A Survey On Thesauri Application In Automatic Natural Language Processing

Ivan Shchitov, Ksenia Lagutina, Nadezhda Lagutina, Ilya Paramonov, Andrey Vasilyev
P.G. Demidov Yaroslavl State University
Yaroslavl, Russia
ivan.shchitov@e-werest.org, lagutinakv@mail.ru, lagutinans@gmail.com, {Ilya.Paramonov,Andrey.Vasilyev}@fruct.org

***Abstract*—This paper is devoted to investigate efficiency of thesauri use in popular natural language processing (NLP) fields: information retrieval and analysis of texts and subject areas. A thesaurus is a natural language resource that models a subject area and can reflect human expert's knowledge in many NLP tasks. The main target of this survey is to determine how much thesauri affect processing quality and where they can provide better performance. We describe studies that use different types of thesauri, discuss contribution of the thesaurus into achieved results, and propose directions for future research in the thesaurus field.**

I. Introduction

In scientific literature the term "thesaurus" is used for different linguistic resources [1]. In tasks of automatic NLP researchers most often use two types of them. The first type is related to specialized information retrieval thesauri describing relationships between terms of a certain subject area. The second type denotes linguistic tools such as WordNet and EuroWordNet that are large general purpose thesauri constructed manually. Their structure represents a hierarchical system of synonym groups of words. Both types have much in common including semantic relationships between terms and hierarchical structure. Therefore, the thesaurus used in an automatic NLP system can be defined as a dictionary, where words or phrases of a natural language are collected and different types of relationships are assigned between them: synonymous, hypernym-hyponymous, associative, etc.

Initially thesauri were designed for use by people, for example, for manual indexing of texts. Nowadays tasks of automatic NLP require automation more and more because of increasing volume of data. When extracting linguistic information from texts, the human expert uses not only a thesaurus or dictionary, but also the knowledge of subject area's peculiarities. In the automatic processing of texts there are no mediator between a text and algorithm's result. There is only an automatic process and a thesaurus that should include both traditional information and knowledge that the expert uses to analyze the text. Thus, the thesaurus intended for automatic processing, should contain much more data about the structure and details of the subject area than traditional thesauri, more terms and relationships between them. Unfortunately, there are a few thesauri that contain an amount of information sufficient for automatic qualified processing of texts. This problem is repeatedly underlined by researchers [2].

Modern linguistic resources, such as general purpose thesauri of the English language, WordNet, and specialized medical thesaurus MeSH, were designed taking into account the described problem, and contain more information than thesauri for manual NLP. There is a large number of works where tasks of processing texts are solved using thesauri, both existing and constructed by researchers. The use of such resources has a different degree of success, and for many tasks thesauri application proved to be effective [3], [4].

Unfortunately, there are no works that generalize experience of use of such linguistic resources. Most of surveys are devoted to methods of solution of separate problems and thesauri application is usually considered as one of applied approaches. In this paper we set a goal to make a step to elimination of this gap and consider the peculiarities of using thesauri in different tasks of automatic NLP and analyze quality of obtained results, involving our own experience in this area [5].

For the survey we chose NLP problems where different thesauri were applied most frequently: query expansion, document indexing, information retrieval, text and sentiment classification, text summarization, and construction of information and expert systems. Our main focus was of numerical estimation of efficiency of thesauri application and determination of the most promising directions for further thesauri application.

The rest of the paper is structured as follows. Section II describes the thesauri use in information retrieval tasks. In Section III we overview studies devoted to thesauri in text analysis. Section IV discusses thesauri application for analysis of a subject area. Conclusion summarizes advantages and disadvantages of thesauri in NLP and proposes directions for further research.

II. Thesauri in information retrieval

Information retrieval is the task of searching documents, including texts, that are relevant to a query inputted by a user. This process always includes preprocessing of texts and user queries. The former concerns document indexing and extraction of keywords, the latter includes preparation of additional queries that help to find more relevant documents. Both tasks require processing of linguistic information and can be solved using a thesaurus. Besides, the thesaurus often represent a structure of a subject area and can be integrated in a core of a search engine.

A. Thesauri in query expansion

To get the most complete search results, queries sometimes need to contain synonyms of the words contained in the initial

query, because the same concept in documents can be reflected by different words [6]. Full text search engines usually return documents containing only words that are specified in the user query and do not take into account close concepts in subject area. For example, in some cases "machine" is "car", "transport", and "vehicle". Using simple search engines users have to list all the synonyms of the term every time when they want to get more information. If the engine uses a general purpose or specialized thesaurus, the words associated with the user's query can be found and added to the query automatically, and then in the search results the system would return both documents that contain the word entered by the user and documents containing related words.

Studies on the topic of query expansion have been held for a long time, and the main pros and cons of this task are already known. There are a lot of works in the scientific literature that apply different types of thesauri to process various test collections of documents related to different subject areas, and different search models and methods for query expansion.

A large group of research is focused on the use of the general purpose thesaurus WordNet. Initially, most of these studies showed significant deterioration or insignificant changes in the quality of information retrieval [7], [8], [9]. The main problem with thesaurus application was that user queries were extended with terms from improper subject areas that have different semantics. To solve this problem, researchers used methods that resolve word sense disambiguation in queries and documents in order to filter out unnecessary terms. This approach allows to get noticeable improvement in the results of the search with a thesaurus.

One of such positive results is described by Liu et al. [10]. The authors process only short queries consisting of two or three words. As a basic model of information retrieval the probabilistic model OKAPI-BM25 is used. The model is complemented with the technique of the search by phrases and query expansion using WordNet. After word sense disambiguation, synonyms, hyponyms, and words from definitions of terms are added to queries. The authors test possibility of query expansion, and also compute the weight of the extension. Then they apply rules that allow to choose best queries basing on results of the previous step. Experiments on TREC9, TREC10, and TREC12 text datasets prove that the developed approach provide the better mean average precision (MAP). For example, for the TREC9 dataset, MAP increased from 18.12 % for the algorithm without thesauri to 26.13 % with it, for the TREC10 — from 18.34 % to 25.74 %, and for one of TREC12 datasets — from 27.79 % to 33.46 %.

Besides the general purpose resource WordNet there are a medical thesaurus Unified Medical Language System (UMLS) that also can be successfully used in query expansion. Lu and Mu [11] combine UMLS, the search engine Lemur Toolkit, and the MEDLINE medical texts. They apply the specialized thesaurus in the similar way as the previous work's authors work with WordNet. The result is increasing precision of information retrieval from 40.45 % (without thesauri) to 43.23 % and recall from 43.54 % to 56.17 %.

The UMLS terms can be mapped by the MetaMap tool. It was originally created for indexing of biomedical texts by UMLS terms and searching by them, but can be applied in query expansion too, for example, by Abdulla et al. [12]. Initial queries are processed using MetaMap that searches for suitable texts among its own dataset. Then the algorithm choose most frequent UMLS terms from found texts and expand the query by them. New queries are sent to the Indri system that process TREC 2006 and 2007 Genomic datasets with which authors conduct basic experiments. As a result, MAP increases from 25.7 % (without thesauri) to 31.1 %.

In addition to manually constructed thesauri, automatically generated ones are also often used to expand queries. Mandala et al. [13] apply several thesauri that were created in different ways. Particularly, they compare three thesauri and their combinations: the manually created thesaurus WordNet, a thesaurus built automatically by a method based on terms co-occurrence, and an automatic thesaurus constructed by a predicate-argument (morpho-syntactic) method. The proposed algorithm weights query terms taking into account thesaurus relationships and uses semantic similarity measures. Terms with the best weights expand queries. Researchers underline that the highest results are achieved by a combination of all three types of thesauri. In experiments authors used the search engine SMART 11.0 and the TREC-7 dataset. For full-text search they increase MAP from 19.76 % without thesauri to 25.73 %, in the case of the combination of all thesauri.

Pérez-Agüera and Araujo [14] generate and use two automatic thesauri in one of steps of a pseudo-feedback method for query expansion. The probabilistic methods of thesauri generation are based on first results that the engine issues to the original user query, and different measures that compute terms co-occurrence. Then as in similar query expansion methods, terms are weighted and filtered by the result. Experiments, conducted with the Lucene search system and EFE94 dataset, show that MAP increased from 40.06 % to 49.64 %, in the cases with and without thesauri. Authors also discovered that the best results are observed for a combination of two automatically constructed thesauri.

The Table I summarizes the best results from described articles. The columns "Texts" and "Queries" contain number of texts in the processed dataset and user queries correspondingly. *P* and *R* mean standard measures precision and recall. The sign "-" means that there is no information about it in the article.

From the table we can see that the algorithm based on the specialized thesaurus UMLS provides the highest MAP. The second from the best measures was achieved by the method that use two automatic thesauri. Thus, specialized thesauri outperforms general purpose ones.

Comparing cases with and without thesauri, we can conclude that in all studies thesauri significantly improve search quality. For example, the maximum increasement by 9 % for MAP is achieved by Pérez-Agüera and Araujo's method that uses two automatic thesauri. Lu and Mu's algorithm increases recall by 13 % using the UMLS thesaurus. Therefore, thesauri prove their efficiency in query expansion.

An important feature is that in almost all studies authors use primarily thesaurus' synonymous and associative relationships. Hierarchical ones and influence of different types of semantic relationships is not studied sufficiently.

TABLE I. THE BEST RESULTS IN QUERY EXPANSION

Article	Thesaurus	Dataset	Texts	Queries	*P*	*R*	MAP
Liu et al. [10]	WordNet	TREC-9, 10, 12	1 trillion	100	-	-	33.46
Lu and Mu [11]	UMLS	MEDLINE	1 033	30	43.23	56.17	56.27
Abdulla et al. [12]	UMLS, MetaMap	TREC-6, 7 Genomic	1 trillion	100	-	-	31.10
Mandala et al. [13]	WordNet, auto	TREC-7	528 155	50	-	-	25.73
Pérez-Agüera and Araujo [14]	auto	EFE94	215 000	-	-	-	49.64

B. Thesauri in document indexing

Document indexing is a task of associating texts with keywords that are used later for retrieval. Application of thesauri in this process is mapping documents by thesaurus terms and their synonyms, hyponyms, hypernyms, associations, etc.

Callejas et al. [15] mark documents using the MetaMap system mentioned in the previous subsection, and terms of the UMLS thesaurus. Then they estimate indexing quality running the Indri search engine. In comparison with the standard keyword-based algorithm that is default for the engine, the proposed method is slightly better, the precision increased from 49 % to 53 %.

El-Haj et al. [3] index social and economic texts from the UK Data Archive database. They apply the HASSET thesaurus created manually for the purpose of indexing this database. The algorithm uses the thesaurus to select terms that are mapped with texts. The words from the input text are compared with the terms of the thesaurus and their synonyms. Next, the terms are filtered according to the frequency of their appearance, the place in the text, and other features. This approach allows to index documents only by frequent terms, and their set is limited by the thesaurus. Experiments on 1353 documents show precision of 47 %, recall of 73 %, and F-measure of 43 %.

Willis and Losee [16] analyse the impact of the thesaurus structure to subject indexing. The authors developed a method for indexing process that include the random walk algorithm. This algorithm takes into account different semantic relationships with different probabilities, thus thesaurus relationships affect indexing results in a more or a less degree. In experiments with four manual thesauri: AGROVOC, HEP, NALT, and MeSH, the authors varied probabilities for narrower terms (hyponyms), broader terms (hypernyms), and associations. The best average precision of 0.1901 is quite low. This result is achieved by different probabilities for different thesauri: three thesauri of four require high probability for associations and low probabilities for other relationships. For the fourth thesaurus the case is the opposite: broader terms influence stronger than others.

The comparative quality of methods from these three article is shown in Table II. The sign *F* in the table marks F-measure, other designations mean the same as in the previous tables.

When comparing the considered methods, we can see that all of them use specialized thesauri. Two methods achieve good precision of 47–52 that demonstrate thesauri efficiency. The last work do not show high results, but it estimates influence of semantic relationships to information retrieval. Although low results of this investigation do not allow to do proper conclusions about impact of different relationships, the idea to apply hierarchical and synonymous relationships differently seems reasonable and perspective for future research.

C. Integration of thesauri into the search model

Thesauri also can be integrated directly into the search model. In this case they become an essential part of an algorithm. Unfortunately, there are few investigations that apply thesauri in such a way.

Shutze et al. [17] consider a method for information retrieval that automatically constructs a thesaurus from the text corpus basing on lexical co-occurrence. The thesaurus contains a set of all words from the corpus where each word is correlated to the “thesaurus vector”. The vector corresponds to words and its attributes represent how many times the words appear together in the text. Authors propose the model based on the constructed thesaurus. They represent documents as the “context vectors” that are the weighted sums of the thesaurus vectors in the context. Users’ queries are represented in a similar way. And then authors determine the correspondence between documents and the user query as the cosine proximity between the each document of the context vectors and the query. The proprosed model have been tested on the ARPA Tipster reference corpus. And the average precision is increased from 0.271 to 0.3 compared to the standard vector search model.

Cao et al. [18] study statistical language model of the information retrieval. They state the model is based on the words independence. But actually they change the model in such a way that the interrelation of words is taken into account. Authors expanded the language model by the word relationships. The relationships are taken from two sources: one is from co-occurences of terms in the dataset and the other is from the well-known WordNet thesaurus. The integrated model has been tested on TREC datasets. As a result, the average precision is increased on 5 %.

Table III displays the best results of information retrieval with the integrated thesaurus. In both cases quality is low, especially in comparison with results from previously reviewed topics. Besides, in both articles authors compare performance of the proposed algorithms with standard ones and algorithms’ results do not significantly differ. We can conclude that the thesaurus becomes a good natural language tool when it is used as a supplementary resource, but not when the search model has it as a part of its core.

III. THESAURI IN TEXT ANALYSIS

Text analysis is a field of computer science that includes a lot of tasks whose goal is to organize text documents. Most of this tasks needs a subject area model that describes area’s

TABLE II. THE BEST RESULTS IN DOCUMENT INDEXING

Article	Thesaurus	Dataset	Texts	*P*	*R*	*F*
Callejas et al. [15]	UMLS, MetaMap	TREC 2012	93 551	52	-	-
El-Haj et al. [3]	HASSET	UK Data Archive	1 353	47	73	43
Willis and Losee [16]	MeSH	NLM-500	500	19	-	-

TABLE III. THE BEST RESULTS IN THESAURI INTEGRATION INTO THE SEARCH MODEL

Article	Thesaurus	Dataset	Texts	Average precision
Schütze and Pedersen [17]	auto	ARPA Tipster	173 000	30.0
Cao et al. [18]	WordNet	Wall Street Journal (TREC)	74 520	26.2

structure and characteristics, and allows to extract linguistic data from raw texts. One of such model is a thesaurus. We chose three text analysis tasks where thesauri are widely applied in recent studies: text and sentiment classification and text summarization.

A. Thesauri in text classification

Topical text classification is the task to divide a set of texts into several classes. Most of methods, that solve this problem, are structured as follows. They compute some features of text's words, generate a vector for each text, and apply a standard classifier like SVM.

A througout study of text classification methods is presented in the survey [19]. Authors investigate more than 100 research papers that were published before 2014. Thesauri are used in three works and everywhere they are combined with mathematical methods that are standard for this field. The main disadvantage of this survey is that it does not describe numerical results of the experiments and does not compare performance of the methods. Therefore, there is a need for the additional research that allows to estimate efficiency of thesaurus application.

We studied four recent works that use a thesaurus as a significant part of a classification algorithm.

The GENIE system [20] and the method based on the character-level ConvNets [21] use a thesaurus only as a dictionary. In both cases authors extract a list of keywords from texts and simply extend this list by thesaurus terms related to words from the initial set. Then these keywords are taken as input for a novel classifier. The first system uses a geographical names dictionary and Eurowordnet thesaurus and increases precision and recall by 20 % comparing with the case without a thesaurus. The second method uses the WordNet thesaurus and decreases the number of errors by 8–15 %.

Nagaraj et al. [22] use popular lexical resources WordNet and Wikipedia. Unlike previous works, authors take into account relationships between terms. They calculate the number of related terms for each word in both thesauri and take it as a vector feature for a standard classifier. The highest F-measure of 85–95 % is achieved on 20NewsGroup and Classic3 datasets that is better then the system that used only Wikipedia, by 10–15 %.

Sanchez-Pi et al. [23] also uses thesaurus relationships. They apply the OpenOffice Brazilian-Portuguese thesaurus as an auxiliary tool for corresponding texts and terms from the ontology that they constructed. Firstly the algorithm search for each word the closest terms in the thesaurus and extract ones that are also contained in the ontology. Then a novel classifier processes texts based on data extracted on the previous step. Precision and recall of the result are higher by 7–10 % in comparison with SVM and equal to 60–70 %. To increase classification quality authors enriched the thesaurus and ontology by additional terms and relationships with expert's help and changed a similarity measure that is used to find corresponding terms and texts. These improvements allow to achieve precision and recall of 96 %. The great results can be explained by expert's work that provide high quality of the used lexical resource.

All best results from the described articles are united in the Table IV. The column "Classes" contains numbers of classes in the processed dataset. "Errors" is the number of total errors of the classifier. Other names and sighs means the same as in previous tables.

The highest measures for text classification are achieved by methods of Nagaraj et al. and Sanchez-Pi et al. In the first case the thesaurus was not a key part of the method, thus it is hard to say how much it affects outcomes. In the second case the thesaurus was improved by experts and it significantly increased result quality.

Summarily, the qualitatively constructed thesaurus can play an important role in text classification methods. Nevertheless, such thesauri require a lot of the human labour. There are several general purpose thesauri like WordNet and Eurowordnet constructed by professional linguists, but they may not be suitable for specific subject areas. In this case the idea of applying automatically constructed thesauri looks prominent.

B. Thesauri in sentiment classification

The main goal of sentiment classification is division of texts, sentences, or other lexical units into several classes by sentiment polarity. Most often the number of classes equals two: positive and negative, sometimes three: positive, negative, and neutral (or objective). This task differs from the text classification due to nature and length of the text under analysis, which can include sentences and word combinations. The base algorithm to perform the task includes vector's features calculation and application of standard classifiers.

A very popular lexical resource for sentiment classification is SentiWordNet. It contains a set of common English terms and vectors with three weights for each term. Weights mean probabilities of different sentiments and mark how likely is

TABLE IV. THE BEST RESULTS IN TOPICAL TEXT CLASSIFICATION

Article	Thesaurus	Dataset	Texts	Classes	*P*	*R*	*F*	Accuracy	Errors
Garrido et al. [20]	Eurowordnet	Spanish news	11 275	5	75	79	77	-	-
Zhang et al. [21]	WordNet	DBPedia	45 000	14	-	-	-	-	1.31
Nagaraj et al. [22]	WordNet, Wikipedia	Classic3	3 891	3	79–98	-	80–98	80–98	-
Sanchez-Pi et al. [23]	modified OpenOffice	OHS	500	2	96	96	87	94	-

that this term is positive, negative or objective. The sum of term's weights equals 1.0.

Although it is difficult to call SentiWordNet a thesaurus, because it does not have any semantic relationships, we include into the survey works that use it as a natural language tool.

Padmaja et al. [24] process long political newspaper articles where they extract negative sentences. Authors compare three methods and all of them on the first step take sentiment weights for terms from SentiWordNet. In result the best precision, recall, and F-measure are 79, 77, and 78 %. In this work SentiWordNet is used only as source of initial information about word's polarities, thus it is hard to estimate its own contribution to the quality of the result.

Hung and Lin [25] complement SentiWordNet terms with additional polarities that depend on a subject area. Firstly the described algorithm finds for each term the weight with the largest absolute value out of the three weights in the SentiWordNet. If this weight corresponds with positive or negative sense then the term has the same polarity. Else the term is neutral, therefore authors compute its sentiment polarity in the following way. Firstly they mark sentences as positive or negative depending on the polarity of terms contained there: the sentence is positive if the sum of its term weights is positive and conversely for negatives. Secondly the neutral word gets the polarity of sentences where it appears more frequently. Finally all terms has the concrete positive or negative polarity weight that are used as features in vectors for the SVM classifier. Experiments on 2000 IMDB texts show accuracy of 74–79 %.

Bollegala et al. [26] automatically construct a thesaurus with associations and their polarity weights. They take as input a set of texts that are already marked as positive or negative. A term sentiment depends on the polarity of the texts where it appears more often. The terms appearing in one sentence are considered to be associated in the thesaurus. Weights of these associations also depend on term frequencies. Further, the classification algorithm calculates weights of each term-text pair using the number and weights of relationships between the term and other terms in the text. This features form text's vectors that are classified by the Maximum Entropy algorithm. Experiments with Amazon reviews show one of the highest results in state-of-the-art: the accuracy is about 78–85 %. The proposed method is positioned as being independent of the subject area, therefore it can be used for any type of reviews.

Almatarneh and Gamallo [4] also generate a sentiment thesaurus fully automatically and prove its efficiency in the classification task. They take as input texts of IMDB reviews and their ratings that are numbers from 0 to 10. The algorithm compute polarities for terms taking into account term frequency and ratings of reviews where the term appears. Then term polarities are used in feature vectors for standard classifiers. Experiments show the F-measure of 76–83 %.

Kamps et al. [27] solve the similar task of adjectives classification by sentiments. They calculate for each term the number of WordNet synonyms between it and several fixed terms whose sentiment polarity is known beforehand, for example, terms "good" and "strong" are uniquely considered as positive and "bad" and "weak" are negative. The difference between such numbers marks term polarity. Accuracy of classification with this lexical resource is quite good: 61–71 %.

Highest results of sentiment classification are contained in the Table V. The column "Units" contains the number and type of natural language elements that are classified. "Auto" in the "Thesaurus" column means that authors generate the thesaurus without the help of an expert. From the table we can see that algorithms with automatically created thesauri outperforms others. One of the main reason for this trend is that such thesaurus models the subject area and describes its particular qualities of term sentiments, thus methods that use them are more finely tuned.

Investigations devoted to the sentiment classification usually take into account only one type of thesaurus relationships or consider all relationships as associations. Unfortunately, significance of different relationships in subsphere of document classification is not studied sufficiently. Thus, more deep using of the thesaurus structure can be a goal for future investigations.

C. Thesauri in text summarization

Text summarization is devoted to automatic generation of short summaries for long texts. Summarization algorithms extract from the text most significant sentences. Thesauri are quite rarely used for solutions of this problem.

The thesaurus WordNet is an integral part of the algorithm for automatic text summarization in the research of Pal et al. [28]. The algorithm compares definitions of WordNet terms with input texts and computes for each text's sentence the number of overlaps with definitions. Finally sentences with the largest measure form the final summary. Experiments show that this algorithm provides high precision and recall of 81–89 % in the cases when texts contain a small number of named entities. Authors state that if a text has a lot of such terms, summarization quality will be less. This is the main limitation of the proposed algorithm, and, probably, can be partly overcomed by replacing the thesaurus with a larger one. It should be noted that this approach assumes the use of the thesaurus only as a dictionary, and the semantic relationships between terms are ignored, thus, their application can be the direction for further investigations.

TABLE V. THE BEST RESULTS IN SENTIMENT TEXT CLASSIFICATION

Article	Thesaurus	Dataset	Units	*P*	*R*	*F*	Accuracy
Padmaja et el. [24]	SentiWordNet	Newspaper articles	513 texts	79	77	78	-
Hung and Lin [25]	SentiWordNet	IMDB	2 000 texts	-	-	-	74–79
Bollegala et al. [26]	auto	Amazon	1 600 texts	-	-	-	85
Almatarneh and Gamallo [4]	auto	IMDB	2 000 texts	81	84	83	-
Kamps et al. [27]	WordNet	General Inquirer	3 000 adjectives	-	-	-	61–71

Sankarasubramaniam et al. [29] compare sentences from input texts with terms from Wikipedia. They propose a text ranking algorithm that creates a sentence-by-term matrix, the values of the cells are numbers of appearances of Wikipedia's terms in the corresponding sentence. Then the algorithm calculates the ranks of sentences using graph methods. Authors experimented with the DUC 2002 dataset of news texts and got precision, recall, and F-measure of 57, 50, and 51 % correspondingly. This approach allows to generate summaries with a large number of proper names and specific terms of a subject area, therefore, it well suits for news.

Sakamoto et al. [30] also propose a graph algorithm for automatic text summarization. It builds a graph with words and sentences as vertices and uses the Japanese general purpose thesaurus to calculate weights of word-word edges in the following way. The more common hyperonyms words have, the smaller a distance between them (or an edge weight) is. Experiments with texts that are related to different topics and do not have the common one, show a quite low quality of the final summaries: average R-Precision of 0.338. Most probably, the main reason for this result is that the algorithm does not take into account the peculiarities of the subject area.

The Table VI describe the best values of standard measures for summarization methods. The highest results are achieved by the algorithm that is based on WordNet as a dictionary and depends on input text's features that was described earlier. Other methods show significantly lower quality. Probably, it can be increased by a more sophisticated application of thesauri similar to techniques used in text and sentiment classification, where thesauri already proved their efficiency.

IV. THESAURI IN ANALYSIS OF A SUBJECT AREA

Analysis of a subject area concerns organizing knowledge about the concrete area, including processing text documents and providing them to users. As well as text analysis, analysis of a subject area requires creation of the area's model that accumulates semantic information. As such a model, a thesaurus can be a natural language tool for methods that process texts and provide navigation on them.

A. Thesauri in information systems

Information systems usually contain large collections of textual data. Thus, such systems should have a good navigation on their texts, so users can quickly find nesessary information. A thesaurus represents a structure of a subject area, therefore, it can be used as a navigation model.

Garcia et al. [31] think that one of the main problems that the the specialized data repositories creators face is how to facilitate access to the digital resources. Authors set a task to find out whether it is possible to make retrieving the digital resources easier and convenient using the information visualization. For this purpose authors developed a tool to visualize the data array. The authors uses 42800 digital resources from the Europeana digital library as the dataset. The sample collection contains concepts and topics of art, culture and heritage defined by the Art and Architecture Thesaurus (AAT). The thesaurus is used as a knowledge representation scheme. Authors select a set of the AAT thesaurus terms to define concepts and use Europeana API for creating connections between metadata and digital resources. Then this part of the thesaurus was integrated into the navigation system based on the hierarchical structure where all terms and connections between them are displayed graphically. Each term corresponds to a set of the digital resources. The only one restriction is the size of the thesaurus' graph. The problem is that the more levels of hierarchy the thesaurus has, the more difficult it is to gain access to the digital resources.

The thesaurus quality was determined by 16 participants with basic knowledge in handling Web applications. The users evaluated the ease and the speed they can find necessary digital resources using visual search interface. According to the test results the participants made scores and the following rating was drawn up.

- 9.4% participants have rated searching as "excellent";
- 53.1% — "good";
- 18.8% — "regular";
- rest of participants have rated searching as "low" and "very low".

We also applied a thesaurus to extract better keywords for information system's texts and improve navigation through the system [5]. We developed a method that combines a known keyword extraction algorithm and our thesaurus-based procedure. The procedure complement the list of keywords selected by the existing algorithm, with related thesaurus terms: associations and hypernyms. The next step is to compute how many texts corresponds with each keyword and remove keywords with minimum scores.

Experiments with the Open Karelia system and general purpose Russian thesaurus RuThes show that precision, recall and F-measure of keyword extraction increase from 57.5 %, 24.1 %, 34.0 % for the algorithm without a thesaurus to 70.0 %, 76.8 %, 72.7 % with it correspondingly. Besides, we estimated efficiency of navigation and chose parameters of the system's graph as quality measures. This graph has texts as vertices and each text pair have a common edge if they have a common keyword. After keyword extraction the graph has 11 connected components in the case without a thesaurus and 1 in the case

TABLE VI. THE BEST RESULTS IN TEXT SUMMARIZATION

Article	Thesaurus	Dataset	Texts	*P*	*R*	*F*	average R-Precision
Pal et al. [28]	WordNet	Articles	50	89	89	89	-
Sankarasubramaniam et al. [29]	Wikipedia	DUC 2002	567	57	50	51	-
Sakamoto et al. [30]	Japanese	Web source	48	-	-	-	33.8

with it. Also each text has 30–50 keywords in average (with a thesaurus) instead of 2–13 (without a thesaurus). Thus, the thesaurus-based method allow to simplify navigation providing more links between texts in the information system.

Summarily, thesauri can significantly improve navigation through the information system that contains a lot of texts. The main reason is that the methods actively use thesaurus relationships and reflect thesaurus structure to the system of navigation. Therefore, for information systems a thesaurus is a good model of a subject area.

B. Thesauri in expert systems

Expert systems also provide users a lot of textual data and, during their construction, researchers face the same issues that appears in creation of information systems: text analysis and navigation on documents. And thesauri can be applied in the similar way.

In the paper [32] Garrido et al. write about an expert system called "Hypatia" that uses the EuroWordNet thesaurus. Hypatia is created to facilitate the work of users with the company's documentation. The thesaurus is used in the system for the document classification and for resolving disambiguation by searching synonyms for termsf. The expert system quality is evaluated by focus group of 100 people (20 of them are experts of working with a documentation, 80 are not). Participants evaluated the system features by several criteria (relevance, usefulness, manageability, novelty) and on a scale from 0 to 10. It should be noted, that part of the expert system worked with the thesaurus received the most low scores compared to rest of modules: 7 from the common users and 8.5 from experts.

Expert system from the paper [33] is used in the information portal to evaluate reactivity of organic molecules in radical reactions. Two automatically constructed thesauri are used in this system and they have the rather specific structure. Thesauri contain tuples of organic molecules and radicals (atoms), indexes to physical and chemical properties: constants of the bond dissociation energy and radical reactions rate. This data is used for the chemical reaction identification and its parameters determining. Thus, thesauri are used as the specific subject area knowledge base. In the paper the system quality evaluation is not given.

We can conclude that thesauri are used in expert systems either as a tool for processing documents, or they represent the structure of a specific subject area. Further there are two cases of the thesaurus quality evaluation. In the first case evaluation can be carried out according to criteria usual for the NLP methods. In the second case the thesaurus can not be separated from the expert system, therefore we can evaluate only work of the system as a whole.

V. CONCLUSION

Analysis of modern studies about thesauri application to NLP problems allows to draw the following conclusions. The use of thesauri in information retrieval systems leads to a slight improvement in the quality of the problem being solved. Nevertheless, the best results of the thesaurus methods in absolute values are comparable with the results of other methods, therefore, they can be successfully used to solve information retrieval problems.

Thesauri show the great results in solution of text analysis tasks: text and sentiment classification, text summarization. Particularly, in these tasks, as well as in most others, the efficiency is increased and the parameters of NLP systems are improved through the application of a specialized thesaurus. Since there are very few such thesauri, it seems promising to study the methods of automatic thesaurus construction for specific subject areas and different languages.

The thesaurus application for the analysis of subject areas is quite successful. In this case the thesaurus becomes an effective model of the area or a part of such a model. Possible future research in this field can concern generation and use of automatic specialized thesauri.

Besides, we want to note some peculiarities of thesaurus details application, especially relationships between terms. In many studies the thesaurus is used only as a dictionary and relationship between terms is not conserned. Nevertheless, research shows that in almost all tasks of NLP, relationships between terms play an important role, and most recent studies use them more actively. However, types of relationships usually are not differentiated and the impact of different semantic relationships on the result quality is not investigated throughly. Thus, this problem represents a broad direction for future research and a possible way to improve NLP quality.

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] A. Kilgarriff and C. Yallop, "What's in a thesaurus?" in *LREC*, 2000, pp. 1371–1379.

[2] N. Loukachevitch and B. Dobrov, "The sociopolitical thesaurus as a resource for automatic document processing in russian," *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, vol. 21, no. 2, pp. 237–262, 2015.

[3] M. El-Haj, L. Balkan, S. Barbalet, L. Bell, and J. Shepherdson, "An experiment in automatic indexing using the HASSET thesaurus," in *Computer Science and Electronic Engineering Conference (CEEC), 2013 5th*. IEEE, 2013, pp. 13–18.

[4] S. Almatarneh and P. Gamallo, "Automatic construction of domain-specific sentiment lexicons for polarity classification," in *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, 2017, pp. 175–182.

[5] 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.

[6] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval*. Cambridge university press Cambridge, 2008, vol. 1, no. 1.

[7] E. M. Voorhees, "Query expansion using lexical-semantic relations," in *SIGIR'94*. Springer, 1994, pp. 61–69.

[8] A. F. Smeaton, F. Kelledy, and R. O'Donnell, "TREC-4 experiments at dublin city university: Thresholding posting lists, query expansion with WordNet and pos tagging of spanish," *Harman [6]*, pp. 373–389, 1995.

[9] E. M. Voorhees, "Using WordNet for text retrieval," *Fellbaum (Fellbaum, 1998)*, pp. 285–303, 1998.

[10] S. Liu, F. Liu, C. Yu, and W. Meng, "An effective approach to document retrieval via utilizing WordNet and recognizing phrases," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2004, pp. 266–272.

[11] K. Lu and X. Mu, "Query expansion using UMLS tools for health information retrieval," *Proceedings of the American Society for Information Science and Technology*, vol. 46, no. 1, pp. 1–16, 2009.

[12] A. A. A. Abdulla, H. Lin, B. Xu, and S. K. Banbhrani, "Improving biomedical information retrieval by linear combinations of different query expansion techniques," *BMC bioinformatics*, vol. 17, no. 7, p. 238, 2016.

[13] R. Mandala, T. Tokunaga, and H. Tanaka, "Query expansion using heterogeneous thesauri," *Information Processing & Management*, vol. 36, no. 3, pp. 361–378, 2000.

[14] J. R. Pérez-Agüera and L. Araujo, "Comparing and combining methods for automatic query expansion," *arXiv preprint arXiv:0804.2057*, 2008.

[15] P. Callejas, A. Miguel, Y. Wang, and H. Fang, "Exploiting domain thesaurus for medical record retrieval," DTIC Document, Tech. Rep., 2012.

[16] C. Willis and R. M. Losee, "A random walk on an ontology: Using thesaurus structure for automatic subject indexing," *Journal of the Association for Information Science and Technology*, vol. 64, no. 7, pp. 1330–1344, 2013.

[17] H. Schütze and J. O. Pedersen, "A cooccurrence-based thesaurus and two applications to information retrieval," in *Intelligent Multimedia Information Retrieval Systems and Management-Volume 1*. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 1994, pp. 266–274.

[18] G. Cao, J.-Y. Nie, and J. Bai, "Integrating word relationships into language models," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2005, pp. 298–305.

[19] R. Jindal, R. Malhotra, and A. Jain, "Techniques for text classification: Literature review and current trends," *Webology*, vol. 12, no. 2, p. 1, 2015.

[20] A. L. Garrido, M. G. Buey, S. Escudero, A. Peiro, S. Ilarri, and E. Mena, "The GENIE system: Classifying documents by combining mixed-techniques," in *International Conference on Web Information Systems and Technologies*. Springer, 2014, pp. 231–246.

[21] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649–657.

[22] R. Nagaraj, V. Thiagarasu, and P. Vijayakumar, "A novel semantic level text classification by combining NLP and Thesaurus concepts," *IOSR Journal of Computer Engineering*, vol. 16, no. 4, pp. 14–26, 2014.

[23] N. Sanchez-Pi, L. Martí, and A. C. B. Garcia, "Improving ontology-based text classification: An occupational health and security application," *Journal of Applied Logic*, vol. 17, pp. 48–58, 2016.

[24] S. Padmaja, S. S. Fatima, and S. Bandu, "Evaluating sentiment analysis methods and identifying scope of negation in newspaper articles," *International Journal of Advanced Research in Artificial Intelligence*, vol. 3, no. 11, pp. 1–6, 2014.

[25] C. Hung and H.-K. Lin, "Using objective words in SentiWordNet to improve sentiment classification for word of mouth," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 47–54, 2013.

[26] D. Bollegala, D. Weir, and J. Carroll, "Cross-domain sentiment classification using a sentiment sensitive thesaurus," *IEEE transactions on knowledge and data engineering*, vol. 25, no. 8, pp. 1719–1731, 2013.

[27] J. Kamps, M. Marx, R. J. Mokken, M. De Rijke *et al.*, "Using WordNet to measure semantic orientations of adjectives." in *LREC*, vol. 4. Citeseer, 2004, pp. 1115–1118.

[28] A. R. Pal and D. Saha, "An approach to automatic text summarization using WordNet," in *Advance Computing Conference (IACC), 2014 IEEE International*. IEEE, 2014, pp. 1169–1173.

[29] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh, "Text summarization using Wikipedia," *Information Processing & Management*, vol. 50, no. 3, pp. 443–461, 2014.

[30] K. Sakamoto, H. Shibuki, T. Mori, and N. Kando, "Fusion of heterogeneous information in graph-based ranking for query-biased summarization." in *GSB@ SIGIR*, 2015, pp. 19–22.

[31] P. A. G. García, S. Sánchez-Alonso, and C. E. M. Marín, "Visualization of information: a proposal to improve the search and access to digital resources in repositories," *Ingeniería e Investigación*, vol. 34, no. 1, pp. 83–89, April 2014.

[32] A. L. Garrido, A. Peiro, and S. Ilarri, "Hypatia: An expert system proposal for documentation departments," in *Intelligent Systems and Informatics (SISY), 2014 IEEE 12th International Symposium on*. IEEE, 2014, pp. 315–320.

[33] V. E. Tumanov, E. S. Amosova, and A. I. Prokhorov, "Developing an expert system integrated into the information portal to evaluate reactivity of organic molecules in radical reactions," in *MATEC Web of Conferences*, vol. 76. EDP Sciences, 2016, p. 04007.