ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 72

IV International Research Conference "Information Technologies in Science, Management, Social Sphere and Medicine"

(ITSMSSM 2017)

Named Entity Recognition in Natural Language Texts obtained through Audio Interfaces

Nail Gafurov
Department of Computation Technologies
Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics
Saint-Petersburg, Russia
nrgafurov@corp.ifmo.ru

Alexey Platonov
Department of Computation Technologies
Saint-Petersburg National Research University of Information Technologies, Mechanics and Optics
Saint-Petersburg, Russia
avplatonov@corp.ifmo.ru

***Abstract*—The article describes the features of named entity recognition from natural language texts in English obtained by audio interfaces and a method for increasing the efficiency of technologies based on machine learning by using a modified data set for training.**

Keywords—named entity recognition; NER; information extraction; facts extraction; natural language processing; NLP; machine learning

I. INTRODUCTION

Increases the number of systems that interact with the user through natural for him methods. One of the most convenient and quite common such methods of interaction is audio interface. The world leaders trends in the development of programming interfaces confirm the relevance of this direction (Siri from Apple, Google app, Alice - voice assistant from Yandex).

Often the shorter instruction is transmitted by voice. But, in the future, in addition to the audio-editors of the text, which should separate the text and editing commands, there are a number of areas for working with large dictated texts. For example, the creation of an IDE (Integrated Development Environment), which will write programs through verbal commands of the programmer.

When processing dictated texts, there are some features: there are not or not enough punctuation marks (quotes, dots, question marks, etc.), there are no uppercase letters to separate proper names. This imposes restrictions on the text processing in many stages. The step of extracting named entities from natural language texts obtained through audio interfaces is the subject of this article.

It is important that the paper does not consider the stage of converting audio information into text.

II. NAMED ENTITY RECOGNITION IN NATURAL LANGUAGE TEXTS

Named entity recognition is the stage that allows to merge these entities into facts in the future. What is the named entity? A named entity is the relationship of some objects and their belonging to a group. For example, "Camomile" - this is the designation of the organization. Vladimir Putney is a designation of some person.

In the paper, the programming language Python 3.4.6, a package of libraries and programs for symbolic and statistical processing of natural language NLTK 3.2.5 (Natural Language Toolkit) using the technology of machine learning, the GBoard application (Google), which allows you to convert audio information into text - a tools for working with natural language texts. In accordance with [1], these tools are one of the most relevant today, the use of machine learning is one of the main approaches to the processing of natural language texts, as shown in [2].

To show the main features of the research object, consider the example. Part of the text from the news article [3] was taken for analysis:

MPs in the Public Accounts Committee criticised HMRC for being "too cautious" in pursuing the "fraudsters".

Amazon and eBay said they were working with HMRC on the issue.

After processing the source text, the obtained result shown in Fig. 1.

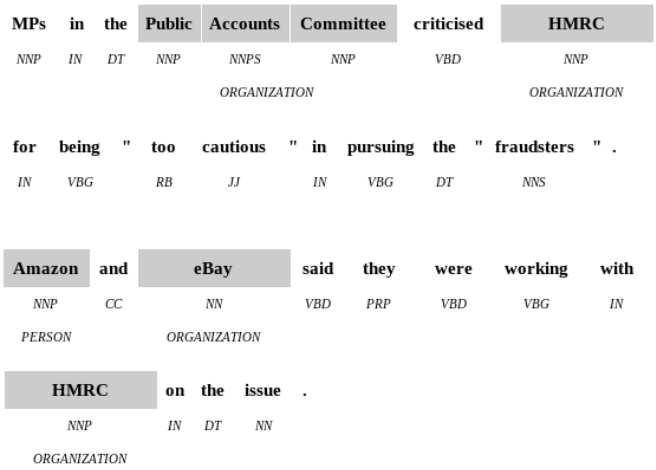

Fig. 1. Original text with marked named entities

Tags:

- CC - Coordinating conjunction;
- DT - Determiner;
- IN - Preposition or subordinating conjunction;
- JJ - Adjective;
- NN - Noun, singular or mass;
- NNP - Proper noun, singular;
- NNPS - Proper noun, plural;
- NNS - Noun, plural;
- PRP - Personal pronoun;
- RB - Adverb;
- VBD - Verb, past tense;
- VBG - Verb, gerund or present participle;
- VBN - Verb, past participle;
- VBP - Verb, non-3rd person singular present.

Common entity types include ORGANIZATION, PERSON, LOCATION, DATE, TIME, MONEY, and GPE (geo-political entity).

Five entities were singled out. The company name Amazon was defined as Person (which is incorrect, but this error is not considered in this paper as not relevant to the main topic).

After dictating this text through the GBoard application, the obtained text is:

Mpsf in the Public Accounts committee criticized hmrc for being too cautious in pursuing the routes stairs Amazon and eBay said they were working with hmrc on the issue

This text is "tagged" and "chanked".

Tagging - morphological marking of the text, the marking of parts of speech and grammatical characteristics of words in the text (corps) with attributing to them the appropriate tags.

The term chunk or chunking is widely used in the theory of learning. A chunk is a generalized term that includes all other terms that denote such combinations of words as collocations, lexical beams [4]. Chunking is the process of identifying and classifying flat, non-overlapping segments of a sentence that constitute the basic non-recursive phrases corresponding to the basic parts of speech found in the majority of grammars of wide coverage [5]. That is, chunking is a method of partial parsing, which consists in splitting the text into syntactically related fragments of the text or chunks (ready-made phrases and parts of sentences or strongly related, indivisible segments).

Tagging and chunking were performed sequentially, in accordance with the general scheme for natural language text processing [6].

The result of processing the text obtained through the audio interface shown in Fig. 2.

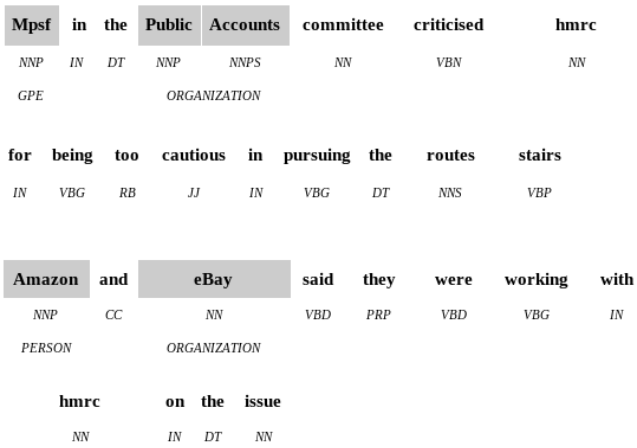

Fig. 2. Text from the audio interface with marked named entities

Four entities were found. One of them is new and the second is not defined.

Since this paper does not consider the stage of converting audio information into text, errors of perception (incorrect word recognition) are not considered. The sample of this error is first word in Fig, 2, which was incorrectly received.

As can be seen, there is a problem with uppercase, punctuation marks was missing. The voice recognition system automatically identified well-known companies names (the ontology approach is probably involved) and this helped in the recognition of entities, but this approach is possible only with well-known companies. Less popular companies can not boast of such a privilege.

III. MACHINE LEARNING IN NAMED ENTITY RECOGNITION

The experiments was carried out on the reference data set CoNLL-2000 [7], in which the tasks of marking and named entities recognition were solved.

In the first experiment, we used the original training data set. Each experiment consists of three tests. A standard test sample from the data set was used as the first test sample. To simulate the situation of obtaining text from the audio interface, two additional modified copies of the specified data set are created: (1) a set with symbols reduced to lowercase for the second test; (2) a set with lowercase and deleted punctuation characters (dots, dots, commas, quotes, question marks, exclamation marks) for the third test.

The results of the accuracy estimation obtained after learning the tagging module and the module of the chunking using the naive Bayesian classifier [8] by training on the original data set (CoNLL-2000) presented in Table 1.

TABLE I. ANALYSIS OF EFFICIENCY AFTER TRAINING ON THE CONLL-2000

Test through	Mark		Result
CoNLL-2000	Tagger accuracy		93,73%
	Chunk Parser score	IOB Accuracy	93,10%
		Precision	88,10%
		Recall	90,80%
		F-Measure	89,40%
corpus without uppercase letters	Tagger accuracy		87,09%
	Chunk Parser score	IOB Accuracy	93,00%
		Precision	87,90%
		Recall	90,60%
		F-Measure	89,20%
corpus without uppercase letters and punctuation marks	Tagger accuracy		82,04%
	Chunk Parser score	IOB Accuracy	89,50%
		Precision	83,60%
		Recall	83,20%
		F-Measure	83,40%

Tagger accuracy - an estimation of accuracy of a marking in comparison with "the gold standard" (definition from [1]). Tags from the sample text are deleted, this text is tagged using a trained system, the obtained results are compared with the original ones and the accuracy is estimated.

Chunk Parser score - evaluation of the accuracy of the chunking module.

IOB (inside, outside, beginning) is the format for labeling tokens in the partitioning task. This standard format is performed by introducing tags to represent the beginning (B) and internal (I) parts of each fragment, and elements that are outside (O) of any chunk [5].

IOB Accuracy is the accuracy of the IOB markup. It indicates how many words are outside the NP groups (proper names).

Precision indicates how many of the elements that we identified are relevant.

Recall indicates how many of the relevant items we identified.

F-Measure is a generalizing estimate, is defined as the average harmonic value between the two previous values: (2 × Precision × Recall) / (Precision + Recall) according [1].

The obtained data show the probability of a significant decrease in the efficiency of named entities recognition from texts obtained through audio interfaces (a test on the corps without uppercase letters and punctuation marks).

A. Increasing efficiency through training by the modified data

To increase efficiency we propose to produce machine learning using a modified data set. First we convert the marked corpus of data to lowercase, we train the system on the received corpus (experiment 2), then punctuation marks are removed from this body and the system is trained again (experiment 3).

The hypothesis is that the system will not take into account the eliminated signs, during training, and learns to named entities recognition without using the register of letters or punctuation marks.

The results of the accuracy estimation obtained after learning the tagging module and the chunker module through corpus of a data set without uppercase letters presented in Table 2.

The obtained results of the accuracy evaluation after the tagging module and the chunker module learning through a set of data without uppercase letters and punctuation marks (dots, dots, commas, quotes, question marks, exclamation marks) presented in Table 3.

The data obtained during the experiments are combined in Table 4.

Tests: (1) a standard data set was processed (CoNLL-2000); (2) the data set without uppercase letters was processed; (3) the data set without uppercase letters and punctuation marks was processed.

Trained modules: (a) modules are trained on a standard data set (CoNLL-2000); (b) the modules are trained on a data set without uppercase letters; (c) the modules are trained on data set without uppercase letters and punctuation marks.

TABLE II. ANALYSIS OF EFFECTIVENESS AFTER TRAINING ON DATA WITH LOWERCASE LETTERS

Test through	Mark		Result
CoNLL-2000	Tagger accuracy		98,27%
	Chunk Parser score	IOB Accuracy	94,20%
		Precision	89,90%
		Recall	92,30%
		F-Measure	91,10%
corpus without uppercase letters	Tagger accuracy		93,38%
	Chunk Parser score	IOB Accuracy	93,90%
		Precision	89,40%
		Recall	91,90%
		F-Measure	90,60%
corpus without uppercase letters and punctuation marks	Tagger accuracy		87,81%
	Chunk Parser score	IOB Accuracy	90,80%
		Precision	85,60%
		Recall	85,00%
		F-Measure	85,30%

TABLE III. ANALYSIS OF EFFECTIVENESS AFTER TRAINING ON DATA WITH LOWERCASE LETTERS AND WITHOUT PUNCTUATION

Test through	Mark		Result
CoNLL-2000	Tagger accuracy		98,27%
	Chunk Parser score	IOB Accuracy	94,20%
		Precision	89,90%
		Recall	92,30%
		F-Measure	91,10%
corpus without uppercase letters	Tagger accuracy		93,38%
	Chunk Parser score	IOB Accuracy	93,90%
		Precision	89,40%
		Recall	91,90%
		F-Measure	90,60%
corpus without uppercase letters and punctuation marks	Tagger accuracy		87,81%
	Chunk Parser score	IOB Accuracy	90,80%
		Precision	85,60%
		Recall	85,00%
		F-Measure	85,30%

TABLE IV. COMPARISON OF EXPERIMENTAL RESULTS

Test	Mark	Trained modules		
		a	b	c
1	Tagger accuracy	93,73%	98,27%	98,27%
	Chunker F-Measure	89,40%	91,10%	91,10%
2	Tagger accuracy	87,09%	93,38%	93,38%
	Chunker F-Measure	89,20%	90,60%	90,60%
3	Tagger accuracy	82,04%	87,81%	87,81%
	Chunker F-Measure	83,40%	85,30%	85,30%

IV. CONCLUSION

The experiments confirmed that the accuracy of the system trained by the proposed method (based on the modified data set) exceeds the accuracy of the system trained on a standard data set widely used for similar tasks.

It is noteworthy that the elimination of punctuation marks in training data did not affect the result. This can be explained by the fact that in English the punctuation marks is not so important as, for example, in the Russian language.

Also the proposed method of training is suitable for processing natural language texts obtained not through the audio interface (marked with punctuation marks and uppercase letters). As can be seen from the results of experiments, the proposed method of learning additionally increased the efficiency of extracting named entities from natural language texts obtained not through audio interfaces.

As a further direction of the research, it is important to check the proposed method in other languages. Especially in languages where punctuation is more important than in English. For example the Russian language.

It will be interesting to compare the results with the methods based on rules, methods based on ontologies, use different neural networks and compare their effectiveness.

As a way to increase the probability of identifying punctuation marks, we consider the possibility of more rigorous consideration of intonation information, pauses in audio interfaces. This will increase the efficiency of the system as a whole. It may be worth adding markup indicating intonational information for analysis.

REFERENCES

[1] Natural Language Toolkit documentation. 2017. http://www.nltk.org Retrieved October 24, 2017.

[2] N. R. Gafurov, "INFORMATION EXTRACTION FROM NATURAL LANGUAGE TEXTS", Collection of Proceedings of the VIII Scientific and Practical Conference of Young Scientists "Computing Systems and Networks (Mayorovsy Readings)", 2017

[3] "Amazon and eBay warned by MPs about VAT fraudsters", BBC, 18 October 2017, URL: http://www.bbc.com/news/business-41658436 (tested: 24.10.2017)

[4] Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. - S.l.: Pearson Education Limited, 1999. 1204 p.

[5] Speech and Language Processing. Daniel Jurafsky & James H. Martin, 2017.

[6] Posvkin R. V., Bessmertny I. A., "Natural language user interface of a dialogue system", Software & Systems, vol. 3, pp. 5-9, August 2016

[7] Erik Tjong Kim Sang. CONLL 2000 CHUNKING DATA. 2010. https://github.com/teropa/nlp/tree/master/resources/corpora/conll2000 Retrieved October 24, 2017.

[8] Christopher D. Manning, Prabhakar Raghavan & Hinrich Schütze, "Introduction to Information Retrieval", Cambridge University Press, 2008