Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media

Patrick Jansson and **Shuhua Liu**
Arcada University of Applied Sciences
Jan-Magnus Janssonin aukio 1, 00560 Helsinki, Finland
{patrick.jansson, shuhua.liu}@arcada.fi

Abstract

This paper reports our participation in the W-NUT 2017 shared task on emerging and rare entity recognition from user generated noisy text such as tweets, online reviews and forum discussions. To accomplish this challenging task, we explore an approach that combines LDA topic modelling with deep learning on word level and character level embeddings. The LDA topic modelling generates topic representation for each tweet which is used as a feature for each word in the tweet. The deep learning components consist of two-layer bidirectional LSTM and a CRF output layer. Our submitted result performed at 39.98 (F1) on entity and 37.77 on surface forms. Our new experiments after submission reached a best performance of 41.81 on entity and 40.57 on surface forms.

1 Introduction

The shared task Emerging and Rare Entity Recognition at the 3rd Workshop on Noisy User-generated Text (W-NUT 2017) takes on the challenge of identifying unusual, previously-unseen entities in noisy texts such as tweets, online reviews and other social discussions (http://noisy-text.github.io/2017/emerging-rare-entities.html). The emergent nature of novel named entities in user generated content and the often very creative natural of their surface forms make the task of automatic detection of such entities particularly difficult. To address such challenges, the shared task organizer prepared training, development and test datasets and provided to the participants. The datasets try to "resemble turbulent data containing few repeated entities, drawn from rapidly-changing text types or sources of non-mainstream entities". Results from the shared task are evaluated using F1 measures on the entities and surface forms found in the test data. It rewards systems at correctly detecting a diverse range of entities rather than only the frequent ones.

Inspired by the work of Limsopatham and Collier (2016, winner of w-nut 2016 shared task on Named Entity Recognition in Twitter), Chiu and Nichols (2016), and Huang et al (2015), we approached this shared task with bidirectional LSTM models (Long Short Term Memory recurrent neural network model) enhanced by CRF output layer, using both character-level and word-level embeddings as inputs. In addition, different from the study of Limsopatham and Collier (2016), we didn't make use of orthographic features of characters but tried to incorporate POS tags as well as document topics extracted from LDA topic modelling as optional inputs to the modelling process. The LDA topic modelling generates topic representation for each tweet which is used as a feature for each word in the tweet.

Our submitted result performed at 39.98 (F1) on entity and 37.77 (F1) on surface forms, using 10% of the combined training and development set for validation. After submission, we continued with more experiments, using data combining the training set and development set in training process, with ground truth available that helps the selection of the results. Our best result reached a performance of 41.81 on entity and 40.57 on surface forms.

2 Data and Preprocessing

The shared task datasets consist of a training set, a development set and a test set. Basic statistics of each data set in shown in Table 1. The shared task focuses on discovering 6 types of target entities and surface forms of: Corporation, Creative-Work, Group, Location, Person and Product (Derczynski et al, 2017).

Entities all

	Training	Dev	Test
# tweets/posts	3394	1009	1287
Tokens total	62730	15733	23394
Entities total	3160	1250	1740
Corporation	267	46	88
Creative-work	346	238	360
Group	414	64	235
Location	793	107	244
Person	995	587	560
Product	345	208	253

Surface

	Training	Dev	Test
Corporation	180	44	79
Creative-work	259	203	292
Group	343	60	188
Location	589	99	174
Person	742	484	476
Product	284	184	200
	2397	1074	1409

Table 1: Dataset overview

In counting surface forms, every "word-label" combination has to be unique and letter case sensitive. When the same word appears twice but as different entities both are counted.

For the stop words removing, we utilized the Stopwords ISO (https://github.com/stopwords-iso/stopwords-en) list. The cutoff value for infrequent terms is set as one when applying LDA modelling.

3 Emerging and Rare Entity Detection from Social Media: Framework and Methods

Our approach to emerging and rare entity detection from social media is illustrated in Figure 1. Our methodology framework consists of the following components: (1) character-level embeddings and bidirectional LSTM modeling; (2) word level embeddings and bidirectional LSTM modelling; (3) LDA topic modelling, POS tags enhanced bidirectional LSTM; (4) fully connected layers, and (5) a CRF (Conditional Random Fields) output layer.

3.1 Character-level Representation

Character-level information was found to be valuable input for named entity recognition from social media (Limsopatham and Collier, 2016; Vosoughi et al, 2016). Chiu and Nichols (2015) found that modelling both the character-level and word-level embeddings within a neural network for named entity recognition helps improve the performance.

In our system, each character is represented as an N dimensional embedding which is learned and adjusted during the training process. The character level representations will then be merged into one M dimensional (50d-200d seems to work well) representation for each word. Character capitalization is kept.

We used 20-dimensional embeddings to represent each character. To learn character-level representations for each word we use a bidirectional LSTM to create a 200-dimensional representation for each word.

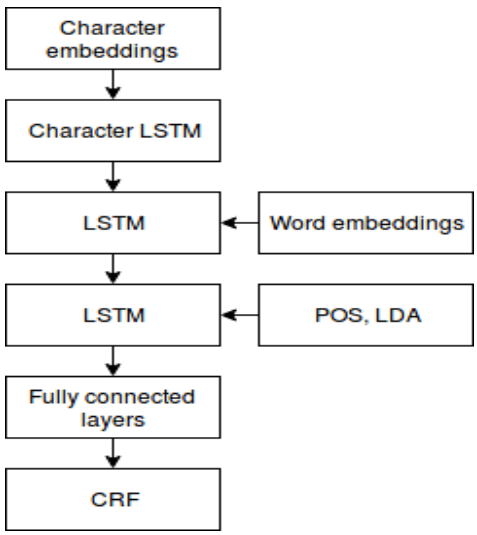

Figure 1: Methodology Framework

3.2 Word Embeddings

Word embeddings are distributed representation of words that offers continuous representations of words and text features such as the linguistic context of words (Mikolv et al, 2013a,

2013b). Word embeddings are the current norm for many text applications as they are found to be able to accurately capture not only syntactic regularities but also (local) semantic regularities of words and phrases (Mikolv et al, 2013a, Hasen et al, 2015; Limsopatham and Collier, 2016; Vosoughi et al, 2016).

Estimation of the word vectors is done using various types of model architectures trained on large corpora. Word2vec (Mikolv et al, 2013a, 2013b) and GloVe (Pennington et al, 2014) are two widely used efficient model architectures and algorithms for learning high quality continuous vector representations of words from huge data sets with billions of words. They have been used to train and create word embeddings that can be applied directly by other applications.

Considering our target source and based on some primitive test, we choose to use 200-dimensional GloVe pre-trained embeddings (Pennington et al, 2014), which was trained on a Twitter corpus with 27 billion tokens and a vocabulary size of 1.2 million.

3.3 POS Tagging

Part of Speech is also an important indicator of named entities, which we would like to include in our model (Huang et al, 2015). GATE Twitter POS tagger (https://gate.ac.uk/wiki/twitter-postagger.html) is used to assign POS tags for each word. POS tags is represented as 50-dimensional trainable embedding.

3.4 LDA Topic Modelling

Topic modeling offers a powerful means for finding hidden thematic structure in large text collections. In topic models, topics are defined as a distribution over a fixed vocabulary of terms and documents are defined as a distribution over topics. LDA topic modeling and its variations represent the most popular methods (Blei et al, 2003; Blei, 2012).

We consider the topic composition of each Tweet or social media post an important indicator of subject domain context, which can be used to complement the local linguistic context of word vector. We make use of topic representation for each tweet derived from LDA modelling as a feature for each word in the Tweet.

We applied the online LDA method by Hoffman et al (2010), implemented in Genism (https://radimrehurek.com/gensim/models/ldamodel.html). It can handily analyze massive document collections or document streams.

When generate topic models, all the three datasets are combined into one corpus, and each entry is treated as a separate document. Each document is cleaned and preprocessed, which includes removing stop-words, punctuation and infrequent terms. An LDA model of 250 topics was trained and used for our system that generated submitted results. Using the model, we get a document level topic for each document, the topic value is then assigned to each word in the document. We also use the model to get a topic for each word in the documents. If the probability that a document or a word belongs to a topic is, the same for each topic, a special token is assigned to it instead of a topic. Each topic token is then assigned to a 250-dimensional embedding, embeddings for document and word-level topics are initialized separately.

3.5 Two-Layer Bidirectional LSTM

Bidirectional LSTM has been shown effective for modelling social media sentences (Huang et al., 2015; Dyer et al., 2015; Limsopatham and Collier, 2016). To learn deep neural models for named entity recognition we adopted a two-layer bidirectional LSTM, followed by two fully connected layers, and a Conditional Random Field (CRF) as an output layer where we maximize the joint likelihood.

For the first LSTM layer, we concatenate the 200-dimensional GloVe word embeddings and the 200-dimensional embeddings for character level representation. For the second layer, we concatenate the output of the first layer with the POS-feature embeddings and LDA-feature embeddings. The LSTM output dimensions are 256 for the first layer and 512 for the second layer.

After the second LSTM layer, we use two fully connected layers at each time step, and feed this representation into the CRF output-layer. The dimensions of the fully connected layers are 128 and 64 for the first and second layer respectively.

Between each layer in the network we applied dropout and batch normalization (Ioffe and Szregedy, 2015). A dropout rate of 0.25 is used for the first two layers of the network (the Character LSTM and the character + word LSTM). For all the other layers of the network, a dropout rate of 0.5 is used.

The fully connected layers are extra hidden layers before the CRF output layer, which allow the models to learn higher level representations

without adding complexity through an extra compositional layer (Rei and Yannakoudakis, 2016).

Conditional random field (CRF) has shown to be one of the most effective methods for named entity recognition in general and in social media (Lafferty et al., 2001; McCallum and Li, 2003; Baldwin et al., 2015; Limsopatham and Collier, 2016). It also helped our system to gain performance in recognizing emerging entities and surface forms.

The deep neural model was implemented using Keras with a TensorFlow backend and Keras community contributions for the CRF implementation. One model is trained for both entity and surface form recognition. Any feature can be included or excluded as needed when running the model.

4 Experiments and Results

In this section, we report two sets of experiments and results. Results from the 1st set of experiments were submitted to the shared task organizer for evaluation. The 2nd set of experiments are done after the submission. Using the ground truth released by the organizer we evaluated the results directly by ourselves. The ground truth being available also helps us in identifying the best model.

4.1 1st Set of Experiments and Submitted Results

To train the model, the training set and the dev sets are merged, of which 10% (in terms of size, about half of the original dev set) are used for validation. We used a batch size of 32 for training, and the RMSprop optimizer with an initial leaning rate of 0.001. The results are shown in Table 2. The results from all participating systems are presented in Table 3 (Derczynski, et al, 2017).

The overall performance of our system reached 39.98 on entities and 37.77 on surface forms. The performance on Person and Location types of entities and surface forms are comparatively better, with F1 score at 55,88 and 47,38 respectively for entities, and F1 score at 53.30 and 42.80 for their surface forms. The system is less effective on identifying Corporation, Product, Creative-work and Group types of entities and surface forms, especially disappointing in terms of recall. For Creative-work and Product type entities, recall only reached 9.86% and 11.02% respectively.

	Accuracy	Precision	Recall	FB1
Entities	94.03%	47.40%	34.57%	39.98
Surface forms		44.94%	32.57%	37.77
	Entity types	Precision	Recall	FB1
	Corporation	19.05%	18.18%	18.60
	Creative-work	31.82%	9.86%	15.05
	Group	38.36%	16.97%	23.53
	Location	44.00%	51.33%	47.38
	Person	58.91%	53.15%	55.88
	Product	31.11%	11.02%	16.28
	Surface forms	Precision	Recall	FB1
	Corporation	20.37%	18.33%	19.30
	Creative-work	32.56%	10.29%	15.64
	Group	35.29%	17.02%	22.97
	Location	39.73%	46.40%	42.80
	Person	56.38%	50.53%	53.30
	Product	31.82%	11.97%	17.39

Table 2: Our submitted results

Team	F (entity)	F (surface)
MIC-CIS	37.06	34.25
Arcada	39.98	37.77
Drexel-CCI	26.30	25.26
SJTU-Adapt	40.42	37.62
FLYTXT	38.35	36.31
SpinningBytes	40.78	39.33
UH Ritual	41.86	40.24

Table 3: Submitted results, all participants

4.2 2nd Set of Experiments and Updated Results

After submission, we continued our modelling work with new training strategies. In terms of the data, all samples of the training set and dev set are used for training the model, which is then directly applied to test set. We also experimented more with different options of the number of topics in LDA topic modelling. We found that incorporating LDA features does have a positive effect on the performance. We used models with topic counts in the range of 20, 50, 150, 250, 350, 450. The results (FB1 value for entity and surface forms) are illustrated in Table 4. The scores are maxima out of

two runs of experiments, where each run goes through all the topic counts.

# Topics	0	20	50	150
Entity	40.63	40.63	41.48	**41.81**
Surface	38.06	38.95	39.68	**40.57**
# Topics	250	350	450	
Entity	41.78	41.66	40.95	
Surface	39.90	39.48	39.29	

Table 4: Performance variation related with number of topics for LDA modelling

When topic number set as 150, breakdown of the performance shows that the system performed best for the more difficult entity types and surface forms, as is shown in Table 5. For Creative-work and Product type entities, recall reached 15.49% and 14.96% respectively. For their surface forms, recall reached 16.18% and 16.24% respectively.

	Accuracy	Precision	Recall	FB1
Entities	94.10%	50.86%	35.50%	41.81
Surface forms		49.55%	34.35%	40.57
	Entity types	Precision	Recall	FB1
	Corporation	31.71%	19.70%	24.30
	Creative-work	37.29%	15.49%	21.89
	Group	40.62%	15.76%	22.71
	Location	49.09%	54.00%	51.43
	Person	61.16%	51.75%	56.06
	Product	31.15%	14.96%	20.21
	Surface forms	Precision	Recall	FB1
	Corporation	31.43%	18.33%	23.16
	Creative-work	37.93%	16.18%	22.68
	Group	40.32%	17.73%	24.63
	Location	47.14%	52.80%	49.81
	Person	60.06%	49.20%	54.09
	Product	32.20%	16.24%	21.59

Table 5: Performance on different types of entities, number of topics for LDA modelling = 150

5 Discussion and Conclusion

In this paper, we reported our participation in the W-NUT 2017 shared task on emerging and rare entity recognition from user generated noisy text. We described our system that leverages the power of LDA topic modelling, POS tags, character-level and word-level embeddings, bidirectional LTSM and CRF. The LDA topic modelling generates topic representation for each tweet or social media post. The deep learning model consists of two-layer bidirectional LSTM, two fully connected layers and a CRF output layer. We make use of topic representation for each tweet derived from LDA modelling as a feature for each word in a tweet or post. The topic composition of each post offers a certain subject domain context that could complement the local linguistic context of word embeddings.

We reported two sets of experiments and results. Results from the 1st set of experiments were submitted to the shared task organizer for evaluation. Our submitted results performed at 39.98 (F1) on entities and 37.77 (F1) on surface forms.

The 2nd set of experiments are done as follow up study after the submission, adopting a different training strategy. Using the ground truth released by the organizer we evaluated the results directly by ourselves. The ground truth being available helped us to identify the best model.

We experimented more with different options of the number of topics in LDA topic modelling. We found that incorporating LDA features does have a positive effect on the performance. The new results reached a best performance of 41.81 on entities and 40.57 on surface forms, with the number of topics set as 150. When the number of topics is set the same as for our submitted results (i.e. 250), the new results showed performance gain as well, reached 41.78 on entities and 39.90 on surface forms.

For future work, it would be interesting to train the LDA model on a larger corpus, to hopefully find a more accurate subject domain context for each tweet or post. It would be useful as well to explore the effects of alternative word embeddings such as fasttext. It would also be interesting to apply our system in identifying city event related entities and surface forms from other social media data.

Acknowledgements

This work is part of the DIGILENS-HKI project on mining community knowledge from social media (http://rdi.arcada.fi/katumetro-digilens-hki). We thank and gratefully acknowledge funding from Helsinki Region Urban Research Program (http://www.helsinki.fi/kaupunkitutkimus/) and Arcada Foundation (tuf.arcada.fi).

References

Blei D, A. Ng and M. I. Jordan. 2003. Latent Dirichlet Allocation. Advances in Neural Information Processing Systems. Pages 601-608.

Blei D. 2012. Probabilistic Topic Models. Communications of the ACM, 55(4):77-84.

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 Proceedings of the Workshop on Noisy User-generated Text (WNUT 2016), Osaka, Japan.

Baldwin Timothy, 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.

Chiu Jason P. C., Eric Nichols, Named Entity Recognition with Bidirectional LSTM-CNNs, TACL2016

Derczynski Leon, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphae ̈l Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. Information Processing & Management, 51(2):32–49.

Derczynski Leon, Eric Nichols, Marieke van Erp, Nut Limsopatham, Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition, in Proceedings of the 3rd Workshop on Noisy, User-generated Text, 2017

Dyer Chris, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. arXiv preprint arXiv:1505.08075.

Godin Frederic, Baptist Vandersmissen, Wesley De Neve and Rik Vande Walle. 2015. Multimedialab@acl w-nut shared task: Named entity recognition for twitter micro posts using distributed word representations. ACL-IJCNLP 2015, page 146.

Hasan Sadid A., Yuan Ling, Joey Liu, Oladimeji Farri, Exploiting Neural Embeddings for Social Media Data Analysis. TREC 2015

Hoffman M, D. Blei and F. Bach. 2010. Online learning for Latent Dirichlet Allocation. Advances in Neural Information Processing Systems 23, 856-864.

Huang Zhiheng, Wei Xu, Kai Yu, Bidirectional LSTM-CRF Models for Sequence Tagging, ArXiv2015

Lafferty John, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the eighteenth international conference on machine learning, ICML, volume 1, pages 282–289.

Ioffe Sergey, Szregedy Christian, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, ArXiv2015

Limsopatham Nut, Nigel Collier, Bidirectional LSTM for Named Entity Recognition in Twitter Messages, 2016

McCallum Andrew and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pages 188–191. Association for Computational Linguistics.

Mikolov Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.

Mikolov Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Proceedings of NIPS, 2013

Pennington Jeffrey, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In EMNLP, pages 1532–1543.

Vosoughi Soroush, Prashanth Vijayaraghavan, Deb Roy, Tweet2Vec: Learning Tweet Embeddings Using Character-level CNN-LSTM Encoder-Decoder, Proceedings of SIGIR 2016, July 17-21, 2016, Pisa, Italy

Zhang Xiang, Junbo Zhao, Yann LeCun. Character-level Convolutional Networks for Text Classification. Advances in Neural Information Processing Systems 28 (NIPS 2015). Poster. Datasets. Code. Errata.

Marek Rei and Helen Yannakoudakis, “Compositional Sequence Labeling Models for Error Detection in Learner Writing”, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 1181–1191, Berlin, Germany, August 7-12, 2016.