IRSTI 20.19.21
UDC 004.912

AN ANALYTICAL STUDY OF MODERN INFORMATION EXTRACTION TECHNOLOGIES AND APPROACHES

A. Y. Nuraliyeva
Kazakh-British Technical University, Almaty, Kazakhstan
nuraliyevaassel@gmail.com
ORCID ID: 0000-0001-6451-3743

Abstract. Due to the massive use of electronic media the amount of unstructured text is increasing tremendously day by day. Many of researchers in machine learning work with that data in order to extract relevant and succinct information for such application areas like biomedical natural language processing, maintaining clinical inventories, providing speech aid to challenged children and machine translation where one convert semantic features of one language to another language. NLP technologies help us to improve our communication, achieve our goals and get results from every interaction. They also help us to overcome personal obstacles and psychological problems. By studying NLP methods correctly, we can achieve our goals in a very satisfactory way and overcome the obstacles we face. This paper covers three scientific papers and aims to provide their approach, main idea, techniques and usefulness. Article can be extremely helpful in academics, researches in natural language processing and also to novice specialists.

Keywords: unstructured text, machine learning, natural language processing, feature extraction, information extraction.

Introduction

A large number of unstructured electronic texts are available online, including news feeds, blogs, email messages, government documents, discussion journals, etc., which can be used to create a variety of content. Natural Language Processing (NLP) is a sub-section of computer science and AI that deals with how computers analyze natural (human) languages. NLP allows the application of machine learning algorithms for text and speech. For example, we can use NLP to create systems such as speech recognition, document summarization, machine translation, spam detection, named entity recognition, question answering, autocomputer, predictive text input, etc. The ultimate goal of NLP is to help computers understand language like we do. It is the driving force behind things like virtual assistants, speech recognition, emotion analysis, automatic text summary, machine translation and much more. Natural language processing (NLP) is the intersection of computing, linguistics and machine learning. The field focuses on communication between computers and humans in natural language, and NLP is about making computers understand and generate human language [1,2].

Information extraction means the automatic extraction of structured information such as entities, relationships between entities, and attributes that describe entities from unstructured sources. Information extraction (IE) is the task of automatically retrieving structured information from unstructured and/or semi-structured machine-readable documents. In most cases, this activity is related to the processing of human language text using the Natural Language Processing (NLP) method [3,4,5].

In addition, extracting the semantic relationships between objects in natural language text is a crucial step in understanding natural language applications that recognize the relationships between objects in unstructured text. In this paper, we will review three scientific papers dealing with important tasks in the field of natural language processing [5,6,7].

In this paper, three scientific works are covered: A Detailed Analysis of Core NLP for Information Extraction by Simran Kaur et al., Large Scaled Relation Extraction with Reinforcement Learning by Shizhu He et al, Attention-Based Convolutional Neural Network for Semantic Relation Extraction by Yatian Shen et al.

Methodology

In the first article, Simran Kaur and Rashmi Agrawal provided a detailed analysis of the

Stanford Core GNP which provides a range of natural language analysis tools and also examined a wide variety of techniques involved in information extraction and the problems they solve. Information extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources. For information extraction, the bootstrapping approach is mainly used. It extracts a large amount of information from unrated seeds and data from texts or other word corpora. Bootstrapping is an extremely powerful approach that extracts good patterns from unstructured language. However, its weak point is its tendency to decrease accuracy over time, since initially there was no tagging. For this algorithm as well, the choice of seeds is crucial for success. In this paper, three approaches to bootstrapping algorithms such as Nomen, Basilisk and Snowball were discussed [8,9].

The first approach essentially consists in checking whether two different algorithms have the same result on the problem or not, which is used to solve the problem of generalized name learning in a biomedical context. Second, the Basilisk algorithm was originally designed to extract semantic lexicons automatically or semi-automatically, including information extraction, answering questions and adding prepositional sentences. Thirdly, in Snowball, he works on the idea of the duality of pattern relations, according to which a good pattern will have good tuples present in it and vice versa by following an alternative approach.

Bootstrapping systems are better suited to natural language processing tasks because of their ability to learn and navigate the syntactically rich, unstructured and extremely complex nature of unstructured natural languages [10].

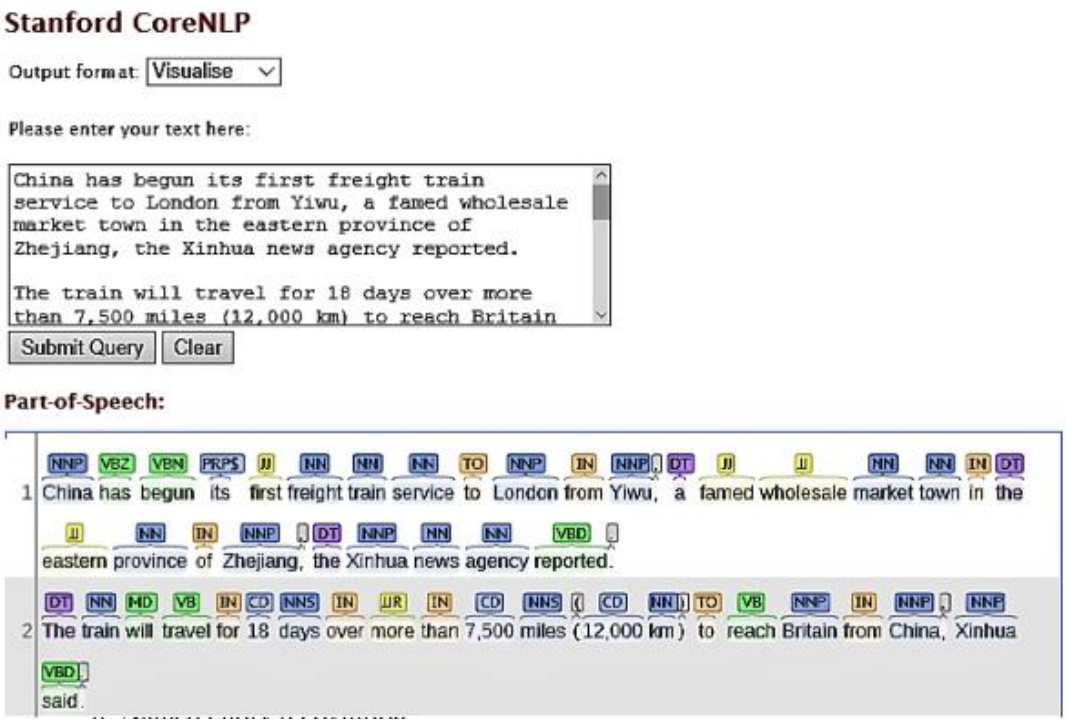

Figure 1 – Standford CoreNLP interface

For basic GNP cases, the Stanford Core GNP tool is excellent because it provides the basic forms of words, their parts of speech, indicates which names refer to the same entities, indicates feeling, extracts special or open class relationships between entity mentions, etc. fig. 1 shows Standford CoreNLP interface. A tool can be run with only two lines of code and can be used for pos tagging, recognizing names, numeric and temporal entities, generating lemmas and relationships between entities using relationship annotators [11].

In the next one, the purpose of sentence relation extraction is to extract facts relevant to the case from the verdicts. In order to extract a largely scalar relationship, the authors used the existing knowledge base to heuristically line up with the texts. In their research using remotely controlled

data, they encountered a problem: sentences in a remotely controlled dataset are not marked directly and not all sentences mentioning a pair of entities may represent a relationship between them. To solve this problem, they proposed a new model with enhanced learning. Model is trying to extract relations from every single sentence while the DSRE models aim at extracting relation of an entity pair from all sentences that mention these two entities (the bag). They dealt with two types of experiments on a publicly published dataset. As a result, the method significantly exceeded the comparative baseline, leading to a 13.36% improvement [12, 13, 14, 15].

As an example of report extraction, the authors presented several suggestions and their report. They supplemented the expected reports with sentences to predict the scholarship report, which, compared to the Golden Scholarship report, was used to determine long-term compensation, and then used it to train the report writer. In our model, we need to integrate the expected sentencing reports into the sack report so that we can compare them with the gold sack report to determine the long-term compensation. They followed the last-minute hypothesis to predict the bag report.

In Shizhu He, Kang Liu, Kang Liu, Jun Zhao, Xiangrong Zeng' novel model, first extracted the report of each sentence independently, then predicted the scholarship report based on the extracted reports and compared it with the gold scholarship report. Finally, they used the result of the comparison to guide the formation of the report extractor. For the report extractor they used the PCNN to represent the sentences because it is easier to implement and more efficient for the calculation. The process is: in the raw sentence input, first it is divided into tokens. Then, each token splits into dense vectors, which will be used as input for convolutional neural networks. Finally, a multi-layer perceptron with softmax is applied to produce the probability of each relation. To reduce the variance and make the training faster and more stable they used Williams' simple algorithm that follows the statistical gradient. In other words, or a batch of data with N bags, the baseline is calculated as the average of all the advantages in batch. While conducting the experiment they noted that the bag prediction relies heavily on the relationship extractor, therefore, the results of the evaluation in a remote supervised relationship extraction task can demonstrate the effectiveness of the model [16, 17].

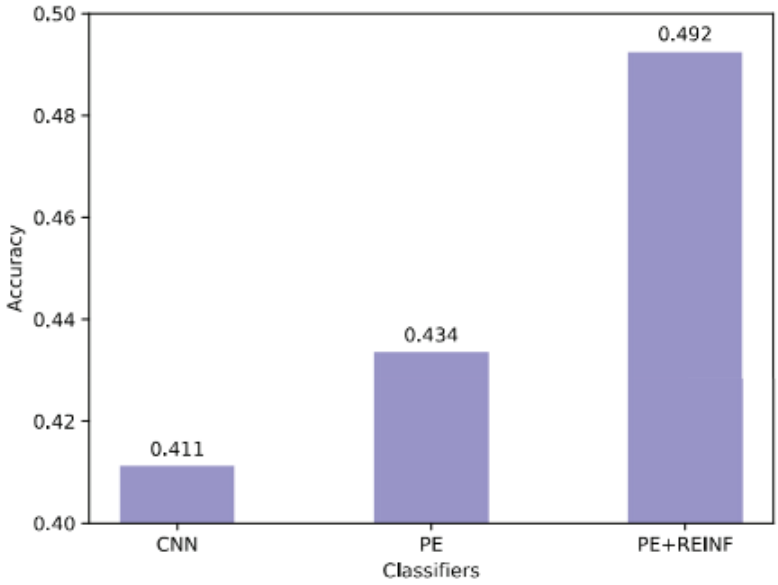

Figure 2 – The accuracy of relation extractor based on CNN, PE and PE+REINF

In fig. 2 one can see, that there is no doubt that their model leans a better relation extractor. As a result, the model achieved an improvement of 19.71% and 13.36% compared to CNN and PE. The improvement was achieved by PE to PE+REINF, which shows that applying reinforcement learning and using remote supervision to guide training can lead to better results.

In the last paper, Yatian Shen, Xuanjing Huang propose a new convolutional neural network model based on attention that makes full use of word embedding, language part tag embedding

and position information embedding. Their word-level attention mechanism is better able to determine which parts of the sentence are more influential than the two entities of interest. The architecture makes it possible to learn some important features from task-specific tagged data, while renouncing the need for external knowledge such as explicit dependency structures. The experiments covered the reference data set of SemEval-2010 Task 8, which showed that the new model performs better than several state-of-the-art neural network models [18, 19, 20].

The authors had a hypothesis if the relevance of words with respect to the target entities is effectively captured, if critical words that determine semantic information can be found. Therefore, they proposed to introduce the mechanism of attention in a neural convolution network (CNN) to extract the words important for the meaning of the sentence and aggregate the representation of those informative words to form a sentence vector. Their process is as follows: given a set of x1, x2, ...xn sentences and two corresponding entities, the model measures the probability of each r relation. In their architecture the extraction of the characteristics is the main component, which is composed of the convolution of the sentences and the selection of the context based on attention. After the feature extraction two types of vectors are generated - the vector of the sentence convolution and the attention-based context vector - for the classification of semantic relations. To obtain the conditional probability they applied a softmax operation on all relation types [21, 22].

Table 1. Score obtained for various sets of features on the test set.

Feature Sets	F1, %
WF	74.5
+pF	80.7
+POSF	82.6
+WA	84.3
+WA+(Lexical Feature)	85.9

Authors performed ablation tests on the four sets of features in tab. 1 to determine which type of features contributed the most. From the results one can observe that their learned position embedding features are effective for relation classification. The F1-score is improved remarkably when position embedding features are added. To evaluate the effectiveness of automatically learned features, authors have choosen six approaches as competitors to their method. All models have adopted the word embedding as a representation, except for SVM. Their network model mainly contains four sets of functions: Word Embedding (WF), Position Embeddings (pF), Part-of-speech tag Embeddings (POSF) and Word Attention (WA). Investments in POS increased F1 by 1.9%, the system achieved an improvement of about 2.3% with the addition of Word Attention. When all the features were combined, they achieved the best result of 85.9%. The bottom portion of the table shows the best combination of all the features.

Conclusion

In conclusion, there is no doubt that natural language processing is an area that covers various issues such as speech recognition, natural language understanding and natural language generation. NLP technologies help us to improve our communication, achieve our goals and get results from every interaction. They also help us to overcome personal obstacles and psychological problems. By studying NLP methods correctly, we can achieve our goals in a very satisfactory way and overcome the obstacles we face.

In the first article, the result generated by bootstrapping algorithm are tokens and enhanced relations between them using Basilisk algorithm which provides gold summary which has high confidence and accuracy in order to improve the results of tagged patterns which would further help in other application areas like biomedical natural language processing, maintaining clinical inventories, providing speech aid to challenged children and machine translation where we convert semantic features of one language to another language.

In the second scientific work, authors' novel method with reinforcement learning results that model outperforms comparative baselines significantly. There are many directions of future work.

Most neural models in relation extraction task are based on convolution neural network and utilize position embeddings as the feature.

In the last paper, authors' an attention-based convolutional neural network architecture for semantic relation extraction model made full use of word embedding, part-of-speech tag embedding and position embedding information. Meanwhile, their word level attention mechanism is able to better determine which parts of the sentence are most influential with respect to the two entities of interest.

Reference:

[1] Collobert R., Weston J., Bottou L., Karlen M., Kavukcuoglu K., and Kuksa P. Natural language processing (almost) from scratch. The Journal of Machine Learning Research, 2011. 12. 2493–2537.

[2] Narasimhan, K., Yala, A., and Barzilay, R. Improving information extraction by acquiring external evidence with reinforcement learning. In Proceedings of EMNLP, (2016). 2355– 2365.

[3] Hoffmann R., Zhang C., Ling X., Zettlemoyer L., and Weld D. S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, (2011). 1. 541–550.

[4] Agichtein, E. and Gravano L.. Snowball: Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries (DL‟00), San Antonio, (2000), Jun., TX.

[5] Agichtein, E., Eskin E., and Gravano L. Combining strategies for extracting relations from text collections. In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD 2000), 2000.

[6] Agichtein, E. and Gravano L. Snowball: Extracting relations from large plain-text collections. In Proceedings of the 5th ACM International Conference on Digital Libraries (DL‟00), San Antonio, Jun., (2000), TX.

[7] Thelen M. and Riloff E. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics, (2002). 10. 214–221.

[8] Kaur, S., and Agarwal, R. A Detailed Analysis of Core NLP for Information Extraction. International Journal of Machine Learning and Networked Collaborative Engineering, 2018. 1(01). 33-47.

[9] Lin W., Yangarber R., and Grishman R. Bootstrapped learning of semantic classes from positive and negative examples. In Proceedings of ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data, (2003). 4(4).

[10] Manning C. D., Surdeanu M., Bauer J., Finkel J., Bethard S. J., and McClosky D. The stanford corenlp natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, (2014). 55–60.

[11] Brin S. Extracting patterns and relations from the World Wide Web. In The World Wide Weband Databases, Berlin: Springer Berlin Heidelberg, (1999). 172–183.

[12] Hoffmann R., Zhang C., Ling X., Zettlemoyer L., and Weld D. S. Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of ACL, (2011). 541–550.

[13] Jiang, X., Wang Q., Li, P., and Wang B. Relation Extraction with Multi-instance Multi-label Convolutional Neural Networks. In Proceedings of COLING, (2016). 1471–1480.

[14] Zeng X., He Sh., Liu K., and Zhao J. Large scaled relation extraction with rein-forcement learning. In Proceedings of AAAI. (2018).

[15] Turian J., Ratinov L., and 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, Association for Computational Linguistics, (2010). 384–394.

[16] Wang L., Cao Zh., Melo G. d., and Liu Zh. Relation classification via multi-level attention

CNNs. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, (2016).

[17] Shen Y. and Huang X. Attention-based convolutional neural network for semantic relation extraction. In COLING, 2016. 2526–2536.

[18] Yin W., Schutze H., Xiang B., and Zhou B. Attention-based convolutional neural network for modeling sentence pairs, 2015.

[19] Ji G., Liu, K., He S., and Zhao J. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In Proceedings of AAAI, (2017). 3060–3066.

[20] Li J., Monroe W., Ritter A., Galley M., Gao J., and Jurafsky D. Deep Reinforcement Learning for Dialogue Generation. In Proceedings of EMNLP, (2016). 1192–1202.

[21] Zeng D., Liu K., Chen Y., and Zhao J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In Proceedings of EMNLP, (2015). 1753–1762.

[22] Nguyen B. and Sameer B. A review of relation extraction. Literature review for Language and Statistics II, 2007.