МРНТИ 28.23.37
УДК 004.023

https://doi.org/10.51889/2021-3.1728-7901.11

SCIENTIFIC NAMED ENTITY RECOGNITION WITH THE HELP OF MODERN METHODS

Yelenov A.M. [1,3], *Jaxylykova A.B.*[1,2*]

[1]*Institute of Information and Computing Technologies, Almaty, Kazakhstan*
[2]*Al-Farabi Kazakh National University, Almaty, Kazakhstan*
[3]*Kazakh-British Technical University, Almaty, Kazakhstan*
**email: aselya17.89@mail.ru*

Abstract

This research focuses on a comparative study of the Named Entity Recognition task for scientific article texts. Natural language processing could be considered as one of the cornerstones in the machine learning area which devotes its attention to the problems connected with the understanding of different natural languages and linguistic analysis. It was already shown that current deep learning techniques have a good performance and accuracy in such areas as image recognition, pattern recognition, computer vision, that could mean that such technology probably would be successful in the neuro-linguistic programming area too and lead to a dramatic increase on the research interest on this topic. For a very long time, quite trivial algorithms have been used in this area, such as support vector machines or various types of regression, basic encoding on text data was also used, which did not provide high results. The following dataset was used to process the experiment models: Dataset Scientific Entity Relation Core. The algorithms used were Long short-term memory, Random Forest Classifier with Conditional Random Fields, and Named-entity recognition with Bidirectional Encoder Representations from Transformers. In the findings, the metrics scores of all models were compared to each other to make a comparison. This research is devoted to the processing of scientific articles, concerning the machine learning area, because the subject is not investigated on enough properly level.The consideration of this task can help machines to understand natural languages better, so that they can solve other neuro-linguistic programming tasks better, enhancing scores in common sense.

Keywords: scientometrics, Bidirectional Encoder Representations from Transformers, transformers, Named-entity recognition, Neuro-linguistic programming, Random Forest Classifier.

Аннотация

А.М. Еленов[1,3], *А.Б. Джаксылыкова*[1,2]*
[1]*Институт информационных и вычислительных технологий, г.Алматы, Казахстан*
[2]*Казахский национальный университет им. аль-Фараби, г.Алматы, Казахстан*
[3]*Казахстанско-Британский технический университет, г.Алматы, Казахстан*

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В НАУКЕ ПРИ ПОМОЩИ СОВРЕМЕННЫХ МЕТОДОВ

Данное исследование посвящено сравнительному изучению задачи распознавания именованных сущностей для текстов научных статей. Обработка естественного языка может рассматриваться как один из краеугольных камней в области машинного обучения, которая уделяет внимание проблемам, связанным с пониманием различных естественных языков и лингвистическим анализом. Уже было показано, что современные методы глубокого обучения обладают хорошей производительностью и точностью в таких областях, как распознавание изображений, распознавание образов, компьютерное зрение и так далее. Что может означать, что такая технология, вероятно, будет успешной и в области нейро-лингвистического программирования и приведет к резкое увеличение исследовательского интереса к этой теме. В течение очень долгого времени в этой области использовались довольно тривиальные алгоритмы, такие как поддержка векторных машин или различные типы регрессии, также использовалось базовое кодирование текстовых данных, что не давало высоких результатов.

Для обработки экспериментальных моделей использовался следующий набор данных: Набор данных ядро связи с научными объектами. Используемые алгоритмы: Долгая краткосрочная память, Классификатор случайного леса с условными случайными полями и распознавание именнованных сущностей с двунаправленным отображением энкодера из транмформеров. В выводах оценки показателей всех моделей сравнивались друг с другом для сравнения. Исследование посвящено обработке научных статей, в области машинного обучения, поскольку данная тема не исследована на достаточном уровне. Рассмотрение этой задачи может помочь машинам лучше понимать естественные языки, чтобы они могли лучше решать другие задачи нейро-лингвистического программирования, повышая оценки в здравом смысле.

Ключевые слова: наукометрия, двунаправленные отображения энкодера из транмформеров, преобразователи, распознование именнованных сущностей, нейро-лингвистическое программирование, классификатор случайных лесов.

Аңдатпа
*А.М. Еленов[1] , А.Б. Жақсылықова[2]**
[1]Ақпараттық және есептеуіш технологиялар институт, Алматы қ., Қазақстан
[1]әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы қ., Қазақстан
[3]Қазақстан-Британ техникалық университеті, Алматы қ., Қазақстан

ҚАЗІРГІ ЗАМАНҒЫ ӘДІСТЕРДІ ҚОЛДАНУ АРҚЫЛЫ ҒЫЛЫМДАҒЫ АТАЛҒАН ЗАТТЫН МАҒЫНАЛАРЫН ТАНУ

Бұл зерттеу ғылыми мақалалар мәтіндері үшін аты аталған тұлғаны тану тапсырмасын салыстырмалы зерттеуге бағытталған. Табиғи тілді өңдеуді әр түрлі табиғи тілдерді түсінуге және лингвистикалық анализге байланысты мәселелерге назар аударатын машиналық оқыту аймағының негізінің бірі ретінде қарастыруға болады. Қазіргі заманғы терең оқыту әдістері суретті тану, үлгіні тану, компьютерлік көру және тағы басқа

Да салаларда жақсы өнімділік пен дәлдікке ие екендігі көрсетілген, бұл мұндай технология нейро-лингвистикалық бағдарламалау саласында да сәтті болуы мүмкін және осы тақырыпқа деген қызығушылықтың күрт артуына әкелуі мүмкін. Ұзақ уақыт бойы бұл салада өте маңызды емес алгоритмдер қолданылды, мысалы, векторлық машиналар немесе регрессияның әртүрлі түрлері, сонымен қатар жоғары нәтиже бермейтін мәтіндік деректердің негізгі кодтауы қолданылды. Эксперименттік модельдерді өңдеу үшін келесі мәліметтер жиынтығы қолданылды: ғылыми объектілермен байланыстың өзегі болып табылатын деректер базасы. Келесі алгоритмдер қолданылды: ұзақ қысқа мерзімді жады, шартты кездейсоқ өрістері бар кездейсоқ орман классификаторы және трансформерлерден энкодерді екі бағытты бейнелеу мен аталған нысандарды тану кездейсоқ орман классификаторы. Алынған нәтижелерде салыстыру үшін барлық модельдердің көрсеткіштерін бағалау бір-бірімен салыстырылды. Бұл зерттеу машиналық оқыту саласына қатысты ғылыми мақалаларды өңдеуге арналған, өйткені бұл тақырып тиісті деңгейде жеткілікті зерттелмеген. Бұл мәселені қарастыру машиналарға табиғи тілдерді жақсы түсінуге көмектеседі, осылайша олар басқа нейро-лингвистикалық бағдарламалау мәселелерін жақсы шеше алады, жалпы бағалауды жоғарылатады.

Түйін сөздер: ғылымиметрия, трансформерлерден энкодерді екі бағытты бейнелеу, трансформерлер, аталған нысандарды тану, нейро-лингвистикалық бағдарламалау, кездейсоқ орман классификаторы.

Introduction

Purpose: The general purpose of this research is to analyze the data in scientific papers in the area of machine learning alongside processing the information from those papers to get a better understanding of natural languages and specifics in this particular area for neural networks. In general, this type of approach could lead us towards improvement and development of the problem-solving capabilities of existing algorithms, thus providing ground for solving other problems that exist in the NLP area and improving their performance.

Relevance: Natural language processing could be considered as one of the cornerstones in the machine learning area which devotes its attention to the problems connected with the understanding of different natural languages and linguistic analysis. It was already shown that current deep learning techniques have a good performance and accuracy in such areas as image recognition, pattern recognition, computer vision, etc., that could mean that such technology probably would be successful in the NLP area too and lead to a dramatic increase on the research interest on this topic. For a very long time, quite trivial algorithms have been used in this area, such as support vector machines or various types of regression, basic encoding on text data was also used, which did not provide high results [1]. Better techniques would be useful for a variety of practical applications, ranging from parsing to question answering tasks or sentiment analysis. In a nutshell, text processing techniques are divided into two categories: statistical and accurate. The accuracy-related approach entails information extraction, which is made up of Part-of-Speech Tagging, Text Segmentation, and Named Entity Recognition. The second technique is based on the estimation of statistics of words in contexts. For example, frequency of occurrence, reverse frequency, word lengths, and what-not [2]. The mentioned rapid growth in the deep learning area created a huge boost to NLP which could be categorized into several parts, the first of them being speech recognition where the main goal is to convert speech to text; next part is devoted to understanding the meaning of the given texts; final part being the text generation, where the neural network could generate different texts on a given topic or a certain keyword. Moreover, you could classify different approaches in NLP, namely syntactic and semantic, where the first approach is about the different construction rules, grammar, and general structure of a sentence, whilst the second approach is about meaning, the general gist of the text.

Significance of the study: Natural language processing task solutions are divided into three categories: first, rule-based, which are thought to be the early ones. However, this kind is still utilized since their proof is reliable. Pattern-matching and parsing, as well as embeddings, are their main major appeals. Typically, these

methods exhibit good performance in restricted tasks; nevertheless, it is impacted by deterioration throughout the generalization phase. As a result, they have a poor degree of accuracy while having a high level of recall metrics. It employs models such as the Linear Classifier, Probability Models, and Likelihood Maximization [3]. These solutions are distinguished by training data using markups, feature engineering, and fitting the model to test data; third, neural networks are analogous to the preceding kind. The difference is that they only learn key characteristics and on raw data, which is original data converted into vector form. RNNs and CNNs are the most frequent examples.

Knowing that people dislike it when someone or something strikes them or that the road might be extremely wet after a rainstorm helps individuals avoid numerous issues in everyday life. This is known as commonsense knowledge, and it is a benefit that individuals have. For many years, however, supplying and teaching this capacity and skill machines was an unfathomable goal. Despite this, for the time being, common sense thinking has become one of the most important jobs, as developers have begun to pay more attention to it. There has recently been a plethora of divergent methods in the field of automation of common sense understanding. One of the most recent efforts was linked to extraction strategies, such that individuals now have enormous graphs of reasoning. Furthermore, many papers were completed that included assimilation to this topic, so that systems such as smart dialogues and agents for answering questions are now more advanced and intelligent. In recent years, developments such as preliminary language training models have led to tools for performing the understanding of human skills. This evolution has to raise the question of whether they must grasp and directly model a good sense from our writings, which are provided in a symbol style. On the other hand, this invention did not give a strong undeniable guarantee that this type of model may create difficult interpretations or that machines can analyze the sophisticated superficial interplay of standards in a simple method. As a result, individuals were faced with the challenge of assessing the correctness and progress of a good sense interpretation. Knowing that people dislike it when someone or something strikes them or that the road might be extremely wet after a rainstorm helps individuals avoid numerous issues in everyday life. This is known as commonsense knowledge, and it is a benefit that individuals have. However, giving and teaching these competence and skill machines was an unfathomable goal for many years. Despite this, common sense thinking has recently become one of the most important jobs, as developers have begun to pay greater attention to it. Nowadays, there have been a plethora of divergent methods in the field of automation of common-sense comprehension. One of the most recent efforts was linked to extraction strategies, such that individuals now have enormous graphs of reasoning. Furthermore, many papers were completed that included assimilation to this issue, so that such systems as smart dialogues and agents for answering questions are now more advanced and intelligent. In recent years, developments such as preliminary language training models have led to tools for performing the understanding of human skills. This evolution has to raise the question of whether they must grasp and directly model a good sense from our writings, which are provided in a symbol style. Contrary, this invention did not give a strong undeniable guarantee that this type of model may create difficult interpretations or that machines can analyze the sophisticated superficial interplay of standards in a simple method. As a result, individuals were faced with the challenge of assessing the correctness and progress of a good sense interpretation. As a result, the necessity for the development of trustworthy texts for testing the machine's skills for modeling good judgment in diverse settings has arisen in activities such as public communications and real-world situations [4]. At the time, there are twelve benchmarks engaged in sound judgment initiatives. The first article represents the strategy for the question-answering systems. The authors suggested a methodology that, in the first phase, transforms each query dedicated to the train so that it has a basic text view. The encoding approach, which may also be utilized in other formats, is employed to carry out the transformation. The collection created as a consequence of this conversion comprises all potential training materials. Following that, the collection of objects with their probabilities is divided into batches. As a result, each batch has the same number of items for each group of the train, regardless of its size [5]. The following method is based on the complicated interplay of two models: the masked language algorithm and the semantic correspondence algorithm. They both contribute to the encoder used in BERT, but the difference is that the entering and outgoing layers are distinct. The authors suggested a solution to the Winograd architectural problem. There is a sequence on the entering layer that comprises a pronoun and a preceding candidate. As a result, the phrase passes through both of the aforementioned levels at the same time. The output of both of them to the encoder, which transforms each word into vectors of embeddings format. The vectors are routed through the same two masking and semantic layers. As a result, there are two data outputs. The first describes the candidate's resemblance to the selected pronoun, while the second illustrates their

connection [6]. Another algorithm is based on transformers as well. The approach's distinguishing characteristic is that the developers offer two types of vectors for the word I in the sequence. The first denotes its meaning, while the second denotes its conditional location, which is reliant on another token j. The authors next present the formula for estimating the value of attention, which falls between these two tokens I and j. As a result, they show the weight of attention between two words as the total of four attention markers in the form of matrices, which take into account content subject-to-subject, subject-to-location, location-to-subject, and location-to-location. This is regarded as a model feature because, in existing techniques, individual matrices of embeddings are used to calculate the bias of the relational location in the process of computing weights of consideration, which is the same as the use of subject-to-subject and subject-to location. The authors argue that all four types of weights are relevant, although competing models only utilize two of them.

The named entity is a physical thing having identifying names, such as persons, geographical locations, businesses, and products. NER aims to classify and assign these name entities in any type of unstructured text to preset categories of entities, which is linked to the information extraction job. The extraction can be useful for a variety of activities, some of which are described below: - Because NER can read messages and provide suitable responses and suggestions, it is essential for developments such as chatbots. This sort of recognition is beneficial for text classification. For instance, consider the development of a news article hierarchy utilizing elements such as entertainment, politics, and sports. - Furthermore, NER enhances text processing semantic methods. Semantic text search engines can be more efficient and precise if they become acquainted with a larger corpus, including its concepts and meaning. - NER extraction is utilized in bigger tasks such as machine translation and question answering [7]. The NER-based method presented by members of the University of Quebec is one example of how this sort of recognition might be useful for modeling good judgment for machines. They offer a method that incorporates common sense information into the named entities, therefore improving NER's capabilities and performance. The effort is connected to the previous solution, which involved the generation of embeddings in the strong collection format. It shows the distributive embeddings of tokens from contexts and ConceptNet research. The collection of embeddings exhibits disparate areas of knowledge and interdependence power. The authors employ architecture, which consists of two major levels. They are Bidirectional Long Short-Term Memory layers and Conditional Random Field layers, which are strengthened by features such as preliminarily learned token embedding layers and dropout layers. Another aspect of the work is the use of representation in the form of characters as well as representation in the form of words. The investigation's major objective is to compare embeddings of words, giving relative information, and the correlation of them with embeddings of words. The investigation's major objective is to compare embeddings of words that offer information in a relative format and their relationship with embeddings of words that present knowledge in a distributive format.

Materials and methodology

There are several datasets linked to NER in various areas: Wikipedia articles, news stories, biomedicine, media, and so on. The dataset utilized in this study is unique in that it comprises 500 abstracts from 12 AI conferences and workshops. SciERC is a new dataset for SemEval 2017 Task 10 and SemEval 2018 Task 7. It adds more coreference connections for cross-sentence linkages, as well as other sorts of entities and relations. Annotations are classified into seven types: Material, Scientific Terms, Method, Task, Metric, and Generic. The developers built a single framework SCIIE to make entity determination and categorization.

The Google Team proposed the following method. It is classified into many kinds based on the size of the vocabulary. BERT, in general, comprises two tasks: masking and next-sentence prediction (NSP). The method in the first task is that the developers attach masks for a specified proportion of tokens and then forecast them. It is required to create a pretrain of the model in such a way that it may examine both sides left to right and right to left. It aids the model in making predictions by considering prior and subsequent data. The authors masked 15% of each sequence randomly throughout their studies. When the i– th word is picked, they replace it with the i– th token appended with the [MASK] token; this happens in 80% in all situations, another 10% for accidental tokens, and the remaining 10% for tokens that remain unmodified. Following that, Ti is used to forecast the true token using the loss of cross-entropy. The drawback of this technique is that the mismatch is generated with pre-trained and fine-tuned information because there is no occurrence of the [MASK] during the second step. To circumvent this problem, the authors do not usually use masking [8].

Another recommended technique, Ours: cross-sentence, was tested on the SciERC dataset. The authors stated the issue in such a way that there is a phrase X composed of n tokens x1, x2,..., xn. Assuming that S = s1, s2,..., sn represents all potential spans that are involved in X based on its length L. Extra markers are indicating the beginning and conclusion of each sequence si, which are STARTi and ENDi. Furthermore, there is an option for establishing a more wonderful depiction by combining the lay across phrases in the text [9].

The next method is being developed to combine many ways into one called SCIIE. The authors attempted to identify and classify scientific items, as well as to establish relationships and resolve coreference across sequences. The model benefits from span products that are placed in a specific context, such as the classifier's characteristics. The activities of sequence level can be quite beneficial by transmitting presentations of span. Specifically, they obtain it from data of coreference across sequence resolution, but without increasing the complexity of the interference [10].

Results and their discussion

As stated at the outset of this study, the primary aim is to conduct research, analyze, and compare the data collected. The models and their scores achieved on the SciERC dataset are shown in Table 2.

Table 2 - The results of training on the SciERC dataset

	Name of the model	*F1*
Exist	*Ours: Cross-sentence*	*71.32*
	BERT Base	*70.11*
	SCIIE	*69.51*
Experimental	*LSTM*	*79.54*
	RANDOM FOREST CLASSIFIER	*73.46*
	RF + CONDITIONAL RANDOM FIELDS CLASSIFIER	*81.06*

The algorithms used to handle the dataset in this study are not the same. They all utilized different approaches. To create a comparison, BERT and Bi - LSTM employ a similar approach of searching for prior and subsequent data to forecast the tag of the present one [11]. However, the characteristics of these approaches vary. The first model employs it during the masking process, whereas the second employs it during the cell state. However, the results vary in that LSTM performed better, suggesting that BERT requires fine-tuning for use on this dataset. The BERT, on the other hand, is utilized in the first model from the table to create a particular representation for the following layer. As a result, BERT and Ours: cross-sentence rely on transformers. Despite this, the outcomes of forecasts varied slightly. BERT had a higher prediction score. Except for RF + CRF and LSTM, RFC had a greater level of score than the other models. Despite this, the results might be regarded as good in comparison to the models studied in this work. RFC employs an algorithm that is distinct from all others. It made use of decision tree ideas. However, the identical one was used in the previous model. Consequently, such a combination received the highest score among the models in the thesis. It denotes the use of decision trees in the development of a prediction process based on the modeling of relationships between goal and input variables. As can be seen, the final model, which is a mixture of two techniques, Random Forest, and Conditional Random Fields Classifier, produces the best results. While the single RF model did not fare well in comparison to the prior one. However, the Long-Short Term Memory network performed well, scoring better than the RFC network.

Conclusion

The research in the field of natural language processing for the job of sound judgment reasoning was conducted in this work. Furthermore, a diverse range of models and their performance in the aforementioned job were investigated. The dataset for this work, which consisted of scientific article texts, was discovered. It is made up of 500 abstracts from 12 different AI conferences. Data tokenization and tagging were applied. On this data, several models were trained and evaluated. More specifically, the three of them, which include LSTM, Random Forest, Random Forest with Conditional Random Fields Classifier, and Random Forest with Conditional Random Fields Classifier. The metric scores were achieved. The outcomes were contrasted and examined.

Acknowledgment

We gratefully acknowledge financial support of Institute of Information and Computational Technologies. (Grant AP09260670, Kazakhstan).

References

1. *Luan, Yi, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. (2018) Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1360.*
2. *C, Olah. (2015) Understanding Lstm Networks, 1–19.*
3. *Atanassova I., Bertin M., Philipp M, (2019), Mining Scientific Papers: Nlp-Enhanced Bibliometrics. Frontiers Research Topics. https://doi.org/10.3389/978-2-88945-964-3.*
4. *Maarten S., Rashkin H., Chen D., Le Bras R., and Choi Y. (2019) Social Iqa: Commonsense Reasoning about Social Interactions. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1454.*
5. *Khashab, Daniel, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi (2020). UNIFIEDQA: Crossing Format Boundaries with a SINGLE QA System. Findings of the Association for Computational Linguistics: EMNLP 2020, https://doi.org/10.18653/v1/2020.findings-emnlp.171.*
6. *Pilault, Jonathan, Amine El hattami, and Christopher Pal. (2021) Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. ICLR 2021.*
7. *Kurama, Vihar (2021). Named Entity RECOGNITION: NLP with NLTK & SPACY. AI Machine Learning Blog. AI & Machine Learning Blog.*
8. *Hong, Zhi, Roselyne Tchoua, Kyle Chard, and Ian Foster. (2020) Sciner: Extracting Named Entities from Scientific Literature. Lecture Notes in Computer Science, 308–21. https://doi.org/10.1007/978-3-030-50417-5_23.*
9. *Färber, Michael, Alexander Albers, and Felix Schüber. (2021) Identifying Used Methods and Datasets in Scientific Publications. SDU@ AAAI.*
10. *Beltagy, Iz, Kyle Lo, and Arman Cohan. (2019) SciBERT: A Pretrained Language Model for Scientific Text. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). https://doi.org/10.18653/v1/d19-1371.*
11. *Brownlee, Jason. (2021) A Gentle Introduction to Long Short-Term Memory Networks by the Experts. Machine Learning Mastery.*