APPLICATION OF RUSSIAN NAMED ENTITY RECOGNITION AND COREFERENCE RESOLUTION IN THE OIL INDUSTRY

A.D. Kulnevich [1, a], V.L. Radishevskii[1], R.A. Chugunov [1], A.A. Shevchuk [2]

[1] *National Research Tomsk Polytechnic University, 30 Lenina avenue, Tomsk, 634050, Russia*

[2] *National Research Tomsk State University, 36 Lenina avenue, Tomsk, 634050, Russia*

E-mail: [a] kulnevich94@mail.ru

This paper describes the application of named entity recognition and coreference resolution algorithms in the oil industry. Oil industry researchers and businesses generate large amounts of content every day. Managing them correctly is very important to get the most use of each article and document. Named entity recognition algorithms can automatically scan entire articles and reveal the most significant people, organizations, and places discussed in them, while coreference resolution combines each entity mention into clusters of mentions. Each cluster represents one entity across one document. These methods allow to simplify the analysis of large numbers of documents and articles for researchers, managers, engineers, etc.

Keywords: natural language processing, named entity recognition. coreference resolution.

1. Introduction

Named entity recognition is a process where an algorithm takes a string of text (sentence or paragraph) as input and identifies relevant nouns (people, places, and organizations) that are mentioned in that string.

In practice, texts often have the same entities mentioned in various ways (anaphora, cataphora, split antecedents, coreferring noun phrases). Coreference resolution algorithms are used to address this problem and to combine all mentions of the same entity into one cluster.

These algorithms may greatly simplify the analysis of documents and articles. For example:

- They can be used to create efficient search engines. If for every search query the algorithm ends up searching all the words in millions of articles, the process will take a lot of time. Instead, if named entity recognition can be run once on all the articles and the relevant entities (tags) associated with each of those articles are stored separately, this could speed up the search process considerably. With this approach, a search term will be matched with only a small list of entities discussed in each article, leading to faster search execution.
- They can be used to improve content recommendation systems. This can be done by extracting entities from a document and recommending other documents that have the most similar entities mentioned in them.
- An online journal or publication site can hold millions of research papers and scholarly articles. There can be hundreds of papers on a single topic with slight modifications. Information search can become complicated. Segregating articles by tags extracted using named entity recognition and coreference resolution can help find the desired article or document.
- They can be used to create ontology objects and object properties.
- They can be used to classify content for news providers: such algorithm can scan entire articles and reveal the most significant people, organizations, and locations discussed in them.
- There are several ways to make the process of customer feedback handling smooth by means of solving named entity recognition tasks.
- They can be used for automatic summarization systems: named entities are the important information of the text and increase the performance of identification of text segments that are further included in summarized data.
- This is especially important for the oil industry for two reasons:
- new technologies, cited in scientific papers, can save millions of dollars daily after implementation;
- thousands of documents are generated in every oil company every day, these documents often require meticulous analysis.

2. Implementation of Named Entity Recognition

There are two main approaches to address the named entity recognition (NER) problem [6]. The first one is based on handcrafted rules, and the other one relies on statistical learning. The rule-based methods are primarily focused on engineering the grammar and syntactic extraction of patterns related to the structure of the language. In this case, laborious tagging of a large number of examples is not required. The downsides of fixed rules are the poor ability to generalize and the inability to learn from examples. As a result, this type of NER systems is costly to develop and maintain. Learning-based systems automatically extract patterns relevant to the NER task from a training set of examples, so they don't require deep language-specific knowledge. This makes it possible to apply the same NER system to different languages without significant changes in architecture.

In this paper, we use a hybrid approach to this task in the Russian language:

- An algorithm based on context-free grammar is used to extract some of the document's entities, keywords, and attributes.
- Another algorithm based on conditional random fields and word vectorization using a pre-trained skip-gram word2vec model for the Russian language and POS tags.

The concept of conditional random fields (CRFs) [1] has been successfully adapted in many sequence labeling problems [2-3]. Even the in deep learning architecture, CRF has been used as a fundamental element in named entity recognition [4-5]. One of the primary advantages of applying a CRF to language processing is that it learns transition factors between hidden variables corresponding to the label of a single word.

We used a hybrid approach to extract entities from texts: extracted entities were merged together, removing the duplicating ones. Entities of the following types were extracted:

- person;
- organization;
- location;
- product;
- event;
- money.

Named entity extraction algorithms used morphological and part-of-speech tags to correctly label entities.

To train and validate models, we used the Dialogue-2016 dataset and additionally labeled documents (newspapers, fiction books, technical documents).

3. Implementation of Coreference Resolution

The coreference resolution algorithm is based on neural network, which is mostly derived from previous work [7]. Some changes were made to improve the results in the Russian coreference resolution task:

- To train a network for the Russian language, we used the Dialogue-2014 dataset.
- LSTM layers in the network have been changed to GRU layers (GRU showed slightly better results during evaluation on test data due to a smaller number of parameters and small dataset).
- Pre-trained Russian word2vec skip-gram vectors, morphology, and POS tags were used as features.
- An extracted named entity tag was added as a feature to help the network find coreferences between the entities extracted by the NER algorithm.

4. Results

Named entity recognition (Table 1) and coreference resolution (Table 2) modules were tested on a holdout subsample of the dataset (randomly selected 10% of data). Metrics for entities were calculated for every word separately. Classes of entities were unbalanced: most of the words in the texts were not parts of entities

Table 1. Entity recognition results on a holdout subsample

	Precision	**Recall**	**F1-score**	**Support**
B-PER	0.78	0.69	0.73	361
I-PER	0.74	0.74	0.74	425
B-ORG	0.76	0.60	0.67	533
I-ORG	0.77	0.73	0.75	562
B-LOC	0.75	0.86	0.80	651
I-LOC	0.76	0.60	0.67	263
B-PROD	0.63	0.65	0.64	752
I-PROD	0.70	0.47	0.56	227
B-DATE	0.92	0.86	0.89	337
I-DATE	0.91	0.97	0.94	420
O	0.99	0.99	0.99	53780
Avg / Total	0.97	0.97	0.97	58311

Table 2. MUC-5 coreference resolution results on a holdout subsample

Metrics	MUC		
	Prec.	Rec.	F1
Our model	71.7	65.2	0.683

5. Application

Entity recognition and coreference resolution models were combined in a single pipeline, which also included document OCR, text preprocessing, and tokenizing. A web service was created, which included a search system based on the Elasticsearch framework. The extracted entities were used in ranging the search output. The system was loaded with oil industry-related documents: scientific articles and business documents. The documents in the search could be viewed with highlighted entities and coreferences. The agglomerative clustering method (using Doc2Vec model for feature extraction) and a simple named entity linking algorithm based on regular expressions were used to recommend similar documents to help the user quickly find relevant documents that are like the current document.

Examples of our web service GUI and processed text can be seen in the Figures 1 and 2 below:

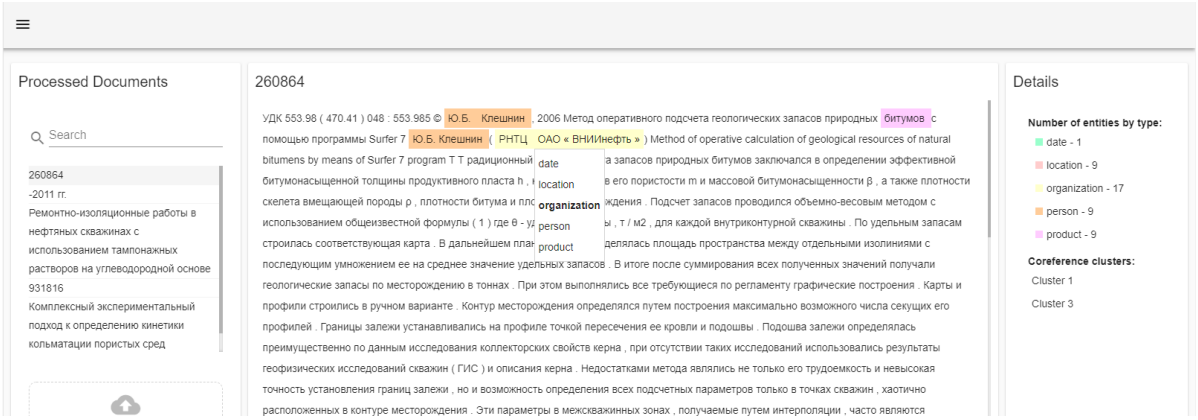

Figure 1. An example of Natural Language Application for Gas and Oil industry

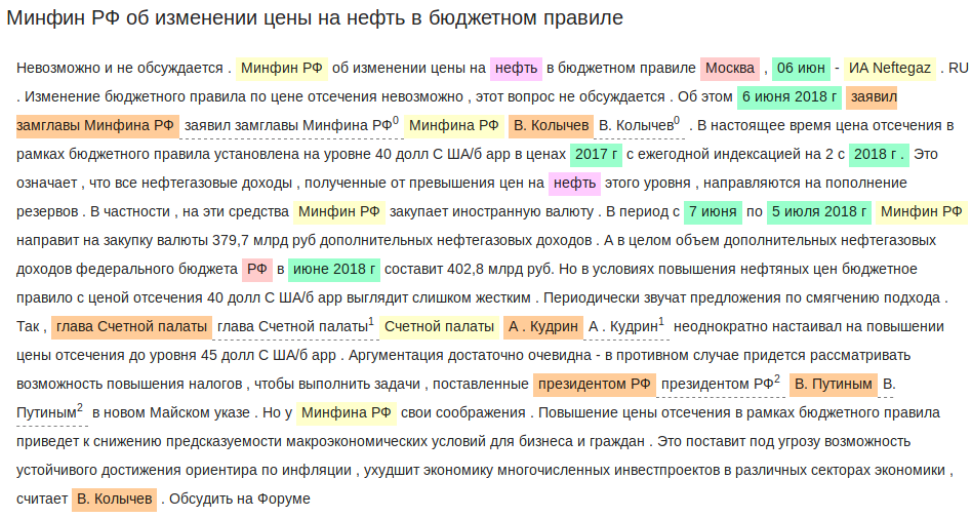

Figure 2. An example of processed text

Altogether, this system considerably improves information search efficiency and document analysis.

6. Conclusion

This article describes application of machine learning algorithms for natural language processing tasks. Named Entity Recognition and Coreference Resolution allows to improve search engines and helps to analyze documents faster. Future work includes optimization of algorithms and addition of summary extraction, Named Entity Linking, recommendation engine based on documents features. All of these features are aimed to optimize the process of text documents analysis.

References

[1] Lafferty J., McCallum A., Pereira F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data // In International Conference on Machine Learning (ICML), 2001, pp. 282–289.

[2] McCallum A. and Li W. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons // Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003, Vol. 4, 2003, pp. 188-191.

[3] Sha F. and Pereira F. Shallow parsing with conditional random fields // Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Vol 1, 2003, pp. 134-141.

[4] Lample G. et al. Neural architectures for named entity recognition // arXiv preprint arXiv:1603.01360, 2016.

[5] Liu Z. et al. Entity recognition from clinical texts via recurrent neural network // BMC Medical Informatics and Decision Making, 2017, Vol. 17, №. 2, p. 67. DOI: 10.1186/s12911-017-0468-7.

[6] Maithilee L. et. al. Approaches to Named Entity Recognition: A Survey // International Journal of Innovative Research in Computer and Communication Engineering, 2015, Vol 3. Issue 12, pp. 12201–12208.

[7] Lee K. et. al. End-to-end Neural Coreference Resolution // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, 2017, pp 188-197.

[8] Anh L. et. al. Application of a Hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition // arXiv preprint arXiv:1709.09686, 2017.