INTELLECTUAL TEXTS PROCESSING IN SOCIO-ECONOMIC APPLICATIONS

S.D. Belov [1,2 a], I.S. Kadochnikov[1,2 b], V.V. Korenkov[1,2 c], P.V.Zrelov[1,2 d]

[1] *Joint Institute for Nuclear Research, 6 Joliot-Curie St, Dubna, Moscow Region, 141980, Russia*

[2] *Plekhanov Russian University of Economics, Stremyanny lane, 36, Moscow, 117997, Russia*

E-mail: [a] belov@jinr.ru, [b] kadivas@jinr.ru, [c] korenkov@jinr.ru, [d] zrelov@jinr.ru

This paper discusses some approaches to the intellectual text analysis in application to automated monitoring of the labour market. The scheme of construction of an analytical system based on Big Data technologies for the labour market is proposed. Were compared the combinations of methods of extracting semantic information about objects and connections between them (for example, from job advertisements) from specialized texts. A system for monitoring of the Russian labour market has been created, and the work is underway to include other countries in the analysis and the system to check foreign counterparty companies using Big Data. The considered approaches and methods can be widely used to extract knowledge from large amounts of texts.

Keywords: text analysis, natural language processing, Big Data, machine learning

Sergey Belov, Ivan Kadochnikov, Vladimir Korenkov, Petr Zrelov

Copyright © 2020 for this paper by its authors. Use
permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-2772/65-71-paper-10.pdf

1. Introduction

Natural language processing is an area of research in computer science and artificial intelligence (AI). Processing usually involves translating natural language into numerical data, through which the computer can obtain information about the world around it. For these purposes, so-called NLP (Natural Language Processing) technologies are being developed. In the 2010s, natural language processing and NLP-based dialog machines (chatbots) became increasingly common. At first, Google search only resembled working with a subject index-a tool that did not require special skills to use. But soon, he became more intelligent and began to understand search queries close to natural language. Then there was auto-completion in smartphones. With the advent of bots like Microsoft's Tau bot, it became clear that NLP bots affect society. Bots began to collect tweets to predict elections' results, and later – to influence these results. Systems for predicting economic trends appeared. Such algorithms started to impact the economy more and more and led to a shift in the public consciousness –people began to use NLP to strengthen the role of "machines" in making their own decisions. Thanks to the rapid flow of unstructured data on politics and Economics, NLP has become an integral tool for political strategists and financiers. Generating an increasing amount of entertainment, advertising, and financial reporting content does not require human involvement. Computer games and virtual worlds contain NLP bots that can communicate with humans. NLP provides a useful information search and uses filtering mechanisms or promotion of individual pages to influence the information consumed by the user. Search is historically the first commercially successful area of an NLP application. Search inspired the increasingly rapid development of NLP algorithms, which refined search technologies based on NLP indexing and prediction techniques.

A search engine can provide more accurate search result if it indexes web pages or document archives in a way that takes the value into account. Autocompletion using NLP has become widespread in search engines and mobile devices. Many word processors, browser plugins, and text editors have built-in spell-checking, grammar, and word matching tools. Some dialog machines (chatbots) use natural language search to find the answer to their interlocutor's message. NLP systems can be used to create short replies in dialog machines, virtual assistants, generate short posts in social networks, and compose longer passages of text. The Associated Press uses NLP-based robot journalists to write entire articles on Finance and sports reports [1]. NLP-based spam filters in early email programs contributed to the fact that in the 1990s, email overtook the telephone and Fax as a communication channel. Chatbots made up approximately 20 % of tweets about the 2016 us presidential election [2, 3]. These bots support points of view that are beneficial to their owners and developers. NLP technologies are used to compile movie reviews and product reviews from stores. The more significant number of studies is a product of the work of Autonomous NLP pipelines that have never been to a movie theater or purchased the product under review.

Other tasks solved with applied NLP systems are machine translation, referencing, annotating and analyzing the text's tonality, categorizing, classifying and clustering the text, building knowledge bases, etc.

2. Classical approach to the text analysis

The most useful opportunities and high quality of text analysis can be obtained by conducting a complete analysis of the text. However, the difficulties that arise when making such a study are such that all the theoretical provisions developed at the moment have not yet been implemented in practice. The main problems here are the complexity of text parsing and the difficulty of creating a full-fledged expert system. For full-fledged work, the text analysis system should be able to analyze the test submitted by the user for input from the point of view of syntax (sentence structure), semantics (concepts used in the text) and pragmatics (the correctness of the use of concepts and the purposes of their use). Next, the system must generate its response in an internal representation suitable for logical inference and synthesize its response in natural language. In General, a system that supports full analysis should contain the following modules. Graphemic analysis - provides the selection of syntactic or structural units from the input text, which can be a linear structure having a single text fragment. Morphological analysis - defines the normal form from which the word form was formed, and a set of parameters attributed to it.

Parsing is the most challenging part of text analysis. Here it is necessary to define the roles of words and their relationships with each other. The result of this step is a set of trees showing such relationships. Semantic analysis performs the analysis of the text "make sense." On the one hand, semantic analysis clarifies connections that post-syntactic analysis could not clarify since many roles are expressed using the means of language and taking into account the meaning of the word. On the other hand, semantic analysis allows you to filter out some word meanings or even whole parsing variants as "semantically disjointed". The semantic analysis stage ends with the study of the input text. If it is necessary to generate a response, such as during a dialogue with the user or when translating documents from a foreign language, the synthesis stages – syntactic, morphological, and graphemic-are- are added to the considered stages. Response generation is, to a different extent, inherent in all types of dialog systems, some types of systems for composing text abstracts, statistical text analysis, and text generation. The response is selected from a specific corpus of texts or generated "on the fly".

3. Semantic text analysis

3.1. Latent semantic analysis

Latent semantic analysis (LSA) [4] maps documents and individual words into the so-called "semantic space", in which all further actions are performed. The following assumptions are made: documents are a set of words, the word order in documents is ignored, only the frequency of occurrence of a word in a document is important; the semantic meaning of a document is determined by a set of words that are usually used together. For example, in stock market reports, the words "Fund", "stock", "dollar" are often found»; each word has a single meaning.

TF-IDF text analysis. TF-IDF analysis is one of the machine learning methods. TF-IDF is a statistical indicator used primarily to evaluate a specific word in the context of the entire document included in the General collection. The term TF / IDF TF literally means the frequency of the term (term frequency), and IDF– the inverted frequency of the document (inverse document frequency). According to the TF / IDF ratio, the weight of a certain word depends directly on how many times it occurs in a particular text. It is inversely dependent on the number of uses of this word in the set of other documents. TF or word frequency is the ratio of the number of occurrences of a particular term to the total set of words in the document under study. This indicator reflects the importance of the word within a specific article/publication. The IDF or inverted document frequency is the inverse of the frequency with which a particular word appears in a document collection. Thanks to this indicator, it is possible to reduce the weight of the most widely used words (prepositions, conjunctions, General terms and concepts). For each term within a certain text base, only one single IDF value is provided. If a word frequently occurs in a document but rarely in all other documents, then this word has a great significance for the document itself. It is also worth noting the implementation of the TF-IDF method in the skLearn library [4], which is easy to implement and has many useful input parameters.

3.2.Vector representation of words

The main idea of the vector space model (VSM) is to represent each text of the collection as a point in a multidimensional space (a vector in a vector space). Close-lying points correspond to semantically similar documents. To represent words in vector format, there are already ready-made methods, one of the most popular is the Word2Vec technology [5] include identifying semantically similar words, searching for typos, and evaluating the importance of words in a query.

One of the most popular applications of neural networks is the construction of word vectors related to distributive semantics: it is believed that the meaning of a word can be understood by the meaning of its context, by surrounding words. Indeed, if we are unfamiliar with a word in a text in a known language, it is possible to guess its meaning in most cases. Word vectors serve as a mathematical model of word meaning: strings in a large matrix "word-context", built on a fairly large body of texts. The" contexts " for a particular word can be adjacent words, words that are part of the same syntactic or semantic construction with the data, and so on. In the cells of such a matrix, frequencies can be recorded (how many times a word has been encountered in a given context), but more often, the coefficient of positive pairwise Mutual Information (PPMI) is used, which shows how unusual the appearance of a word in a particular context was. Such matrices can be used quite successfully for clustering words or for searching for words that are close in meaning to the desired word.

As was later shown, Word2vec is nothing more than a factorization of the word-context matrix with PPMI weights.

4. Machine learning for NLP

Natural language processing includes speech recognition and generation, classification, extraction of knowledge from texts and other actions aimed at understanding texts to fill knowledge bases, form answers to questions and conduct a dialogue.

Significant progress in the field of natural language processing technologies is largely due to machine learning. In machine learning, a special place belongs to classification algorithms in text processing tasks: spam filtering, sorting documents by topic, and selecting named entities. The field of thematic modeling has emerged, in which documents are considered to be the product of a certain probabilistic process and consist of a mixture of topics. In computational linguistics, the definition of parts of speech has become highly accurate thanks to statistical methods such as hidden Markov chains and maximum entropy models.

Neural networks allow to find hidden connections and patterns in texts, but these connections cannot be represented explicitly. Firstly, the use of neural networks significantly improves the quality of solving some standard problems of text classification and sequences, reduces the complexity when working directly with texts, and thirdly, allows you to solve new problems (for example, create chatbots).

Neural network technologies have radically changed the work with text data. If earlier each element of the text (letter, word or sentence) had to be described with the help of many features of different nature (morphological, syntactic, semantic, etc.), now in many tasks the need for complex descriptions disappears. Theorists and practitioners of neural network technologies often talk about" representation learning " – in a raw text, divided only into words and sentences. The neural network can find dependencies and patterns and independently create a feature space. Unfortunately, in such a space, a person will not understand anything – during training, the neural network puts each element of the text in accordance with one vector consisting of certain numbers representing the detected "deep" relationships. When working with text, the emphasis shifts from constructing a subset of features and searching for external knowledge bases to selecting data sources and marking up texts for subsequent neural network training, which requires significantly more data than standard methods.

The use of deep learning methods, due to the progress in the field of high-performance systems and the emergence of large amounts of data used for training, made it possible to eliminate the work on creating features for machine learning, providing the possibility of simultaneous training in the selection of features and training directly to the task itself. Thanks to new algorithms and approaches, including deep learning, the speed of grammatical analysis has increased. In addition, almost all leading algorithms and models have become widely available to researchers. Probably the most famous work in the field of deep learning for NLP was the already mentioned Word2vec algorithm. Now it is customary to refer Word2vec to distri-butive semantics, and not to deep learning, but the initial impetus for the creation of This model was the use of a neural network. In addition, it turned out that Word2vec vectors serve as a convenient representation of the meaning of a word, which can be fed to the input of deep neural networks used for text classification.

4.1. Representation of words

In common approach, the words of a sentence are treated as elements of a set of words from a dictionary, and serious difficulties arise: if you take into account various colloquial forms, for example, technical jargon, the volume of words, even in English, becomes too large, and the compilation of a complete semantic dictionary for a wide range of applications is a very time – consuming task.

The Word2vec algorithm is included in many standard machine learning packages and is trained in high-quality representations of words on large unmarked corpora (a variety of different texts on different topics, written in different genres and styles). In contrast to traditional word representations, we use a neural probability model of language – each word is represented by a vector of real numbers in the space of relatively small dimension (compared to the size of a complete dictionary), for example, a dimension of 300 measurements. Vectors are initialized with random values. In the process of learning, for the word chosen vector, the max-but similar (in the case of this algorithm as a scalar product) on vectors of other words that occur in similar contexts. A small window of preceding and following words is taken as context (for example, five words before, five words after). This approach gives exciting results. First, the close words (as far as the scalar product is concerned) are indeed often semantically close. Second, it turns out that many interesting relations for natural language processing are encoded in vectors. It turned out that for such vectors, one can define arithmetic operations of addition and subtraction (meanings). As an illustration, the following example is often used: if you take the vector of the word "Paris" from the vector of the word "France" and add the vector "Russia", you get a vector very close to the vector "Moscow", if you add the vector "Italy" instead of the vector "Russia", the result is the vector "Rome"-the relation "capital" is encoded in word vectors. Another example is defined by the equation: "king-man + woman = Queen".

As you know, most natural language processing methods successfully use only word representations, ignoring the syntax and semantics that can be inferred from the syntactic structure of sentences. Such a model of text representation is called BOW (bag of words) – a simple set of words without taking into account their order. For example, in the case of vector representations, it is possible to cluster the word vectors of the corpus on which the model is trained, and use such clusters for simple classification problems. But if the task is to extract better semantic representations, then you will need text processing tools that work with the syntactic structure of sentences or at least do not ignore the word order in the sentence. For example, if you want to analyze reviews on social media about hotels left by their customers, you can find this sentence: "the Hotel is nice, but the bar is smoky." Without analyzing the sentence structure, we will not be able to understand which word each adjective refers to.

4.2. Representation of sentences

Deep learning methods provide an opportunity for a different approach to working with sentences-modeling a sentence as a sequence of vectors obtained by the Word2vec method and using it in machine learning algorithms.

Recurrent neural networks, which accept a single word in a vector representation and have several internal levels at the input, and build a classifier at the output, perfectly cope with this task.

Sequence classification is a task in which each word must be assigned a single label: morphological analysis (each word is assigned a part of speech), extraction of named entities (determining whether each word is part of a person's name, geographical name, etc.), and so on. When classifying sequences, methods are used to take into account the context of the word: if the previous word is part of the person's name, then the current word may also be part of the name, but it is unlikely to be part of the organization's name. Recurrent neural networks help to implement this requirement in practice, expanding the idea of language models proposed at the end of the last century. The classical language model predicts the probability that the word i will occur after the word i-1. Language models can also be used to predict the next word: which word is most likely to occur after a given word?

Recurrent neural networks have proven themselves well in solving various tasks, from language modeling to machine translation, but this class of networks has a significant drawback – they use only the order of words in a sentence, and they cannot be forced to work with grammatical structures obtained by traditional tools. In fact, recurrent networks have to "learn" the language's grammar from scratch for each task.

In addition, the recurrent network does not build representations for intermediate phrases, so recursive neural networks are used for tasks that require high-quality representations of various phrases that make up sentences. In training, a recursive network can learn to make qualitative representations for complete sentences and all sentence phrases. Simultaneously, the neural network can weaken the effect of grammatical errors, especially affecting the task on which the recursive neural network is trained. Thus, we obtain a measure of semantic proximity for both words and all phrases in a sentence.

5. Examples of texts analysis in applied socio-economic studies

5.1. Labour market analysis

Recently, the prospects of "digitalization" of economic processes have been actively discussed. This is an extremely difficult task that has no solution in the framework of traditional methods. The prospects for their qualitative development in the article are illustrated by the example of using Big Data analytics and text mining to assess the labor force needs of regional labor markets. There is also an important question of studying the interaction between labour market and professional education system [6]. The problem was solved using the automated information system developed by the authors for monitoring the compliance of employers' personnel needs with the level of specialist training. The information base for collecting information was open sources. The presented system creates additional opportunities for identifying qualitative and quantitative relations between the education sector and the labor market. It is aimed at a wide range of users: authorities and administrations of regions and municipalities; management of universities, companies, recruitment agencies; graduates and university graduates.

Basic information about the state of the labor market is obtained by analyzing the database of collected vacancies. To obtain correct statistics, it is necessary to solve, first of all, the following tasks:

- Seeking for duplicate vacancies. Even if you use one source, job ads can be duplicated, but if you use multiple sources, such checks are necessary.
- Classification of vacancies by branches of professional activity.
- Analysis of the job offer content, analysis of individual requirements for skills and competencies.

The need to delete identical vacancies is connected with the fact that the sample we use consists of uploading data from several sources, and on each of the sources the same vacancy can be republished repeatedly with some time interval. In order that the data of the same vacancy were not processed several times, it was decided to implement search of identical and similar vacancies with further removal of duplicates. Despite a direct comparison under the name of employer, job title and address, it is necessary to take into account the fact that the name of the position and the content of jobs may change if re-published or the information could be just written on a slightly different way.

Previously, to compare the meaning of text fields, the method of comparing the vector representation of texts in semantic space was used (using the word2vec approach). Further, to make the analysis more specific, it was necessary to distinguish words and expressions characteristic of certain professions and fields of activity. For this purpose, the statistical indicator TF-IDF [4] (term frequency - inverted document frequency) was used, which is mainly used to assess the importance (weight) of a particular word (term) in the context of the entire document included in the general collection (base).

Due to the data from hh.ru and superjob.ru are already structured, it can be used as training data for a kind of multi-label classification [5]. That is, initially there is a sample with about one million marked data and it is possible to operate on it. The next step is to extract only the data necessary for classification. These are the duties, requirements, as they contain basic information about the job and a list of professional areas and specializations to which the job belongs. After the preprocessing: removal of stop words, tokenization and lemmatization of the text, there is everything necessary for further classification of the vacancy. Job offers than were classified against professional areas and required competencies. For the classification, it was trained and used a neural network implementation from the scikitlearn library. When jobs are classified, it is, moreover, easier to find identical records in the database.

5.2. Checking foreign counterparty companies using Big Data

The project [7] aims to create a database of companies and company data and an automated analytical system based on this data. The development of the system will allow credit institutions to obtain information about the links between companies, to carry out a policy of "Know your customer" - to identify the final beneficiaries, to assess risks, to identify relationships between customers. It could the need of banks to fulfill the requirements on national authorities, laws on offshore tax evasion and FATCA, the recommendations of the Group of development of financial measures of struggle against money-laundering (FATF), the Basel Committee on banking supervision. For the moment, there are some projects like OpenCorporates [8] having global databases of companies collected from many jurisdictions.

But at the same they don't cover neither all the national registries, nor other useful data sources (courts, customs, press, etc.). Also, the existing services have rather sketchy abilities on searching for relations between companies, which are not always direct. The project we present is about to overcome main of these deficiencies. Number of companies worldwide is more than 150 million. Having company information from many sources, there is no other reasonable way to process it using Big Data technologies. In the research we use such technologies along with machine learning and graph databases.

To identify the affiliation of the companies, in addition to direct comparison of relationships through the founders and owners, the analysis of indirect signs is used. We consider companies that have a coincidence in several positions. First, the fragments of the name, officers, founders, registration address, contact information, owners, subsidiaries, historical ties, similarity of the names and company profiles, etc. in addition, it uses the previously found relations.

Discovered information about those or other links of the companies is stored in a graph database, entries in which both the company and the other object types (officers, founders, registration address, contact information). This approach allows for more flexible link analysis and complex search queries. For the analysis and storage of the revealed connections the graph base Neo4j [9] is used. This database also allows to visualize the graph links using built-in tools.

6. Conclusion

In the paper, we have discussed some approaches to the intellectual text analysis in application to automated monitoring of the labour market. The scheme of construction of an analytical system based on Big Data technologies for the labour market is proposed. Were compared the combinations of methods of extracting semantic information about objects and connections between them (for example, from job advertisements) from specialized texts. A system for monitoring of the Russian labour market has been created, and the work is underway to include other countries in the analysis and the system to check foreign counterparty companies using Big Data. The considered approaches and methods can be widely used to extract knowledge from large amounts of texts.

Acknowledgement

The study was carried out at the expense of the Russian Science Foundation grant (project No. 19-71-30008).

References

[1] AP's 'robot journalists' are writing their own stories now // The Verge, 25 January 2015 — AVAILABLE AT: www.theverge.com/2015/1/29/7939067/ap-journalism-automation-robots-financial-reporting.

[2] The New York Times, 18 October 2016 — Available at: www.nytimes.com/2016/11/18/technology/automatedpro-trumpbots-overwhelmed-pro-clinton-messages-researchers-say.html

[3] MIT Technology Review, November 2016 — Available at: www.technologyreview.com/s/602817/how-the-bot-y-politicinfluenced-this-election/.

[4] Mark Needham. scikit-learn: TF/IDF and cosine similarity for computer science papers — Available at: https://markhneedham.com/blog/2016/07/27/scitkit-learn-tfidf-and-cosine-similarity-for-computer-science-papers/.

[5] Tomas Mikolov et. al. Efficient Estimation of Word Representations in Vector Space, arxiv.org. Available at: http://arxiv.org/pdf/1301.3781.pdf.

[6] Belov S.D. et al., CEUR Workshop Proceedings, 2019, vol. 2507, pp. 469–472

[7] Badalov L.A.et al., Checking foreign counterparty companies using Big Data, CEUR Workshop Proceedings, 2018, vol. 2267, pp. 523–527

[8] OpenCorporates: The Open Database Of The Corporate World — Available at: https://opencorporates.com/

[9] Neo4j graph database. Available at: https://neo4j.com