A Reverse Approach to Named Entity Extraction and Linking in Microposts*

Kara Greenfield, Rajmonda Caceres, Michael Coury, Kelly Geyer, Youngjune Gwon, Jason Matterer, Alyssa Mensch, Cem Sahin, Olga Simek

MIT Lincoln Laboratory, 244 Wood St, Lexington MA, United States

{kara.greenfield, rajmonda.caceres, michael.coury, kelly.geyer, gyj, jason.matterer, alyssa.mensch, cem.sahin, osimek}@ll.mit.edu

ABSTRACT

In this paper, we present a pipeline for named entity extraction and linking that is designed specifically for noisy, grammatically inconsistent domains where traditional named entity techniques perform poorly. Our approach leverages a large knowledge base to improve entity recognition, while maintaining the use of traditional NER to identify mentions that are not co-referent with any entities in the knowledge base.

Keywords

Named entity recognition; entity linking; twitter; DBpedia, social media

1. INTRODUCTION

This paper describes the MIT Lincoln Laboratory submission to the Named Entity Extraction and Linking (NEEL) challenge at #Microposts2016 [1]. While named entity recognition is a well-studied problem in traditional natural language processing domains such as newswire, maintaining high precision and recall when adapting it to micropost genres continues to prove difficult [2]. In traditional named entity extraction and linking systems, named entity recognition is done before entity linking and clustering. Any misses in the named entity recognition aren't recoverable by later steps in the pipeline.

In this system, we build upon the work developed in [3], leveraging the existence of a knowledge base which contains entities corresponding to many of the named mentions we wish to extract thus allowing us to reduce our reliance on named entity recognition. Our end-to-end system has parallel pipelines for those entity mentions that are linkable to the database and those which are not linkable.

2. SYSTEM ARCHITECTURE

Our overall system architecture is shown in Figure 1. For entities which are in the knowledge base (DBpedia), we began by hand-curating an ontology mapping from the DBpedia class ontology to the named entity ontology that is being used in the NEEL evaluation (Person, Organization, Location, Fictional Character, Thing, Product, Event).

For each DBpedia entry that mapped to one of the named entity classes of interest, we generated a set of candidate names for that entity which correspond to ways in which an author might reference that entity when writing a micropost. We then searched the tweets for those candidate names. Finally, we down-selected from the found instances of candidate names, resolving overlaps and false alarms in the candidate name generation.

We fused several named entity recognition systems in order to extract named entity mentions that do not have corresponding entities in DBpedia. We filtered out any named mentions that were previously identified as linked named entity mentions, leaving a set of typed NIL named entity mentions. We then applied clustering to the NIL mentions.

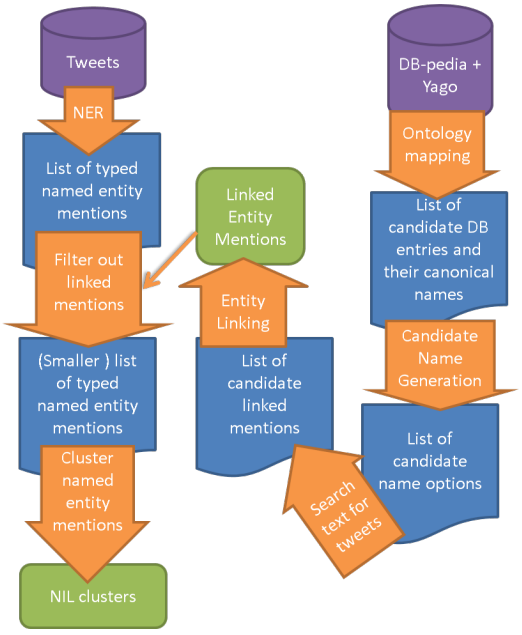

Figure 1 System Architecture

3. SYSTEM COMPONENTS

3.1 Ontology Mapping

Our goal for the ontology mapping was to have as high of a recall for each of the entity types as possible, simultaneously optimizing

*This work was sponsored by the Defense Advanced Research Projects Agency under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Published as part of the #Microposts2016 Workshop proceedings, available online as CEUR Vol-1691 (http://ceur-ws.org/Vol-1691)

#Microposts2016, Apr 11th, 2016, Montréal, Canada.

for precision only so much as to avoid computational bottlenecks in later steps in the pipeline. We experienced high variance between entity types in the degree of difficulty of manually creating the ontology mapping. As seen in Table 1, this resulted in vastly different levels of recall for the different entity types. Our mapping contained 100% of the linked Person entities in the dev set, but only 11% of the Fictional Character entities. In future work, we would like to explore either automating or crowdsourcing a more comprehensive ontology mapping.

Table 1 Ontology Mapping Recall

Entity Type	Recall
Person	1
Organization	.6364
Location	.8667
Product	.8889
Thing	.5
Fictional Character	.1111
Event	.5

3.2 Candidate Name Generation

In writing microposts, authors are constrained in the number of characters that they can write. This has led to the development of authors shortening their words (often as much as possible) while maintaining understandability by a human reader. Spelling mistakes and the existence of multiple standard spellings of named entities are two means by which variation in mention spelling can occur, but in the micropost genre, deliberate shortened alternate spellings are a much more common form of spelling variation. In order to address this, we examined the mentions in all of the named entity classes of interest and attempted to identify rules by which authors shorten entity names. We then applied these rules to all of the entities in our mapped ontology in order to generate candidate name spellings.

Authors use different rules when shortening a name depending on the context: using the name as part of plain text versus using the name as part of a hash-tag or at-mention. The main difference is that entity mentions which are hash-tags or at-mentions often contain the characters from descriptive words in addition to characters from the canonical form of the entity name as the text of the at-mention or hash-tag. We found that authors follow different rules depending on what type of entity the mention is. For example, abbreviating the canonical form of a Person entity is very common, but abbreviating a Thing entity is very rare. On the other hand, the canonical forms of Location entities are often partially abbreviated (i.e. abbreviating only the words which occur after a comma in the canonical spelling). Our candidate name generation computes various abbreviations and shortenings of the canonical name.

Table 2 Candidate Name Generation Recall

Entity Type	Recall
Person	.8961
Organization	.32
Location	.5625
Product	.4273
Thing	0
Fictional Character	.1538
Event	0

Finally, events are often written very differently from their canonical spellings, rendering candidate name generation a poor choice for this entity type. In future work, we would like to train an event nugget detector on the micropost genre in order to extract the Event entities. Our system was unable to correctly generate candidate names for any of the Thing mentions that were included in our ontology mapping, although the candidate generation did work for many of the Thing mentions that were not included in the ontology.

3.3 Linkable Mention Detection

We searched all of the tweets for all of our generated candidate mentions. Search results were limited to mentions which were either bound on both ends by white space, punctuation, or the beginning / end of the tweet or which were part of an at-mention or hash-tag. For results that were part of an at-mention or hash-tag, we expanded the returned result to encompass the entire at-mention or hash-tag.

3.4 Entity Linking

We experimented with two methods of entity linking. The first method was a random forest trained on several features of each (mention, entity) pair. The features used were: COMMONNESS, IDF_{anchor}, TEN, TCN, $TF_{sentence}$, $TF_{paragraph}$, and REDIRECT [4]. The random forest classifier attempts to detect whether or not a given mention corresponds to a given entity. We then perform consistency resolution in order to assure that each mentions resolves to at most a single entity. Results can be seen in Table 5.

We also experimented with leveraging AIDA [5] for entity linking. This method was able to correctly recall 25% of the Location mentions and 26% of the Person mentions, but did not perform well on the other entity types. We hypothesize that this is due to a combination of cascaded performance degradation from earlier steps in the pipeline and the fact that the current version of AIDA is based off of an older version of DBpedia, which doesn't contain more recent entities.

3.5 Named Entity Recognition

We experimented with several different named entity recognition systems: Stanford NER [6], MITIE [7], twitter_nlp [8], and TwitIE [9]. For MITIE, we used both the off-the-shelf model and a model that was custom trained on the NEEL training data (for all of the NEEL entity types); the custom training improved F1 scores on all entity types. Ultimately we fused the results from all of the systems by applying a majority vote. The results presented in Table 3 are in the format: precision; recall; F1.

Table 3 Named Entity Recognition Precision, Recall, and F1

NER System	Person	Location	Organization
Stanford	.84; .27; .41	.81; .76; .78	.57; .12; .2
MITIE	.48; .1; .17	.33; .18; .24	.1; .06; .06
MITIE (trained on NEEL data)	.78; .5; .61	.29; .24; .26	.33; .15; .21
Twitter_NLP	.56; .08; .14	.5; .18; .26	.5; .06; .11
TwitIE	.41; .06; .11	.5; .29; .37	.62; .15; .25
Fused System	.72; .67; .69	.44; .65; .52	.19; .18; .19

Even with considering multiple state of the art named entity recognition systems and in-domain training, performance on the

micropost genre is low. In future work, we would like to experiment with more advanced methods of system fusion and bootstrapping in order to gain a much larger in-domain training corpus.

3.6 Entity Clustering

We use the normalized Damerau–Levenshtein (DL) distance metric [10] to find the similarity between two unlinked entities. This metric helps us create clusters that are spelling-error tolerant, while at the same time capturing slight local words variations often observed in microposts.

As an alternative method, we used the Brown clusters produced by Percy Liang's implementation [11] of the Brown clustering algorithm [12] on 56,345,753 English tweets, as described in [13]. Mentions that belonged to the same Brown cluster were clustered together.

Table 4 gives the results on our NIL entity clustering task. We report performance scores with gold standard named entity mentions. Since the NIL entity clustering step is the last step in our system, we expect propagated errors from the other tasks to have the biggest impact here. Of note is that the small number of mentions in the evaluation dev set means that these numbers may not be representative of algorithm performance on a larger corpus.

In future work, we would like to experiment with word embedding based methods for clustering. We performed some early exploration into this line of research, but more work is needed into how to map between different word embeddings.

Table 4 Mention_CEAF of Clustering Algorithms

Clustering Method	Gold Standard NER mentions (NIL and non-NIL)
Damerau-Levenshtein	.587
Brown	.531

4. Experimental Results

Our top performing systems on the dev data used a random forest for entity linking and either Brown clustering or Damerau-Levenshtein clustering for clustering the NIL mentions. While Brown Clustering and Damerau-Levenshtein clustering returned slightly different clusters when run on the dev set, the mention_ceaf was the same for both methods. Results are shown below.

Table 5 Overall System Results

Metric	Precision	Recall	F1
strong typed mention match	.587	.287	.386
strong link match	.799	.418	.549
mention ceaf	.375	.766	.504

5. CONCLUSIONS

In this paper, we described the MIT Lincoln Laboratory submission to the NEEL 2016 challenge. In this work, we have expanded upon the linking first approach to named entity extraction and linking first developed in [3]. We introduced methods of candidate name generation which are specifically tailored to microposts. We also experimented with multiple approaches to named entity recognition, entity linking, and entity clustering and presented comparisons of the performance of the different methods.

6. ACKKNOWLEDGEMENTS

We would like to thank Bernadette Johnson and Joseph Campbell for their ongoing support and guidance. We would also like to thank Michael Yee and Arjun Majumdar for their support with MITIE.

7. REFERENCES

[1] G. Rizzo, M. van Erp, J. Plu, and R. Troncy. Making Sense of Microposts (#Microposts2016) Named Entity rEcognition and Linking (NEEL) Challenge. in *#Microposts2016*, pp. 50–59, 2016.

[2] A. Ritter, S. Clark and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," in *EMNLP '11*, 2011.

[3] I. Yamada, H. Takeda and Y. Takefuji, "An End-to-End Entity Linking Approach for Tweets," in *#Microposts2015*, 2015.

[4] E. Meij, W. Weerkamp and M. de Rijke, "Adding semantics to microblog posts," in *Proceedings of the fifth ACM international conference on Web search and data mining*, 2012.

[5] J. Hoffart, M. A. Yosef, I. Bordino, H. Furstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater and G. Weikum, "Robust Disambiguation of Named Entities in Text," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[6] J. R. Finkel, T. Grenager and C. Manning, "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling," in *Proceedings of the 43rd Annual Meeting of the aAssociation for Computational Linguistics (ACL 2005)*, 2005.

[7] D. King, "MITLL/MITIE," [Online]. Available: https://github.com/mit-nlp/MITIE.

[8] A. Ritter, S. Clark, Mausam and O. Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," in *EMNLP*, 2011.

[9] K. Bontcheva, L. Derczynski, A. Funk, M. A. Greenwood, D. Maynard and N. Aswani, "TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text," in *Proceedings of the International Conference on Recent Advasnces in Natural Language Processing, ACL*, 2013.

[10] G. V. Bard, Spelling-error Tolerant, Order-independent Pass-phrases via the Damerau-Levenshtein String-edit Distance Metric, vol. 68, Ballarat: Proceedings of the 5th Australian Symposiu on ACSW Frontiers, 2007, pp. 117--124.

[11] P. Liang, "Semi-supervised Learning for Natural Language," Massachusetts Institute of Technology, 2005.

[12] P. F. Brown, "Class-based n-gram models of natural language," *Computational linguistics,* vol. 18, no. 4, pp. 467-479, 1992.

[13] O. Owoputi, B. O'Connor, C. Dyer, K. Gimpel and N. Schnelder, "Part-of-speech tagging for Twitter: Word clusters and other advances," in *School of Computer Science*, 2012.