Analysis of Relation Extraction Methods for Automatic Generation of Specialized Thesauri: Prospect of Hybrid Methods

Ksenia Lagutina, Eldar Mamedov, Nadezhda Lagutina, Ilya Paramonov, Ivan Shchitov
P.G. Demidov Yaroslavl State University, Yaroslavl, Russia
lagutinakv@mail.ru, eldar.mamedov@e-werest.org, lagutinans@gmail.com,
Ilya.Paramonov@fruct.org, ivan.shchitov@e-werest.org

Abstract—**The paper is devoted to analysis of methods that can be used for automatic generation of specialized thesauri. The authors developed a test bench that allows to estimate most popular methods for relation extraction that constitute the main part of such generation. On the basis of experiments conducted on the test bench the idea of hybrid thesaurus generation methods that combine the algorithms showed the best performance was proposed. Its efficiency was illustrated by creation of the thesaurus for the medical domain with its subsequent estimation on the test bench.**

I. Introduction

A specialized thesaurus is a set of terms from a certain professional area that are linked by named semantic relationships [1]. Main applications of such thesauri are query expansion and document indexing in information retrieval, text classification and summarization, and other fields of text processing. A specialized thesaurus is one of the way to model a concrete subject area. Unlike general purpose thesauri it contains a lot of specific words or phrases and relationships between them.

Specialized thesauri in open access exist for a very small number of subject areas because their construction requires a lot of hard work including processing of a large number of text documents and evaluation of quality of the result thesaurus. Such evaluation is often performed by an expert and takes a long time. Also the expert can take part in other stages of thesaurus construction, for example, in estimation of intermediate results that makes a process of thesaurus generation expensive. Therefore, the topical task is to develop methods that maximally automate the process of specialized thesaurus generation, improve the quality of automatically generated thesauri, and minimize the participation of the expert.

In this article we attempted to systematize data on algorithms that can be used for automatic generation of specialized thesauri. We developed a test bench that allows to perform and estimate most popular methods for relation extraction that constitute the main part of such generation. On the basis of experiments conducted on the test bench we proposed an idea of hybrid thesaurus generation methods that combine results of the algorithms showed the best performance. The efficiency of the hybrid approach was illustrated by creation of the thesaurus for the medical domain.

The paper is structured as follows. In Section II we describe general principles of automatic construction of specialized thesauri. Section III overviews articles related to extraction of thesaurus terms and relationships. Section IV describes the proposed test bench for thesaurus generation and evaluation. In Section V we provide results of our experiments with standard methods for relationship extraction conducted on the test bench. In Section VI we introduce two hybrid methods and find out that they provide the same quality and better connectivity of the resulted thesaurus than the standard approaches.

II. Automatic thesaurus generation

Thesaurus terms and relationships can be automatically extracted from a structured or unstructured text corpus. The base algorithm for automatic thesaurus generation consists of two stages:

1) recognition of terms or concepts;
2) extraction of semantic relationships between terms.

The term recognition step is usually comes to extraction of keyphrases from a text corpus that can be in-turn decomposed to the following sub-steps [2]:

1) Candidate selection—search of phrases that satisfy some semantic and/or statistical criteria.
2) Feature computation—construction of feature vectors that contain numerical characteristics of candidate phrases.
3) Keyphrase selection—comparison of feature vectors and selection of candidates with the best characteristics.

The detailed description of efficient term recognition methods is provided in the next section.

After extraction of terms they are linked together by semantic relationships that can be divided into two categories: vertical and horizontal.

Vertical or hierarchical relationships are hyponym-hypernym and whole-part relations. The former means that one term (hypernym) represents a class and the other term (hyponym) represents this class member. For example, “food habits” is a hyponym for “feeding behavior”. A whole-part relation arises when one concept consists from several parts and the other represents one of its part. For example, “brain” is a part of “central nervous system”.

Most used horizontal thesaurus relationships are synonymy and associative relations, and lexical variants.

Synonyms are different terms with the same meanings in a wide range of contexts. For example, "antibiotic A23187" and "calcimycin" are synonyms.

Lexical variants are different word forms or combinations for the same expression. For example, the term "Abelson murine leukemia virus" in MeSH thesaurus has three combinations: "Abelson Leukemia Virus", "Leukemia Virus, Abelson" and "Virus, Abelson Leukemia".

Associative relations do not have a precise definition. Usually the terms are considered as associated when they are related conceptually but do not represent horizontal or synonym relationships [1]. For example, "thylakoids" and "photosynthesis" are related but they are neither synonyms nor constitute a hypernym and hyponym pair.

III. Related work

All keyphrase extraction approaches can be divided into two categories: supervised and unsupervised, according to the method that is used for the keyphrase selection step of the base algorithm mentioned in the previous section.

Supervised methods train on texts with manually chosen keyphrases and classify candidates as keyphrases and non-keyphrases. One of most effective supervised algorithm is Maui [3]. It selects noun phrases as candidates and computes TF*IDF, the number of phrase occurrences in the training set and other numerical features for each candidate. Then Maui applies bagged decision trees classifier.

Unsupervised methods often use graph-based ranking methods for keyphrase selection. TextRank [4] is a well-known unsupervised algorithm that builds a graph with candidate phrases as nodes and uses the co-occurrence relation between phrases for edge creation. Then it calculates candidate scores that depend on phrase links in the graph.

In [5] authors proposed the Topical PageRank algorithm that is an improved TextRank version. Before the keyphrase selection stage it groups candidate keyphrases by topics of the corresponding text. Then Topical PageRank applies TextRank for each topic. This algorithm shows one of the best result among all unsupervised keyphrase selection approaches [6].

Most popular and effective methods for keyphrase extraction were investigated in detail in our previous paper [7].

The second step of thesaurus generation is semantic relationships construction. Most of the methods related to this step establish only a single type of relations between terms pairs. Such methods can be divided into two categories: statistical and semantic.

Statistical methods find related terms by constructing word feature vectors and calculating similarity measures between them. High measure value means a strong relation between terms. Usually such methods extract associative or synonym links.

Ferret [8] analyzes in detail most used semantic similarity measures and feature weighting functions. The best results are obtained with cosine and pointwise mutual information measures.

To improve method precision similarity measures can be combined. The algorithm for synonym extraction [9] chooses most frequent key terms from the text and computes two measures: CBoW and Skip-gram. Then it sums measure values and links term pairs with the highest ones. Such approach produces a more qualitative result then many individual measures.

One of the most effective approaches for association extraction is LSA algorithm [10]. It divides texts into paragraphs and treats each paragraph as a separate document. Then it creates a matrix with terms and documents as rows and columns. Each matrix cell contains a number of appearances of a concrete term in a particular document. The method computes singular value decomposition for the matrix and applies cosine measure to term vectors that are actually rows of result matrix.

Semantic methods find concrete vertical or horizontal relations between terms using syntactic patterns and rules, thesauri and dictionaries. These approaches are often supplemented with statistical or clustering algorithms.

Most effective semantic algorithms use structured text corpora where each text has a strong structure that allows to extract using simple syntactic rules. Espinosa-Anke et al. [11] propose a classifier that selects hypernyms by a combination of syntactic and clustering methods using the text corpus that is a set of word definitions from Wikipedia. The method finds statistical and semantic term features and performs clustering by the conditional random fields algorithm.

The Hearst's method [12] is very efficient for unstructured text corpus. It extracts hypernyms and hyponyms applying specific lexico-syntactic patterns to all sentences of texts.

Only few papers describe how to fully automatically generate a thesaurus with several relationships categories. All of them use structured text corpus from specialized web resources.

The method [13] generates a thesaurus using texts from several sites. The algorithm corresponds each page with one particular term and allocates semantic relations from hyperlinks. Links to index pages determine hyponym-hypernym relationships, all the other ones—associations.

The algorithm [14] extracts information about terms and their relationships from Wikipedia where each page is a description for one term. As the previous method, it analyzes hierarchical and horizontal hyperlinks and transformates them into semantic relations.

The main disadvantage of approaches from previous two articles consists in the fact that they require a text resource with a strong structure and same type of pages, therefore they depend on the text corpus and can be used for a small number of subject areas that have text corpora with the same structure.

IV. Test bench for thesaurus generation and evaluation

A. Data processing stages

The proposed test bench automatically generates the specialized thesaurus from the unstructured text corpus and evaluates quality of its terms and relationships. The test bench performs the input corpus processing in the following stages:

1) Thesaurus term selection using well-known keyphrase extraction algorithm;
2) Construction of associative relations by the LSA algorithm;
3) Construction of hierarchical hyponym-hypernym relations using statistical and semantic approaches, as well as their combinations;
4) Construction of synonym relations by different statistical algorithms;
5) Filtration of terms without relationships;
6) Evaluation of thesaurus terms and relationships.

In our experiments we vary only methods for 3rd and 4th stages because we focus on automatic thesaurus relationship extraction. The final step of thesaurus construction is filtration of redundant terms. We remove all terms that do not have any relations, because they are needless in all scenarios of thesaurus usage.

B. Invariable stages

In this work we do not investigate term extraction algorithms, therefore we have chosen one of them for all experiments and treated the thesaurus term selection stage as invariable.

For term extraction we use a well-known graph algorithm TextRank [4]. Unsupervised methods do not require texts with manually extracted keyphrases for training, so they fit well with our goal to maximally automate thesaurus construction. Although one of the best unsupervised algorithm is Topical PageRank [5], we can successfully replace it by TextRank, because Topical PageRank determines text topics and repeats TextRank for each of them. In this research all the texts had the same topic, so TextRank was good enough for our task.

The next step of thesaurus generation is construction of associations between terms. For this part of the test bench we used the LSA algorithm [10] mentioned in the previous section. It is suitable for unstructured texts and does not require expert's work. Results of this algorithm are used in the course of extraction of the other semantic relations, particularly for comparison of term contexts. Due to the marginal importance of associative relations for specialized thesauri we do not vary algorithms for this step and use a single algorithm.

C. Variable stages

For hyponym-hypernym relationship construction we compared four methods that use:

- morpho-syntactic rules;
- lexico-syntactic patterns;
- general-purpose thesaurus WordNet;
- measurement of term information quantity.

The simplest algorithm for hierarchical relation extraction is applying of morpho-syntactic rules to concrete words or phrases. We used a single rule: if a first phrase is a part a second phrase the first one is a hyponym of the second one [15].

Lexico-syntactic patterns are applied to each statement that contains several terms. This method checks whether the statement contains terms placed near a specific combination of words. In our research we used the patterns from [12]:

HYPERN such as HYPON [, HYPON]* [and/or HYPON]
HYPERN as HYPON [, HYPON]* [and/or HYPON]
HYPON [, HYPON]* [,] and/or other HYPERN
HYPERN[,] including/especialy HYPON [, HYPON] * [and/or HYPON]

Here HYPERN and HYPON means that terms on corresponding places are hypernyms and hyponyms respectively, [] means an optional term, []* means zero or more repetition of the term in square brackets.

The third used method adds a hierarchical link between terms in the resulted specialized thesaurus if they are linked as a hyponym and hypernym in the general-purpose thesaurus WordNet [16].

The last tested method [17] allows to extract both hierarchical and synonym relations. It calculates the amount of information of terms using the formula

$$IC = -\log_2 p(t),$$

where $p(t)$ is number of occurences of the term in a large text corpus divided by the number of terms in the corpus. Such definition means that more frequent term are less informative in the text corpus and vice versa [?]. Then, the terms are considered as terms with close amount of information if their frequency differ from each other no more than a certain cutoff value, otherwise they are the terms with different amount of information. The cutoff value is calculated as follows:

$$\text{Cutoff value} = \frac{\text{Total number of words in text}}{\text{Number of different words in text}}$$

If two terms have the same context and close amounts of information then they are considered as synonyms. If they have the same context and different quantity of information then they form a hypernym-hyponym pair. To determine the terms with similar context we used the association relations determined by the LSA method at the corresponding step of the test bench.

Besides, we use Levenshtein distance for synonym selection [18]. Small Levenshtein distance value means that terms are similar that it used to find cognate words and lexical variants of thesaurus terms.

D. Thesaurus evaluation

To estimate quality of the resulted thesaurus we use the following metrics:

- number of extracted terms and relationships of different types (synonyms, hyponyms etc.);
- number of connected components, isolated vertices, vertices in the largest component that describe connectivity of the thesaurus graph;
- precision of extracted terms, hierarchical and synonym relations; recall of hierarchical and synonym relations.

Numbers of extracted terms and relationships are the most simplest and coarse metrics that allow to estimate thesaurus size. The total number of terms should not be small, the number of horizontal relations and hypernyms should be more than the number of terms because in practice almost all terms have at least one hypernym and several associations or synonyms. Violation of these rules means that the quality of the resulted thesaurus is low.

Estimation of the thesaurus connectivity is inspired by our previous research [7] and justified by the fact that for most of the thesaurus's applications navigation over relations is crucial. Particularly, we calculate characteristics of the graph where terms of the thesaurus are vertices and relationships between terms are edges. If the thesaurus has only small connected components it usually means that it does not reflect the structure of the subject area. Also the thesaurus should not have isolated terms because terms without relationships cannot be used in practice.

Precision (P) and recall (R) are the most popular metrics that allow to estimate quality of thesaurus terms and relationships. They are calculated as follows:

$$P = \frac{|D_{\text{rel}} \cap D_{\text{extr}}|}{|D_{\text{extr}}|},$$

$$R = \frac{|D_{\text{rel}} \cap D_{\text{extr}}|}{|D_{\text{rel}}|},$$

where D_{rel} is the number of all relevant terms or relationships, D_{extr} is the number of all extracted terms or relationships.

V. Performance of standard methods

For evaluation we used well-known MEDLINE text collection (http://ir.dcs.gla.ac.uk/resources/test_collections/medl/) that contains 1033 articles from medical journals and intended to evaluate information retrieval methods. We used this collection as a corpus for extraction of terms and relationships for a specialized thesaurus construction.

The frequency of words for the method involving the quantity of information is determined using Stedman's Medical Dictionary—http://stedmansonline.com/public/LearnMore.aspx?resourceID=Medical).

All the mentioned methods to extract term relationships was implemented in Python programming language with the use of several natural language processing libraries, such as Gensim (https://radimrehurek.com/gensim/), NLTK (http://www.nltk.org/), Pattern (https://pypi.python.org/pypi/Pattern), and python-Levenshtein (https://pypi.python.org/pypi/python-Levenshtein).

The results on statistical values of automatically constructed thesauri are shown in the first four rows of Table I. The largest number of extracted terms, largest number of extracted hypernyms, and greatest connectivity of thesaurus graph was achieved to the method based on using WordNet relationships. The lexico-syntactic patterns method extracted the smallest number of hypernym relations, whereas the largest number of extracted synonyms refers to the method that uses quality of information metrics.

To evaluate the quality of the constructed thesauri we compared them with well-known biomedical thesaurus MeSH (https://www.nlm.nih.gov/mesh/), particularly, we calculated how many terms and relations match with MeSH and how many do not match. The results presented in Table II show what the largest number of matched terms and hypernyms were achieved when using WordNet. The largest number of matched synonyms was extracted using the method based on amount of information of terms. However, these methods also extracted the largest number of corresponding relations that do not match with MeSH thesaurus.

To get the results on quality of the tested methods more clear we calculated precision and recall of extracted terms and relations in comparison with MeSH thesaurus. The results are shown in Table III (first 4 rows).

Precision of extracted terms is about the same for all methods and approximately equals 39 %. The Levenshtein distance method for synonym extraction showed very high precision and equals about 75–85 % but recall for this method is rather low (11–14 %). The method based on amount of information showed bad results for synonym extraction—both its precision and recall are about 15 %.

The quality of hypernym extraction in all our experiments generally turned out pretty low. The most accurate method is the one based on morpho-syntactic rules. Its precision is 24 %, which is several times higher than precision of the other methods. Highest recall (19.7 %) refers to the method involving WordNet. The other methods show much lower results for this metric. Note that in our experiments the lexico-syntactic patterns method showed the worst results and it did not found any hypernyms matched with corresponding relations from MeSH thesaurus.

The method that uses WordNet found many correct relationships but its main disadvantage lies in the fact that the general-purpose thesaurus does not contain all the terms from the specific area and many relationships cannot be found this way.

The methods using morpho-syntactic rules and lexico-syntactic patterns allow to find all semantic relations that are clearly indicated in the text but they become too restricted for unstructured texts and therefore cannot find many relationships for the thesaurus.

Statistical algorithms generate a large number of different term relations. However, they also find incorrect relations more frequently than semantic approaches.

The LSA algorithm found several times more relationships between terms than all the other methods, and the number of associations is about twice more than the number of terms, so LSA provides high connectivity of the thesaurus graph. The quality of the extracted associations cannot be estimated without an expert because MeSH is the only medical thesaurus available in open access and it does not contain associations. That is why we did not calculate precision and recall of the LSA results.

Summarily, the achieved results corroborated the thesis that existing methods are good in extraction of certain types of relations, not all of them at once, and suggested us the idea of hybrid methods described in the following section.

TABLE I. STATISTICAL VALUES OF AUTOMATICALLY CONSTRUCTED THESAURI: t—NUMBER OF EXTRACTED TERM, h—NUMBER OF HYPERNYMS, s—NUMBER OF SYNONYMS, a—NUMBER OF ASSOCIATIONS. C—NUMBER OF CONNECTED COMPONENTS IN THE THESAURUS GRAPH (GRAPH OF RELATIONS BETWEEN THESAURUS TERMS), $d_{\max}$—MAXIMUM DEGREE OF VERTICES IN THE GREATEST COMPONENT

Method	t	h	s	a	C	$d_{\max}$
Morpho-syntactic rules	2090	237	48	4716	83	1867
Amount of information	2105	346	753	3883	84	1881
WordNet relationships	2397	1570	61	4835	35	2306
Lexico-syntactic patterns	2167	102	63	4918	86	1936
Hybrid method I	2433	2188	709	3696	37	2350
Hybrid method II	2397	2096	48	4343	45	2275

TABLE II. COMPARISON OF AUTOMATICALLY CONSTRUCTED THESAURI WITH MESH THESAURUS: T—NUMBER OF MATCHED TERMS, $\bar{T}$—NUMBER OF NOT MATCHED TERMS, S—NUMBER OF MATCHED SYNONYMS, $\bar{S}$—NUMBER OF NOT MATCHED SYNONYMS, SM—NUMBER OF SYNONYMS RELATIONS BETWEEN EXTRACTED TERMS FOUND FROM MESH, H—NUMBER OF MATCHED HYPERNYMS, $\bar{H}$—NUMBER OF NOT MATCHED HYPERNYMS, HM—NUMBER OF HYPERNYM RELATIONS BETWEEN EXTRACTED TERMS FOUND FROM MESH

Method	T	$\bar{T}$	S	$\bar{S}$	SM	H	$\bar{H}$	HM
Morpho-syntactic rules	810	1280	23	4	210	12	38	111
Amount of information	810	1295	31	211	204	8	104	96
WordNet relationships	920	1477	24	6	245	25	277	127
Lexico-syntactic patterns	846	1321	25	8	224	0	33	106
Hybrid method I	944	1489	27	205	258	30	436	130
Hybrid method II	928	1469	26	3	252	30	400	127

TABLE III. PRECISION (P) AND RECALL (R) OF AUTOMATICALLY CONSTRUCTED THESAURI TERMS AND RELATIONS (S—SYNOMYMS, H—HYPERNYMS, T—ALL TERMS) IN COMPARISON WITH MESH THESAURUS

Method	T_P	S_P	S_R	H_P	H_R
Morpho-syntactic rules	38.8	85.2	11.0	24.0	10.8
Amount of information	38.5	12.8	15.2	7.1	8.3
WordNet relationships	38.4	80.0	13.9	8.3	19.7
Lexico-syntactic patterns	39.0	75.8	11.2	0.0	0.0
Hybrid method I	38.8	11.6	10.5	6.4	23.0
Hybrid method II	38.7	89.7	10.3	7.0	23.6

VI. IDEA OF HYBRID METHODS AND THEIR PERFORMANCE

The idea of hybrid method of thesaurus construction consists in combination of using several methods of relation extraction, different ones for different relations. Presumably such a method should improve results in some aspects.

In our research selection of proper methods for certain relations was made on the best results of experiments presented in the previous section. Particularly, we implemented the hybrid method that contains the following steps:

1) thesaurus terms extraction;
2) extraction of associations using LSA method;
3) extraction of hypernyms using the method based on WordNet;
4) extraction of hypernyms using lexico-syntactic patterns;
5) extraction of hypernyms using morpho-syntactic rules;
6) extraction of synonyms using Levenshtein distance;
7) extraction of hypernyms and synonyms using the method based on quantity of information;
8) filtration of isolated terms.

For this hybrid method we performed the same experiments as the ones described in previous section. We calculated the statistical values for the constructed thesaurus and estimated its quality by comparison with MeSH thesaurus. The results of these experiments are shown in the row “Hybrid method I” of Tables I–III.

From Table I we can see that the hybrid method extracts the largest number of terms and largest number of hypernym relations. Number of synonyms and associations are about the same level as in case of the method based on quantity of information, which is a good result in comparison with other methods. Also thesaurus constructed by the hybrid method has well-connected thesaurus graph: It has 37 connected components with the largest one comprising 2350 terms.

In comparison with MeSH thesaurus (see Table II) it is obvious that the hybrid method found more matched terms and hypernym relations. Again, the number of synonyms is about the same level as in case of the method based on quantity of information. At the same time, the number of not matched terms and relations has increased. Especially it concerns to hypernym relations. Indeed results from Table III shows that precision of hypernym extraction has decreased to 6.4 % but recall for this method has considerably increased up to 23 %. Precision of extracted terms remains at the same level.

From the previous experiments with standard methods we found out that the lexico-syntactic patterns method showed negative results on our corpus and did not extract any hypernym relation matched with MeSH thesaurus. Also, the method based on quantity of information for synonym extraction led to significant deterioration of precision. Such considerations led us to the idea of excluding the steps with lexico-syntactic patterns and synonym extraction with the use of the method based on the quantify of information.

As a result of the experiments with the modified hybrid method (the last rows of Tables I–III) we established that this method extracted a slightly number of terms and relations but increased quality of the constructed thesaurus. Precision of synonym extraction increased to 89.7 %, which is the best result over all the tested methods. Recall of synonym extraction remains at the same level and equals 10.3 %. Precision of hypernym extraction became 7 % and recall became 23.6 %, which is better than the results of previous hybrid method.

In comparison with the other mentioned methods the results of the last hybrid method are better in most aspects. In synonym extraction it showed highest precision with a little lower recall. In hypernym extraction it showed highest recall and approximately the same precision as the other methods. And generally, this method constructed the thesaurus with the largest number of terms and relations.

VII. Conclusion

In this paper we analyzed several relation extraction methods that can be used for automatic generation of specialized thesauri. Particularly, we proposed a test bench for thesauri generation and estimation. It extracts terms using the unsupervised TextRank method, finds associative relations by the LSA algorithm, and constructs hierarchical and synonym relationships using different statistical and semantic methods.

To evaluate performance of the methods we involved several statistical metrics, metrics of the thesaurus connectivity, as well as metrics based on the comparison of the result with the existing specialized thesaurus. The achieved results corroborated the thesis that existing methods are good in extraction of certain types of relations, not all of them at once, and suggested us the idea of hybrid methods that consist in combination of using several methods of relation extraction, different ones for different relations. Selection of proper methods for certain relations was made on the best results of experiments with particular methods.

We found out that hybrid methods extract more terms and different relationships than standard methods separately and construct more connective thesaurus graph that can be an indication of more complete model of the subject area and can be useful for applications. Besides, they raise recall of hyponym-hypernym relations and leave precision of extracted terms and recall of synonym relations on the same level.

However the discovered effect of improvement can be dependent on the corpus in use or subject area, the idea of hybrid methods looks prominent in the perspective of automatic generation of specialized thesauri, which is a topical task due to possible economy of expert's efforts. Further investigation of the features of hybrid methods and creation of approaches for automatic assessment of particular methods for particular corpora with subsequent synthesis of hybrid methods providing robust results looks like a good direction for future research in this area.

Acknowledgment

The research was supported by the grant of the President of Russian Federation for state support of young Russian scientists (project MK-5456.2016.9).

References

[1] J. Aitchison, A. Gilchrist, and D. Bawden, *Thesaurus construction and use: a practical manual.* Psychology Press, 2000.

[2] N. Astrakhantsev, D. Fedorenko, and D. Y. Turdakov, "Methods for automatic term recognition in domain-specific text collections: A survey," *Programming and Computer Software*, vol. 41, no. 6, pp. 336–349, 2015.

[3] O. Medelyan, E. Frank, and I. H. Witten, "Human-competitive tagging using automatic keyphrase extraction," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, vol. 3. Association for Computational Linguistics, 2009, pp. 1318–1327.

[4] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in *Proceedings of EMNLP.* Association for Computational Linguistics, 2004, pp. 404–411.

[5] Z. Liu, W. Huang, Y. Zheng, and M. Sun, "Automatic keyphrase extraction via topic decomposition," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, 2010, pp. 366–376.

[6] K. S. Hasan and V. Ng, "Automatic keyphrase extraction: A survey of the state of the art," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* Association for Computational Linguistics, 2014, pp. 1262–1273.

[7] I. Paramonov, K. Lagutina, E. Mamedov, and N. Lagutina, "Thesaurus-based method of increasing text-via-keyphrase graph connectivity during keyphrase extraction for e-tourism applications," in *International Conference on Knowledge Engineering and the Semantic Web.* Springer, 2016, pp. 129–141.

[8] O. Ferret, "Testing semantic similarity measures for extracting synonyms from a corpus." in *LREC*, vol. 10, 2010, pp. 3338–3343.

[9] A. Leeuwenberg, M. Vela, J. Dehdari, and J. van Genabith, "A minimally supervised approach for synonym extraction with word embeddings," *The Prague Bulletin of Mathematical Linguistics*, vol. 105, no. 1, pp. 111–142, 2016.

[10] P. Wiemer-Hastings, K. Wiemer-Hastings, and A. Graesser, "Latent semantic analysis," in *Proceedings of the 16th international joint conference on Artificial intelligence.* Citeseer, 2004, pp. 1–14.

[11] L. Espinosa-Anke, F. Ronzano, and H. Saggion, "Hypernym extraction: combining machine-learning and dependency grammar," in *International Conference on Intelligent Text Processing and Computational Linguistics.* Springer, 2015, pp. 372–383.

[12] M. P. Oakes, "Using Hearst's Rules for the automatic acquisition of hyponyms for mining a pharmaceutical corpus." in *RANLP Text Mining Workshop*, vol. 5, 2005, pp. 63–67.

[13] Z. Chen, S. Liu, L. Wenyin, G. Pu, and W.-Y. Ma, "Building a web thesaurus from web link structure," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* ACM, 2003, pp. 48–55.

[14] K. Nakayama, T. Hara, and S. Nishio, "Wikipedia mining for an association web thesaurus construction," in *International Conference on Web Information Systems Engineering.* Springer, 2007, pp. 322–334.

[15] E. Lefever, M. Van de Kauter, and V. Hoste, "Evaluation of automatic hypernym extraction from technical corpora in english and dutch," in *9th International Conference on Language Resources and Evaluation (LREC).* European Language Resources Association (ELRA), 2014, pp. 490–497.

[16] B. Verginica, "Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora," in *Proceedings of First Central European Student Conference in Linguistics*, 2006.

[17] E. Mozzherina, "Avtomaticheskoe postroenie ontologii po kollektsii tekstovyh dokumentov [Automatic creation of ontology from collection of text documents]," in *Digital Libraries: Advanced Methods and Technologies (RCDL-2011)*, 2011, pp. 293–298, (in Russian).

[18] S.-Y. Noh, S. Kim, and C. Jung, “A lightweight program similarity detection model using XML and Levenshtein distance.” in *FECS*. Citeseer, 2006, pp. 3–9.