Results of the WNUT16 Named Entity Recognition Shared Task

Benjamin Strauss, Bethany E. Toma, Alan Ritter, Marie-Catherine de Marneffe, Wei Xu
The Ohio State University OH, USA
{strauss.105,toma.18,ritter.1492,demarneffe.1,xu.1265}@osu.edu

Abstract

This paper presents the results of the Twitter Named Entity Recognition shared task associated with W-NUT 2016: a named entity tagging task with 10 teams participating. We outline the shared task, annotation process and dataset statistics, and provide a high-level overview of the participating systems for each shared task.

1 Introduction

The increasing flood of user-generated text on social media has created an enormous opportunity for new data analysis techniques to extract and aggregate information about breaking news (Ritter et al., 2012), disease outbreaks (Paul and Dredze, 2011), natural disasters (Neubig et al., 2011), cyber-attacks (Ritter et al., 2015) and more. Named entity recognition is an important first step in most information extraction pipelines. However, performance of state-of-the-art NER on social media still lags behind well edited text genres. This motivates the need for continued research, in addition to new datasets and tools adapted to this noisy text genre.

In this paper, we present the development and evaluation of a shared task on named entity recognition in Twitter, which was held at the 2nd Workshop on Noisy User-generated Text (W-NUT 2016) and attracted 10 participating teams, 7 of which described their approach in peer-reviewed papers. This is a re-run of the NER task from W-NUT 2015, with an updated test set. The new test set consists of tweets annotated with named entities from a later time period than the data from 2015. The new test set developed for the 2016 iteration of the task consists of 3,856 tweets; this roughly doubles the amount of annotated data available for Twitter named entity recognition. We hope this new data will help to advance research on named entity recognition in noisy text.

A major development as compared to 2015 is the increased use of neural network methods by participants. Several teams, including the winning team, CambridgeLTL, used bidirectional LSTMs. Other teams achieved competitive performance by integrating a broad range of linguistic and knowledge-based features using conditional random fields (e.g., NTNU) or learning to search methods (Talos).

Another new development for 2016 was the inclusion of small amounts of domain-specific data into the test set. The motivation was to test whether Twitter named entity taggers targeting general-domain suffer a drop in performance when applied to tweets on specific types of events. For this purpose we annotated 350 tweets related to cyber-attacks and 500 related to mass shooting events. Note that no data from these domains was specifically included in the training or development data.

In the following sections, we describe details of the task including training and development datasets in addition to the newly annotated test data for 2016. We briefly summarize the systems developed by selected teams, and conclude with results.

2 Named Entity Recognition over Twitter

Named entity recognition is a crucial component in many information extraction pipelines. However, the majority of available NER tools were developed for newswire text and these tools perform poorly on informal text genres such as Twitter. While performance on named entity recognition in newswire is

Proceedings of the 2nd Workshop on Noisy User-generated Text,
pages 138–144, Osaka, Japan, December 11 2016.

quite high[1] (Tjong Kim Sang and De Meulder, 2003), state-of-the-art performance on Twitter data lags far behind.

The diverse and noisy style of user-generated content presents serious challenges. For instance tweets, unlike edited newswire text, contain numerous nonstandard spellings, abbreviations, unreliable capitalization, etc. Because of these issues, off-the-shelf named entity recognition tools tuned for newswire suffer a severe performance degradation when applied to noisy Twitter data. But tweets often contain more up-to-date information than news, in addition the increased volume of text offers opportunities to exploit redundancy of information which is very beneficial for information extraction (Downey et al., 2005). To exploit the opportunities for information extraction on top of social media, there is a crucial need for in-domain annotated data to train and evaluate named entity recognition systems on this noisy style of text.

Twitter processing has the additional challenge that the language people use on Twitter changes over time (Dredze et al., 2010; Fromreide et al., 2014). The previous edition of this task (Baldwin et al., 2015) addressed this issue by evaluating on a test set collected from a later time period then the training and development data. This year we take a similar approach, providing a new test dataset of tweets gathered from 2016. In addition to enabling research on adapting named entity recognition to new language over time, we hope this new dataset will be useful for adapting future Twitter named entity recognition systems, improving their performance on up-to-date data.

Additionally, this year we address the issue of topic distribution by including evaluation data from two specific domains (cybersecurity events and mass shootings) along with general domain data. Both the time period and topic selection of the evaluation data were not announced to participants until the (unannotated) test data was released at the beginning of the evaluation period. Teams had 7 days to submit their results on the test data, which were subsequently scored and gold annotations were released to participants. Evaluating the NER systems on these domains specific Twitter data provides information about possible system weakness.

2.1 Training and Development Data

The training and development data for our task was taken from prior work on Twitter NER (Ritter et al., 2011; Baldwin et al., 2015), which distinguishes 10 different named entity types (see Table 1 for the set of types). The training data was created from the union of the training and development data from the 2015 task (Baldwin et al., 2015).

The data was split into 2,394 annotated tweets for training and 1,000 as a development set. We also provided an additional 425 annotated tweets from the 2015 development data set (Baldwin et al., 2015).

2.2 Test Data Annotation

The data set we created for testing is new for this shared task. We collected general Twitter data and domain specific Twitter data. In real world situations people want to run taggers on a specific subset of the twitter stream. To simulate these situations we collected and tested on two example domains. The domain specific data sets are mass shooting events and cybersecurity events. General domain test data was randomly sampled from December 2014 through February 2015. Shooting domain data was collected by searching for tweets that are referring to a shooting event. To collect tweets from the mass shooting domain used www.gunviolencearchive.org's mass shooting event database for information about shooting events; including date, location, victim, and perpetrators of shooting events. Using this information, we searched for tweets that occur on the same day as a mass shooting and include the key word "shooting" and the location of the shooting event. The shooting domain contains 8,963 tokens with 751 phrases. Computer hacking events were found by searching for tweets including the keyword "breach". The breach domain contains 5,537 tokens with 603 phrases.

The additional data annotated this year was completed by a single annotator instructed to follow the annotation guidelines of the prior annotations. The annotator was presented with a set of simple guidelines[2] that cover common ambiguous cases and was also instructed to refer to the September 2010 data

[1] For example, the Stanford named entity tagger (Finkel et al., 2005) achieves an F1 score of 0.86 on the CoNLL data set.

[2] `http://bit.ly/1FSP6i2`

for reference (Ritter et al., 2011). The BRAT tool[3] was used for annotation. Figure 1 is a screenshot of the interface presented to the annotators. To ensure that the new annotations were consistent with the earlier annotations, 100 tweets were annotated in both tasks to calculate agreement. The new annotator proved a high agreement with the old data set with a F1 score of 67.67.

Table 1 presents the count of each of the 10 named entity types labeled by the annotators in the training, development and test sets created for this shared task.

	Train	Dev	Test
company	171	39	621
facility	104	38	253
geo-loc	276	116	882
movie	34	15	34
musicartist	55	41	191
other	225	132	584
person	449	171	482
product	97	37	246
sportsteam	51	70	147
tvshow	34	2	33
Total	1496	661	3473

Table 1: Named entity type counts in the train, development and test sets.

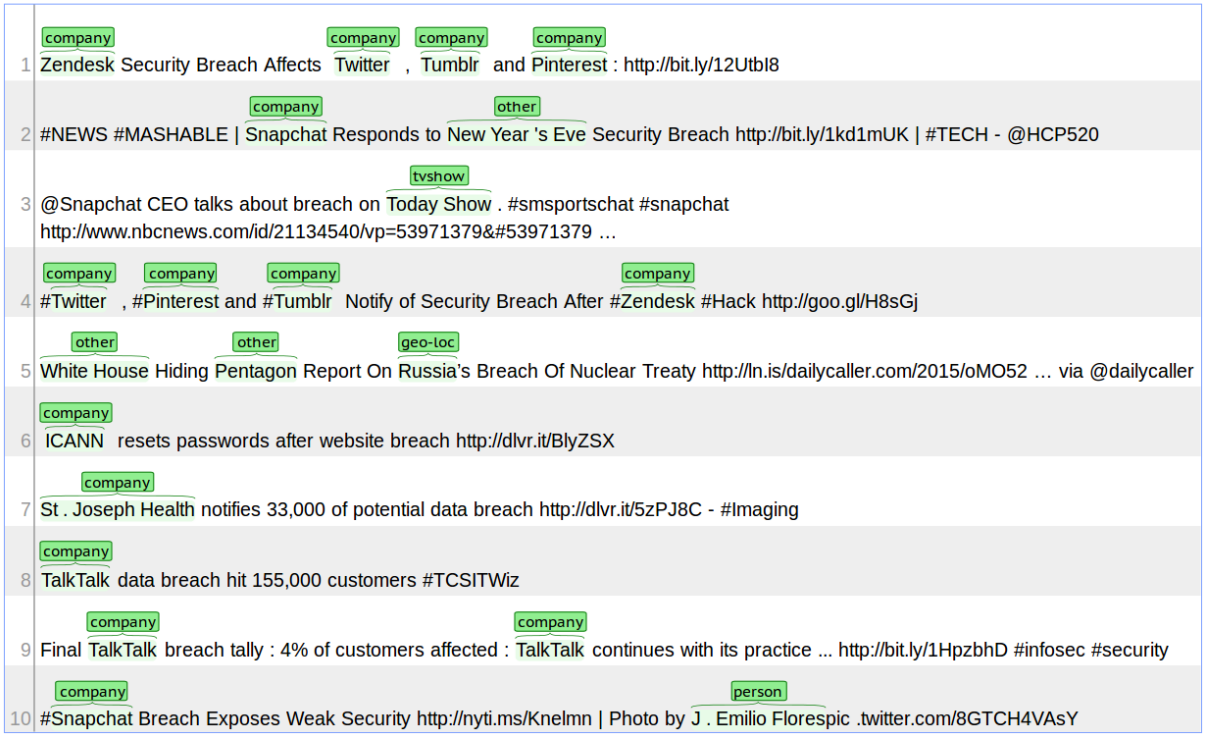

Figure 1: Annotation interface.

[3] http://brat.nlplab.org/

Team ID	Affiliation
CambridgeLTL	University of Cambridge
Talos	Viseo R&D
akora	University of Manchester
NTNU	Indian Institute of Technology Patna
ASU	Ain Shams University, Cairo, Egypt
DeepNNNER	Honda Research Institute Japan
DeepER	University of Illinois at Urbana-Champaign
hjpwhu	Wuhan University
UQAM-NTL	Université du Québec à Montréal
LIOX	The Hong Kong Polytechnic University

Table 2: Team ID and affiliation of the named entity recognition shared task participants.

	POS	Orthographic	Gazetteers	Brown clustering	Word embedding	ML
BASELINE	–	✓	✓	–	–	CRFsuite
CambridgeLTL	–	–	–	–	–	LSTM
akora	–	–	–	–	–	LSTM
NTNU	✓	✓	✓	–	–	CRF
Talos	✓	✓	✓	✓	GloVe	L2S
DeepNNNER	–	–	–	–	Multiple	LSTM-CNN
ASU	–	–	✓	✓	–	LSTM
UQAM-NTL	✓	✓	✓	–	–	CRF

Table 3: Features and machine learning approach taken by each team.

2.3 System Descriptions

This section briefly describes the approach taken by each team. Overall we noticed different trends between the types of systems submitted this year and last year. The most notable change is the use of LSTM-based systems. Four of the seven submissions were LSTM-based as opposed to zero submissions last year. The previous year Conditional Random Fields was the most popular ML technique for extracting named entities.

CambridgeLTL (Limsopatham and Collier, 2016) The system uses bidirectional LSTM to automatically induce and leverage orthographic features for performing Named Entity Recognition in Twitter messages.

akora (Kurt Junshean Espinosa and Ananiadou, 2016) This system uses bidirectional LSTM networks and exploits weakly annotated data to bootstrap sparse entity types.

NTNU (Sikdar and Gambäck, 2016) This system is based on classification using Conditional Random Fields, a supervised machine learning approach. The system utilizes a large feature set developed specifically for the task, with eight types of features based on actual characters and token internal data, five types of features built through context and chunk information, and five types of features based on lexicon-type information such as stop word matching, word frequencies, and entries in the shared task lexicon and Babelfy (Moro et al., 2014).

Talos (Ioannis Partalas and Kalitvianski, 2016) The system uses three types of features: lexical and morpho-syntactic features, contextual enrichment features using Linked Open Data, and features based on distributed representation of words. The system also exploits words clustering to enhance performance. The learning algorithm was solved by using Learning to search (L2S) that resembles a reinforcement learning algorithm.

DeepNNNER (Dugas and Nichols, 2016) The system uses a bidirectional LSTM-CNN model with word embedding trained on a large scale Web corpus. Additionally, the system uses automatically constructed lexicons with a partial matching algorithm and text normalization to handle the large vocabulary problem in Web texts.

	Precision	Recall	F1
CambridgeLTL	60.77	46.07	52.41
Talos	58.51	38.12	46.16
akora	51.70	39.48	44.77
NTNU	53.19	32.13	40.06
ASU	40.58	37.58	39.02
DeepNNNER	54.97	28.16	37.24
DeepER	45.40	31.15	36.95
hjpwhu	48.90	28.76	36.22
UQAM-NTL	40.73	23.52	29.82
LIOX	40.15	12.69	19.26

Table 4: Results segmenting and categorizing entities into 10 types.

	Precision	Recall	F1
CambridgeLTL	73.49	59.72	65.89
NTNU	64.18	62.28	63.22
Talos	70.53	52.58	60.24
akora	64.75	54.28	59.05
ASU	57.55	52.98	55.17
DeepER	63.17	43.31	51.38
DeepNNNER	70.66	36.14	47.82
hjpwhu	63.00	37.06	46.66
UQAM-NTL	53.21	37.95	44.30
LIOX	58.18	31.33	40.73

Table 5: Results on segmentation only (no types)

	Acc	P	R	F1
CambridgeLTL	90.57	69.75	51.24	59.08
Talos	89.36	60.49	41.13	48.96
akora	88.42	54.21	36.32	43.50
hjpwhu	88.21	59.79	28.86	38.93
ASU	87.76	42.22	32.84	36.94
NTNU	87.72	51.89	27.36	35.83
DeepNNNER	87.66	62.88	23.88	34.62
DeepER	84.32	40.23	22.89	29.18
UQAM-NTL	85.64	37.97	16.75	23.25
LIOX	84.41	30.08	6.63	10.87

Table 6: Results for the Cyber domain data on segmenting and categorizing entities into 10 types.

	Acc	P	R	F1
CambridgeLTL	93.00	66.25	56.72	61.12
Talos	92.03	68.53	49.00	57.14
DeepER	91.96	64.01	51.40	57.02
akora	91.54	58.89	49.40	53.73
NTNU	91.14	61.36	42.08	49.92
DeepNNNER	91.22	59.88	41.15	48.78
hjpwhu	90.83	53.71	41.41	46.77
ASU	90.74	45.40	47.94	46.63
UQAM-NTL	89.38	45.80	33.42	38.65
LIOX	88.35	55.77	23.17	32.74

Table 7: Results for the Shooting domain on segmenting and categorizing entities into 10 types.

ASU (Michel Naim Gerguis and Gerguis, 2016) The system shows an experimental study on using word embeddings, Brown clusters, part-of-speech tags, shape features, gazetteers, and local context to create a feature representation along with a set of experiments for the network design. A Wikipedia-based classifier framework was adopted to extract lists of fine-grained entities out of few input examples to be used as gazetteers. The model uses the LSTM algorithm to learn a NE classifier from the feature representation.

UQAM-NTL (Ngoc Tan LE and Sadat, 2016) The system is based on supervised machine learning and trained with a sequential labeling algorithm, using Conditional Random Fields to learn a classifier for Twitter NE extraction. The model uses 6 different categories of features including (1) orthographic, (2) lexical and (3) syntactic features as well as (4) part-of-speech tags, (5) polysemy count and (6) longest n-gram length in order to create a feature representation.

3 Summary

In this paper, we presented a shared task for Named Entity Recognition in Twitter data. We detailed the task setup and datasets used in the respective shared tasks, and also outlined the approach taken by the participating systems. The shared task included larger data sets than prior shared task (Baldwin et al., 2015). The evaluation data included new tweets collected from 2016. First, we are able to draw stronger conclusions about the true potential of different approaches in the latest Twitter data. Second, through analyzing the results of the participating systems, we are able to suggest potential research directions for both future shared tasks and noisy text processing in general.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. IIS-1464128. Alan Ritter and Benjamin Strauss are supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C-0114. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, or the U.S. Government.

References

Timothy Baldwin, Young-Bum Kim, Marie Catherine de Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP*, 126:2015.

Doug Downey, Oren Etzioni, and Stephen Soderland. 2005. A probabilistic model of redundancy in information extraction. In *Proceedings of the 19th international joint conference on Artificial intelligence*.

Mark Dredze, Tim Oates, and Christine Piatko. 2010. We're not in kansas anymore: detecting domain changes in streams. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 585–595. Association for Computational Linguistics.

Fabrice Dugas and Eric Nichols. 2016. Deepnnner: Applying blstmcnns and extended lexicons to named entity recognition in tweets. In *proceedings of W-NUT 2016*, Osaka, Japan.

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 on Association for Computational Linguistics*, ACL '05, pages 363–370, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hege Fromreide, Dirk Hovy, and Anders Søgaard. 2014. Crowdsourcing and annotating ner for twitter# drift. *European language resources distribution agency*.

Nadia Derbas Ioannis Partalas, Cédric Lopez and Ruslan Kalitvianski. 2016. Learning to search for recognizing named entities in twitter. In *proceedings of W-NUT 2016*, Osaka, Japan.

Riza Theresa BatistaNavarro Kurt Junshean Espinosa and Sophia Ananiadou. 2016. Learning to recognise named entities in tweets by exploiting weakly labelled data. In *proceedings of W-NUT 2016*, Osaka, Japan.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional lstm for named entity recognition in twitter messages. In *proceedings of W-NUT 2016*, Osaka, Japan.

M. Watheq El-Kharashi Michel Naim Gerguis, Cherif Salama and Michel Naim Gerguis. 2016. Asu: An experimental study on applying deep learning in twitter named entity recognition. In *proceedings of W-NUT 2016*, Osaka, Japan.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Graham Neubig, Yuichiroh Matsubayashi, Masato Hagiwara, and Koji Murakami. 2011. Safety information mining-what can nlp do in a disaster-. In *IJCNLP*.

Fatma Mallek Ngoc Tan LE and Fatiha Sadat. 2016. Uqam-ntl: Named entity recognition in twitter messages. In *proceedings of W-NUT 2016*, Osaka, Japan.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *ICWSM*.

Alan Ritter, Sam Clark, Oren Etzioni, et al. 2011. Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Mausam Ritter, Alan, Oren Etzioni, and Sam Clark. 2012. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*.

Alan Ritter, Evan Wright, William Casey, and Tom Mitchell. 2015. Weakly supervised extraction of computer security events from twitter. In *Proceedings of the 24th International Conference on World Wide Web*.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Feature-rich twitter named entity recognition and classification. In *proceedings of W-NUT 2016*, Osaka, Japan.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.