Identification of the Author's Idea Based on the Modified TextRank Method

Yuliia Hlavcheva, Olga Kanishcheva, Maryna Vovk and Maksym Glavchev

National technical University "KhPI", 2 Kyrpychova str., Kharkiv, 61002, Ukraine

Abstract

Taking into account the significant rate of new information volume formation, including scientific, semantic analysis of the text continues to be a relevant area of research. The results of this paper on extraction the author's idea can be applied to identify features of intellectual plagiarism to promote academic integrity. To identify the author's idea as the main content of the text, the authors use semantic and graph-based methods. The paper proposes a method for identifying the author's idea based on the modified TextRank algorithm. This method takes into account the pronominal anaphoric connections between sentences, allows to form a more complete description of the semantic relationships between sentences in the text. An experiment was carried out on scientific texts in the Ukrainian language, which confirmed an increase in the number of semantic links between sentences in comparison with the simple TextRank algorithm, which affects the weight of sentences and their order in the abstract.

Keywords 1

Extractive summarization, text summarization, TextRank, academic plagiarism, similarity of ideas, sentence similarity.

1. Introduction

Identification of the author's idea is a complex scientific task. The author's idea is the main content of the text. Thus, the task of automatic abstracting is close to the task of identifying the author's idea. For the task of automatic abstracting it is necessary to obtain a brief description of the main content of the document. Natural language processing methods are widely used for text abstracting, which reduce the original amount of text compared to the input and highlight only important information from the original text. Different statistical, graph and deep learning methods are used for abstracting [1, 2]. Among the popular statistical methods are TF-IDF, TextRank, PageRank, Latent Dirichlet Allocation (LDA), and others.

Articles [3, 4, 5] describe the use of graph methods. The document under study is represented as a graph in which the vertices are sentences and the arcs are the connections between sentences.

The authors [3] used a modified PageRank algorithm. According to this algorithm, each node (sentence) of the graph has an initial score, which is formed based on the number of nouns in this sentence. According to the authors [3], more nouns in a sentence mean that it contains more information, so it is the nouns that are used as the initial rank of the sentence. The summary of the above modified PageRank algorithm includes sentences that contain the most information and are well semantically related.

The authors proposed and investigated the use of the modified TextRank method in the article [5]. It is based on the PageRank algorithm. The proposed method forms a graph with vertices-sentences, which takes into account the similarities between the two sentences. Modified inverse sentence

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine

EMAIL: yuliia.hlavcheva@khpi.edu.ua (Yu. Hlavcheva); kanichshevaolga@gmail.com (O. Kanishcheva); Maryna.Vovk@khpi.edu.ua (M. Vovk); maksym.glavchev@khpi.edu.ua (M. Glavchev)

ORCID: 0000-0001-7991-5411 (Yu. Hlavcheva); 0000-0002-9035-1765 (O. Kanishcheva); 0000-0003-4119-5441 (M. Vovk); 0000-0001-9670-9118 (M. Glavchev)

frequency-cosine similarity is used to give different weightage to different words in the sentence, whereas traditional cosine similarity treats the words equally. The graph is formed sparsely and is divided into different clusters from the position that the sentences in the cluster are similar to each other, and the sentences from different clusters have differences. The authors [5] demonstrate the effectiveness of the proposed method of abstracting.

Comparative analysis of TF-IDF, TextRank, Latent Dirichlet Allocation methods is presented in the publication [4]. To determine the best of them, the study was conducted on three different data sets. To evaluate the performance of the methods, the F-measure indicator and ROUGE and Recall indicators were used as a criterion of accuracy. In general, TextRank showed better results compared to TF-IDF and LDA.

A detailed review of abstracting technologies is described in review publications [4, 6, 7, 8].

Although each new study increases the quality of the results, there are many aspects that affect the quality of the abstract and which are not taken into account in modern methods of automatic abstracting. The review [6] indicates that without the use of natural language processing methods in the generated summary, the semantic integrity of the content may be violated, it may be unbalanced. The authors of the review note that one of the important points of the study is to determine the optimal correct weight of individual elements (vertices (sentences), arcs (connections between sentences)). It depends on the quality of the summary.

In general, all abstraction methods are divided into two major classes: Extractive Summarization and Abstractive Summarization [6, 9, 10].

To determine the elements of academic plagiarism (similarity of ideas), it makes sens to focus on the study of extractive methods of abstracting. Their main task is to extract the most significant fragments of the input text. Abstract methods also form an abstract based on grammar, semantic rules, etc., which allows the program to generate new text other than the input. Thus, the class of abstraction methods is not suitable for determining the signs of academic plagiarism, because it makes additional "noise" when forming a new text.

Text abstracting of high quality is characterized by the following components:

- relevance – the abstract contains the most important and relevant information, the selected sentences should be closely related to the main content of the source;
- content coverage – the abstract should cover as many important aspects of the source document and should minimize the loss of information in the review process;
- variety – the abstract should be short and contain as few minor ("extra") sentences, ie two sentences with the same meaning should not be selected when forming the abstract;
- resume length – usually determined by the user. In our case, the optimal length of the abstract depends on the text characteristics (type, size, etc.) and can be determined by experiments.

Thus, to create an abstract, it is necessary to select a subset of sentences from each studied academic document so that the created abstract contains the main idea of the document and meets the above requirements.

Each researched academic document is divided into a subset of sentences. The weight of a sentence is determined based on the analysis of the words used in it. In this case, the same sets of words are used in several sentences from the document. This feature affects the weight of words, sentences, connections between sentences and must be taken into account when forming an abstract.

In this paper, we consider the application of automatic abstracting algorithms for the problem of identifying the author's idea. This is a relevant area, as it can be used to detect elements of academic plagiarism. After all, "academic plagiarism is not reduced to textual coincidences, but may also relate to incorrect borrowing of facts, hypotheses, numerical data, methods, illustrations, formulas, models, program codes, etc.", including ideas [11]. The most difficult to detect is the plagiarism of ideas, or intellectual (hidden) academic plagiarism.

2. Modification of the TextRank algorithm taking into account the pronoun anaphoric connections

The authors proposed a method of identifying the author's idea based on a modified method (algorithm) TextRank. This method takes into account the pronoun anaphoric connections between sentences, which allows forming a more complete description of the semantic connections between sentences in the text.

2.1. Modified graph model TextRank

As a basic model, the graph model TextRank was used to obtain an abstract of the document proposed in the article [12]. Here is a brief description. We have an undirected graph of the document:

$$G(d) = G(V, E, f, e),$$

where V – nodes / vertices of the graph, E – the edges of the graph, f – knot weight, e – edge weight.

The elements of the undirected graph have the following content: V – sentence, E – the relationship between sentences, e – the value of semantic similarity between sentences, and f – the weight of the node, which is calculated by the principles of the PageRank algorithm [13, 14]. We describe by the formula the value of the weight of the arc between the two vertices *i* and *j* of the graph *G (d)*:

$$e_{ij} = \begin{cases} e_{ij}, < i,j > \in e \\ 0, < i,j > \notin e \end{cases},$$

where e_{ij} – the value of the arc weight, i, j – vertices of an undirected graph $G(d)$. If there is a connection, the weight of the arc is used in the calculations, and if the sentences are not connected, then this value is zero.

In the original work [12], the relationship between two sentences is defined as the number of common tokens between the lexical representations of two sentences. In this study, the connection is calculated taking into account the greater number of semantic connections based on anaphoric links between sentences.

In order to avoid the dominance of long sentences, the normalization factor is used. It determines the ratio of the number of common terms to the length of each sentence [12]. Thus, if we have two sentences S_i and S_j, and N_i is a set of words in a sentence, then the sentence will look like this:

$$S_i = W_1^i, W_2^i, \ldots, W_{Ni}^i.$$

According to the studied modified method, the similarity of S_i and S_j is defined as:

$$Similarity(S_i, S_j) = \frac{|W_k | W_K \in S_i \& W_k \in S_j|}{\log(|S_i|) + \log(|S_j|)} + W_a,$$

where W_a – coefficient that reflects the semantic similarity of sentences based on anaphoric references between sentences.

Thus, the result is an undirected graph that has weighted arcs and weighted nodes. Schematically, it is presented in the figure 1.

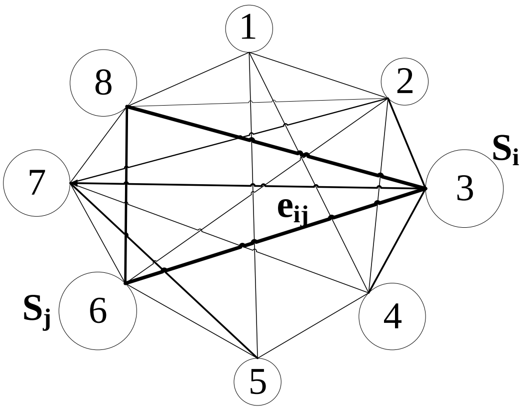

Figure 1: Schematic representation of the document graph

To weigh the graph nodes in the TextRank algorithm, an iterative approach is used, according to the PageRank algorithm [12, 13]. In the first iteration, the nodes are assigned random numbers (0-10 is recommended). Thus, the graph consists of sentences with random weights connected by edges.

The weight of the edges depends on their similarity. The formula for calculating the weight of the node rank is as follows:

$$WS(V_i) = (1 - d) + d * \sum_{V_j \in I_n(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j),$$

where WS – sentence weight (V_i), d – damping factor$(d) = 0{,}85$.

The damping factor is a value calculated by Google engineers in their own PageRank system. It ensures that the weights of their nodes are ultimately reduced to a single value. The damping factor can be any from 0 to 1, but 0.85 is generally recommended [13].

At each subsequent iteration, a new calculated rank of one sentence is used the calculation must be repeated for each sentence. The "true" weight of each sentence is found with each subsequent iteration. The recalculation stops when the "acceptable error rate" reaches the accepted value (TextRank uses the "acceptable error rate" of 0.0001) [13]. This error rate is calculated by subtracting the weight of the sentence before and after the recalculation for each sentence. When the value is 0.0001, the scales are close enough to their "true" weight.

The $tf\text{-}idf$ metric and the cosine measure were used to implement the TextRank method. We form a term-sentence model and a sentence similarity matrix, namely a content-graph join model. Let the input document be represented by a set of sentences $D = \{s_1, s_2, \dots, s_n\}$, where n denotes the number of sentences, s_i denotes the i-th sentence in *D*. In order to form the matrices "sentence-term" and "sentence similarity", each of the sentences should be represented as a vector.

A standard Vector Space Model (SVM) is used using the "Bag of Words" approach, which represents the text units of a document as vectors in a single vector space. To weigh and determine the weight of terms, special metrics are usually used based on any important property. The most common and popular of them is the metric $tf\text{-}idf$ [4]. This metric combines local and global term weighting:

$$W_{iv} = tf_{iv} \times isf_{iv} = freg_{iv} \times log\left(\frac{n}{n_v}\right),$$

where weight W of the term t_v in the sentence s is determined by multiplying the term weight t_v in the sentence sv and the total weight of the term t_v.

According to VSM, the sentence is represented as a weight vector $S_i = [w_{i1}, w_{i2}, \dots, w_{im}]$, where w_{iv} – term weight t in the sentence s_i. The calculation of the angle between two vectors $S_i = [w_{i1}, w_{i2}, \dots, w_{im}]$ and $S_j = [w_{j1}, w_{j2}, \dots, w_{jm}]$ can be obtained as an Euclidean product:

$$(s_i, s_j) = |s_i| \times |s_j| \times \cos\alpha.$$

Thus, the degree of closeness between two sentences is calculated as:

$$sim(s_i, s_j) = \cos\alpha = \frac{(s_i, s_j)}{|s_i| \times |s_j|} = \frac{\sum_{l=1}^{m} w_{il} w_{jl}}{\sqrt{\sum_{l=1}^{m} w_{il}^2 \times \sum_{l=1}^{m} w_{jl}^2}}, i, j = 1,2, \dots, n.$$

The resulting matrix of "sentence similarity" describes the similarity of the presented sentences as objects in Euclidean space. The columns and rows of the matrix are sentences, and their intersection displays the value of the similarity of sentences.

The abstract formation is based on the ranking of sentences. The ranking of sentences in the abstract is based on the nodes weight. The abstract is usually much smaller than the main text. The number of sentences selected for the abstract depends on the user's settings, namely the desired length.

2.2. Defining semantic relations between sentences based on pronoun anaphora

Analysis of scientific texts has shown that they contain semantic relations between sentences of different types. Semantic relations can be divided into three varieties [15]:

- texts with parallel (remote) relations;
- texts with serial (chain) relations;
- texts with connecting links or with a combined relation.

In parallel relations, the sentences are equal. Parallel relation is the use of sentences in which the same word order, the same grammatical forms of the sentence members. The main means of implementing parallel relations is syntactic parallelism. This is when the same or similar construction of sentences, which is often expressed in the same sequence of words, and the unity of temporal forms of verbs-predicates (predicates) [15].

Example of using parallel communication in scientific texts:

«Ступінь розвитку Web-простору ***буде визначатися*** *технологіями роботи з величезним обсягом інформації, що накопились в Інтернет. Web наступного покоління* ***буде характеризуватися*** *переходом від мережі документів до мережі даних, що при необхідності агрегуються в семантично зв'язані документи за допомогою Web-сервісів».*

("The degree of Web-space development will be determined by the technology of working with a huge amount of information accumulated on the Internet. The next-generation Web will be characterized by the transition from a network of documents to a network of data, which, if necessary, are aggregated into semantically related documents using Web-services".)

Serial or chain relations exist because the complement of the previous sentence becomes the subject in the next sentence. The structural form of this relation is as follows: "complement-subject". Other models of the sentence structure are also widely used: "subject-complement", "complement-object", "subject-subject".

The syntactic essence of the chain relation is expressed in these syntactic models, in the syntactic relations between neighboring members of the sentence. This is the internal, structural side of the chain relation. There are ways to embody syntactic relations in a serial relation [15]: lexical repetition, synonyms, indicative words, personal pronouns, pronoun adverbs, conjunctions, verbal omission, etc.

According to mentioned above, we can distinguish: 1) chain relation by lexical repetition, 2) chain synonymous relation, 3) chain pronoun relation.

Pronouns are thought to combine sentences more closely than repetition or synonymous vocabulary. Chain pronouns are extremely diverse. Personal pronouns, personal pronouns in the sense of possessive and actually possessive, indicative pronouns take part in their organization [16].

Some styles of chain relations are especially distinctive to scientific texts. In academic works, we observe a clear sequence and close relation of separate parts of the text, separate sentences, where each subsequent one is connected with the previous one. Presenting the material, the author consistently moves from one stage of reasoning to another. And this method is most consistent with the chain relation.

Thus, we distinguish the following means of relation: similar words, pronouns, adverbs, numerals, and other means, repetition of words, words that indicate the sequence of content development. Based on this, the following types of anaphora are distinguished: pronoun, noun, adverb, and zero [17]. One of the most commonly used relations is a relation between the anaphoric pronoun and the antecedent. "Anaphoric pronouns are pronouns that refer to some word or phrase (antecedent) of this text, the semantic meaning of which they reflect" [18].

For example, *«Найбільш вираженим у плані динаміки є, безперечно,* ***сегмент*** *інформації у вигляді новин. З одного боку,* ***він*** *має найвищий рівень оновлення, а з іншого боку -* ***у ньому*** *генеруються і поширюються насправді великі обсяги даних».*

("The most pronounced in terms of dynamics is, of course, the segment of information in the form of news. On the one hand, it has the highest level of updating, and on the other hand - it generates and distributes really large amounts of data.")

We single out personal anaphoric pronouns for research. third-person pronouns are most often present in scientific texts of all personal pronouns.

Thus, we investigate the definition of anaphoric relation by the following selected types of constructions.

The construction of the first type with the noun in the singular with identical features for identification: *«Фінансово-економічна* ***криза*** *має протилежні форми прояву у суспільстві. З одного боку,* ***вона*** *оновлює механізми господарювання, а з іншого призводить до зростання соціальної напруженості у суспільстві».* ("The financial and economic crisis has opposite forms of manifestation in society. On the one hand, it renews the mechanisms of management, and on the other hand, it leads to an increase in social tensions in society.") The antecedent "***криза***" and the anaphora "***вона***" have identical morphological characteristics: singular, feminine, nominative case.

The construction of the second type with a noun in the singular with distinctive features for identification: *«Експериментальна **робота** проводилася протягом 2006-2010 років. Основними завданнями **її** було визначення структурної моделі управлінської компетентності для інженерів-керівників електромашинобудування».* (“Experimental work was carried out during 2006-2010. Its main tasks were to determine the structural model of managerial competence for engineers-managers of electrical engineering.”) The antecedent "***робота***" and the anaphora "***її***" have identical morphological characteristics: singular, feminine. The difference for this pair is the grammatical case.

The construction of the third type with a noun in the plural with identical features for identification: *«Це означає, що на наших ринках господарюють зарубіжні **компанії**. І саме **вони**, за наш рахунок, вкладають гроші у власний розвиток науки, техніки, створюють додаткові робочі місця». (*“This means that our markets are managed by foreign companies. And they, at our expense, invest money in their own development of science and technology, create additional jobs.") The antecedent "***компанії***" and the anaphora "***вони***" have identical morphological characteristics: plural, nominative case.

The construction of the fourth type with nouns in the plural with signs for identification that differ: *«Процеси докорінної зміни соціально-виробничих відносин, що відбуваються в нашому суспільстві, не обминають і технічні **університети**. Суспільство вимагає від **них** підготовки компетентних фахівців у спеціальній і психолого-педагогічній галузях, зокрема сформованості управлінської компетентності».* (“The processes of the radical change of social and industrial relations taking place in our society do not bypass technical universities. Society requires them to train competent specialists in special and psychological and pedagogical fields, in particular the formation of managerial competence.”) The antecedent "***університети***" and the anaphora "***них***" have the same characteristic, they are plural and differ in cases.

When solving an anaphora, it is important to identify, based on syntactic and morphological information, the characteristic features by which the antecedent is identified in relation to the anaphora. Antecedent and anaphora should have similar characteristics. 90% of antecedents are in the same or previous sentence with anaphora [19]. In our case, the anaphora and the antecedent must be in different sentences.

To solve the anaphora (identification of the pair anaphora - antecedent) different methods are used: system analysis, construction of the classifier, machine learning algorithms, and others. Approaches to solving this problem are described in the following publications [17, 20, 21, 22].

The authors of this article conducted a study of the solution of anaphoric references for structures of the first type. The article [23] describes the process of identification of the pair anaphora - antecedent. Using the mathematical apparatus of the algebra of finite predicates, a logical network is constructed to identify an anaphoric relation based on seven features.

To improve the process of determining the semantic similarity of sentences informative features are optimized. To identify the chain relation for all of the above types of structures four features were used. They are listed in Table 1.

Table 1
Signs of a chain relation

№	Name of attribute	Notation	Attribute value
1	Part of speech is a noun	$X_1 = \{x_1^1, x_1^2, x_1^3\}$	x_1^1 – noun x_1^2 – other part of speech x_1^3 – not specified
2	Grammatical gender	$X_2 = \{x_2^1, x_2^2, x_2^3, x_2^4\}$	x_2^1 – masculine gender x_2^2 – feminine gender x_2^3 – neutral gender x_2^4 – not specified
3	Number	$X_3 = \{x_3^1, x_3^2, x_3^3\}$	x_3^1 – singular, x_3^2 – plural, x_3^3 – not specified

№	Name of attribute	Notation	Attribute value
4	Grammatical case	$X_4 = \{x_4^1, x_4^2, x_4^3, x_4^4,\ x_4^5, x_4^6, x_4^7\}$	x_4^1 – Nominative, x_4^2 – Genitive, x_4^3 – Dative, x_4^4 – Accusative, x_4^5 – Ablative, x_4^6 – Prepositional, x_4^7– not specified

The text is a set of sentences. In the course of our algorithm, first, every two consecutive sentences (in pairs) are checked for the presence of a semantic relation. It should be remembered that the anaphora is in the second sentence of the pair, and the antecedent is in the first one.

To determine the potential anaphora in the second sentence of the pair is a search for personal pronouns «він», «вона», «воно», «вони» ("he", "she", "it", "they"), and their declension forms. If there is a potential anaphora, then to determine the potential antecedent in the first of a couple of sentences search for words related to nouns (X_1).

For potential anaphora and antecedent, morphological features (X_2, X_3, X_4) are determined and checked for compliance. If the morphological features coincide (according to a certain type of construction) then the anaphora-antecedent pair is identified and the semantic relation is confirmed.

For example:

«<...>

*1. **Оцінювання** ефективності зовнішньої реклами – не пряме і досить складне. (Evaluating the effectiveness of outdoor advertising is not direct and quite complex.)*

*2. **Воно** виконується шляхом визначення кількості потенційних рекламних контактів через оцінювання потенційної аудиторії конкретного місця знаходження реклами.(It is performed by determining the number of potential advertising contacts through the evaluation of the potential audience of a particular location of advertising.)*

<...>».

In the second of a couple of sentences, the third-person pronoun for the role of a potential anaphora is identified as «воно» ("it").

The first sentence in a pair identifies nouns for the role of potential antecedents – «реклами», ("advertising"), «оцінювання» ("evaluation").

After determining and comparing morphological features, we can identify a pair of anaphora – antecedent: «воно» – «оцінювання» ("it" is "evaluation").

Therefore, we have a confirmed semantic relation between these sentences. For the description of our model, we used the mathematical apparatus of the algebra of finite predicates and comparative identification. This mathematics was used to describe the process of chain identification based of these features.

Thus, semantic relations are determined based on anaphoric references, which could not be detected based on tf-idf. Newly defined semantic relations are taken into account when forming the "sentence similarity" matrix, and when calculating the value of the relation between sentences.

3. Experiments and Results

The research is conducted on our own scientific text corpus in the Ukrainian language. The source of data is the repository of the National Technical University "Kharkiv Polytechnic Institute" (http://repository.kpi.kharkov.ua) and the portal of scientific publications of the National Technical University "Lviv Polytechnic" (http://science.lpnu.ua/uk). The total number of authors is 32, the number of documents (individual publications) is 271 (Table 2).

5.5% of the total words in the text corpus are pronouns (23255 out of 415565). These data indicate the frequent use of pronouns. Anaphoric pronouns most often include personal pronouns of the third-person, indicative, inverse, relative pronouns [24, 25]. The presence of pronouns in the Ukrainian language documents of the text corpus is presented in Table 3.

In this research, we analyze not all document genres, but only scientific. For scientific texts, some style of chain links is especially characteristic. In academic works, we meet with a clear sequence and close connection of separate parts of the text, separate sentences, where each subsequent one is connected with the previous one. Presenting the material, the author sequentially moves from one stage of reasoning to another. In addition, in this way it is most consistent with the chain link.

Table 2
The main statistical corpus indicators

Name of attribute	Attribute value
Number of authors	32
Number of documents	271
Total size (tokens)	415565
The average number of tokens in the document	1533
The total number of sentences	24743
The average number of sentences in a document	91
The total number of pronouns	23255
The average number of pronouns in the document	86

Table 3
Types of pronouns and their number (text corpus, Ukrainian language)

Type of pronoun	Number
Indicative	12581
Personal	5319
Relative	4322
Defining	1025
Appropriative	8

Another feature of the structure of the scientific language is that the chain connection of sentences is carried out, as a rule, at the place of their connection. It is especially important to emphasize the position of the repeating member of the sentence at the beginning of the next sentence. Thanks to this, continuity and consistency of reasoning is achieved. Each time at the beginning of a new sentence, the opinion seems to return to the main element of the previous sentence, which becomes the starting point for the development of thought in a new sentence.

Figure 2 shows how many pronouns are contained in our corpus, which allow us to organize the connection between sentences and continue the author's idea. It 2 shows the most common pronouns (TOP-10) which were found in our corpus. Pronouns from the TOP-10 make up 40% of the total number of pronouns.

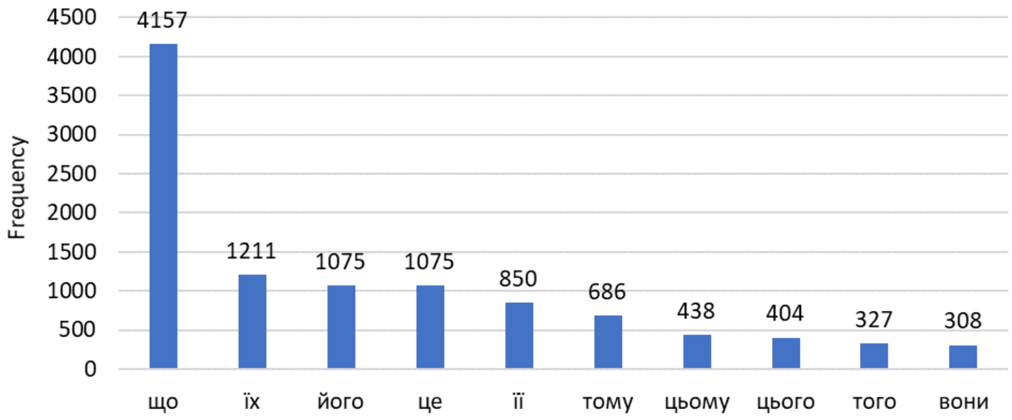

Figure 2: The most common pronouns (TOP-10)

Let us consider examples of sentences from our corpus, between which semantic relations were established using the algebra of finite predicates and comparative identification.

*«**Електроенергія** є основною галуззю національної економіки, стабільність якої має особливе значення для розвитку країни. **Вона** впливає не тільки на розвиток національної економіки, але і на територіальну організацію продуктивних сил». ("Electricity is the main sector of the national economy, the stability of which is of particular importance for the development of the country. It affects not only the development of the national economy, but also the territorial organization of productive forces".)*

Our method identified the pronoun *"Вона"* and found the corresponding antecedent *"Електроенергія"*.

Experiments have shown that semantic relationships were found in our corpus using the pronoun anaphora with an accuracy of 96%. Cases, where the model doesn't work, are related to the syntactic features of the sentences, errors in the morphological analyzer, and the distance between the anaphora and the antecedent.

Let's analyze a fragment of the original article (Doc 1), which consists of about 1463 words and 103 sentences. The fragment contains 58 third-person pronouns in Doc 1. The statistics are presented in Table 4.

Table 4

Statistics of third-person pronouns in Document 1

Pronoun	Document 1	Pronoun	Document 1	Pronoun	Document 1
вона (she)	3	йому (him)	1	нього (Genitive or Accusative case of he)	1
вони (they)	3	нею (Ablative case of she)	1	ній (Prep. case of she)	2
воно (it)	3	неї (Genitive or Accusative case of she)	1	їх (Genitive or Accusative case of they)	10
він (he)	4	ними (Ablative case of they)	2	її (her)	11
його (his)	6	них (Genitive, Accusative or Prep. case of they)	10		

16 of the total number of pronouns belong to our model. Thus, according to the overall result, the value of semantic similarity increases for 16 pairs of sentences.

The number of defined sentence pairs with anaphoric reference (additional semantic relations) to the total number of sentences in the text for 5 documents is presented in Table 5.

Table 5

Statistics of sentence pairs with anaphoric references to the total number of sentences

Document	Number of words	Number of sentences	Number of third person pronouns	Pairs of sentences with a solved anaphora
Document 1	1463	103	58	16
Document 2	2602	580	55	7
Document 3	1516	195	32	8
Document 4	2605	706	39	5
Document 5	1650	314	45	18

Although it can be seen from the Table 5 that there are not so many such pairs of sentences in relation to the total number of sentences, however, these connections qualitatively affect the determination of the similarity of two sentences, which improves the process of obtaining an abstract of the document.

In this paper, we proposed method identified semantic relations between sentences that were not identified by other methods. This affects the ranking of sentences in the abstract.

For the final and detailed analysis of the results and their comparison with existing methods, it is planned to prepare a special text corpus. Calculating the accuracy of the method of the author's idea determining also requires a special corpus, which will contain not only texts but also their main abstracts (ideas). Due to the current lack of such a corpus, this is a task for further research.

4. Conclusions and Recommendations

The task of automatic abstracting is still relevant and not completely solved. In our work, we believe that the abstract demonstrates the main content of the author's text and doesn’t contain "noise", because it consists of the most important sentences of the text. Therefore, the proposed method can be used to identify features of intellectual latent plagiarism and provide the expert with additional information about the "idea" of the text for analysis.

The presented modified TextRank method takes into account anaphoric references with third-person pronouns. This is an important aspect to determine all the semantic relations between sentences in the text when forming an abstract. In future works it is planned to determine the value of the modified TextRank method accuracy. To do this, a special text corpus will be prepared, focused on this task.

Thus, it can be concluded that the modification of the TextRank graph method described in the article allows obtaining a document abstract, which takes into account a greater number of semantic relations between sentences, compared to the simple TextRank method. Due to the solution of anaphora, the use of predicate algebra and predicate operations has demonstrated a successful application for determining the pronoun anaphora within the TextRank method.

5. References

[1] P. G. Magdum, S. Rathi, A Survey on Deep Learning-Based Automatic Text Summarization Models, in: Chiplunkar N., Fukao T. (Eds.), Advances in Artificial Intelligence and Data Engineering. Advances in Intelligent Systems and Computing, Springer, Singapore, vol 1133, 2021. doi:10.1007/978-981-15-3514-7_30.

[2] Kouris Panagiotis, Georgios Alexandridis, Andreas Stafylopatis, Abstractive Text Summarization Based on Deep Learning and Semantic Content Generalization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistic, 2019, pp. 5082–5092.

[3] Elbarougy Reda, Gamal Behery, Akram El Khatib, Extractive arabic text summarization using modified PageRank algorithm, Egyptian Informatics Journal 21(2) (2020) 73–81.

[4] Rani Ujjwal, Karambir Bidhan, Comparative Assessment of Extractive Summarization: TextRank, TF-IDF and LDA, Journal of Scientific Research 65(1) (2021). doi:10.37398/JSR.2021.650140.

[5] Mallick Chirantana, Ajit Kumar Das, Madhurima Dutta, Asit Kumar Das, Apurba Sarkar, Graph-based text summarization using modified TextRank, in: Soft computing in data analytics, Springer, Singapore, 2019, pp. 137–146. doi:10.1007/978-981-13-0514-6_14.

[6] Gupta Vishal, Gurpreet Singh Lehal, A survey of text summarization extractive techniques, Journal of emerging technologies in web intelligence 2(3) (2010) 258–268. doi:10.4304/jetwi.2.3.258-268.

[7] Tas Oguzhan, Farzad Kiyani, A survey automatic text summarization, PressAcademia Procedia 5(1) (2007) 205–213.

[8] Mahajani Abhishek, Vinay Pandya, Isaac Maria, Deepak Sharma, A comprehensive survey on extractive and abstractive techniques for text summarization, Ambient Communications and Computer Systems 904 (2019) 339–351. doi:10.1007/978-981-13-5934-7_31.

[9] B. S. Prakash, K. V. Sanjeev, R. Prakash, K. Chandrasekaran, M. V. Rathnamma, V. V. Ramana, Review of Techniques for Automatic Text Summarization, in: Proceedings of the Third International Conference on Computational Intelligence and Informatics, Springer, Singapore, 2020, pp. 557–565. doi:10.1007/978-981-15-1480-7_47.

[10] V. Soni, L. Kumar, A. K. Singh, M. Kumar, Text Summarization: An Extractive Approach, Soft Computing: Theories and Applications, Springer, Singapore, 2020, pp. 629–637. doi:10.1007/978-981-15-4032-5_57.

[11] Do pytannya unyknennya problem i pomylok u praktykakh zabezpechennya akademichnoyi dobrochesnosti, Lyst Ministerstva osvity i nauky Ukrayiny vid 20.05.2020 № 1/9-263, 2020. URL: https://mon.gov.ua/ua/npa/do-pitannya-uniknennya-problem-i-pomilok-u-praktikah-zabezpechennya-akademichnoyi-dobrochesnosti.

[12] R. Mihalcea, Graph-based ranking algorithms for sentence extraction, applied to text summarization, in: Proceedings of the ACL interactive poster and demonstration sessions, 2004, pp. 170–173.

[13] R. Mihalcea, P. Tarau, Textrank: Bringing order into text, in: Proceedings of the 2004 conference on empirical methods in natural language processing, 2004, pp. 404–411.

[14] M. Bianchini, M. Gori, F. Scarselli, Inside pagerank, ACM Transactions on Internet Technology (TOIT) 5(1) (2005) 92–128.

[15] A. I. Vavilenkova, Syntez lohiko-linhvistychnykh modeley rechen' pryrodnoyi movy yak zasib pobudovy zmistovnoyi modeli tekstu, Systemy pidtrymky pryynyattya rishen', Teoriya i praktyka, Kyyiv, IPMMS NANU, 2013, pp. 49–51.

[16] Ukrayins'kyy pravopys. URL: http://pravopys.mova.info/pravopys.aspx?SectionID=1457.

[17] T. H. Voznyuk, Pobudova klasyfikatora dlya vyrishennya zaymennykovoyi anafory na osnovi tenzornoyi modeli, Visnyk Kyyivs'koho natsional'noho universytetu imeni Tarasa Shevchenka, Seriya: Fizyko-matematychni nauky 2 (2015) 113–116.

[18] T. H. Voznyuk, Zastosuvannya keruyuchoho prostoru syntaksychnykh struktur pryrodnomovnykh tekstiv dlya vyrishennya problemy anafory, Visnyk Kyyivs'koho natsional'noho universytetu imeni Tarasa Shevchenka, Seriya: Fizyko-matematychni nauky 2 (2014) 100–103.

[19] J. R. Hobbs, Resolving pronoun references, Lingua 44(4) (1978) 311–338.

[20] Rhea Sukthanker, Soujanya Poria, Erik Cambria, Ramkumar Thirunavukarasu, Anaphora and coreference resolution: A review, Information Fusion 59 (2020) 139-162. doi:10.1016/j.inffus.2020.01.010.

[21] V. Yu. Dudnyk, Vykorystannya systemnoho analizu dlya rozv'yazku anafory pryrodomovnykh tekstiv dlya ukrayins'koyi movy, Naukovi notatky 65 (2019) 67–73.

[22] N. Herysh, Semantychni vidnoshennya v anaforychnykh konstruktsiyakh imennyk–zaymennyk, Naukovyy visnyk Skhidnoyevropeys'koho natsional'noho universytetu imeni Lesi Ukrayinky, Filolohichni nauky, Movoznavstvo 5 (2014) 151–155.

[23] Y. Hlavcheva, O. Kanishcheva, M. Vovk, Identifying Semantic Relations Between Sentences by Solving an Anaphora, Information Processing Systems 3(162) (2020) 36–43. doi:10.30748/soi.2020.162.04.

[24] O. A. Bakun, Sehmentni konstruktsiyi z nazyvnym uyavlennya: semantychnyy analiz, Naukovyy chasopys Natsional'noho pedahohichnoho universytetu imeni M. P. Drahomanova, Seriya 10: Problemy hramatyky i leksykolohiyi ukrayins'koyi movy 7 (2011) 118–122.

[25] Lynhvystycheskyy entsyklopedycheskyy slovar', in: V. N. Yartseva (Ed.), 2nd. ed., Moskva, 2002.