Applicability of Automatically Generated Thesauri to Text Classification in Specific Domains

Ksenia Lagutina, Ivan Shchitov, Nadezhda Lagutina, Ilya Paramonov

P.G. Demidov Yaroslavl State University,
Sovetskaya Str. 14, 150003, Yaroslavl, Russia
lagutinakv@mail.ru, ivan.shchitov@e-werest.org,
lagutinans@gmail.com, ilya.paramonov@fruct.org

Abstract. The paper is devoted to comparison of the quality of text classification with the use of manually and automatically generated thesauri. For this purpose, the authors applied the BM25 algorithm with word features based on thesaurus's relations between terms. The experiments, conducted with text corpora from three specific domains (medicine, economics, and sport), showed that using an automatically generated thesaurus provides nearly the same classification quality as the manually created one. These results make the authors' approach promising for text classification in many specific domains where no thesaurus is available, as it allows to avoid consumption of high amount of resources for manual thesaurus creation.

Keywords: text classification, specialized thesaurus, BM25, automatic thesaurus generation.

1 Introduction

Automated topical text classification aims to order unstructured corpora and assigns raw texts to pre-defined main topics. The most popular classification algorithms are developed and studied primarily for processing news, reviews, and spam [1,5]. These texts mostly contain common words, so their classification does not depend on a domain.

Conversely, domain-specific texts usually have non-trivial structure, many terms, and other peculiar properties that complicate text mining. One of the way to simplify processing and extract semantic information is the thesaurus use. A specialized thesaurus contains almost all domain's terms and reflects semantic relations between them [2]. Studies show that embedding a thesaurus into text processing algorithms increases classification quality significantly [4,11].

However, construction of high-quality thesauri that can be used for text mining, requires much expert's time, therefore the number of thesauri in open access is limited. An alternative for human-made thesauri is automatically created ones. They are less popular in text classification than both manually created specialized thesauri and general purpose thesauri. Also, the difference between results achieved using thesauri of different types is rarely investigated in text mining papers.

The goal of the authors' research was to compare efficiency of manually and automatically generated specialized thesauri in application to the topical text classification. Each thesaurus was used in conjunction with BM25 algorithm to take into account relations between classes and words from texts. The experiments revealed potential of both types of thesauri in use. Moreover, this comparison allowed to figure out whether the automatically created thesauri can be used in domains that do not have a thesaurus yet.

The paper is structured as follows. Section 2 overviews the state-of-the-art in text classification with thesauri. Section 3 explains the main steps of the classification method and details of the thesaurus embedding. Section 4 describes text corpora and the evaluation procedure used in experiments. Section 5 provides numerical results of experiments and explains them. The conclusion summarizes the paper, shortly describes key results and discusses future research.

2 Related work

Almost all existing algorithms that classify texts by topics have the same main steps:

1. Extraction of keyword candidates from texts.
2. Computation of statistical and linguistic features of the candidates.
3. Ranking candidates using different functions and machine learning algorithms.
4. Choice of candidates with the best scores as text classes.

A thesaurus can be embedded into such algorithms in several ways. The first one is to extend the candidate list by thesaurus terms related to keywords from the initial set. This approach is implemented in the GENIE system [4]. It classifies news texts using a geographical names dictionary and Eurowordnet thesaurus. Experiments showed that such an approach allows to increase precision and recall by 20 % comparing with the same algorithm without a thesaurus. Another example is the method that combines the thesaurus with character-level ConvNets [13]. Testing on news, reviews, DBPedia, and Yahoo Answers corpora showed that using the WordNet thesaurus decreases the number of errors by 8–15 %.

An important feature of such methods is the usage of a general purpose thesaurus. It fits well for news, reviews, and other texts with common words, but may work worse with documents from a concrete domain containing many specific terms [10].

The second way to use the thesaurus is to compute additional features for candidates. Nagaraj et al. [9] proposed a method that finds semantic relations between words in WordNet or Wikipedia and counts their number for each word.

The best results was shown on 20NewsGroup and Classic3 subsets, about 85–95 %; F-measure for the business Reuters-21578 corpus was lower, from 60 to 90 % depending on the training set size.

The system of review classification [3] divides reviews into two groups: positive and negative. The solution of this task requires sentiment analysis, so Bollegala et al. automatically created a specialized thesaurus with terms that express human emotions and opinions. The classification algorithm computes the number of thesaurus relations for each keyword in the text and adds this score to the feature vector. As a result, accuracy raised from 60–70 % to 70–80 %. Such high scores were reached because the thesaurus was built for a specific domain taking into account its features and peculiarities.

Summarily, the usage of a thesaurus allows to improve quality of text classification, and results are better if the thesaurus contains information about semantic relations between the words from the texts' domain. For text classification both manually and automatically generated thesauri can be used, but comparison of the classification quality reachable with the use of these two classes of thesauri is not properly presented in modern research.

3 Method of text classification using automatically generated thesaurus

3.1 Thesaurus creation

The main goal of the authors' research was to examine how well text classification can be performed with the use of automatically created specialized thesauri. Such thesauri were generated from several domain-specific corpora with the use of the algorithm from the authors' previous work [7]. It includes the following steps of text processing:

1. Term extraction using TextRank algorithm.
2. Construction of different semantic relations (associations, hyponym—hypernyms, and synonyms) with the use of the combination of statistical and semantic methods.
3. Isolated term removal.

This approach provides a thesaurus with the large number of semantic relations between terms. Comparison with the existing thesaurus showed that automatically constructed one has quite good recall and very high precision for synonyms.

The main advantage of the algorithm is that it processes data fully automatically and does not require any expert's work, so the thesaurus can be quickly constructed for any domain. Moreover, according to the authors' previous results, such a thesaurus usually has enough terms and relations of quite good quality. Therefore, it can provide additional information for methods that solve text mining problems.

3.2 Text classification

After construction the specialized thesauri were applied to the text classification task.

In the authors' experiment text classification is performed by the existing existing unsupervised classification algorithm BM25 [6] modified to use the thesaurus. This algorithm takes as an input a corpora, where texts are not assigned with topics, a list of classes, and an automatically constructed thesaurus from the previous section. It contains the following steps:

1. Extract all words from texts and compute their frequency.
2. For each class create the list of related thesaurus terms.
3. Rank the text-class pairs using the terms frequency and the BM25 algorithm.

Firstly, the algorithm extracts individual words from texts, builds the inverted index for each text document, and calculate word frequency.

Secondly, the thesaurus is browsed for the closest related terms of the following classes: synonyms, hyponyms, and hyperonyms of the first order. Associations are not included because they often link terms that can indicate different topic, for example "blood clot" and "heart". On the contrary, synonyms, hyponyms, and hyperonyms obviously reflect semantic relations between words from a common topic, so they can help to juxtapose texts with their classes better.

Finally, the BM25 algorithm is applied to the text-class pairs for ranking each document by the query terms. This algorithm is a popular approach for many text mining and information retrieval tasks, including text classification [12]. Also it does not require any additional parameters and training with treated samples, so it can be easily extended by a thesaurus. These features make the algorithm suitable for the authors' research.

4 Evaluation procedure and used text corpora

In the classification experiments the authors used three text corpora from different domains:

- PubMed text corpus (https://www.nlm.nih.gov/databases/download/pubmed_medline.html). It has 63 classes and 1000 medical articles with 154 850 words.
- Reuters text corpus (http://www.daviddlewis.com/resources/testcollections/reuters21578/). It has 15 classes and 1534 articles from economics domain with 294 813 words.
- BBCSport text corpus (http://mlg.ucd.ie/datasets/bbc.html). It has 5 classes and 737 texts about sport with 253 667 words.

Thesauri were generated automatically based on the chosen corpora. To compare the classification quality when using automatically and manually generated thesaurus the authors used the well-known MeSH thesaurus (https://www.nlm.nih.gov/mesh/meshhome.html) and STW thesaurus (http://zbw.

`eu/stw/version/latest/about.en.html`) for medical and economics domains correspondingly.

The tool for extracting keywords from a text corpus is based on the Topical PageRank keyword extraction algorithm [8]. It was implemented by the authors in Python programming language. The tool takes a file with a text corpus as an input parameter, reads all of texts, extracts keywords, and writes them to separate files.

The text classification algorithm is also implemented as a tool in Python. It contains three modules: parser, query processor, ranking function. The parser module reads and parses files with a list of classes and a set of texts to create the data structures for further calculations. The query processor takes each class from the list and scores the documents based on the class terms using the ranking function. The ranking function is an implementation of the BM25 algorithm.

Algorithm's outcomes were compared with results of manual classification that are provided with each corpus. For evaluation the authors chose most popular quality measures: micro-average precision, recall, F-measure, and accuracy. The precision is the fraction of documents actually belonging to given classes among all documents that the algorithm assigned to classes. The recall is the fraction of documents found by the algorithm that are belong to given classes among all documents of classes. The F-measure is the harmonic mean of the precision and recall. The accuracy is the fraction of the retrieved documents for which the classifier made a correct decision.

5 Results

The results of experiments achieved with the use of the text corpora described in the previous section are contained in Table 1 and Table 2.

Table 1 displays results in absolute values. The first column of the table contains the name of text corpora, the second one—the type of used thesauri. "Auto" means the automatically generated thesaurus based on the text corpus. The third and fourth columns display the number of texts in the corpus and the number of predefined classes. For example, texts from the PubMed corpus are classified by the following classes: toxicity, pharmacology, immunology, surgery, et al. The fifth and sixth columns display the number of text-class pairs found using the algorithm described in Section 3.2 and correct correspondence between texts and classes given in advance correspondingly.

These results show that the automatically generated thesaurus allows to find more text-class pairs compared to the manually constructed one. Comparing the original number of text-class pairs (the last column of Table 1) with the others, we can see the following trend. The algorithm with manual thesaurus leaves some texts unclassified or juxtaposes texts with smaller number of classes than it should be. On the contrary, the generated thesaurus provides too many pairs for PubMed and Reuters corpora, but for the BBCSport corpus this number is close to the original.

Table 1. Classification quality of the algorithm with manually and automatically generated thesauri in the absolute values

Corpus	Thesaurus	Total number of		Number of texts classified	
		texts	classes	automatically	manually
PubMed	Auto	1000	63	5057	3146
PubMed	MeSH	1000	63	1392	3146
Reuters	Auto	1543	15	4158	1835
Reuters	STW	1543	15	1545	1835
BBCSport	Auto	737	5	877	737

Table 2 displays results in relative values. The results show that the experiments with the automatically generated thesauri have greater recall and smaller precision than the experiments with the thesauri constructed manually. It happens because the algorithm with the generated thesaurus finds too many texts for classes. From one hand, it allows to find more right answers. From the other hand, it provides a singnificant number of redundant text-class pairs. This tendency is particularly takes part for the Reuters corpus. However, the experiment with the BBCSport corpus shows that the generated thesaurus can provide quite high precision for some texts (about 40 % against 33 and 12.5 % for the other domains).

Table 2. Classification quality of the algorithm with manually and automatically generated thesauri in the relative values

Corpus	Thesaurus	Precision	Recall	F-measure	Accuracy
PubMed	Auto	12.5	20.1	15.4	89.0
PubMed	MeSH	26.0	11.5	15.9	93.9
Reuters	Auto	32.9	74.6	45.7	85.9
Reuters	STW	55.8	47.0	51.0	92.8
BBCSport	Auto	39.9	47.5	43.4	75.2

The F-measure and accuracy differ slightly for algorithms with both thesaurus types, so the difference between them is not essential.

Summarily, the results allow to conclude that the automatically created thesaurus can be successfully applied to the classification task. Meanwhile, the approach with the automatically generated thesaurus has the following advantages.

Firstly, it is suitable for different domains. Experiments for medicine and economics domains were conducted with the same parameters of thesaurus construction and embedding and the results showed the same trend for all measures. Therefore, we can expect similar classification quality for other domains.

Secondly, the thesaurus construction goes fully automatically and does not require the expert's labour and parameters adjustment. So the thesaurus can be created and applied easily for any corpus.

Nevertheless, generated thesauri do not provide enough high precision and accuracy of results in some cases, therefore the research of their use should be continued.

6 Conclusion and future research

In this paper the authors compared quality of text classification when using manually and automatically created thesauri. The experiments showed that using manual thesauri provides higher precision and accuracy, but automatical ones allow to find more right classes and generally are not significantly behind.

The algorithm was used for two different domains with both thesaurus kinds and the result's trend was the same. So classification of corpora from other domains would probably have the similar quality.

The results of the research allows to assert that the automatically generated thesaurus is unconditionally applicable for text classification. The authors experimented with the sport news text corpus, where there is no specialized thesaurus, and got quite high measures about 40–70 %.

The future research concerns varying of thesaurus properties and classification algorithm's parameters.

The important thesaurus features are semantic relations. When thesauri are applied to information retrieval and text mining, the least used relations are associations. They are dissimilar and often have different semantics. So, the associations can be classified into several subtypes that can be differently used in text classification. This can be important for the case when the algorithm with an automatically generated thesaurus provides low precision, because it processes a large number of thesaurus' relations and extracts too many false positive text-class pairs. Probably, discrimination of the relations would allow to eliminate redundant text classes.

The algorithm's parameters are weights of words in the scoring function. During the text analysis step the algorithm calculates features for classes and keywords taking into account thesaurus's relations between them. The significance of various relations can be differentiated depending on the domain. If the scoring function would assign different weights for different relations, the algorithm would classify texts using the particular domain's features that can lead it to make smaller number of mistakes.

Summarily, the further investigation should find out types of thesaurus relations that reflect links between classes and texts better for a particular domain. It would allows to vary parameters of the classification algorithm and classify texts more precisely.

Acknowledgements. 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. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining text data, pp. 163–222. Springer (2012)
2. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus construction and use: a practical manual. Psychology Press (2000)
3. Bollegala, D., Weir, D., Carroll, J.: Cross-domain sentiment classification using a sentiment sensitive thesaurus. IEEE transactions on knowledge and data engineering 25(8), 1719–1731 (2013)
4. Garrido, A.L., Buey, M.G., Escudero, S., Peiro, A., Ilarri, S., Mena, E.: The GENIE system: Classifying documents by combining mixed-techniques. In: International Conference on Web Information Systems and Technologies. pp. 231–246. Springer (2014)
5. Jindal, R., Malhotra, R., Jain, A.: Techniques for text classification: Literature review and current trends. Webology 12(2), 1 (2015)
6. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: development and comparative experiments. In: Information Processing and Management. vol. 36, pp. 809–840. London, UK (2000)
7. Lagutina, K., Mamedov, E., Lagutina, N., Paramonov, I., Shchitov, I.: Analysis of relation extraction methods for automatic generation of specialized thesauri: Prospect of hybrid methods. In: Proceedings of the 19th Conference of Open Innovations Association FRUCT. Jyvaskyla, Finland, 7-11 November 2016. pp. 138–144 (2016)
8. Liu, Z., Huang, W., Zheng, Y., Sun, M.: Automatic keyphrase extraction via topic decomposition. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 366–376. Association for Computational Linguistics (2010)
9. Nagaraj, R., Thiagarasu, V., Vijayakumar, P.: A novel semantic level text classification by combining NLP and Thesaurus concepts. IOSR Journal of Computer Engineering 16(4), 14–26 (2014)
10. Nam, J., Kim, J., Mencía, E.L., Gurevych, I., Fürnkranz, J.: Large-scale multi-label text classification—revisiting neural networks. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 437–452. Springer (2014)
11. Sanchez-Pi, N., Martí, L., Garcia, A.C.B.: Improving ontology-based text classification: An occupational health and security application. Journal of Applied Logic 17, 48–58 (2016)
12. Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., Rebholz-Schuhmann, D.: MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 25(11), 1412–1418 (2009)
13. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in neural information processing systems. pp. 649–657 (2015)