A Comparative Study on Generalizability of Information Extraction Models on Protest News

Erkan Başar[1,2], Simge Ekiz[1], and Antal van den Bosch[2]

[1] FloodTags,
Binckhorstlaan 36, M2.11, 2516 BE, The Hague, The Netherlands
{basar,sekiz}@floodtags.com
[2] Centre for Language Studies, Radboud University,
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
a.vandenbosch@let.ru.nl

Abstract. Information Extraction applications can help social scientists to obtain necessary information to understand the reasons behind certain social dynamics. Many recent state-of-the-art information extraction approaches are based on supervised machine learning which can recognize information that has similar patterns with previously shown ones. Recognizing relevant information with never-shown patterns, however, is still a challenging task. In this study, we design a Recurrent Neural Network (RNN) architecture employing ELMo embeddings and Residual Bidirectional Long-Short Term Memory layers to overcome this challenge in the context of CLEF 2019 ProtestNews shared task. Furthermore, we train a classical Conditional Random Fields (CRF) model as our strong baseline to display a contrast between a state-of-the-art classical machine learning approach and a recent neural network method both in performance and in generalizability. We show that RNN model outperforms classical CRF model and shows a better promise on generalizability.

Keywords: information extraction · recurrent neural networks · conditional random fields · word embeddings.

1 Introduction

Social science studies can benefit from analyzing and comparing protest event information from multiple countries to understand the reasons behind certain social dynamics such as emerging welfare regimes. Although, online mass media agencies report the major incidents as soon as in a day, manually collecting such amounts of data is time-taking, expensive and hard to maintain. Automating the process with natural language processing (NLP) allows us to harness such information on a large scale with a high speed.

Many recent state-of-the-art NLP approaches are based on supervised machine learning technique where the machine discovers patterns on manually prepared gold-standard data to learn how to detect relevant information. With the recent developments, it is possible to teach information extraction models to detect information that has similar patterns to the training data at reasonable levels. However, generating a model that can correctly recognize relevant data with never-observed patterns is relatively a challenging task. Generalizability of a model especially affects the results when used on cross-cultural text. Even though the text is written in the same language on the model trained, the cultural differences can affect the context and the language usage such as the word choices and ordering. This issue can cause semantic bias[3], potentially neglect the model, render it useless, and thus, eventually force us to generate models specific to cultures and regions.

CLEF organizers prepared the ProtestNews Task 3:Event information extraction shared task[14], stressing both the importance of automatically extracting protest event information from news and the impact of the generalizability in natural language processing applications. The aim of the shared task is to develop generalizable NLP tools that is robust enough to be used regardless of their data source with a focus on detecting and extracting relevant information on protest news.

We design a Recurrent Neural Network (RNN) using Residual Bidirectional Long Short-Term Memory (BiLSTM) layers with pretrained Embeddings for Language Models (ELMo) word embeddings[20] and a classical Conditional Random Fields (CRF)[15] based machine learning model. We accept the CRF based model as a strong baseline and compare both the performances and the generalizabilities of two models. We display a contrast between one of the classical machine learning approaches and one of the latest methods as well as we aim for this automation to be as accurate as possible, knowing that our methods will probably not be as precise as human annotation.

2 Background

Information Extraction (IE) is the task of automatically extracting structured information from unstructured text and studied as part of the natural language processing area [6]. Previous studies demonstrated that end-to-end information extraction systems can be developed to analyze news media data by employing probabilistic approaches [23, 10]. In the ProtestNews shared task, the event information extraction is designed as an entity sequence labelling task. The goal of the sequence labelling can be described as labelling the sequences of relevant words in text by a single categorical class.

A classical machine learning framework for labelling sequential data is Linear-chain Conditional Random Fields [15]. CRF prevents the label bias problem that occurs when there is an uncertainty in the previous tag of the sequence [19]. The strength of the CRF is also coming from the ability of dealing with the arbitrary, overlapping features of the input [7]. CRF is accepted as one of the state-of-the-

art approaches and CRF-based models are applied to sequence labelling tasks[24, 8] including some of the state-of-the-art named entity recognition tools[9].

In the recent years, studies have shown promise with Recurrent Neural Networks [16]. Long-Short Term Memories (LSTM) [11] based RNNs can learn temporal dependence between the sequences and also when to forget them. Moreover, Bidirectional LSTM [13] based networks are reading the data once from the beginning to the end and once from the end to the beginning and this making them learn stronger relations. Furthermore, neural networks started to be used in feature extraction to generate word representation that is effective in supervised sequence labelling problems [22]. In common state-of-the-art approaches, word representations are generated over individual tokens [17]. ELMo [20] is one of the recent and state-of-the-art word representation network that generates embeddings over the entire word sequences instead of individual tokens, providing an advantage in sequence based tasks.

3 Dataset

In this study, we only use the dataset provided by the ProtestNews shared task organizers in our experiments. The provided data is already separated as training, development, respectively containing 250 and 36 English news sentences from newspapers published in India. In order to test the generalizability of the models, two separate unlabelled test data is provided. The main test set contains 80 English news sentences from India, and the secondary test set contains 39 English news sentences from China. As shown in Figure 1, the average sentence lengths both in training and test sets are around 50. However, while the maximum length in the training data is 440, test data contained two sentences with lengths 579 and 643.

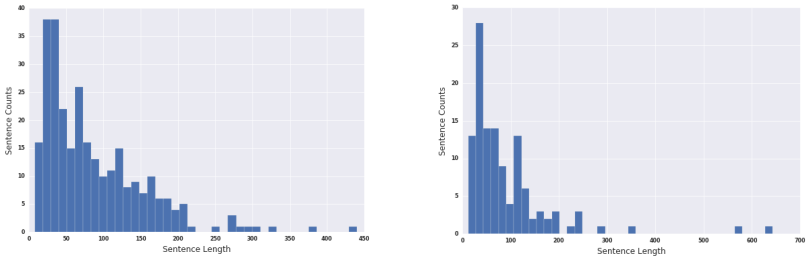

Fig. 1: The distribution of sentence lengths in training data (on the left) and in the combined test sets (on the right)

Likewise, the whole dataset is pre-tokenized and shared in standard CONLL format. Moreover, there are 8 different classes in the dataset named as *participant, trigger, loc, place, etime, fname, organizer*, and *target* given in beginning-inside-outside (BIO) tagging format. The sample sizes of the classes over the

complete entities varies as shown in Figure 2. While the *trigger* class is used 970 times while *place* class used 318 times and the least used class is *fname* with 128 labels.

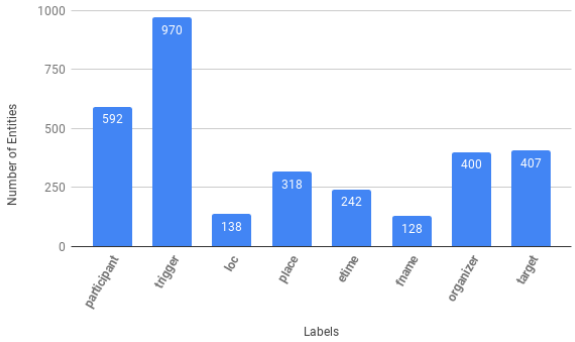

Fig. 2: The distribution of the label usage in training data. The graph shows a count of the labels per complete entity, although the data received in BIO tagging format. The uninformative *outside* class is disregarded.

4 Methods

Hereunder, we describe the two models we trained for the task and evaluation process.

4.1 Conditional Random Fields Algorithm

We employ Conditional Random Fields by using the Python binding of CRFSuite library [18]. The CRF model is trained on the given training data and the hyperparameters are optimized on the given development set by using Random Search method [2], especially to optimize the regularization parameters known as "C" parameters.

In classical supervised machine learning, the features have a great impact on the classification accuracy. For the feature extraction, we use a sliding window with a length of 5 tokens meaning that if we are extracting features of a token at position i, the sliding window will contain token_{i-2}, token_{i-1}, token_{i}, token_{i+1}, and token_{i+2}. Besides getting the features of individual tokens in the sliding window independently, we also include bi-grams and tri-grams of the tokens and most of the features.

The features extracted to train the CRF model is as listed below;

– Context: Token of the focus and its surroundings in a sliding window and n-gram combinations.

– Lemmas: Lemmatized version of the focus and its surroundings in a sliding window and n-gram combinations, obtained by using spaCy [12] NLP tool.
– Orthographic Types: The orthographic types of the tokens in a sliding window and n-gram combinations.
– Part-of-Speech: Part-of-speech tags of the tokens in a sliding window and n-gram combinations, obtained by using spaCy [12] NLP tool.
– Temporal Tags: Temporal tags of the tokens in a sliding window and n-gram combinations, obtained by using HeidelTime temporal tagging tool [21].
– Named Entities: Named entity tags of the tokens in a sliding window and n-gram combinations, obtained by using spaCy [12] NLP tool.
– isCapital: A boolean value indicates whether the first letter of the token_i is capitalised or not.

4.2 Recurrent Neural Network Approach

We propose a Recurrent Neural Network architecture briefly consists of the ELMo word embeddings network, two residual BiLSTM layers and a time distributed dense layer with softmax activation at the end, as shown in Figure 3.

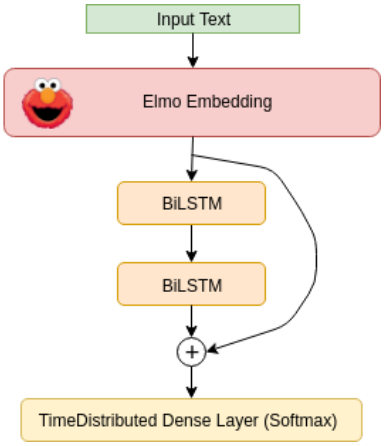

Fig. 3: Architecture of the proposed network consists of ELMo word embeddings network and two residual BiLSTM layers.

We use the ELMo network pretrained on 1 Billion Word Benchmark [4] and distributed in Tensorflow Hub[1]. Thus, ELMo is left out in our training process and used only as a feature extraction layer. ELMo embeddings are generated over the entire word sequences, unlike the other popular embeddings. Consequently the input layer of our network requires tokenized sentences as sequences. We set

[1] Tensorflow Hub: https://www.tensorflow.org/hub, Accessed on 24/05/2019

ELMo to return word embeddings of 1024 dimensions for an inputted sentence sequence. Then, the embeddings are fed to the residually connected BiLSTM layers, each with 512 units and 0.2 dropout rate.

We use Adam optimizer with the learning rate of 0.001 to find the global minimum for the sparse categorical cross-entropy loss function. We observe the relation between epochs and training and validation losses. Thus, we initially train our network for 30 epochs. After the 10^{th} epoch, however, we do not observe any difference in the validation accuracy. The learning rate, dropout rate, units, and dimensions are decided heuristically.

Implementation of the network and the loss function is done in Keras (v2.2.4) [5] with the Tensorflow (v1.14.1) [1] back-end.

4.3 Evaluation

The evaluation of the algorithms are done in the submission platform of the ProtestNews shared task [14]. The evaluation metrics used are precision, recall and f1-score.

5 Results

On the primary test set, we observe a macro average f1-score of 37.64 with CRF-based model and 55.75 with RNN model, as shown in Figure 4. However, we also see that CRF displays a higher precision score than RNN while RNN outperforms CRF model on the recall. Thus, RNN model displays a more balanced performance than CRF. On the secondary test set, the performance of CRF model significantly drops and RNN shows dominance over CRF at each scoring.

In Figure 5, we observe that both of the models has a performance loss when tested on the secondary model. We see that the performance decrease of the CRF based method is highly visible on the second test set. However, RNN based method compensate this performance drop better.

As an overall, our best performing model gives 55.75 f1-score on average.

6 Conclusions

In this study we have proposed a Recurrent Neural Network based approach to extract information on protest news. Furthermore, we have built a classical Conditional Random Fields as our strong baseline to display a contrast between a classical machine learning approach and a more recent method both in performance and in generalizability.

We have seen that the Recurrent Neural Network based model significantly outperforms Conditional Random Fields based approach both in the primary set and the secondary set. On a setup to test generalizability of each model, we have shown that the CRF based model demonstrates our initial claims with

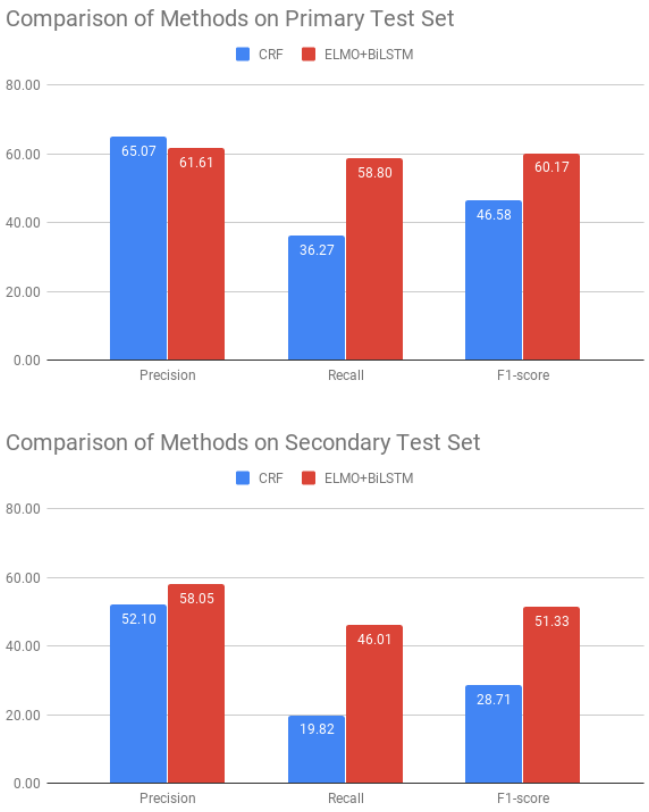

Fig. 4: Results of proposed methods on both test sets. India-English test set referred as primary test set. China-English test set referred as secondary test set

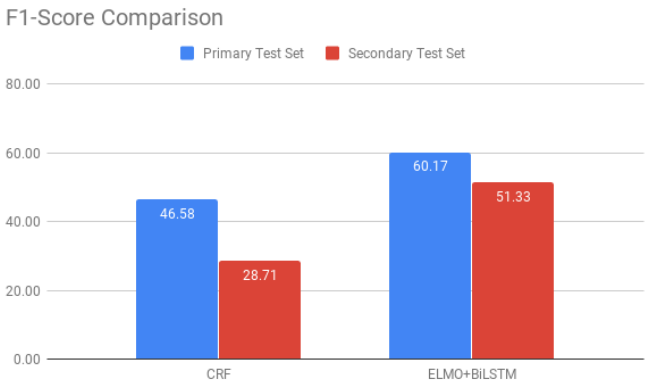

Fig. 5: F1-score comparisons of CRF and RNN models on generalizability

generalizability problem. While both of the models lost performance on the secondary test set, we conclude that RNN based approach shows a better promise on generalizability, compared to the CRF method.

In future studies, we would like to include character embedding, and evaluate whether including character-level information is going to improve the results of RNN based model.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015), http://tensorflow.org/, software available from tensorflow.org
2. Bergstra, J., Bengio, Y.: Random search for hyper-parameter optimization. Journal of Machine Learning Research **13**(Feb), 281–305 (2012)
3. Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (2017)
4. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P.: One billion word benchmark for measuring progress in statistical language modeling. Computing Research Repository (CoRR) **abs/1312.3005**, 1–6 (2013)
5. Chollet, F., et al.: Keras. https://keras.io (2015)
6. Cowie, J., Lehnert, W.: Information extraction. Commun. ACM **39**(1), 80–91 (Jan 1996). https://doi.org/10.1145/234173.234209, http://doi.acm.org/10.1145/234173.234209
7. Culotta, A., McCallum, A.: Confidence estimation for information extraction. In: Proceedings of HLT-NAACL 2004: Short Papers. pp. 109–112. Association for Computational Linguistics (2004)
8. Cuong, N.V., Chandrasekaran, M.K., Kan, M.Y., Lee, W.S.: Scholarly document information extraction using extensible features for efficient higher order semi-crfs. In: Proceedings of the 15th ACM/IEEE-CS joint conference on digital libraries. pp. 61–64. ACM (2015)
9. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005)
10. Gupta, D., Strötgen, J., Berberich, K.: Eventminer: Mining events from annotated documents. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval. pp. 261–270. ACM (2016)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (Nov 1997). https://doi.org/10.1162/neco.1997.9.8.1735, http://dx.doi.org/10.1162/neco.1997.9.8.1735
12. Honnibal, M., Montani, I.: spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)

13. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. CoRR **abs/1508.01991** (2015), http://arxiv.org/abs/1508.01991
14. Hürriyetoğlu, A., Yörük, E., Yüret, D., Yoltar, Ç., Gürel, B., Duruşan, F., Mutlu, O.: A task set proposal for automatic protest information collection across multiple countries. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) Advances in Information Retrieval. pp. 316–323. Springer International Publishing, Cham (2019)
15. Lafferty, J., McCallum, A., Pereira, F., et al.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning, ICML. vol. 1, pp. 282–289 (2001)
16. Li, B., Liu, T., Zhao, Z., Du, X.: Attention-based recurrent neural network for sequence labeling. In: Cai, Y., Ishikawa, Y., Xu, J. (eds.) Web and Big Data. pp. 340–348. Springer International Publishing, Cham (2018)
17. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 3111–3119. NIPS'13, Curran Associates Inc., USA (2013), http://dl.acm.org/citation.cfm?id=2999792.2999959
18. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007), http://www.chokkan.org/software/crfsuite/
19. Peng, F., McCallum, A.: Information extraction from research papers using conditional random fields. Information processing & management **42**(4), 963–979 (2006)
20. Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1202, https://www.aclweb.org/anthology/N18-1202
21. Strötgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Language Resources and Evaluation **47**(2), 269–298 (2013). https://doi.org/10.1007/s10579-012-9179-y
22. Turian, J., Ratinov, L., Bengio, Y.: Word representations: A simple and general method for semi-supervised learning. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. pp. 384–394. ACL '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org/citation.cfm?id=1858681.1858721
23. Vossen, P., Agerri, R., Aldabe, I., Cybulska, A., van Erp, M., Fokkens, A., Laparra, E., Minard, A.L., Aprosio, A.P., Rigau, G., Rospocher, M., Segers, R.: Newsreader: Using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. Special Issue Knowledge-Based Systems, Elsevier (2016). https://doi.org/dx.doi.org/10.1016/j.knosys.2016.07.013, http://www.sciencedirect.com/science/article/pii/S0950705116302271
24. Wu, F., Weld, D.S.: Open information extraction using wikipedia. In: Proceedings of the 48th annual meeting of the association for computational linguistics. pp. 118–127. Association for Computational Linguistics (2010)