On the Quality of Annotations with Controlled Vocabularies

Heidelinde Hobel and Artem Revenko(✉)

Semantic Web Company, Vienna, Austria
{h.hobel,a.revenko}@semantic-web.at

Abstract. Corpus analysis and controlled vocabularies can benefit from each other in different ways. Usually, a controlled vocabulary is assumed to be in place and is used for improving the processing of a corpus. However, in practice the controlled vocabularies may be not available or domain experts may be not satisfied with their quality. In this work we investigate how one could measure how well a controlled vocabulary fits a corpus. For this purpose we find all the occurrences of the concepts from a controlled vocabulary (in form of a thesaurus) in each document of the corpus. After that we try to estimate the density of information in documents through the keywords and compare it with the number of concepts used for annotations. The introduced approach is tested with a financial thesaurus and corpora of financial news.

Keywords: Controlled vocabulary · Thesaurus · Corpus analysis · Keywords extraction · Annotation

1 Introduction

The interaction between corpora and controlled vocabularies has been investigated for a long time. With today's shift to semantic computing, the classical interplay between controlled vocabularies and corpora has intensified itself even further. The research goes in both directions: improving the processing of the corpus using controlled vocabularies (for example, query expansion [10] or word sense disambiguation based on thesauri [6]) and improving the controlled vocabularies using corpus analysis [1,2]. We focus on the second direction. As application areas are created faster than the general semantic economy can keep up with, targeted and automated improvements of thesauri are of utmost importance. On the one hand, many authors agree that so far one cannot rely on a completely automatic construction or even extension of controlled vocabularies [7,8]. On the other hand, the construction of a controlled vocabulary by an expert without the aid of the corpus analysis is also hardly possible. Therefore, the most reasonable scenario is to enable the expert to curate the construction process based on the results of the corpus analysis.

We investigate the scenario when thesauri are used as a source for concepts, and all the concepts found in the documents are used as annotations.

A. Satsiou et al. (Eds.): IFIN and ISEM 2016, LNCS 10078, pp. 98–114, 2016.
DOI: 10.1007/978-3-319-50237-3_4

This paper is a preliminary report on our work in progress. This research is carried out in frames of the PROFIT[1] project.

Contribution. We introduce a measure of the quality of annotations. This measure is applicable to a pair of a thesaurus and a corpus; the value of this measure describes how well the two elements fit each others.

Structure of Paper. This paper is structured as follows. In Sect. 2 we introduce the reader to the topic of controlled vocabularies and their use for annotations. In Sect. 3, we describe our proposed approach for measuring the quality of annotations. We outline our performed case study in Sect. 4 and present our results in Sect. 4.4. We conclude the paper in Sect. 5 and present an outlook of future work in Sect. 6.

2 Controlled Vocabularies

Let a *label* L be a word or a phrase (a sequence of words). Here we understand *word* in a broad sense, i.e. it may be an acronym or even an arbitrary sequence of symbols of a chosen language. Let a *concept* C denote a set of labels. Usually we represent a concept by one of its labels that is chosen in advance (preferred label). A *controlled vocabulary* $\mathbb{V}$ is a set of concepts. We say that a *thesaurus* is a controlled vocabulary with additional binary relations [4] between concepts:

broader B a transitive relation, cycles are not allowed;
narrower N is an inverse of B;
related R an asymmetric relation.

Note that any thesaurus is a controlled vocabulary.

The concepts as defined here capture some features of a concept defined in SKOS [3]. We may consider SKOS thesauri as an instantiation of the thesauri defined here.

We understand *annotation* as a metadata attached to data; for example, a comment to a text or a description of an image. One may come up with different possible options for introducing annotations. For example, techniques for summarization of texts [11] or various techniques of text mining [15] offer multiple methods for introducing annotations. In this work we consider only the annotations that are done with a controlled vocabulary, i.e. only the concepts from a fixed controlled vocabulary can be used in annotations. Therefore, an annotation with a controlled vocabulary is a set of concepts from the controlled vocabulary. Among the advantages of making annotations with controlled vocabulary are the following:

[1] projectprofit.eu.

Consistency. Even if the annotations are done by different annotators (humans or algorithms) they are still comparable without any further assumptions or knowledge;

Translation. Though we cannot manipulate the annotated data, the annotations themselves may be translated, moreover, the annotator and the consumer may not even speak a common language;

Control over Focus. With the control over the controlled vocabulary one has the control over the focus of annotations, hence controlling the individual aspects of the data that require attention;

Alternative Labels. Thanks to multiple labels that may be introduced for a concept one may discover the same entity in different annotations even if the entity is represented by different synonyms.

Moreover, one may introduce additional knowledge about the concepts in the vocabulary in order to be able to perform advanced operations with annotations. For example, if one introduce a proximity measure over distinct concepts than one may compute finer similarity measure over annotations. Therefore, we may say that annotations with controlled vocabularies may be seen as a proxy for performing operations over data.

Though different types of data are important for the applications, the textual data offers unique chance to improve the thesaurus from the data. Moreover, in this case the annotations and the data are in the same format. The current state-of-the-art in combined approaches of thesauri and corpora allows to annotate texts in the corpora according to the described concepts in the thesauri and to employ phrase extraction from the corpora to identify new concepts for the thesauri.

3 Method

Goal. In this paper, we study an approach to assess and measure the quality of fit between a thesaurus and a corpus. The fit is understood in the sense of suitability of a given thesaurus to annotate a given corpus. The annotations are done automatically through finding the concepts from the thesaurus in the corpus.

We approach the goal via measuring the number of annotations found in each text. However, this number should correlate with the amount of information contained in the text. One way to assess the amount of information is to find keywords on a corpus level and check how many of those are contained in the text [13,14].

3.1 Keywords

We utilize different methods to find keywords.

Mutual Information. Mutual information provides information about mutual dependence of variables and can be used to estimate if two or more consecutive words in a text should be considered a compound term that is formed of these words [9]. The idea is that if words are independent then they will occur together just by chance, but if they are observed together more often than expected then they are dependent. The definition of the mutual information score is as follows:

$$\mathrm{MI} = \log_2 \frac{P(t_{12})}{P(t_1)P(t_2)},\ P(t_{12}) = \frac{f_{12}}{n_b},\ P(t_i) = \frac{f_i}{n_s}, \tag{1}$$

where f_{12} is the total number of occurrences for the bigram t_{12}, n_b is the number of bigrams in the corpus, f_i is the total number of occurrences for the i-th word in the bigram, and n_s is the number of single words in the corpus. Analogously, the MI for n words can be calculated as follows:

$$\mathrm{MI} = \log_2 \frac{P(t_{1,\ldots,n})}{P(t_1)\ldots P(t_n)}. \tag{2}$$

Content Score. The idea of the content score is that terms that do not appear in most of the documents of a collection but when they appear in a document they appear a number of times are relevant to define the content of such documents. These terms are potentially relevant for thesaurus construction and one simple way to test to which degree a term falls in this category is to use a Poisson distribution. The idea is to predict based on this distribution the document frequency *df* based on the total frequency f of term i. If the predicted number is higher then the terms are accumulated in a lower number of documents than one would expect based on a random model. The difference between observed frequency and predicted one is what indicates (in an indirect way) the content character of a term. The probability that a term i appears in at least one document in a corpus is defined as follows:

$$p(\geq 1; \lambda_i) = 1 - e^{-\lambda_i},\ \lambda_i = \frac{f_i}{n_d}, \tag{3}$$

where f_i is the total frequency of the term i in the whole corpus and n_d is the number of documents in the corpus. The inverse document frequency (IDF) of the term i is then defined as follows:

$$\mathrm{IDF}_i = \log_2(\frac{n_d}{df_i}), \tag{4}$$

where df_i is the document frequency of term i. The residual IDF (RIDF) for term i that is used to express the difference between predicted and observed document frequency:

$$\mathrm{RIDF}_i = \mathrm{IDF}_i - \log_2(p(\geq 1; \lambda_i)), \tag{5}$$

Final Score. As the final score the combination of the mutual information, the residual IDF, and the concept frequency is used. We filter the keywords according to this score and take different thresholds. As follows from the results represented in Sect. 4.4 the behavior is not dependent on the exact threshold. We use the obtained number of keywords for each text and the sum of the scores of the keywords as the measure of information contained in this text.

4 Case Study

In this section we describe the experimental setup, the used data, and the result of the experiment.

4.1 Data

We ran our experiment on

Thesaurus STW Economics[2];
Corpora The financial corpora extracted from investing.com.

Corpus. The chosen corpus for our analysis consists of 39,157 articles obtained from the financial news website investing.com. The articles are divided into three categories:

eur-usd-news[3] 19,119 articles;
crude-oil-news[4] 14,204 articles;
eu-stoxx50-news[5] 5,834 articles.

We collected the news articles by a customized parser that starts at the overview pages and dives into the subpages, storing the title as well as the full text. The corpora contains high quality financial articles and has been approved by financial experts as being representative corpora for the considered field. The news articles span from 2009 till 2016.

Thesaurus. STW Thesaurus for Economics [5,12] is a controlled, structured vocabulary for subject indexing and retrieval of economics literature. The vocabulary covers all economics-related subject areas and, on a broader level, the most important related subjects (e.g. social sciences). In total, STW contains about 6,000 descriptors (key words) and about 20,000 non-descriptors (search terms). All descriptors are bi-lingual, German and English (Table 1).

In frames of PROFIT project the STW thesaurus was extended. The figures for the extended thesaurus can be found in Table 2. This extension provides us

[2] http://zbw.eu/stw/.
[3] http://www.investing.com/currencies/eur-usd-news.
[4] http://www.investing.com/commodities/crude-oil-news.
[5] http://www.investing.com/indices/eu-stoxx50-news.

Table 1. STW thesaurus for economics statistics: original

Number of concepts	6521
Number of broader/narrower relations	15892
Number of related relations	21008

Table 2. STW thesaurus for economics statistics: extended

Number of concepts	6831
Number of broader/narrower relations	16174
Number of related relations	21008

with a unique opportunity to test the introduced measure. Since the extension was performed based on the corpus analysis and suggested keywords and was done by an expert in the domain we may suppose that the extension is sound and improves the fit between the corpus and the thesaurus.

4.2 Technology Stack

Our semantic technology stack includes term extraction, term scoring, and concept tagging. In the process of term extraction and concept tagging we lemmatize the tokens to improve the accuracy of the methods. We use PoolParty semantic suite[6] to perform all the mentioned tasks.

4.3 Workflow

The proposed workflow (see Fig. 1) consists of four main steps:

1. A web crawler automatically collects articles provided by financial news websites.
2. PoolParty is utilized to extract concepts, terms, and scores.
3. The numbers of found concepts and terms are obtained.

The results of this workflow are directly used to support decision making of experts.

4.4 Results

In this section, we present the results of the described experiment.

[6] https://www.poolparty.biz.

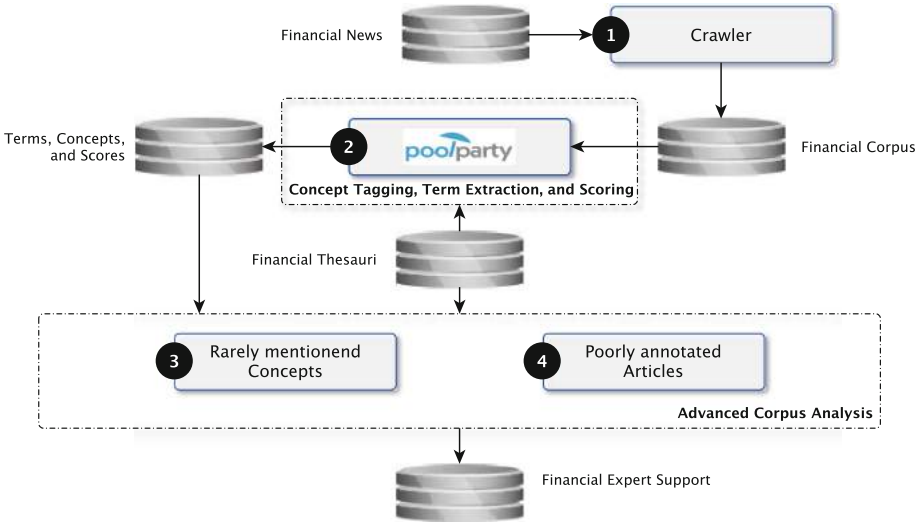
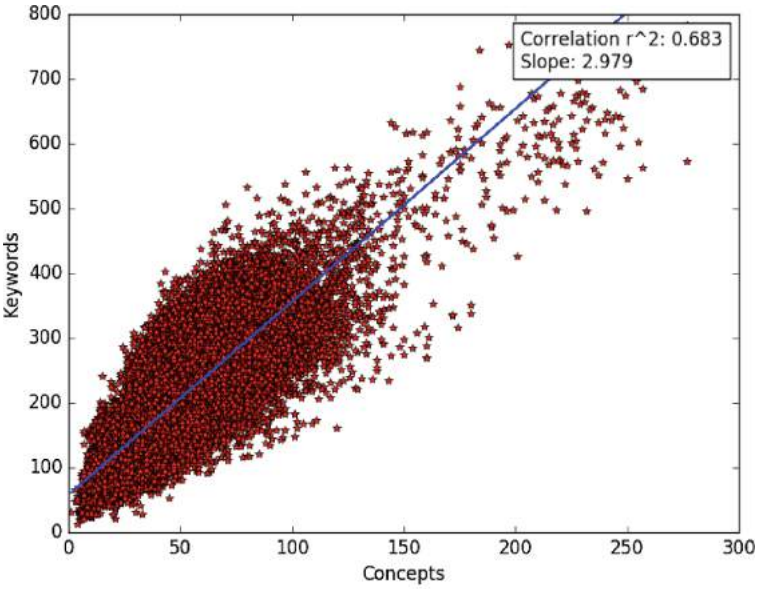


Fig. 1. Processing workflow

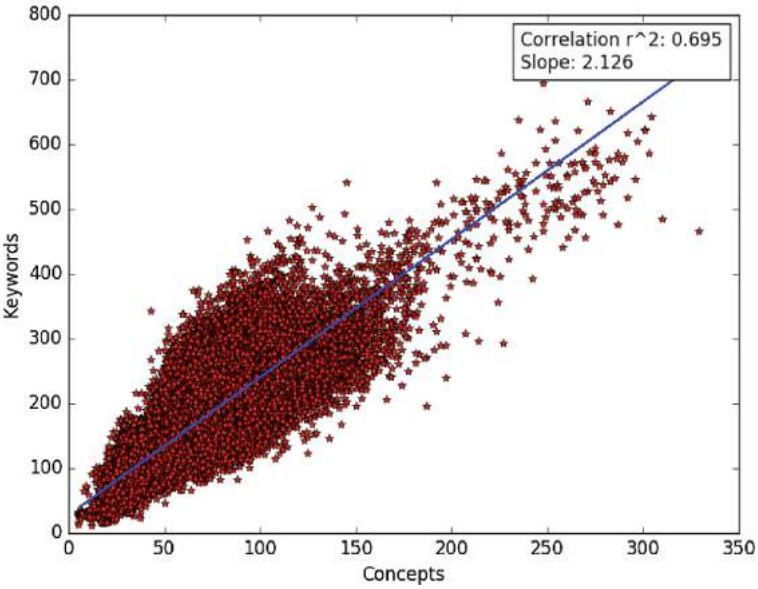
Counting Keywords. We start with comparing the number of keywords with scores above some threshold and a number of extracted concepts. In Figs. 2, 3, 4, 5. In the upper right corner of each figure the correlation coefficient R^2 and the slope of the line are presented. For all thresholds the overall picture remains the same, the correlation coefficient for the extended thesaurus is slightly larger than the same coefficient for the original thesaurus. The difference in the slope of the line is more significant.

Counting Sums of Scores of Keywords. Next we investigate the dependency between the sum of the scores of the keywords above certain threshold and the number of extracted concepts. The correlation coefficients for the extended thesaurus are larger; this indicates that the extended thesaurus is better suitable for the annotation of the given corpora.

It is worth noting that the number of extracted keywords and, hence, the sums of scores are significantly smaller for the extended thesaurus as some of the terms became new concepts and are not counted anymore (Figs. 6, 7, 8 and 9).

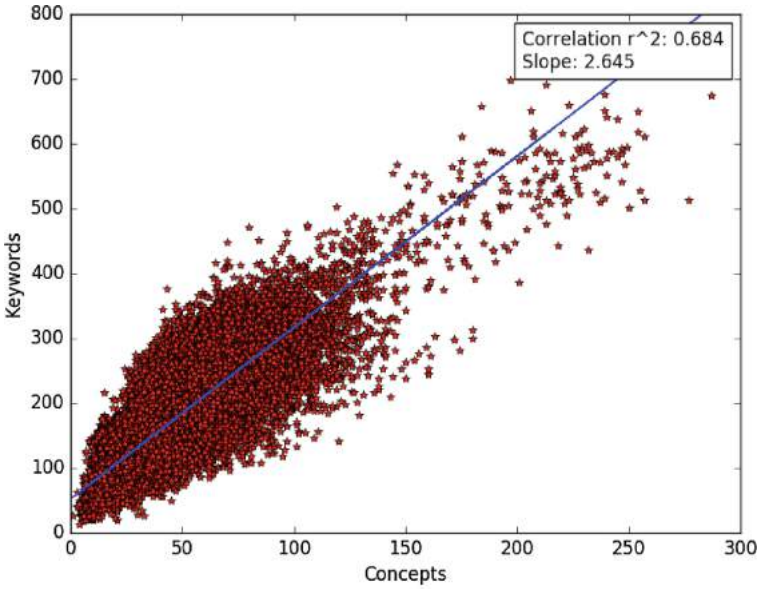


(a) Original thesaurus

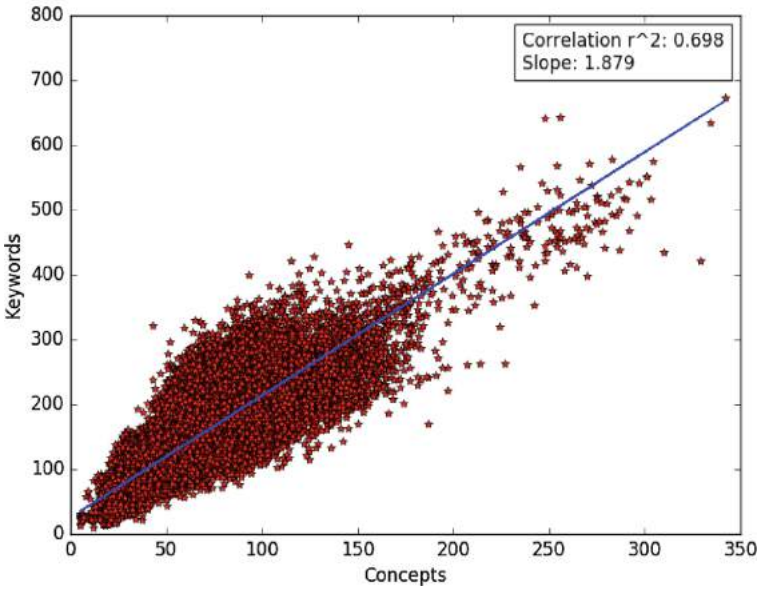


(b) Extended thesaurus

Fig. 2. Number of keywords vs number of concepts, threshold = 1

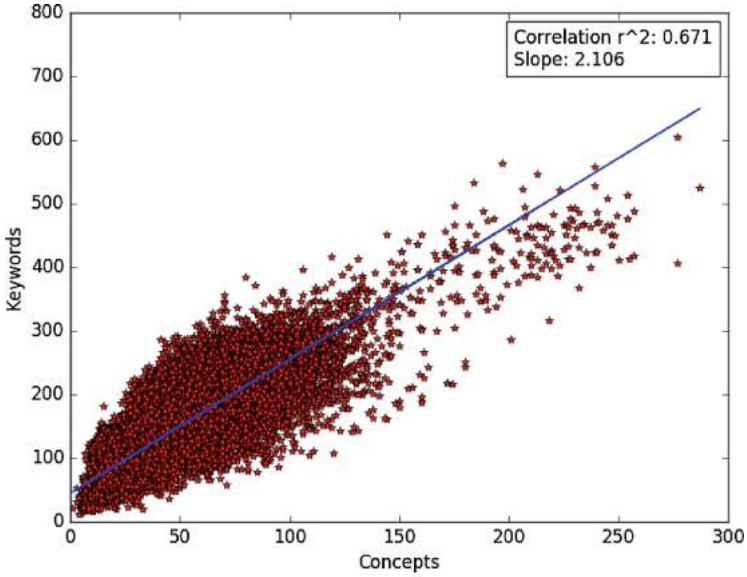


(a) Original thesaurus

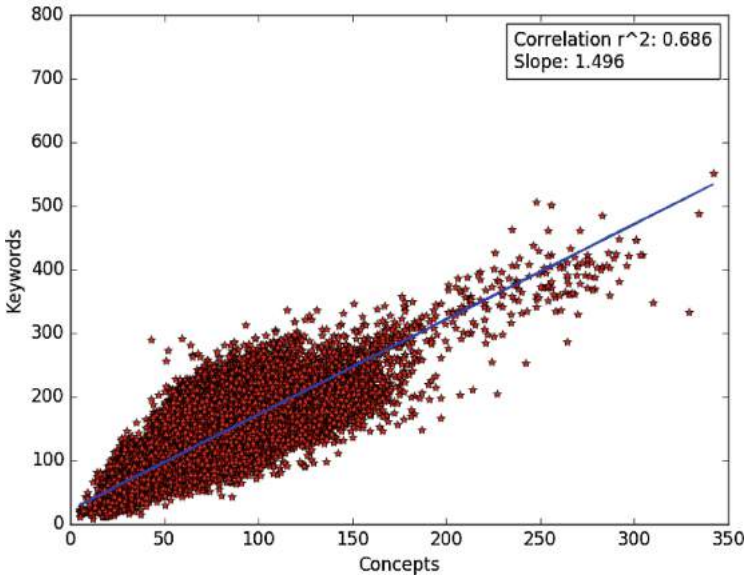


(b) Extended thesaurus

Fig. 3. Number of keywords vs number of concepts, threshold = 2

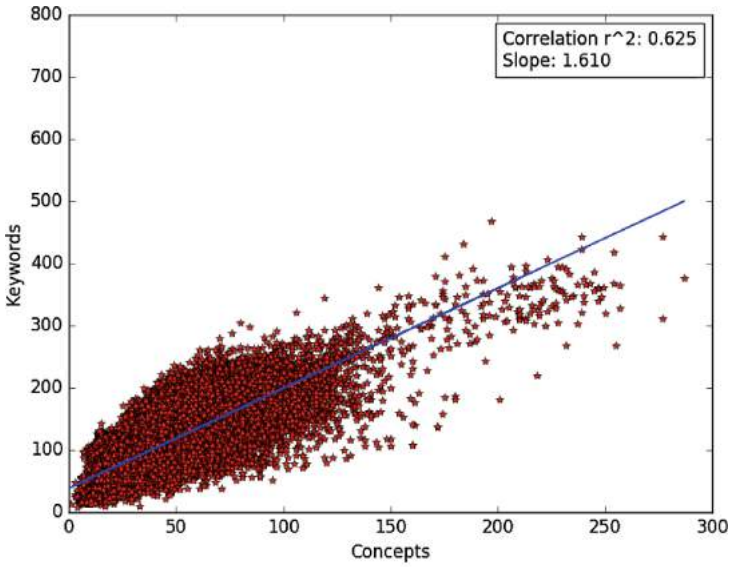


(a) Original thesaurus

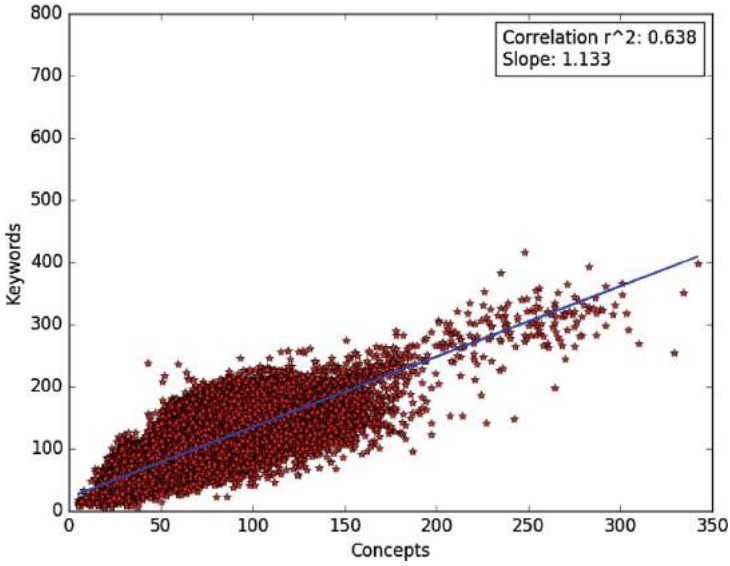


(b) Extended thesaurus

Fig. 4. Number of keywords vs number of concepts, threshold = 5

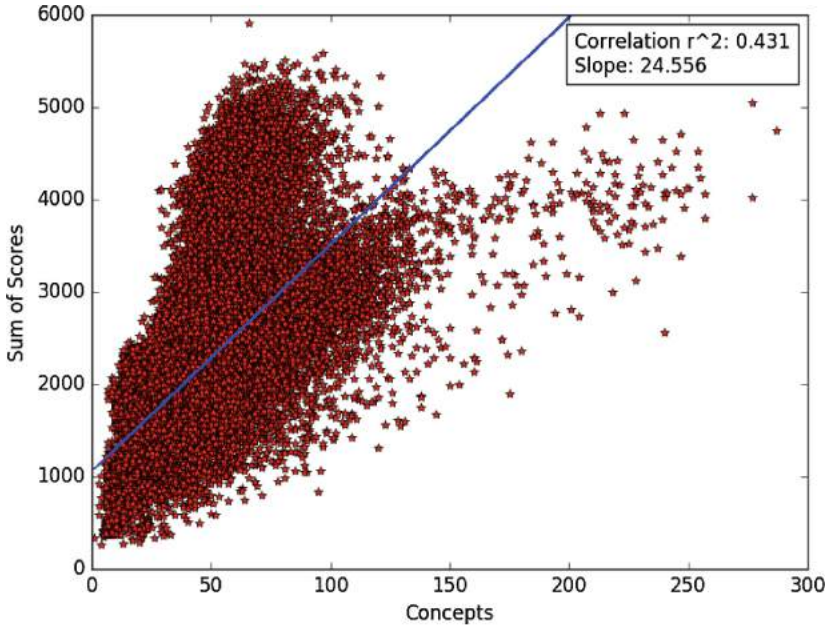


(a) Original thesaurus

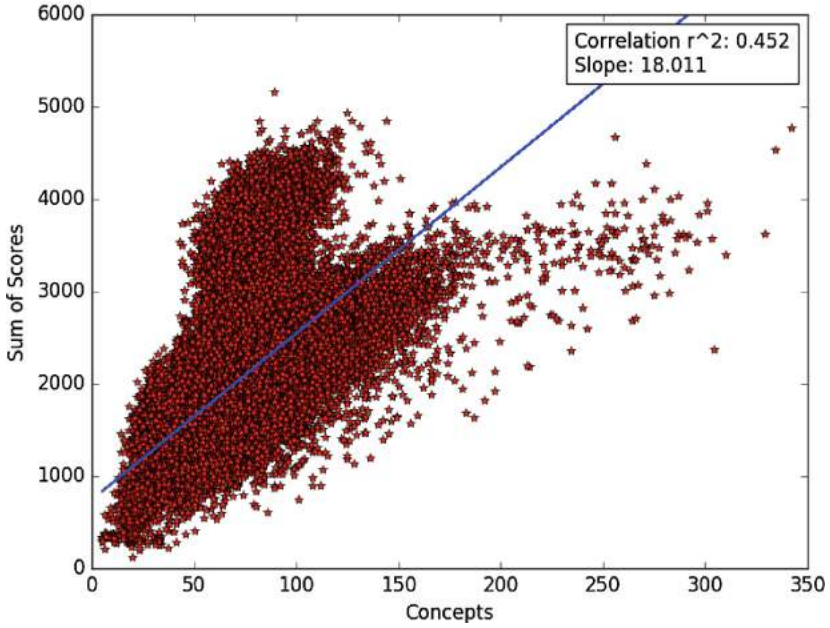


(b) Extended thesaurus

Fig. 5. Number of keywords vs number of concepts, threshold = 10

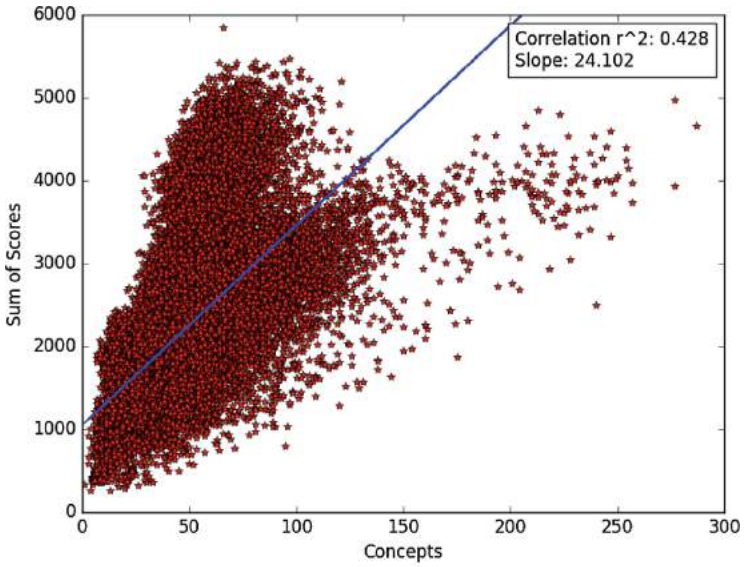


(a) Original thesaurus

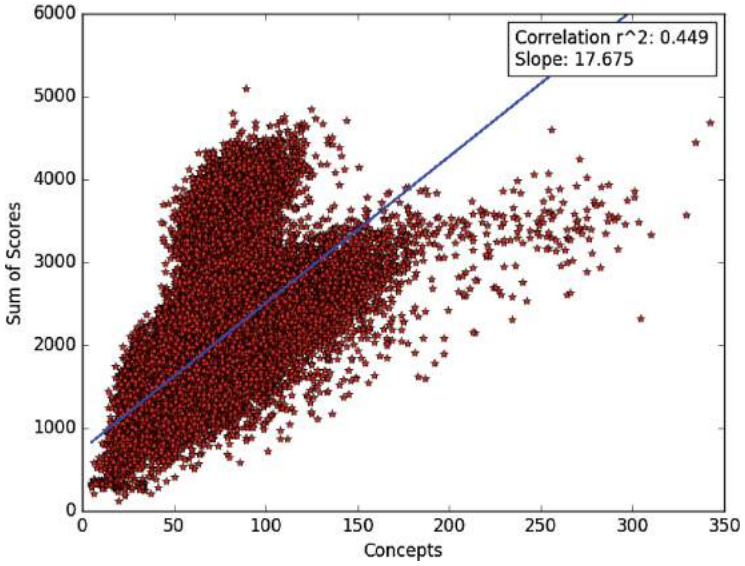


(b) Extended thesaurus

Fig. 6. Sum of scores of keywords vs number of concepts, threshold = 1

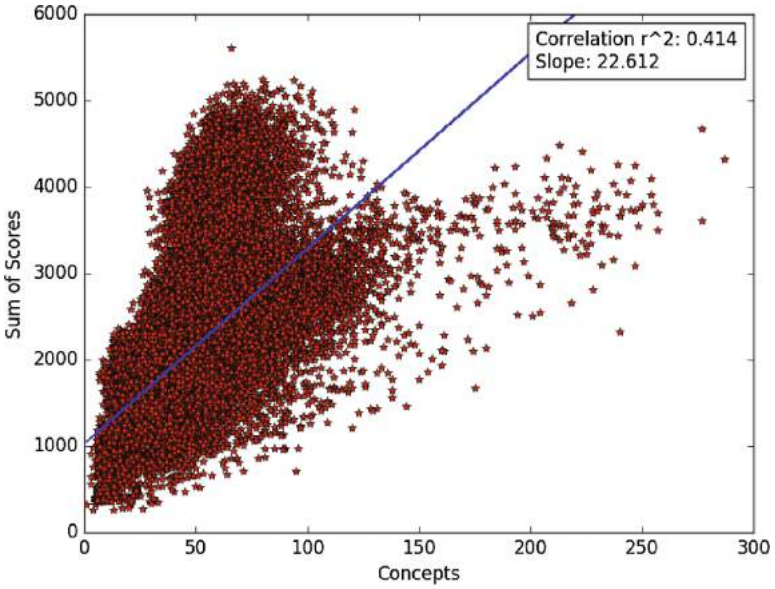


(a) Original thesaurus

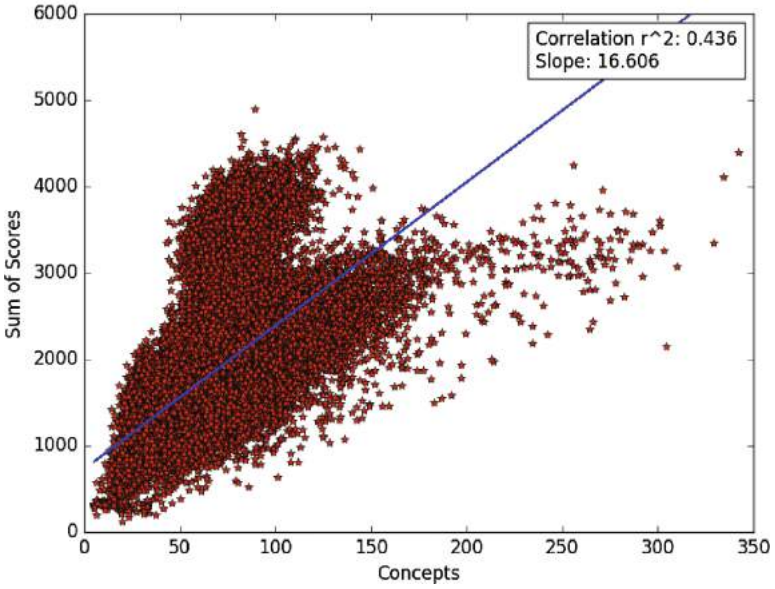


(b) Extended thesaurus

Fig. 7. Sum of scores of keywords vs number of concepts, threshold = 2

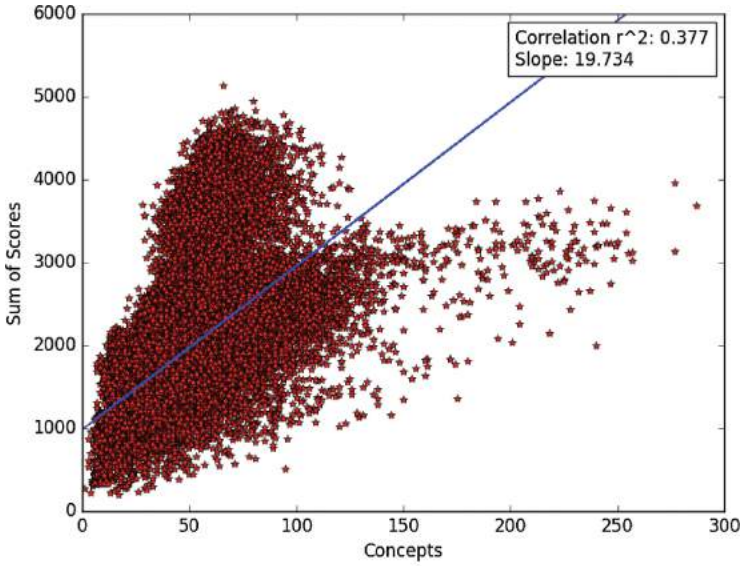


(a) Original thesaurus

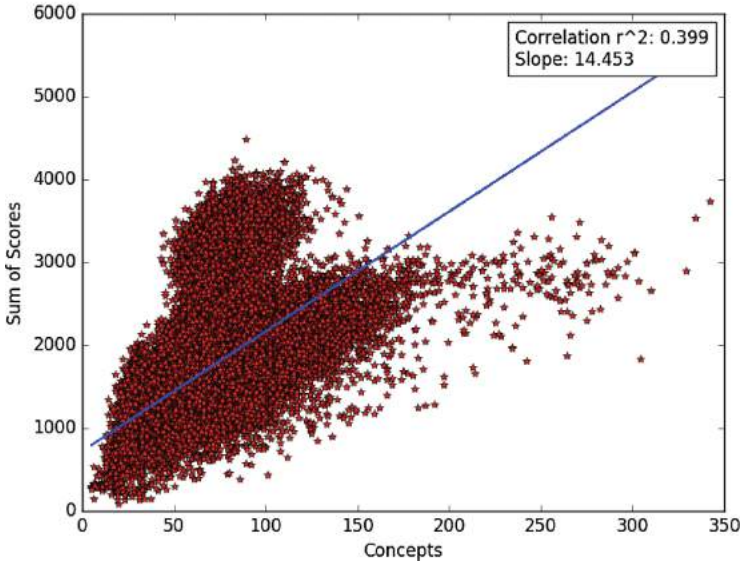


(b) Extended thesaurus

Fig. 8. Sum of scores of keywords vs number of concepts, threshold = 5



(a) Original thesaurus



(b) Extended thesaurus

Fig. 9. Sum of scores of keywords vs number of concepts, threshold = 10

5 Conclusion

Without automated support in the classical interplay between thesauri and corpora, a domain expert may use only his intuition in order to judge if a thesaurus is ready for being used in corpus analysis. We introduce a measure to assist the expert in deciding when to stop the development of thesaurus. The experimental evaluation shows promising results of the proposed approach to evaluate the fit of thesauri to the corpora. The presented approach is part of on-going work but we believe it provides a good foundation for future improvements.

6 Future Work

1. Investigate the plots of the number of extracted concepts vs the length of the document (in symbols and in words);
2. Investigate other measures of the keywords (for example, pure RIDF);
3. Investigate the plots for each topical corpus independently (eur/usd, oil prices, eustoxx);
4. Introduce not only quantitative measure of the annotation (number of concepts), but also a qualitative one (the relations between concepts);
5. Weight the concept occurrences using, e.g., IDF;
6. Investigate the possibility to find incorrect annotation due to, e.g., incorrect disambiguation;
7. Evaluate the results with further corpora and thesauri.

Acknowledgements. We would like to thank Ioannis Pragidis for his work on improving the thesaurus, pointing us to the relevant data, and sharing his deep expertize in the subject domain.

References

1. Ahmad, K., Tariq, M., Vrusias, B., Handy, C.: Corpus-based thesaurus construction for image retrieval in specialist domains. In: Sebastiani, F. (ed.) ECIR 2003. LNCS, vol. 2633, pp. 502–510. Springer, Heidelberg (2003). doi:10.1007/3-540-36618-0_36
2. Aussenac-Gilles, N., Biébow, B., Szulman, S.: Revisiting ontology design: a method based on corpus analysis. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 172–188. Springer, Heidelberg (2000). doi:10.1007/3-540-39967-4_13
3. Bechhofer, S., Miles, A.: Skos simple knowledge organization system reference. In: W3C recommendation, W3C (2009)
4. Birkhoff, G.: Lattice Theory, 3rd edn. Am. Math. Soc., Providence (1967)
5. Borst, T., Neubert, J.: Case study: publishing stw thesaurus for economics as linked open data. In: W3C Semantic Web Use Cases and Case Studies (2009)
6. Jimeno-Yepes, A.J., Aronson, A.R.: Knowledge-based biomedical word sense disambiguation: comparison of approaches. BMC Bioinform. **11**(1), 1–12 (2010)
7. Kacfah Emani, C.: Automatic detection and semantic formalisation of business rules. In: Presutti, V., d'Amato, C., Gandon, F., d'Aquin, M., Staab, S., Tordai, A. (eds.) ESWC 2014. LNCS, vol. 8465, pp. 834–844. Springer, Heidelberg (2014). doi:10.1007/978-3-319-07443-6_57

8. Levy, F., Guisse, A., Nazarenko, A., Omrane, N., Szulman, S.: An environment for the joint management of written policies and business rules. In: 2010 22nd IEEE International Conference on Tools with Artificial Intelligence, vol. 2, pp. 142–149, October 2010
9. Magerman, D.M., Marcus, M.P.: Parsing a natural language using mutual information statistics. In: AAAI, vol. 90, pp. 984–989 (1990)
10. Mandala, R., Tokunaga, T., Tanaka, H.: Combining multiple evidence from different types of thesaurus for query expansion. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 191–197. ACM (1999)
11. Mani, I., Maybury, M.T.: Advances in Automatic Text Summarization, vol. 293. MIT Press, Cambridge (1999)
12. Neubert, J.: Bringing the "thesaurus for economics" on to the web of linked data. In: LDOW, 25964 (2009)
13. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. In: Berry, M.W., Kogan, J. (eds.) Text Mining, pp. 1–20. Wiley, New York (2010)
14. Shah, P.K., Perez-Iratxeta, C., Bork, P., Andrade, M.A.: Information extraction from full text scientific articles: where are the keywords? BMC Bioinform. **4**(1), 1 (2003)
15. Tan, A.-H., et al.: Text mining: the state of the art and the challenges. In: Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases, vol. 8, pp. 65–70 (1999)