Methodology for extracting of key words and phrases and building directed weighted networks of terms with using Part-of-speech tagging

© Dmytro Lande [1,2][0000-0003-3945-1178], © Oleh Dmytrenko [1][0000-0001-8501-5313]

[1] Institute for Information Recording of National Academy of Sciences of Ukraine, Kyiv, Ukraine
[2] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
dwlande@gmail.com dmytrenko.o@gmail.com

Abstract. Today, the rapid globalization of the information space leads to the rise of huge arrays of text data on information resources, including unstructured data. Therefore, developing new and improving existing methods and techniques for finding necessary and relevant information from this text data is important.

This article is devoted to solving an urgent and important task related to conceptualization and further formalization in the form of a network of terms of unstructured data contained in thematic information flows distributed on the Internet.

This work proposes a new method for extracting of key words and phrases from thematic information flows and a new method for determining the directions of links between nodes in undirected networks of terms with using Part-of-speech tagging. An idea of determining the weighted values of links between nodes in the directed network of terms. Also, the holistic methodology of computerized text corpora processing and building the directed weighted networks of terms (of key words and phrases) that extracted with using a previous words' classification process into parts of speech, which is based on the phrase syntactic context – Part-of-speech tagging, are presented. Based on PoS tagging a statistical terms weighting is applied as the next step. The proposed methodology is tested on the example of a children's allegorical story-tale, "The Little Prince" by Antoine de Saint-Exupéry. Applying the proposed method, the key terms were extracted and the directed weighted network of words and phrases related to single key concepts in the studied text was built.

Keywords: Text Corpus, Natural Language Processing, Part-of-Speech (PoS) Tagging, Terminological Ontology, Network of Terms

1 Problem statement

With the beginning a rapid development of IT-technologies and globalization of an information space the huge amount of text data is produced on information resources every day. Of course, the part of unstructured data is growing among these data. This

fact complicates the search for necessary and relevant information. So, the huge volumes of information flows and dynamic text arrays, which are related to some problem subject domain, determine the relevance of the process of data conceptualization and their subsequent formalization in the form of a certain ontological model. Therefore, it is important to develop new and improve existing methods that are used to solve the task described above.

Many tasks arise when working with textual information flows lie at the intersection between mathematical sciences and linguistics. This fact opens wide opportunities to apply a powerful mathematical and linguistic theory. For example, the application of knowledge in the field of discrete mathematics makes it possible to present a text data in the form of a network model that is convenient and effective to use. In terms of a complex network, the texts of a certain thematic orientation can be presented in the form of a network of words and phrases that are connected by a formal semantic connection. The network built from key terms (hereinafter network of terms) is one of the forms of this network model. In this network, the nodes are related with the single key terms of some subject domain, and the edges correspond to the links between these terms. Analysis of such networks can be a basis for decisions making in chosen problem subject domain

But while building the network of terms, the identification and extraction the key object (the key words and phrases) are open and unsolved problem. Due to the sparsity of text data and complex semantics of natural language, the determining of the syntax and semantic connections between nodes that correspond to the terms in the text, and the determining the direction of these connections (links) and their weight values are also open problem of conceptualization. The automatization of processes described above and their further visualization are no less important.

The aim of this work is to propose a new method for extracting the key terms of thematic text corpus and determining the directions of links between nodes in the undirected network of key words and phrases to build the terminological ontologies in the form of a directed weighted network of terms. Further these networks can be used to make constructive conclusions about the network structure and its parameters, and on this basis make effective decisions in the accordingly considered problem subject domains that are related to the texts.

2 Main approaches to natural language processing

Text data is a part of natural language. While using a natural language many problems arise first of all connected with ambiguity, non-compositionality and self-application.

This is due to the natural language contains a different forms of words (word forms that have a common basis) and linguistic expressions that are used for expressing different content; thus, in specific situation the meaning of these word will be depend on the context. This language is also called an Inflected Language. Non-compositionality is caused by the lack of rules in natural language that allows to determine the accurate meaning of a complex statement without knowing its context but knowing the meaning

of all other components of words in the statement. This is because in the statement some phrases can be interpreted ambiguously.

While building terminological ontologies of subject domains based on a certain thematic text document [2] the elements of this formal scheme, terms (words and phrases), that are used as concept names and accompany a chosen subject domain must follow the principle of unambiguity. In other words, a word that used as a name must be the name of only one object if this name is singular; a phrase must be a general name for all objects of one class if this phrase is a general name.

In this work, the most common approaches for preprocessing of text data such as tokenization and stop-word removal are used.

Tokenization is used for preliminary lexical analysis and segmentation the text on elementary units (tokens). As independent unit, a token is some form of word. The token is also considered in conjunction with its possible forms and meanings. The tokenization is usually the initial stage of text processing because makes it possible to work with the word as a separate entity while knowing its context [3].

To remove from the text, in particular, all preposition (for example, „an“, „the“ etc.), a stop-word removal is used. The stop-words can be considered as a source of the noise. Prepositions are common stop words particles, exclamations, conjunctions, adverbs, pronouns, introductory word, numbers from 0 to 9 (single digits), other frequently used, auxiliary and independent parts of speech, symbols, punctuation marks. Relatively recently, this list has been supplemented by sequences of characters as www, com, HTTP, etc. that such frequently used on the Internet.

All mentioned above pre-processing methods can be easily applied to different types of texts. It can be done using the standard Python libraries such as Python NLP (Natural Language Processing), in particular, NLTK [4].

Besides that, the Part-of-Speech tagging approach (or just tagging) (fig. 1) is proposed to use to extrapolation of syntax and structure of text. Tagging is usually the next step of natural language processing. Tagging is used after tokenization, and is to refer the word in the text (corpus) to a certain part of speech. This step is based on both a definition of the word and a context of the word. In other words, it is based on the connection of the word with adjacent and related words in a phrase, sentence, or paragraph. Also, Part-of-Speech tagging is one of the main and basic components of almost any NLP task. The collection of tags assigned to each word in the sentence is used for this task. PoS tagging can be used for word indexing, information retrieval and also for many other applications. PoS tagging can be especially useful if some words or tokens can have multiple tags. And most importantly, tagging simplifies the context that refers to some subject domain.

Parts of speech are also known as word classes or lexical categories (which are based on the syntactic context of a phrase). Then, we tag each word according to its lexical category using the above method of classifying words by parts of speech.

E. Brill's parser [5], which uses rule-based algorithms is one of the first and most widely used an English tagger for parts of speech. In addition to a group of rules-based algorithms, there are also stochastic algorithms.

Fig. 1. Example of Part-of-Speech tagging [6].

To extract keywords from the text it is necessary to assign them a certain numerical assessment (in other words, a statistical indicator of importance). In [7] it was shown that it is effective to use the global frequency of the term (GTF - Global Term Frequency) when working with a text corpus. GTF is determined by the ratio of the total number of occurrences of a term in all corpus documents to the total number of terms in corpus documents. GTF shows how important the word is in the global context. It was shown that, in contrast to the usual statistical indicator TF-IDF [8], the proposed assessment of the importance of terms allows to more effectively find information-important elements of the text when working with a text corpus of a predetermined theme in which the information-important term occurs in almost every document of the corpus.

3 Methodology

The building of a directed network of terms is made within every separate sentence.

In this work, separate functions such as “word_tokenize” and “pos_tag” of a specialized Python add-in, the module NLTK (Natural Language Toolkit that is an open-source library), are used to automatic tokenization and Part-of-Speech tagging to assign the tag to every word, accordingly.

Also in this paper, in addition to the standard sets of stop words, which are available by reference [9], [10], it is proposed to use a list of stop words formed by experts within the research subject domain.

The method for determining of key words and phrases, and also directions of links in the undirected network of terms proposed in this work is based on using the process of classification of words on parts of speech and the corresponding tagging of parts of speech (Part-of-Speech tagging). The practice research shows [4] that the most used parts of speech in English are an article (an abbreviated version is DT), sing or mass noun (NN), plural noun (NNS), personal pronoun (PR), verb base form (VB), adjectives (JJ) and adverb (RB). Also in this work, it is considered phrases that have the form “NN1 NN2”, “JJ1 JJ2”, “JJ1 JJ2 NN”, “JJ1 JJ2 NN1 NN2” and which may be important. Although the articles, prepositions (IN), conjunction or coordinating (CC), single verbs, adverbs and pronouns are stop-words, but the phrases that have the form “VV_1 to VV_2”, “NN_1 IN/CC NN_2”, “JJ_1 IN/CC JJ_2”, “JJ NN_1 IN/CC NN_2”, “JJ_1 IN/CC JJ_2 NN”, “JJ_1 JJ_2 NN_1 IN/CC NN_2”, “JJ_1 IN/CC JJ_2 NN_1 IN/CC NN_2” may be key. After

forming the above described terms and arranging them in a certain order (a sequence where phrases with more number of words are placed before phrases and words that are part of them is formed), single stop words are removed (individual articles, prepositions, conjunctions, some verbs, adverbs and pronouns).

Then, with the help of the global frequency of the term GTF, the idea of which is described above, the statistical weighing of words and phrases that is a part of the sequence formed at the previous stage is carried out.

The so-called tuple is formed for each word, in the order of its occurrence in the text. Each element of the tuple consists of three values: the first one is a term (a word or phrase); the second is a tag that is assigned to a word depending on its belonging to a certain part of speech; the last one is numerical value of a GTF.

It should be noted that the GTF is calculated considering two previous values – the word or phrase and part of speech to which this word or phrase relates. The number of similar tuples in the whole text, which normalized by the total number of formed terms in this text defines the value of the third element.

At the next step, it is proposed to determine the undirected links between terms in the text. For this goal, the Horizontal Visibility Graph algorithm (HVG algorithm) that transform time series into a visibility graphs is used [11]. In our case, the time series is a sequence of numerical GTF values formed in the previous step. Next, we show the main idea of the mentioned HVG algorithm. Two nodes t_i and t_j, which correspond to the elements x_i and x_j of the time series are in horizontal visibility if and only if $x_k < \min(x_i; x_j)$ for all t_k where $t_i < t_k < t_j$.. In our case, the sequence t_i, i=1,..,n is a sequence of words within a sentence (n is the number of words left in the sentence after the above-described pre-processing). HVG allows you to build network structures based on texts in which numerical weight values are somehow assigned to individual words or phrases.

If there is an undirected link, determined by the above algorithm, between the nodes from t_i to t_j of the time series then:

- it is proposed to determine the link from node t_i to t_j, if in a sentence the word (not a phrase) to which the node t_i corresponds occurs earlier than the term (word or phrase) to which the node t_j corresponds;
- it is proposed to determine the link from node t_j to t_i, if in a sentence the phrase (not a word) to which the node t_j corresponds occurs earlier than the term (word or phrase) to which the node t_i corresponds (fig. 2).

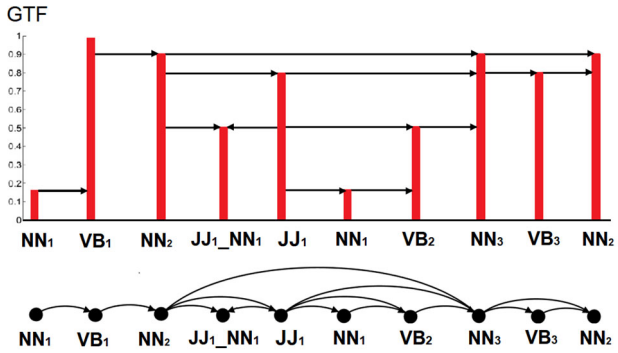

Fig. 2. Example building a directed network of terms.

Given the above-described principle of forming the sequence of the terms and the proposed rules for determining links, it can be noticed that words and phrases will be the part of the corresponding phrases that have more words. In other words, a significant part of phrases with more words is only an extension of the corresponding phrases and words. A similar principle of the building of directed networks of words, the building of networks of natural hierarchies of terms, proposed in the work [12]. In work mentioned above, the directed network of words and phrases is built on the principle of going into the term into its corresponding phrase.

The weight of links between nodes of the directed network of terms is determined by the principle proposed in the work [2]. The main idea of this principle is that the nodes that are corresponded to the same term of the directed network built in the previous stage are combined into a single ("glued"). And the number of the same-directed links between the corresponded nodes determines the weight of the links between these nodes.

4 Result of research

The proposed methodology for computerized processing of text corpus was tested on the example of a children's allegorical story-tale, "The Little Prince" by Antoine de Saint-Exupéry.

According to the methodology proposed above, the selected text document was processed and key terms were identified (Table 1).

If we arrange all the key terms in descending order of their numerical value GTF, then the graph (Fig. 3) shows Zipf's law [13].

Table 1. Top 20 key terms and their numerical GTF values for the text “The Little Prince”.

№	Term	Tag	GTF
1	little	JJ	0.045
2	prince	NN	0.033
3	little_prince	JJ_NN	0.0328
4	planet	NN	0.012
5	flower	NN	0.0094
6	time	NN	0.008
7	good	JJ	0.0077
8	stars	NNS	0.0072
9	fox	NN	0.0061
10	sheep	NN	0.0061
11	man	NN	0.0059
12	morning	NN	0.0054
13	king	NN	0.0049
14	men	NNS	0.0038
15	geographer	NN	0.0038
16	life	NN	0.0037
17	earth	NN	0.0037
18	replied	VB	0.0033
19	water	NN	0.0033
20	great	JJ	0.0033

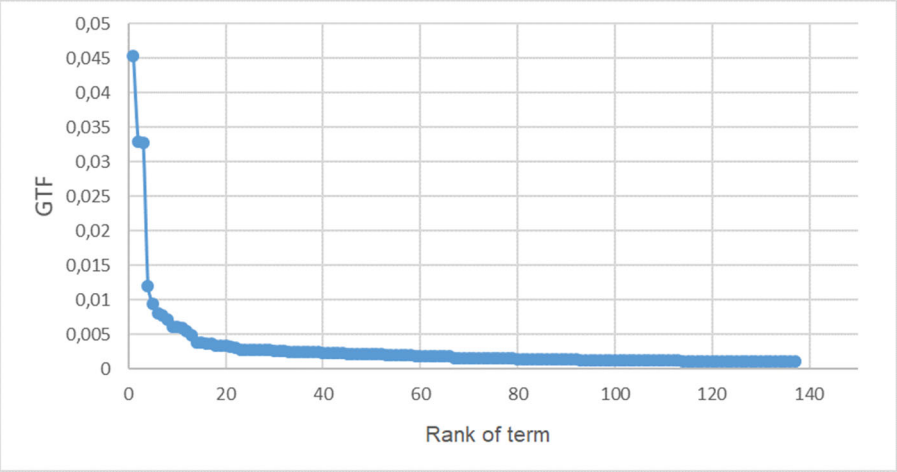

Fig. 3. Graph of Zipf's law for the key terms of the text "The Little Prince".

The obtained directed weighted network of words and phrases was visualized using the Gephi software [14]. This software package was applied not only for modelling and visualization but also for analysis built network. Figure 2 presents the results of the proposed methodology. The links with a weight value that equal 1 and nodes with a zero input and output degree were removed for the built network.

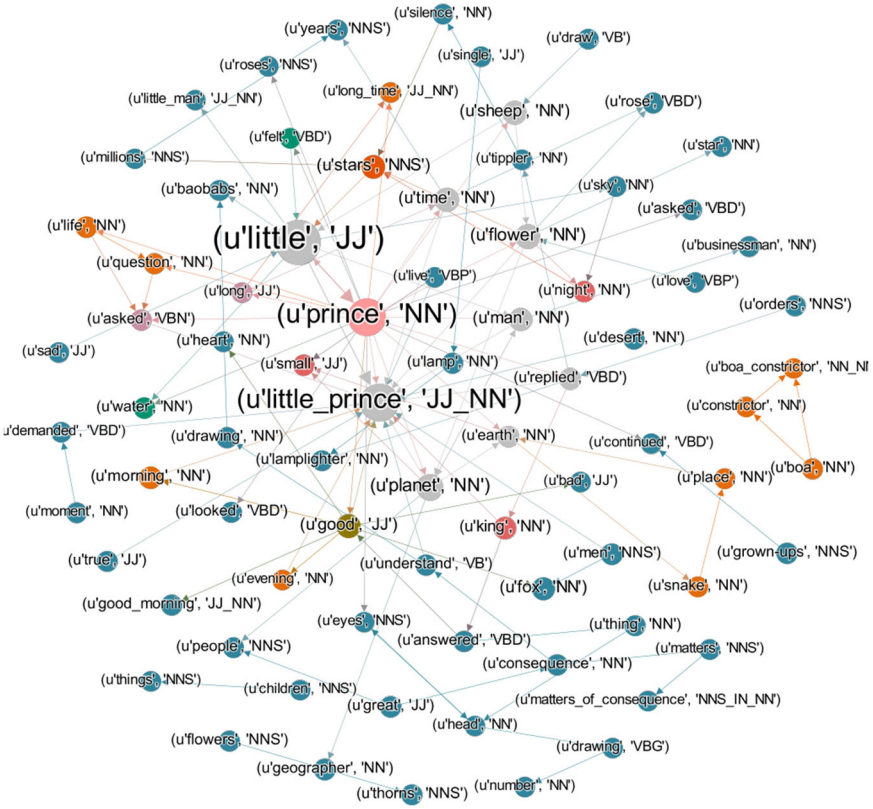

Fig. 4. The directed weighted network of terms built for the tale "The Little Prince" (node labels contain the term and its corresponding tag).

Also, the following parameters of the built network were obtained applying the Gephi software: the number of nodes is 79; the number of links is 117; the average degree is 1.48; the average path length is 3.74; the average clustering coefficient is 0.012; the network density is 0.019; the number of connected components is 4.

The list of the most important links between the corresponding nodes in the network of terms is presented in Table 2.

Table 2. Top 13 most important links for the tale “The Little Prince.

№	Source node	Target node	Weight of link
1	little	little_prince	188
2	little	prince	188
3	good	morning	21
4	good	good_morning	14
5	glass	globe	14
6	glass	glass_globe	14
7	conceited	conceited_man	14
8	fresh	fresh_water	12
9	asked	little_prince	11
10	prince	planet	10
11	prince	little_prince	10
12	matters	consequence	10
13	long	long_time	9
14	long	time	9
15	matters	matters_of_consequence	9
16	little	little_man	8
17	little	man	8
18	prince	little	8
19	boa	boa_constrictor	8
20	boa	constrictor	8

5 Conclusion

In this work, a new method for extracting of key words and phrases from thematic information flows and a new method for determining the directions of links between nodes in undirected networks of terms with using Part-of-speech tagging were proposed. Also, the holistic methodology of computerized text corpora processing and building the directed weighted networks of terms (of key words and phrases) is presented. Using previous words' classification process into parts of speech (Part-of-speech tagging) the key words were extracted. The proposed methodology for computerized processing of text corpus was tested on the example of a children’s allegorical story-tale, “The Little Prince” by Antoine de Saint-Exupéry. The most important links between the corresponded nodes in the network of terms corresponding to certain key concepts in the considered text were revealed after analyzing the results of the methodology. The terms such as “little”, “prince” and “little_prince” turned out the key within the proposed ontological model. These terms also correspond to the name of the considered text document. As expected, the most important links between the key terms are "little → little_prince" and "little → prince".

References

1. Nikonenko A.O.: Review of computer-linguistic methods of natural language texts processing. Artificial Intelligence. № 3, 174-181 (2011). (in Russian)
2. Lande D.V., Dmytrenko O.O.: Creating the Directed Weighted Network of Terms Based on Analysis of Text Corpora. 2020 IEEE 2nd International Conference on System Analysis & Intelligent Computing (SAIC) (Kyiv, 5-9 Oct. 2020).
   DOI: doi.org/10.1109/SAIC51296.2020.9239182
3. Manning, C. D., Raghavan, P., & Schütze, H.: An Introduction to Information Retrieval. Cambridge University Press, 22–36 (2009).
4. Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'-Reilly Media (2009). ISBN 0-596-51649-5
5. Brill. E.: A simple rule-based part of speech tagger. In Proceedings of the third conference on Applied natural language processing (ANLC '92). Association for Computational Linguistics, Stroudsburg, PA, USA, 152-155 (1992).
   DOI: doi:10.3115/974499.974526
6. Extract Custom Keywords using NLTK POS tagger in python. https://thinkinfi.com/extract-custom-keywords-using-nltk-pos-tagger-in-python/. Accessed 24 Oct 2020
7. Lande, D., Dmytrenko, O., Radziievska, O.: Determining the Directions of Links in Undirected Networks of Terms. In: CEUR Workshop Proceedings (ceur-ws.org). Vol-2577 urn:nbn:de:0074-2318-4. Selected Papers of the XIX International Scientific and Practical Conference "Information Technologies and Security" (ITS 2019), vol. 2577, 132-145. (2019). ISSN 1613-0073 [http://ceur-ws.org/Vol-2577/paper11.pdf]
8. Ramos, J.: Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning. vol. 242, 133-142 (2003).
9. Google Code Archive: Stop-words. https://code.google.com/archive/p/stop-words/downloads/. Accessed 24 Oct 2020
10. Text Fixer: Common English Words List. http://www.textfixer.com/tutorials/commonenglishwords.php. Accessed 24 Oct 2020
11. Luque, B., Lacasa, L., Ballesteros, F., & Luque, J.: Horizontal visibility graphs: Exact results for random time series. Physical Review E, 80(4), (2009).
    DOI: doi.org/10.1103/PhysRevE.80.046103.
12. Lande, D. V., Snarskii, A. A., Yagunova, E. V., & Pronoza, E. V.: The use of horizontal visibility graphs to identify the words that define the informational structure of a text. In: 2014 12th Mexican International Conference on Artificial Intelligence, pp. 209-215 (2014).
    DOI: doi.org/10.1109/MICAI.2013.33
13. Li, Wentian.: Random texts exhibit Zipf's-law-like word frequency distribution. IEEE Transactions on information theory. 38.6, 1842-1845 (1992).
14. Gephi. https://gephi.org. Accessed. 02 Dec 2020