

Development of an algorithm for facts extraction from the texts of biomedical articles

Derevianchenko S.^{1*}, Demenkov P.²

¹ Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia

² Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

* sod97@yandex.ru

Key words: facts extraction, text mining, natural language processing

Nowadays a huge amount of new text information appears every day. This information published by billions of different sources. Today one of the most important tasks of information processing is to extract the important and useful part of it from the entire amount. A popular and useful approach is facts extraction from texts.

A fact in the text is structured information about some objects found in this text and their relationship to each other. Thus, facts extraction from texts is the task of searching for structured information in unstructured text data. Facts extraction it's one of Text Mining tasks. In this work is presented a new algorithm for solving the problem of facts extraction from texts. Main characteristics of algorithm are:

- Unsupervised learning (the algorithm is trained on unlabeled data corpora).
- Easy adaptation to new topics (the quality and efficiency of the algorithm are evaluated on texts from biomedical topics. However, it's assumed that the quality remains the same using the algorithm in a new domain).
- Low requirements for computing resources (the ability to train and inference the algorithm on medium-power computers, for example, conventional PCs).

The main idea of the algorithm is the interaction of Latent Dirichlet allocation model and parsing of sentences' dependency trees. Using of these components allows the algorithm to meet all the requirements stated above.

The presented algorithm is characterized by high quality of facts extraction and a low percentage of errors. Its evaluation metrics on the test dataset:

Type of facts	Percent of extracted facts, %	Percent of sentences with incorrect extracted facts, %
Expression regulation	67.1	4.3
Expression upregulation	73.8	5.4
Upregulation	68.4	9.6
Regulation	60.3	4.2
Transport regulation	65.2	4.5

Presented algorithm preserves its quality in other domains. This is its metrics on the Web news dataset:

Type of facts	Percent of extracted facts, %	Percent of sentences with incorrect extracted facts, %
News	62	1
Sport news	64	2

Acknowledgements: Work was supported by the budget project No. FWNR-2022-0020.