5
(Step 4) # of candidates misclassified	239

Table 6. Analysis of false negatives

mentions 'Annie Lenox' (presumably the Scottish singer-songwriter) whose name is actually spelt 'Annie Len**n**ox'.

6 Conclusion

The key idea in our approach is to use the Web as a source of documents that are generally longer and better structured than tweets. This enables us to use standard NLP tools without having to redesign or retrain them. Since NER-tagged training data from the micropost domain is a scarce resource, this is an advantage. Significance tests show that our results are comparable to the state of the art.

As mentioned in the preceding section, however, our approach is based on the assumption that NEs in tweets are correctly written. Our immediate goal in future work would be to handle spelling errors / variations. One obvious way to do this would be to leverage the "Did you mean" feature provided by Google (note that this feature is *not* available via GSA). It may also be possible to handle spelling errors using a dictionary-based spelling correction algorithm that uses the Google n-gram dataset [1] as a lexicon. We would also like to explore the possibility of using our method to create labelled data that may in turn be used to train a more direct approach. This would eventually enable us to avoid the use of the GSA as a black box.

References

1. http://googleresearch.blogspot.in/2006/08/all-our-n-gram-are-belong-to-you.html
2. http://oak.dcs.shef.ac.uk/msm2013/
3. Basave, A.E.C., Rowe, M., Stankovic, M., Dadzie, A.S. (eds.): Proc. Concept Extraction Challenge at the 3rd Workshop on Making Sense of Microposts (#MSM2013): Big things come in small packages. CEUR Workshop Proceedings (May 2013), http://ceur-ws.org/Vol-1019
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A framework and graphical development environment for robust NLP tools and applications. In: Proc. 40th Ann. Meeting of the ACL (July 2002), http://gate.ac.uk/sale/acl02/acl-main.pdf
5. Dlugolinsky, S., Krammer, P., Ciglan, M., Laclavik, M.: MSM2013 IE Challenge: Annotowatch. vol. 1019, pp. 21–26 (2013), In [3].
6. Downey, D., Broadhead, M., Etzioni, O.: Locating complex named entities in web text. In: Proc. 20th IJCAI. pp. 2733–2739. IJCAI'07 (2007), http://dl.acm.org/citation.cfm?id=1625275.1625715
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proc. 43rd Ann. Meeting of the ACL. pp. 363–370. ACL (2005), http://dx.doi.org/10.3115/1219840.1219885
8. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: 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: short papers - Volume 2. pp. 42–47. HLT '11, Association for Computational Linguistics (2011), http://dl.acm.org/citation.cfm?id=2002736.2002747
9. Habib, M.B., van Keulen, M., Zhu, Z.: Concept extraction challenge: University of twente at #msm2013. vol. 1019, pp. 17–20 (2013), In [3].
10. Jennex, M.E., de Walle, B.V. (eds.): International Journal of Information Systems for Crisis Response and Management (IJISCRAM). IGI Global (Est 2009)
11. Kelly, R.: Twitter study – August 2009 (2009), http://www.pearanalytics.com/blog/wp-content/uploads/2010/05/Twitter-Study-August-2009.pdf
12. Ratinov, L., Roth, D.: Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning. pp. 147–155. CoNLL '09, Association for Computational Linguistics (2009), http://dl.acm.org/citation.cfm?id=1596374.1596399
13. Ratinov, L., Roth, D., Downey, D., Anderson, M.: Local and global algorithms for disambiguation to wikipedia. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1375–1384. HLT '11, Association for Computational Linguistics (2011), http://dl.acm.org/citation.cfm?id=2002472.2002642
14. Ritter, A., Clark, S., Mausam, Etzioni, O.: Named entity recognition in tweets: An experimental study. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1524–1534. EMNLP '11, Association for Computational Linguistics (2011), http://dl.acm.org/citation.cfm?id=2145432.2145595
15. Sachidanandan, S., Sambaturu, P., Karlapalem, K.: NERTUW: Named entity recognition on tweets using Wikipedia. vol. 1019, pp. 67–70 (2013), In [3].
16. Vieweg, S., Hughes, A.L., Starbird, K., Palen, L.: Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1079–1088. CHI '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1753326.1753486
17. Welch, B.L.: The generalization of student's problem when several different population varlances are involved. Biometrika 34(1-2), 28–35 (1947), http://biomet.oxfordjournals.org/content/34/1-2/28.short