Information Technology for Extraction of Coherent Text Fragments from Scientometric Databases

Svitlana Petrasova [a], Nina Khairova [a]

[a]*National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., Kharkiv, 61002, Ukraine*

Abstract

The paper considers the information technology for extraction of Ukrainian and English coherent text fragments from scientometric systems. The up-to-date methods of Data Mining, in particular, statistical models of Information Extraction and Machine Learning for identifying text fragments are analyzed. Based on intelligent data processing tools, the technology will make it possible to determine common information spaces of authors' scientific interaction due to identification of statistically significant collocations that display the topic of texts. The technology includes the distributive-statistical model and Natural Language Processing tools for identification and extraction coherent text fragments. The distributional semantic model MI is applied at the stage of probable collocation identification. Based on POS-tagging formalism, regular expressions, developed in accordance with the grammar of a particular language, allow extracting grammatically correct constructions, i.e. potential substantive, verb and adjective collocations. To identify coherence between these text fragments, dependency parsing is applied. The advantage of the developed technology is that both the grammatical structure and frequency of collocations are taken into account. The corpora of Ukrainian and English scientific texts are built on the basis of abstracts from articles indexed in Google Scholar and ScienceDirect scientometric databases. The effectiveness of the developed technology is assessed and exceeds the results of analogs. The use of the proposed technology could improve the quality of natural language processing. The solution to the problem of automatic extraction of coherent text fragments can be employed to monitor the development of scientific directions, extend research fronts, identify texts of the same domain, extract facts, etc.

Keywords

Scientometric database, coherent text fragment, collocation, Data Mining, distribution-statistical model, POS-tagging, dependency parsing

1. Introduction

In recent years, the interest in scientometric research is growing due to the development of scientometric databases. Modern scientometric systems generate statistics characterizing the dynamics of indicators of activity and influence of scientists' research. Thus, identification of research fronts [1], key publications, and their authors allows monitoring the development of scientific directions and science in general.

Therefore, the main purpose of scientometric research is to provide an objective assessment of the development of scientific areas, their relevance, and the laws of formation of information spaces of scientific interaction or research fronts.

The fronts of scientific research are generally determined on the basis of explicit criteria such as keywords [2], citation [3], co-citation [4], etc. However, in most cases, losing some knowledge, they

IntelITSIS'2021: 2nd International Workshop on Intelligent Information Technologies and Systems of Information Security, March 24–26, 2021, Khmelnytskyi, Ukraine
EMAIL: svetapetrasova@gmail.com (S. Petrasova); nina_khajrova@yahoo.com (N. Khairova)
ORCID: 0000-0001-6011-135X(S. Petrasova); 0000-0002-9826-0286 (N. Khairova)

are insufficient in the development of information support for libraries, electronic catalogs, computer bibliography, systems for automated import of documents, etc. Therefore, it is necessary to apply novel technologies for extraction of an implicitly expressed connection between text fragments from scientometric databases that will make it possible to provide relevant search and access to research works performed on the similar topics.

Such short text fragments, inter alia, can be collocations that represent a non-random syntactic and semantic combination of two or more lexemes and provide more specific semantic information than certain words. So, automatic extraction of collocations that display the topic of text documents can be an additional tool to identify common information spaces of authors' scientific interaction.

2. Related Works

Currently, documents, as online multimedia information with hyperlinks, have become a means of monitoring, influencing and communicating. One of the ways of storing such unstructured data, namely text documents, is a scientometric system that is a bibliographic and abstract database with tools for checking the citation of articles published in scientific journals. As the result, it is possible to identify the state-of-the-art directions of scientific research.

However, to form information spaces of scientific communities adequately, the level of automation of unstructured data (text) processing needs increasing [5], in particular, by solving the problems of intelligent analysis.

In general, intelligent analysis of data or Data Mining is defined as a decision support process based on searching hidden patterns (information patterns). Data Mining technologies are one of the most promising tools to extract valuable knowledge from massive amounts of data, discovering structures, relations, and interconnectedness of data [6]. Data Mining methods are at the integration of several areas (Fig. 1): Data Mining, Web Mining, Machine Learning, Information Retrieval, Information Extraction, etc.

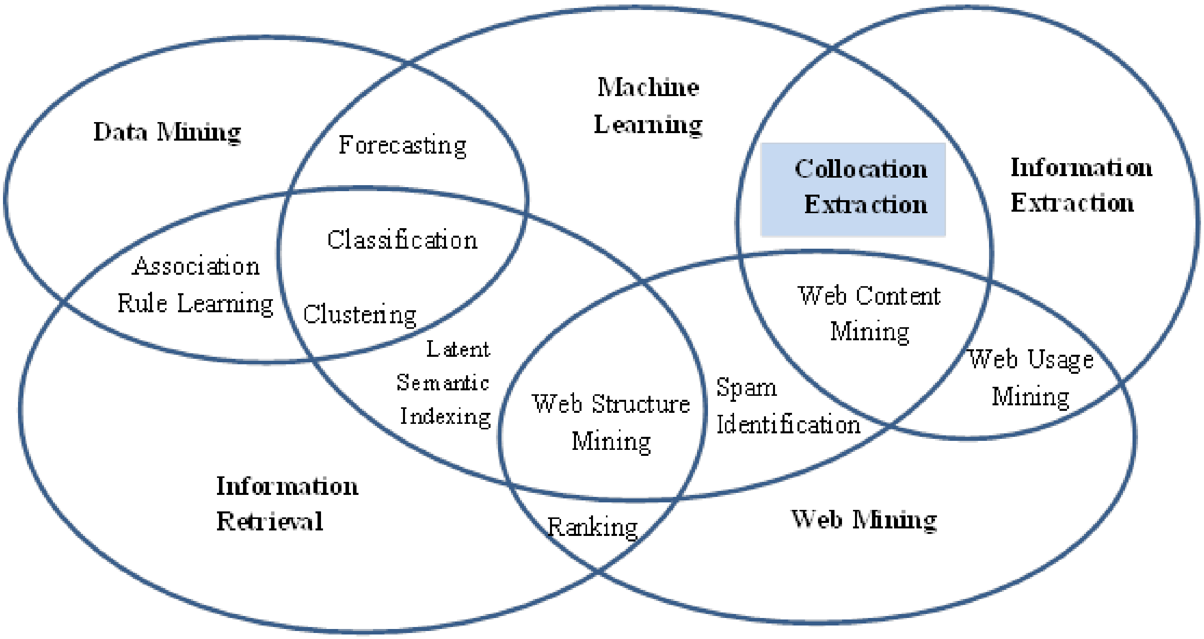


Figure 1: Data Mining Methods

Thus, we can say that the challenging task of extraction of hidden information patterns such as collocations is highlighted at the intersection of Information Extraction and Machine Learning methods.

To identify collocations, distributive-statistical models are considered to be most applicable, the sense of which is the statistical analysis of the probable co-occurrence of variables (lexemes) in large data streams. Traditionally, these types of algorithms [7, 8] use means of mathematical statistics and algebra, without linguistic information.

Statistical metrics or association measures are based on the frequencies of collocates (words) included in collocations to calculate the stability of lexical units. In total, there are more than 80 measures to assess the strength of connectivity. The most commonly used measures are MI, PMI, t-score, log-likelihood, probability coefficient, Pearson's chi-squared test and others. However, statistical methods extract noisy data and ignore syntactic links between words.

At the same time, the analysis of Data Mining methods and statistical models of distributive semantics shows that collocation extraction requires the use of morphological and syntactic tools in addition to statistical models. In this case, identification of collocations allows us not only to take into account the probability of co-occurrence of collocates, but also to formalize grammatical dependencies between the main and dependent components of text fragments.

To solve the problem, we propose the productive combination of a statistical measure, in particular, the model of mutual information MI that compares dependent context-related frequencies with independent ones, and Natural Language Processing tools based on the Universal Dependency (UD) formalism [9], describing coherence between components of collocations.

3. Information Technology

Forming common information spaces of scientific interaction of authors, the developed technology for extraction of Ukrainian and English coherent text fragments includes four stages (Fig. 2).

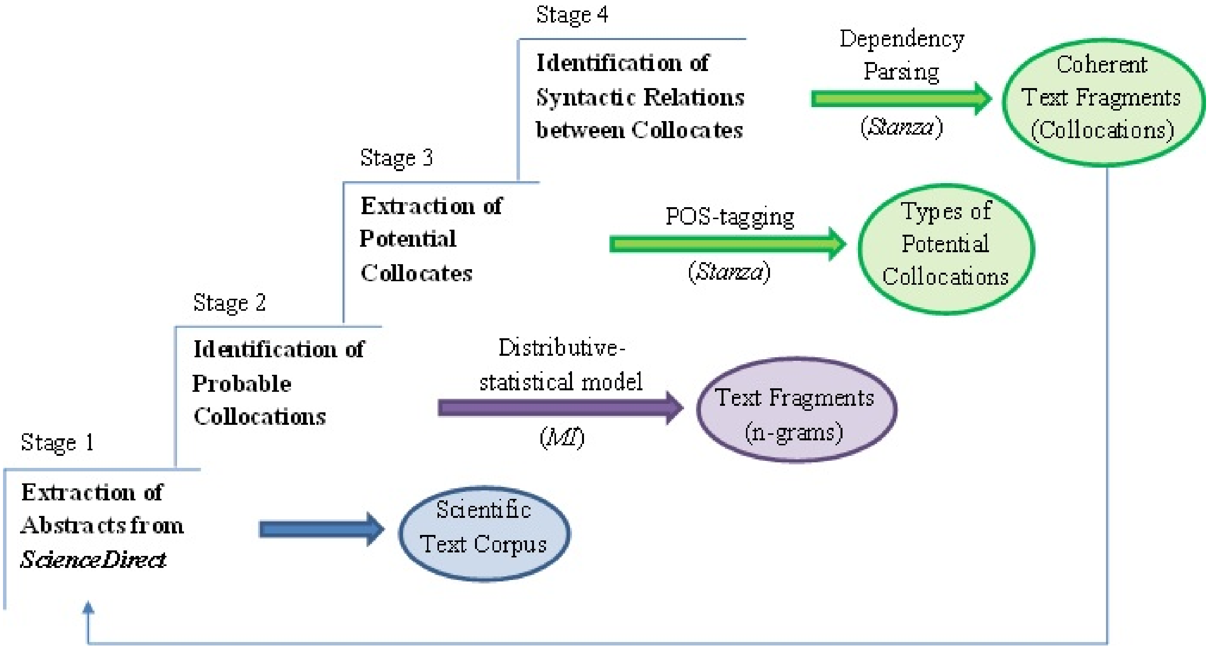

Stage 4
Identification of Syntactic Relations between Collocates
Dependency Parsing
(Stanza)
Coherent Text Fragments (Collocations)
Stage 3
Extraction of Potential Collocates
POS-tagging
(Stanza)
Types of Potential Collocations
Stage 2
Identification of Probable Collocations
Distributive-statistical model
(MI)
Text Fragments (n-grams)
Stage 1
Extraction of Abstracts from ScienceDirect
Scientific Text Corpus

Figure 2: Stages of the Developed Technology

At the first stage of extraction of abstracts from articles indexed in Google Scholar (a freely accessible scientometric database) and ScienceDirect (a database of scientific publications of the Dutch publisher Elsevier), the scientific text corpora were designed. Each corpus contains 350 abstracts of research articles on Artificial Intelligence in the Ukrainian and English languages over a period of 2017-2019. The dataset is of 70 000 (Ukrainian) and 65 000 (English) words.

Since the research on collocations in Natural Language Processing focuses on two related tasks: collocation identification and collocation extraction [10], collocation identification is implemented at the second stage of our technology (to discover the collocation tokens in the corpus) while collocation extraction is carried out at the third stage of the technology (to find the collocation types).

Consequently, based on the model of distributive semantics MI, we identify potential candidates for the main and dependent components of collocations at the second stage. The metric makes it possible to single out rare colocations and identify terminology and other constructions where frequency rates of collocates are very small:

$$MI(n,c) = log_2 \frac{f(n,c) \times N}{f(n) \times f(c)}, \tag{1}$$

where n is a main word; c is a collocate; f (n, c) is the frequency of the main word n in pairs with the collocate c; f(n), f(c) are absolute (independent) frequencies of the main word n and the collocate c in the corpus; N is the total number of words in the corpus.

The result of MI computation is represented by the list of statistically significant bigrams and trigrams, probable collocations, that display the topic of processed texts identified in corpora.

The next stages are devoted to establishing the linguistic correctness of the identified candidates, namely identification of the types of potential collocations and syntactic dependency between collocates via Stanza part-of-speech (POS) tagging and dependency parsing.

Stanza [11] is a Python Natural Language Processing toolkit that features a language-agnostic fully neural pipeline for text analysis, including, in particular, part-of-speech tagging, dependency parsing, using the UD formalism.

The application of Stanza POS-tagging was conducted to extract 3 types of collocations: substantive, verb and adjective. According to the corpus-oriented approach, these types of collocations represent significant text fragments that are most often found in corpora [12].

In this way, at the third stage, we define the following POS-tags in accord with the main types of collocations:

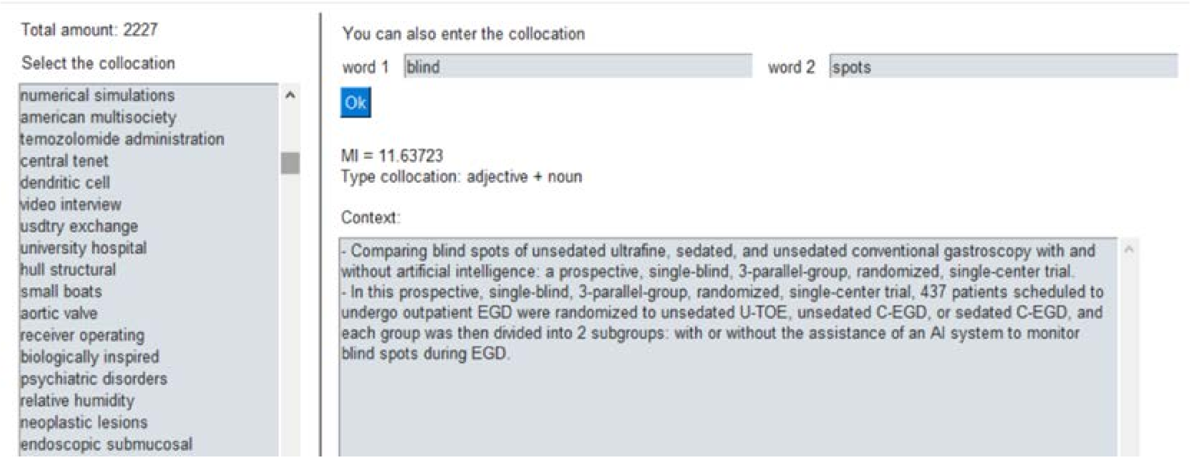
1. Substantive collocations:
   - <NOUN> <NOUN> (e.g. input process, scrambling of data, “налаштування пристрою” - device setting);
   - <ADJ> <NOUN> (e.g. user-centered approach, “програмне забезпечення” - software);
2. Verb collocations:
   - <VERB> <NOUN> (e.g. to meet requirements, “нести відповідальність” - to bear responsibility);
   - <VERB> <ADP> <NOUN> (e.g. to be caused by changes, “використовувати в описах” - to use in manuals);
   - <ADV> <VERB> (e.g. to process iteratively, “уважно ознайомитися” - to see thoroughly);
3. Adjective collocations:
   - <ADV> <ADJ> (e.g. highly accurate, “абсолютно симетричний” - absolutely symmetric).

Using Stanza DepparseProcessor at the final stage of the technology, each sentence in the output is parsed into the UD structure. As the result, syntactic dependency relations between extracted collocates are defined, representing coherence of the text fragments, i.e. collocations:

- nmod: nominal modifier shows nominal dependents of another noun or a noun phrase, corresponds to a genitive complement (in Ukrainian). The nmod relation is used in Substantive collocations (<NOUN> <NOUN>);
- compound: compound relation is used for noun compounds (in English) in Substantive collocations (<NOUN> <NOUN>);
- amod: adjectival modifier of a noun serves to modify the noun. The relation is applied in Substantive collocations (<ADJ> <NOUN>);
- obj: object of a verb denotes the entity acted upon, the object is marked by the accusative case (in Ukrainian). This core argument of a verb is represented in Verb collocations (<VERB> <NOUN>);
- obl: oblique nominal relation functionally corresponds to the prepositional construction in the double object construction, as well as temporal and locational modifiers or nominal modifiers for the agent of a passive verb. It occurs in Verb collocations with prepositions (<VERB> <ADP> <NOUN>);
- advmod: adverbial modifier functions like adverbs and serves to modify predicates in Verb collocations (<ADV> <VERB>) or modifier words like adjectives in Adjective collocations (<ADV> <ADJ>).

Thus, collocations are considered as coherent text fragments if grammatical characteristics of collocates, identified via MI model, satisfy POS-tagging and syntactic dependency parsing features. The collocations extracted from scientific text corpora allow not only identifying the missing combinations of words in Ukrainian and English linguistic sources, but also calculating the statistical indicators of their stability.

The developed technology is implemented as software applications to extract English (Fig. 3) and Ukrainian (Fig. 4) coherent text fragments from large data streams. Software implementation allows you to download text corpora, run POS-tagging via Stanza POSProcessor and dependency parsing of texts via Stanza DepparseProcessor, previously processed by the TokenizeProcessor, MWTProcessor, and LemmaProcessor, and display all the extracted collocations with examples of their contexts in descending order of MI coefficient values.



Total amount: 2227

Select the collocation

numerical simulations
american multisociety
temozolomide administration
central tenet
dendritic cell
video interview
usdtry exchange
university hospital
hull structural
small boats
aortic valve
receiver operating
biologically inspired
psychiatric disorders
relative humidity
neoplastic lesions
endoscopic submucosal

You can also enter the collocation

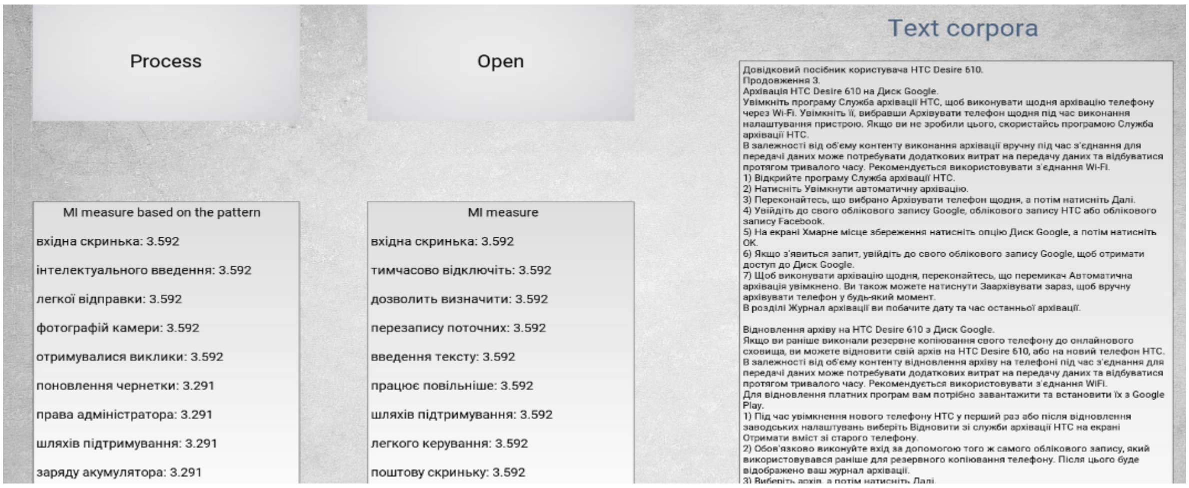
word 1 blind word 2 spots

MI = 11.63723
Type collocation: adjective + noun

Context:

- Comparing blind spots of unsedated ultrafine, sedated, and unsedated conventional gastroscopy with and without artificial intelligence: a prospective, single-blind, 3-parallel-group, randomized, single-center trial.
- In this prospective, single-blind, 3-parallel-group, randomized, single-center trial, 437 patients scheduled to undergo outpatient EGD were randomized to unsedated U-TOE, unsedated C-EGD, or sedated C-EGD, and each group was then divided into 2 subgroups: with or without the assistance of an AI system to monitor blind spots during EGD.

Figure 3: Software Implementation for English Collocation Extraction



MI measure based on the pattern

вхідна скринька: 3.592
інтелектуального введення: 3.592
легкої відправки: 3.592
фотографій камери: 3.592
отримувалися виклики: 3.592
поновлення чернетки: 3.291
права адміністратора: 3.291
шляхів підтримування: 3.291
заряду акумулятора: 3.291

MI measure

вхідна скринька: 3.592
тимчасово відключіть: 3.592
дозволить визначити: 3.592
перезапису поточних: 3.592
введення тексту: 3.592
працює повільніше: 3.592
шляхів підтримування: 3.592
легкого керування: 3.592
поштову скриньку: 3.592

Довідковий посібник користувача HTC Desire 610.
Продовження 3.
Архівація HTC Desire 610 на Диск Google.
Увімкніть програму Служба архівації HTC, щоб виконувати щодня архівацію телефону через Wi-Fi. Увімкніть її, вибравши Архівувати телефон щодня під час виконання налаштування пристрою. Якщо ви не зробили цього, скористайсь програмою Служба архівації HTC.
В залежності від об'єму контенту виконання архівації вручну під час з'єднання для передачі даних може потребувати додаткових витрат на передачу даних та відбуватися протягом тривалого часу. Рекомендується використовувати з'єднання Wi-Fi.
1) Відкрийте програму Служба архівації HTC.
2) Натисніть Увімкнути автоматичну архівацію.
3) Переконайтесь, що вибрано Архівувати телефон щодня, а потім натисніть Далі.
4) Увійдіть до свого облікового запису Google, облікового запису HTC або облікового запису Facebook.
5) На екрані Хмарне місце збереження натисніть опцію Диск Google, а потім натисніть OK.
6) Якщо з'явиться запит, увійдіть до свого облікового запису Google, щоб отримати доступ до Диск Google.
7) Щоб виконувати архівацію щодня, переконайтесь, що перемикач Автоматична архівація увімкнено. Ви також можете натиснути Заархівувати зараз, щоб вручну архівувати телефон у будь-який момент.
В розділі Журнал архівації ви побачите дату та час останньої архівації.

Відновлення архіву на HTC Desire 610 з Диск Google.
Якщо ви раніше виконали резервне копіювання свого телефону до онлайнового сховища, ви можете відновити свій архів на HTC Desire 610, або на новий телефон HTC.
В залежності від об'єму контенту відновлення архіву на телефоні під час з'єднання для передачі даних може потребувати додаткових витрат на передачу даних та відбуватися протягом тривалого часу. Рекомендується використовувати з'єднання WiFi.
Для відновлення платних програм вам потрібно завантажити та встановити їх з Google Play.
1) Під час увімкнення нового телефону HTC у перший раз або після відновлення заводських налаштувань виберіть Відновити зі служби архівації HTC на екрані Отримати вміст зі старого телефону.
2) Обов'язково виконуйте вхід за допомогою того ж самого облікового запису, який використовувався раніше для резервного копіювання телефону. Після цього буде відображено ваш журнал архівації.
3) Виберіть архів, а потім натисніть Далі.

Figure 4: Software Implementation for Ukrainian Collocation Extraction

Running the software program, coherent text fragments are identified that display the topic of scientific texts and form a common information space (a scientific front) of researchers' interaction. Applying the identification results subsequently makes it possible to get new abstracts of the same topic and extend scientific fronts.

4. Experimental Results

To assess the effectiveness of the developed technology and carry out comparative analysis with existing technologies for extracting collocations from large data streams, the precision score was calculated.

Table 1 shows the precision values calculated for three types of collocations in Ukrainian and English. The reason of relatively low results for certain types of Ukrainian or English collocations might be due to mistakes of POS-tagging and UD parsing. As our technology identifies a set of possible grammatical characteristics of collocates, it considerably depends on the result of parsing. For example, in the Ukrainian sentence “В описах використовують налаштування пристрою.” (Manuals apply settings of a device.), Stanza UD parser determined the syntactic dependency relation between words “використовують” (apply) and “налаштування” (settings) as nsubj: nominal subject whereas it should be obj: object (an object of a verb). Consequently, these mistakes are based on morphological or/and syntactic ambiguity, especially in Ukrainian, that is unavoidable and affects the precision of the final result.

Table 1

Comparison of precision scores for different types of collocations extracted from corpora

Types of Collocations	Precision	
	Ukrainian Corpus	English Corpus
Substantive Collocations (NOUN NOUN)	0.88	0.92
Substantive Collocations (ADJ NOUN)	0.91	0.97
Verb Collocations (VERB NOUN)	0.85	0.92
Verb Collocations (VERB ADP NOUN)	0.79	0.85
Verb Collocations (ADV VERB)	0.74	0.89
Adjective Collocations (ADV ADJ)	0.83	0.91

To compare the obtained results with another existing software implementation dealing with a similar problem, the web application Sketch Engine [13] was chosen. This application includes the function of identification of word sketches, i.e. typical phrases that are defined, on the one hand, by syntax that limits the combination of words in a particular language, and on the other one by the probability connected to the frequency of usage of words. Unfortunately, Sketch Engine does not support corpora in the Ukrainian language.

Table 2 shows the comparative analysis of the precision coefficient, determined by the ratio of the number of relevant decisions made by the system (Correctly Extracted Collocations) to the total number of collocations extracted by the system (All the Extracted Collocations). The expert evaluated the extracted collocation as relevant if its morphological and syntactic dependencies were identified correctly and false otherwise.

Table 2

Comparison of different technologies for collocation extraction

Comparison Items	Our Technology		Sketch Engine
	Ukrainian Corpus	English Corpus	English Corpus
All the Extracted Collocations	1672	2227	910
Correctly Extracted Collocations	1388	2027	382
Precision	0.83	0.91	0.42

As a result of the comparative analysis, we can see that the precision coefficient of the developed technology significantly exceeds the same coefficient for the Sketch Engine text analysis software.

5. Conclusions

Based on the statistical model of distributive semantics and Natural Language Processing tools, the information technology is proposed for determining common information spaces of authors’ scientific interaction due to identification of statistically significant collocations that display the topic of texts. The effectiveness of the technology is assessed and exceeds the results of analogs.

The developed software implementation will improve the quality of collocation extraction, and can also be proposed as an advanced model for existing text processing systems. In instance, it can be applied to monitor the development of scientific directions, extend research fronts, classify texts of the same topic, extract facts, etc.

In future studies, we intend to broaden the scope of our research and focus on a more complex challenging problem of distant multi-gram collocations extraction. Additionally, our further work will extend the domain of the texts studied. In prospect, we intend to spread our dataset for free access to fulfil similar approaches.

6. References

[1] C. King, Research Fronts The Hottest Areas in Science, 2016. URL: http://stateofinnovation.com/research-fronts-2016-the-hottest-areas-in-science

[2] L. Wen, Ya. Lu, Hui Li, S. Long, J. Li, Detecting of Research Front Topic in Artificial Intelligence Based on SciVal, in: Proceedings of the 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM2020). Association for Computing Machinery, New York, USA, 2020, pp. 145–149. doi: https://doi.org/10.1145/3421766.342179

[3] H. Sasaki, B. Fugetsu, I. Sakata, Emerging Scientific Field Detection Using Citation Networks and Topic Models—A Case Study of the Nanocarbon Field, in: Applied System Innovation. 2020, 3(3):40. doi: https://doi.org/10.3390/asi3030040

[4] J. Yun, S. Ahn, J. Young Lee, Return to basics: Clustering of scientific literature using structural information, Journal of Informetrics, 14 4 (2020). doi: https://doi.org/10.1016/j.joi.2020.101099.

[5] S.V. Petrasova, N.F. Khairova, Using a Technology for Identification of Semantically Connected Text Elements to Determine a Common Information Space, in: Cybernetics and Systems Analysis, 2017, 53(1), pp. 115–124. doi: https://doi.org/10.1007/s10559-017-9912-z

[6] Ya. Zhao, Ch. Zhang, Yi. Zhang, Z. Wang, J. Li, A review of data mining technologies in building energy systems: Load prediction, pattern identification, fault detection and diagnosis, in: Energy and Built Environment, 1 2 (2020) 149-164. doi: https://doi.org/10.1016/j.enbenv.2019.11.003

[7] A. Lenci, Distributional Models of Word Meaning, in: Annual Review of Linguistics, 2018, 4. pp. 151-171. doi: https://doi.org/10.1146/annurev-linguistics-030514-125254

[8] A. Dinu, L. Dinu, I. Sorodoc, Aggregation methods for efficient collocation detection, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, 2014, pp. 4041–4045.

[9] M.-C. de Marneffe, T. Dozat, N. Silveira, K. Haverinen, F. Ginter, J. Nivre, Ch.D. Manning, Universal Stanford Dependencies: A cross-linguistic typology, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 2014, pp. 4585–4592.

[10] X. Liu, D. Huang, Zh. Yin, F. Ren, Recognition of Collocation Frames from Sentences, in: IEICE Transactions on Information and Systems, 2019, E102.D(3), pp. 620-627. doi: https://doi.org/10.1587/TRANSINF.2018EDP7255

[11] P. Qi, Yu. Zhang, Yu. Zhang, J. Bolton, Ch.D. Manning, Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2020, pp. 101–108. doi: 10.18653/v1/2020.acl-demos.14

[12] S. Petrasova, N. Khairova, W. Lewoniewski, Building the semantic similarity model for social network data streams, in: Proceedings of 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), Lviv, 2018, pp. 21-24, doi: 10.1109/DSMP.2018.8478480

[13] A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, V. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel, The Sketch Engine: ten years on, in: Lexicography, ASIALEX, 2014, 1(1): 7–36. doi: 10.1007/s40607-014-0009-9