A Feature-based Ensemble Approach to Recognition of Emerging and Rare Named Entities

Utpal Kumar Sikdar
R & D Department
Flytxt
Trivandrum, Kerala, India
utpal.sikdar@flytxt.com

Björn Gambäck
Department of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway
gamback@ntnu.no

Abstract

Detecting previously unseen named entities in text is a challenging task. The paper describes how three initial classifier models were built using Conditional Random Fields (CRFs), Support Vector Machines (SVMs) and a Long Short-Term Memory (LSTM) recurrent neural network. The outputs of these three classifiers were then used as features to train another CRF classifier working as an ensemble.

5-fold cross-validation based on training and development data for the emerging and rare named entity recognition shared task showed precision, recall and F_1-score of 66.87%, 46.75% and 54.97%, respectively. For surface form evaluation, the CRF ensemble-based system achieved precision, recall and F_1 scores of 65.18%, 45.20% and 53.30%. When applied to unseen test data, the model reached 47.92% precision, 31.97% recall and 38.55% F_1-score for entity level evaluation, with the corresponding surface form evaluation values of 44.91%, 30.47% and 36.31%.

1 Introduction

The recognition of named entities is inherently complicated by the fact that new names emerge constantly and productively. This is particularly true for social media text and for other texts that are written in a more informal manner, where the issue is further complicated by a higher degree of misspellings as well as different types of unconventional spellings; on social media such as Twitter, abbreviated forms of words are common, as are merging of multiple words, special symbols and characters inserted into the words, etc.

Several approaches to Twitter named entity extraction have been explored, but it is still a challenging task due to noisiness of the texts. Liu et al. (2011) proposed a semi-supervised learning framework to identify Twitter names, using a k-Nearest Neighbors (kNN) approach to label names and taking these labels as an input feature to a Conditional Random Fields (CRF) classifier, achieving almost 80% accuracy on their own annotated data. Ritter et al. (2011) proposed a supervised model based on Labeled LDA (Ramage et al., 2009), and also showed part-of-speech and chunk information to be important components in Twitter named identification. Li et al. (2012) introduced an unsupervised Twitter named entity extraction strategy based on dynamic programming.

The present work addresses emerging and rare entity recognition. The first Twitter named entity shared task was organized at the ACL 2015 workshop on noisy user-generated text (Baldwin et al., 2015), with two subtasks: Twitter named entity identification and classification of those named entities into ten different types. Of the eight systems participating in the first workshop, the best (Yamada et al., 2015) achieved an F_1 score of 70.63% for Twitter name identification and 56.41% for classification, by combining supervised machine learning with high quality knowledge obtained from several open knowledge bases such as Wikipedia. Another team, (Akhtar et al., 2015) used a strategy based on differential evolution, getting F_1 scores of 56.81% for the identification task and 39.84% for classification.

A second shared task on Twitter Named Entity recognition was organized at COLING in 2016 (Strauss et al., 2016). The best placed system (Limsopatham and Collier, 2016) used a bi-directional LSTM (Long Short-Term Memory) neural network model, and achieved 52.41% and 65.89% F_1-scores on entity level and segmentation

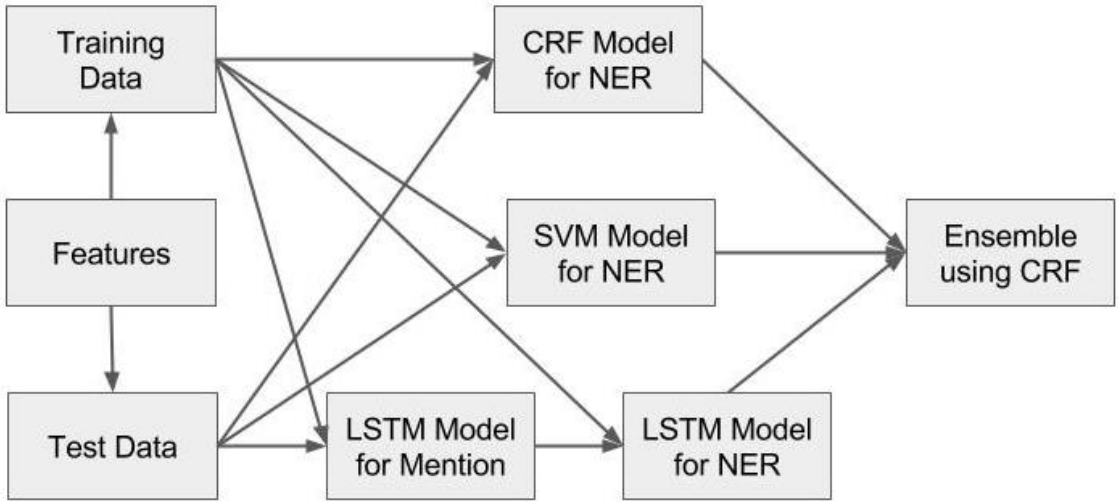

Figure 1: Overall system architecture

level evaluation, respectively. A system based on Conditional Random Fields (CRFs) and a range of features (Sikdar and Gambäck, 2016) achieved the best recall at segmentation level evaluation, and the second best F_1-score (63.22%).

A related shared task on Twitter named entity recognition and linking (NEEL) to the DBpedia database was held in conjunction with the 2016 WWW conference (Cano et al., 2016). Five teams particpated, with the best system (Waitelonis and Sack, 2016) achieving recall, precision and F-scores of 49.4%, 45.3% and 47.3%. In that system, each token was mapped to gazetteers developed from DBpedia. Tokens that were not nouns or did not match stop words were discarded.

The present paper outlines an ensemble-based machine learning approach to the identification and classification of rare and emerging named entities. Here the classification categories are Person, Location, Corporation, Product, Creative-work and Group. A Conditional Random Fields (Lafferty et al., 2001) classifier was trained using the outputs from three other classifiers as features, with those classifiers in turn being built using three different learning strategies: CRFs, Support Vector Machines (SVMs), and a deep learning based Long Short-Term Memory (LSTM) recurrent neural network. The rest of the paper is organized as follows: The named entity identification methodology and the different features used are introduced in Section 2. Results are presented and discussed in Section 3, while Section 4 addresses future work and concludes.

2 Name Recognition Methodology

The named entity recognition method is divided into two steps. In the first step, three classifiers are built to recognize named entities using different features from the unstructured text. In the second step, the outputs from the three classifiers are considered as three features and used to train a CRF classifier working as an ensemble learner, to produce the final named entity recognition. The system architecture is shown in Figure 1.

2.1 CRF-based Named Entity Recognition

The Conditional Random Fields Named Entity Recognition model was implemented using the C++ based CRF++ package[1], which allows for fast training by utilizing L-BFGS (Liu and Nocedal, 1989), a limited memory quasi-Newton algorithm for large scale numerical optimization. The CRF classifier was trained with L2 regularization and a range of features:

- local context (-3 to +2)[2],
- part-of-speech information,
- chunk information,
- suffix and prefix characters (-4, +4), and
- word frequency,

together with a number of Boolean flags, namely, is-word-length < 5, is-followed-by-special-character ('@' or '#'), is-stop-word, is-all-upper-case, is-all-digit, is-alpha-and-digit-together, and is-last-word.

[1] https://taku910.github.io/crfpp/

[2] Here '-' and '+' indicate the number of preceding and following words in the context window, respectively.

2.2 SVM-based Named Entity Recognition

Since Support Vector Machines previously have been successfully utilized to recognize named entities in formal text, e.g. by Isozaki and Kazawa (2002), a classifier was built using the C++ based SVM package Yamcha[3] with polynomial kernel and default settings. The same features as for the CRF model were used to train the SVMs.

2.3 LSTM-based Named Entity Recognition

The proposed deep learning based name entity recognition model consists of two Long Short-Term Memory recurrent neural network (Hochreiter and Schmidhuber, 1997), a model which was also successfully used by Lample et al. (2016) to achieve state-of-the-art named entity recognition results in formal texts. The first LSTM identifies the boundaries of a named entity (called mention) and this mention is then used as one of the features for named entity recognition in the second LSTM.

For identifying mentions, two binary features, is-start-with-capital-letter and is-all-upper-case, were extracted together with the following:

- word shape-1, a length 6 one-hot vector containing the following six binary flags: upper case, lower case, digit, '@' symbol, '#' symbol, and other characters,
- word shape-2, a length 39 one-hot vector consisting of the 26 letters of the English alphabet converted to lower case, together with the ten digits, the two symbols '@' and '#', and one spot for other characters, and
- a word2vec pre-trained vector of length 150,

Tweets were collected from the W-NUT 2016 shared task,[4] the 2016 NEEL challenge,[5] and the W-NUT 2017 workshop datasets to build the word2vec model (Mikolov et al., 2013a,b). The skip-gram approach was used with negative sampling and a context window of 5. All features were then concatenated into one vector and fed to the first LSTM network for mention recognition.

After a mention had been identified, it was used as one of the features for recognition of named entities in the second LSTM model, which as features together with word-shape-1 and word-shape-2 (as above) utilized three Boolean flags (is-mention, is-start-with-capital-letter, and is-all-upper-case), and GloVe (Pennington et al., 2014), a pre-trained Twitter word vector (here a GloVe vector of dimension 100 was selected).

These features were concatenated to train an LSTM model for 50 epochs with a batch size of 256. The network was set up as consisting of two hidden layers with 256 hidden units.

[3] http://chasen.org/~taku/software/yamcha/

[4] http://noisy-text.github.io/2016/

[5] http://microposts2016.seas.upenn.edu/challenge.html

Data set	tweets	named entities
Training	3,394	1,975
Development	1,009	833
Test	1,287	1,041

Table 1: Twitter dataset statistics

2.4 A Named Entity Recognition Ensemble

In the second step, the outputs of the above three classifiers were considered as input features to a CRF classifier, which was trained using these three features together with the previous and next two context words. Note that this final CRF classifier being used a selector in the ensemble thus does not cover all features of the CRF classifier described above (Section 2.1), but only utilizes the context and the three classifiers' outputs as features.

An ensemble based on using majority voting was also tested, which selected the output of one of the classifiers at random, in case they all produced different outputs. The results of the voting-based ensemble improved on the CRF and SVM models, but turned out worse than the LSTM model. However, the ensemble using a Conditional Random Field model to select among the classifier outputs improved results over the board.

3 Experiments

The experiments were based on the datasets provided by the organizers of the W-NUT 2017 shared task on emerging and rare named entity recognition (Derczynski et al., 2017). The statistics of the datasets are shown in Table 1.

3.1 Results

For the experiments, the development data was merged with the training data, and a 5-fold cross-validation was executed. The CRF-based classifier model produced the precision, recall and F_1 values of 51.79%, 45.51% and 48.31%, respectively. The LSTM model performed better when compared to the CRF-based model with respect to recall and F_1-score, achieving precision, recall and

System	Entity			Surface form		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	51.79	45.51	48.31	47.25	42.02	44.48
SVM	48.99	44.87	46.65	44.56	41.64	43.05
LSTM	51.58	51.33	51.37	47.21	47.94	47.57
Ensemble	66.87	46.75	54.97	65.18	45.20	53.30

Table 2: 5-fold cross-validated results on combined development and training data

System	Entity			Surface form		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score
CRF	40.75	28.17	33.32	38.53	27.43	32.05
SVM	34.46	29.38	31.72	32.58	28.69	30.51
LSTM	39.83	30.86	34.78	37.52	29.11	32.78
Ensemble	47.92	31.97	38.55	44.91	30.47	36.31

Table 3: Performance on the unseen test data (5-fold cross-validated)

F_1 values of 51.58%, 51.33% and 51.37%. However, as shown in Table 2, the CRF-ensemble approach outperformed all the other models with respect to F_1-score. For surface evaluation, a similar behaviour could be observed, with the ensemble model achieving the highest F_1-score at 53.30%.

The different classifiers were also applied to the unseen test data and produced similar results after 5-fold cross-validation, with the ensemble approach achieving the best F_1-score compared to all other models, as can be seen in Table 3. The CRF ensemble's named entity precision, recall and F_1-score on the test data were 47.92%, 31.97% and 38.55%, respectively. For surface form evaluation, the ensemble system achieved 44.91% precision, 30.47% recall and 36.31% F_1-score.

Table 4 compares our results (FLYTXT) to the other systems participating in the shared task, with the FLYTXT ensemble-based system placing in 5th position in the final ranking on both named entity and surface form evaluation.

3.2 Error Analysis

The system suffers from poor recall, with the model only finding 720 of 1079 named entities in the test data. The system also classified many identified named entities wrongly, and in total correctly identified 345 named entities. This may be due to almost all named entities present in the test data being unknown and fairly dissimilar to the ones appearing in the training data.

4 Conclusion

This paper has proposed an ensemble-based system for Twitter named entity identification and classification. A range of different features was developed to extract Twitter names from the tweets. Three initial classifiers were built, one for CRF-based named entity extraction, one utilizing SVMs, and one based on a deep learner (LSTM). The ensemble utilized a CRF classifier taking the output of the other three models as input.

In the future, we will analyse the errors in more detail and aim to use external resources (e.g., DBpedia and Wikipedia) to reduce the misclassification of the tokens, as well as to identify more entities from the tweets. We will also try to generate more models and later ensemble these model to improve the system performance.

Team	Entity	Surface form
UH-RiTUAL	41.86	40.24
SpinningBytes	40.78	39.33
SJTU-Adapt	40.42	37.62
Arcada	39.98	37.77
FLYTXT	38.35	36.31
MIC-CIS	37.06	34.25
Drexel-CCI	26.30	25.26

Table 4: Comparison of system results (F_1 scores)

References

Md Shad Akhtar, Utpal Kumar Sikdar, and Asif Ekbal. 2015. IITP: Multiobjective differential evolution based Twitter named entity recognition. In (Xu et al., 2015), pages 106–110.

Timothy Baldwin, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In (Xu et al., 2015), pages 126–135.

Amparo E. Cano, Daniel Preoţiuc-Pietro, Danica Radovanović, Katrin Weller, and Aba-Sah Dadzie. 2016. #Microposts2016 — 6th workshop on 'making sense of microposts'. In *Proceedings of the 25th World Wide Web Conference (WWW'16)*. ACM, Montréal, Canada, pages 1041–1042.

Leon Derczynski, Eric Nichols, Marieke van Erp, and Nut Limsopatham. 2017. Results of the WNUT2017 shared task on novel and emerging entity recognition. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text*. ACL, EMNLP 2017, Copenhagen, Denmark.

Bo Han, Alan Ritter, Leon Derczynski, Wei Xu, and Timothy Baldwin, editors. 2016. *Proceedings of the 2nd Workshop on Noisy User-generated Text*. ACL, 26th COLING, Osaka, Japan.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Hideki Isozaki and Hideto Kazawa. 2002. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th International Conference on Computational linguistics*. ACL, Taipei, Taiwan, volume 1. Paper 54.

John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*. IMIS, Williamstown, MA, USA, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. ACL, San Diego, CA, USA, pages 260–270.

Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and Bu-Sung Lee. 2012. TwiNER: Named entity recognition in targeted Twitter stream. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, Portland, OR, USA, pages 721–730.

Nut Limsopatham and Nigel Collier. 2016. Bidirectional LSTM for named entity recognition in Twitter messages. In (Han et al., 2016), pages 145–152.

Dong C. Liu and Jorge Nocedal. 1989. On the limited memory BFGS method for large scale optimization. *Mathematical Programming* 45(1):503–528.

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*. ACL, Portland, OR, USA, pages 359–367.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*. Curran Associates, Red Hook, NY, USA, pages 3111–3119.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *The 2014 Conference on Empirical Methods in Natural Language Processing*. ACL, Doha, Qatar, pages 1532–1543.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. ACL, Singapore, pages 248–256.

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*. ACL, Edinburgh, Scotland, UK, pages 1524–1534.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Feature-rich Twitter named entity recognition and classification. In (Han et al., 2016), pages 164–170.

Benjamin Strauss, Bethany E. Toma, Alan Ritter, Marie Catherine de Marneffe, and Wei Xu. 2016. Results of the WNUT16 named entity recognition shared task. In (Han et al., 2016), pages 138–144.

Jörg Waitelonis and Harald Sack. 2016. Named entity linking in #tweets with KEA. In *Proceedings of 6th Workshop on Making Sense of Microposts*. CEUR, Montréal, Canada, pages 61–63.

Wei Xu, Bo Han, and Alan Ritter, editors. 2015. *Proceedings of the ACL 2015 Workshop on Noisy User-generated Text*. Beijing, China.

Ikuya Yamada, Hideaki Takeda, and Yoshiyasu Takefuji. 2015. Enhancing named entity recognition in Twitter messages using entity linking. In (Xu et al., 2015), pages 136–140.