Axel-Cyrille Ngonga Ngomo · Petr Křemen (Eds.)

Knowledge Engineering and Semantic Web

7th International Conference, KESW 2016
Prague, Czech Republic, September 21–23, 2016
Proceedings

Springer

Editors
Axel-Cyrille Ngonga Ngomo
Leipzig University
Leipzig
Germany

Petr Křemen
Czech Technical University in Prague
Prague
Czech Republic

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-319-45879-3 ISBN 978-3-319-45880-9 (eBook)
DOI 10.1007/978-3-319-45880-9

Library of Congress Control Number: 2016949634

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

A New Operationalization of Contrastive Term Extraction Approach Based on Recognition of Both Representative and Specific Terms

Aliya Nugumanova[1(✉)], Igor Bessmertny[2], Yerzhan Baiburin[4], and Madina Mansurova[3]

[1] D. Serikbayev East Kazakhstan State Technical University, Ust-Kamenogorsk, Kazakhstan
yalisha@yandex.kz

[2] Saint Petersburg National Research University of ITMO, Saint Petersburg, Russia
igor_bessmertny@gmail.com

[3] Al-Farabi Kazakh National University, Almaty, Kazakhstan
mansurova01@mail.ru

[4] East Kazakhstan State University, Ust-Kamenogorsk, Kazakhstan
ebaiburin@gmail.com

Abstract. A contrastive approach to term extraction is an extensive class of methods based on the assumption that the words frequently occurring within a domain and rarely beyond it are most likely terms. The disadvantage of this approach is a great number of type II errors – false negatives. The cause of these errors is in the idea of contrastive selection when the most representative high frequent terms are extracted from the texts and rare terms are discarded. In this work, we propose a new operationalization of the contrastive approach, which supports the capture of both high frequent and low frequent domain terms. Proposed operationalization reduces the number of false negatives. The experiments performed on the texts of the subject domain "Geology" show promising of proposed approach.

Keywords: Contrastive term extraction · Termhood · Mutual information · LSA

1 Introduction

At present, in the field of information extraction there are numerous methods aimed at automated extraction of knowledge structures from natural language texts [1–4]. Among the knowledge structures being extracted, the simplest ones are lists of terms and the most complex ones are domain thesauri and ontologies. All these structure are designed for setting explicit specifications of subject domain to eliminate uncertainty and ambiguity in the knowledge exchange between humans and applications.

In this work, we focus on extraction of simple but valuable knowledge structures – lists of single word terms. Like the authors of [5], we consider domain terms to be words used by experts to describe conceptual apparatus of the domain. Lists of terms are used everywhere when it is necessary to convey in a structured compressed form

A.-C. Ngonga Ngomo and P. Křemen (Eds.): KESW 2016, CCIS 649, pp. 103–118, 2016.
DOI: 10.1007/978-3-319-45880-9_9

the semantics of a text, topic or domain. Apart from its original value, lists of terms are used as building material for more complex structures, for example, ontologies. Therefore, the significance of automated term extraction can hardly be overestimated.

The process of automated term extraction is very often accompanied by the process of validation. Validation is usually carried out by domain experts in manual or semi-automatic mode and allows confirming or disproving the terminological status of the words being extracted. Therefore, the authors of [6] think it to be more correct to call the words being extracted not terms but terms-candidates from which, in the process of validation, true domain terms will be selected. As a rule, to support the process of validation, the output list of terms-candidates are ranked by the degree of their termhood – an index defining how much a term-candidate corresponds to the conceptual apparatus of the domain [6, 7]. Sometimes, the process of ranking replaces the process of validation, in this case the candidates which termhood values exceed the specified threshold are deemed as terms.

The problem of appropriate measuring of termhood is central in the task of term extraction. In [8], the authors present a detailed review of the corresponding methods and conclude that most of them measure the termhood heuristically: on the basis of assumptions about the character of distribution of terms in texts. The popularity of heuristic approaches can be accounted for the conditions of uncertainty, which accompany the process of terms search. Both the terms themselves and features allowing to recognize terms are unknown. In work "Mathematical discovery", G. Pólya noted that a clearly formulated problem must precisely specify the condition, which has to be satisfied by the unknown [9]. If such conditions cannot be specified, it is expedient to use heuristic methods.

In this work, we study a popular heuristic approach based on contrastive term extraction [8, 10]. This approach uses an idea that the words, which often occur in the domain texts and seldom in the texts of other domains are most likely terms [8]. Like all heuristics, the contrastive method suffers from a grave shortcoming, such as the loss of recall. It makes a great number of the second type errors, i.e. false negatives. The cause of errors is in the idea of contrastive selection when the most representative (high frequent) words are extracted from texts and rare words are discarded. According to the opinion of the authors [11], the practice of exclusion of rare words is usual for the information retrieval tasks but it is not always useful for term extraction. As an example, the authors refer to the collection of medicinal abstracts in which rare terms denoting side effects due to administration of medicinal preparations make up 68 %. Exclusion of these rare terms from consideration would result in great losses, when specifying a subject domain.

The aim of this work is to overcome the above-mentioned shortcoming, i.e. to enhance the recall of contrastive term extraction based on the separated capture of rare and representative terms. For this, we use mutual information criterion described in [12] in detail. For representative terms extraction we use average weighted mutual information, and for rare terms extraction we use pointwise mutual information. However, we recognize that mutual information has own drawbacks. It makes errors of the first type, i.e. false positives. The cause of the errors is in the presence of the so-called conjugate words, i.e. words that "accompany" the domain, but not related to it directly. In [13] as an example quoted the Gulf and Kuwait words, which are closely

related to the domain "Oil" due to their frequent occurrence in texts associated with this domain. However, in fact they are not terms.

Therefore, in this work we have to adjust the estimates obtained with the help-means of mutual information involving a special technique based on the analysis of the causes of false positives. The structure of the work is as follows. Section 2 presents the review of related work dealing with the method of contrastive term extraction. Section 3 contains the necessary theoretical information advancing our approach. Section 4 describes the method being proposed by us allowing to away with the existing drawbacks of the contrastive approach. Section 5 presents the case-study on term extraction from the texts of the subject domain "Geology". In Sect. 5, we formulate conclusions and present a plan for further investigations.

2 Related Work

A contrastive approach is a general name of the methods revealing terms from the point of their different occurrence within and beyond a subject domain. All these methods are governed by general idea of defining the termhood of words based on comparison of their distribution in two collections: target domain and alternative. As an alternative collection, we may use either a contrastive collection, i.e. formed from the texts of different domains, or general collection, i.e. formed from the texts which do not refer to any domain [14]. The differences between contrastive methods are only in the ways of operationalization of the idea in their basis. In this case, we speak about operationalization of such a fuzzy concept as termhood by means of measurable indicators or procedures based on the contrastive approach.

One of the first works concerning contrastive term extraction is [15]. To evaluate the termhood, its authors introduce a new intuitively perceived measure called "weirdness". The weirdness is calculated for each term-candidate and is the ratio of the frequency of term occurrence in a target collection to the frequency of occurrence in a general collection. For usual words, the weirdness formula restores values close to 1, and for terms – values much more exceeding 1 as in this case the denominator of the formula is close to 0. In their later works, for example, in [16], the authors present a modified variant of the formula as the initial formula develops singularity when the denominator turns 0.

In [17], the idea of contrastive term evaluation is formed as not one but two statements: (1) the words more rarely used in the target collection must have a lower value; (2) the words more frequently used in the target collection must have a higher value, bit with reserve that they do not often occur in the contrastive collection or in a limited set of texts of the target collection. The authors operationalize these statements in the form of metrics called by them relevance:

$$Relevance = \frac{1}{log_2(2 + \frac{f_{SL} * N^t_{SL}}{f_{GL}})} \quad (1)$$

where, f_{GL} and f_{SL} are relative frequencies of the word in the target and contrastive collections, respectively; N^t_{SL} is a relative number of texts of the target collection in which this word occurs. The reduced metrics copes well with extraction of representative terms but artificially decreases the value of rare terms.

In [18], the authors propose a somewhat other method for evaluation of termhood on the basis of the contrastive approach. The method is based on the well-known TF-IDF formula according to which the weight of the word in the document is the higher, the higher the frequency of its occurrence in this document and the lower its dispersion in the whole collection. In the new formula, which the authors call "term frequency-inverse domain frequency" they evaluate the weight of the word not in the document but in the target collection. According to the new formula, the weight of the word is the higher, the higher the relative frequency of its usage in target collection and the lower its relative dispersion in all collections:

$$TF \cdot IDF = TF(t, D) \cdot IDF(t) = \frac{n_{t,D}}{\sum_k n_{k,D}} \cdot log\left(\frac{|TS|}{|\{d : t \in d\}|}\right) \quad (2)$$

where, $n_{t,D}$ is the number of the word t entry into the target collection D, $\sum_k n_{k,D}$ is the sum of all words entries into the target collection D, |TS| is the number of documents in all used collections, $|\{d : t \in d\}|$ is the number of all documents into which the word t enters at least once. Thus, the authors consider all words with high concentration within a narrow subset of documents to be terms. For a definite part of terms it is undoubtedly a correct approach but for rare terms it is of title use.

The authors of [19] evaluate the termhood on the basis of formula TF-IDF, too. They call their variant of this formula a contrastive weight and define it as a measure, which is the higher, the higher the frequency of the word usage in the target collection and the lower the relative frequency of its usage in contrastive collections:

$$Contrastive\ Weight = TF(t, D) \cdot IDF(t) = \log\left(f_t^D\right) \cdot \log\left(\frac{F_{TC}}{\sum_j f_t^j}\right) \quad (3)$$

where, f_t^D is the frequency of the word usage in the target collection, $\sum_j f_t^j$ is the sum of frequencies of usages of all words in all collections including the target one. As the authors themselves note, the contrastive weight evaluates the termhood of words significantly better than pure frequencies, however, the efficiency of the method determined with the help of F-measure, according to their words, is not evident.

In [20], the formula of contrastive weight undergoes a critical evaluation. As the authors note, the contrastive weight and similar metrics evaluate in fact not the reference of terms to a domain but their prevalence. To rectify the mentioned drawback, the authors evaluate the termhood on the basis of two indices: DP (domain prevalence) and DT (domain tendency). The formula for calculation of DP is in fact a subduced variant of the contrastive weight (3). Its high value indicates the prevalent distribution of the word in the target collection compared to distribution of the word in the target collection compared to distribution of other words in this collection:

$$DP = log_{10}(f_t^D + 10) \cdot log_{10}\left(\frac{F_{TC}}{f_t^D + f_t^{\bar{D}}} + 10\right) \quad (4)$$

where, f_t^D and $f_t^{D^-}$ are frequencies of the word in the target and contrastive collections, respectively, $F_{TC} = \sum_j f_j^D + \sum_j f_j^{D^-}$ are sums of frequencies of all terms-candidates usage in the target and contrastive collections, respectively.

The formula for calculation of DT is a subdued variant of the weirdness formula, i.e. it penalizes the words, which often occur in the contrastive collections. Its high value indicates the prevalent distribution of the word in the target collection compared to its distribution in the contrastive collections:

$$DT = log_2\left(\frac{f_t^D + 1}{f_t^{\bar{D}} + 1} + 1\right) \quad (5)$$

The measures of DP and DT are combined into one general index called discriminative weight DW. It is the product of DP and DT measures. According to the authors' opinion, the discriminative weight possesses a differentiating ability. It pushes up the terms which often occur in the target collection and rarely in the contrastive collections. However, the indices DT and DP correlate with each other quite well. For example, in our experiments, the correlation values of these indices made up 0.71–0.82. To reveal the nature of this correlation, we divided all terms-candidates into 4 not intersecting groups depending on DT and DP: (1) the values of DT and DP are lower than average; (2) the value of DT is lower than average and the value of DP is not lower than average; (3) the values of DT is not lower than average and the values of DP is lower than average; (4) the values of DT and DP are not lower than average. Both expert evaluations and evaluations based on formula (5) showed the same result: with few exceptions, only candidates from the third and fourth groups can be recognized as terms, this corresponding to high values of DT index. Such result indicates high informational content of DT index and redundancy of DP index.

The reservation "with few exceptions" is not coincidental. Validation of the proposed method demonstrates the fact that in the area of low values of termhood, among usual words, there occur terms which we call rare terms. These are terms, which occur 1–2 times in the target collection and not a single time in the contrastive collection. Extraction of such terms requires the use of more distinctive instruments of differentiation.

Not only is the work [20] distinguished by the use of at once several indices for evaluation of termhood. In [21] for this purpose, at once 3 indices are used: domain pertinence DP, domain consensus DC and lexical cohesion LC, assigned for evaluation of cohesion of verbose terms. As a result, the total evaluation of the word termhood in the target collection *Di* is formed from the linear combination of the three enumerated measures. The measure of pertinence DR is the measure of weirdness generalized for the case of a contrastive collections set:

$$DR(t, Di) = \frac{freq(t, Di)}{\max\limits_{j}(freq(t, Dj)} \quad (6)$$

The measure of consensus DC allows to take into account the distribution of words in separate documents. It is defined via normalized frequencies φ_k of the word occurrence in the documents of target collection Di and is the higher, the more uniformly the word is distributed in these collections:

$$DC(t, Di) = -\sum\nolimits_{d_k \in Di} \phi_k log \phi_k \quad (7)$$

Introducing the measure of consensus, the authors emphasize its high importance. According to their opinion, the terms which frequently occur in a great number of documents of target collection must be evaluated higher than the terms which frequently occur in a restricted number of documents. It should be noted that this statement is completely antagonistic to heuristic used in [18]. The authors of [22] also note this interesting fact illustrating a wide interpretation of the notion 'termhood' and high uncertainty in the choice of criteria of termhood.

In [23], termhood is defined not via the ratio of the word usage frequency in the target and contrastive collections but via the difference of their ranks. The rank of the word in the collection is its position in the dictionary compiled from all words of the collection sorted out according to the increase in the word usage frequencies. The index of the termhood expressed via the difference of word ranks in the target D and contrastive G collections has the form:

$$thd(t, D) = \frac{rank(t, D)}{|V(D)|} - \frac{rank(t, G)}{|V(G)|}$$

where $V(D)$ and $V(G)$ are vocabularies of the corresponding collections. The index has values from -1 to 1, the value 1 corresponds to the case when the word has the highest rank in the target collection and a zero rank in the contrastive collection. Ranking of words allows to draw up the most representative words. However, as the authors note, their approach does not allow to extract only terms.

The last work we would like to note in this number is the [24]. It also develops the idea of penalties and rewards which was introduced into the basic construction of TF-IDF formula and proposes a new variant of this formula called "term frequency – disjoint corpora frequency":

$$TF \cdot DCF = \frac{f_t^D}{\prod_{g \in G} 1 + log(1 + f_t^g)}$$

where f_t^D and f_t^g are frequencies of this word occurrences in target and contrastive collections respectively, G is a set of all contrastive collections.

The authors prove on a series of experiments that their formula is the best in precision of terms extraction compared to the values proposed in works [18, 23] and

a number of other works. They justify the use of the product in the denominator of the formula by the fact that penalty much increase in a geometrical progression for each use of the word in the next in turn contrastive collection. According to the opinion of the authors, the termhood of words which are used few times in a great number of contrastive collections must be evaluated lower than that of the words which are used many times but a small number of contrastive collections.

Thus, in this review we considered 8 most interesting contrastive methods of the termhood value operationalization. All these methods are heuristic, i.e. based on assumptions concerning the character of terms distribution in target and contrastive collections. A comparative analysis of these statements show that there take place both coincidences of positions of different authors and grave divergences, this indicating the presence of unsolved problems in this field.

3 Theoretical Background

The authors of [25] note that the term extraction methods based on heuristic assumptions are often criticized for the absence of theoretical strictness. The authors say that such criticism becomes evident when simple but important questions on the methods of operationalization of these or those heuristics are put, for example "Why are different bases of logarithms used in metrics?" or "Why combination of two weights in metric is based on their product but not their sum?"

However, popularity of using similar ad-hoc metrics can be easily explained. Termhood is a notion which is rather difficult to be formalized and it is easier to express it with the help of heuristics captured of a set of termhood aspects only those which are the most evident and available for formalization. Meanwhile, in the field of automated extraction of terms there have been developed strict statistical criteria based on mathematical grounds of information theory and probability theory.

Mutual information refer to the number of such criteria [12]. The notion of mutual information goes back to the more fundamental and more general notion of information theory – informational divergence, also known as the distance of Kullback–Leibler or relative entropy. Informational divergence is an asymmetric measure of the distance between two discrete distributions $P = \{p_i\}$ and $Q = \{q_i\}$:

$$D(P||Q) = \sum_i p_i log\left(\frac{p_i}{q_i}\right) \tag{10}$$

here and further, the base of logarithms is taken as standard value 2. As a rule, one of the distributions being compared is 'true' (observable) and the second one is expected (under test). Therefore, informational divergence may be treated as the measure of how much the "true" distribution diverges with the expected, approximate one. Mutual information is a particular case of informational divergence when distribution P is a joint distribution of two random discrete values X and Y and distribution Q is the product of marginal distributions of these random values [26]:

$$MI(X,Y) = D(P(X,Y)||P(X) \times P(Y)\} = \sum_{i,j} p(x_i,y_j) \cdot log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) \quad (11)$$

If random values X and Y are independent, the probability of their joint distribution is equal to the product of probabilities of their marginal distributions $p(x_i,y_j) = p(x_i)p(y_j)$, then $log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) = log1 = 0$. Hence, mutual information of these values is equal to 0. Intuitively, this can be explained as follows: if two random values are independent, the appearance of one of them does not give any information in regard to the appearance of the other. Correspondingly, mutual information can be treated as the measure of correlation of these values. The notion of mutual information determined for two random values X,Y is closely related to the pointwise mutual information defined for a concrete pair of outcomes (x,y) of these random values:

$$PMI(x,y) = log\frac{p(x,y)}{p(x) \cdot p(y)} \quad (12)$$

Having compared formulas (11) and (12), one may note that mutual information is an average weighted estimation of pointwise mutual information values on all pairs of random values X and Y outcomes:

$$MI(X,Y) = \sum_{i,j} p(x_i,y_j) \cdot PMI(x_i,y_j) \quad (13)$$

The obtained formulas (12), (13) fit well to the problem of contrastive term extraction. To derive the fit formulas, the authors of [12] introduce designations as is shown in Table 1, and on the basis of this table they evaluate distribution probabilities of two random values: X is the presence of the word in the document; Y is the ratio of the document to the subject domain (see Tables 2, 3 and 4).

For each of the four possible outcomes presented in Table 4, its own formula of pointwise mutual information is derived on the basis of formula (12). Of basic interest is the formula for the outcome $t \wedge d$:

$$PMI\left(t \wedge d\right) = log\frac{p(t \wedge d)}{p(t)p(d)} = log\frac{A/N}{((A+B)/N) \times ((A+C)/N)} = log\frac{A \times N}{(A+B) \times (A+C)} \quad (14)$$

Table 1. The contingency table describing the distribution of words in collections

Number of documents	In target collection	In contrastive collection	Total
Containing this word	A	B	$A+B$
Not containing this word	C	D	$C+D$
Total	$A+C$	$B+D$	$N=A+B+C+D$

Table 2. Marginal distribution of random variable X

Outcomes	Probability
t : the document contains the word	$p(t) = (A+B)/N$
$\bar{t}$: the document does not contain the word	$p(\bar{t}) = (C+D)/N$
Σ:	$p(t)+p(\bar{t}) = 1$

Table 3. Marginal distribution of random variable Y

Outcomes	Probability
d : the document refers to the subject domain	$p(d) = (A+C)/N$
$\bar{d}$: the document does not refer to the subject domain	$p(\bar{d}) = (B+D)/N$
Σ:	$p(d)+p(\bar{d}) = 1$

Table 4. Joint distribution of random variables X, Y

Outcomes	Probability
$t \bigwedge d$: the document contain the word and refers to the subject domain	$p(t \bigwedge d) = A/N$
$\bar{t} \bigwedge d$: the document does not contain the word and refers to the subject domain	$p(\bar{t} \bigwedge d) = C/N$
$t \bigwedge \bar{d}$: the document contain the word and does not refer to the subject domain	$p(t \bigwedge \bar{d}) = B/N$
$\bar{t} \bigwedge \bar{d}$: the document does not contain the word and does not refer to the subject domain	$p(\bar{t} \bigwedge \bar{d}) = D/N$
Σ:	1

The given formula allows to evaluate the amount of information carried by the fact of the presence of this word in the document of the subject domain [27]. The formulas of point wise mutual information for the other three outcomes have the following form:

$$PMI\left(\bar{t} \bigwedge d\right) = log \frac{p(\bar{t} \bigwedge d)}{p(\bar{t})p(d)} = log \frac{C \times N}{(C+D) \times (A+C)} \tag{15}$$

$$PMI\left(t \bigwedge \bar{d}\right) = log \frac{p(t \bigwedge \bar{d})}{p(t)p(\bar{d})} = log \frac{B \times N}{(A+B) \times (B+D)} \tag{16}$$

$$PMI\left(\bar{t} \bigwedge \bar{d}\right) = log \frac{p(\bar{t} \bigwedge \bar{d})}{p(\bar{t})p(\bar{d})} = log \frac{D \times N}{(C+D) \times (B+D)} \tag{17}$$

The formula of average weighted mutual information is derived on the basis of formula (13) and obtained values of PMI similarly:

$$MI = \frac{A}{N} \log \frac{A \times N}{(A+B)(A+C)} + \frac{C}{N} log \frac{C \times N}{(C+D)(A+C)} + \frac{B}{N} \log \frac{B \times N}{(A+B)(B+D)} + \frac{D}{N} \log \frac{D \times N}{(C+D)(B+D)} \tag{18}$$

Despite the fact that both formulas of mutual information (14) and (18) evaluate the amount of information, which is carried by the word on the subject domain, there is a principal difference between them, which is best illustrated by the known expression "briller par son absence".

In other words, average weighted mutual information evaluates the relation between the word and subject domain taking into account not only the probability of its presence in target collection (outcome $t \wedge d$) but also absence (outcome $\bar{t} \wedge d$) as well as probabilities of it presence and absence in contrastive collection (outcomes $t \wedge \bar{d}$ and $\bar{t} \wedge \bar{d}$, respectively). The work [12] states the following important fact which is the effect of the mentioned difference: pointwise mutual information is biased to the side of rare terms, while average – weighted mutual information normalizes bias on account of using weights.

Let us consider on account of what there takes place the bias of estimations obtained with the help of pointwise mutual information, i.e. let us determine under what conditions formula (18) has a maximum. In the reduced formula, values N and $A + C$ (the total number of documents and the number of documents of target collection, respectively) are constants as they do not depend on distribution of words. Therefore, they can be ignored and function $log\left(\frac{A}{A+B}\right) = log\left(\frac{1}{1+B/A}\right)$ must be considered. As we speak about logarithms by base 2 (increasing function), the function aspire to maximum when B/A aspire to minimum. The expression B/A reaches minimum only at $B = 0$, i.e. when the word never occurs in contrastive collection. It does not matter what A equals to, i.e. to it is not important whether the word occurs in target collection 50 or 10 times, evaluation of these words will be the same. In all other cases, when B is not equal to 0, the value of A does not effect the termhood evaluation, only the ratio B/A matters. For example, if the first word occurs 50 times in the target collection and 10 times in contrastive collection and the second word occurs 10 times in target and once in contrastive collection, then the termhood of the second word will be evaluated higher, since $B_1/A_1 = 10/50 > B_2/A_2 = 1/10$.

4 Proposed Approach

We conditionally divide all the words related to the domain into 2 classes. We refer the frequently used, representative words of the domain to the first class and rarely used, specific words of the domain – to the second class. For example, in the field of geological sciences the words "geology", "rock", "crust" refer to the first class, and the words "breccia", "feldspar", "columbite" – to the second class.

We consider that, since the words referring to these two classes are quite differently distributed in the domain, it is necessary to use not one but two criteria to extract them. In particular, as we showed in the previous section, to extract words of the first class the average weighted mutual information fits well and for extraction of the second class words – pointwise mutual information (see Table 5).

To our mind, the main drawback of the methods considered in Section "Related work" is that they aim to combine several differently oriented metrics in one criterion uniting them by multiplication or addition. Thereby, the mentioned methods try to find

Table 5. The criteria used for term extraction

Class	Representative words	Specific words
Used criteria	The average weighted mutual information *MI*	Pointwise mutual information *PMI*
Condition of selection into class	$A > 0$, $MI > MI_{crit}$	$MI \leq MI_{crit}$, $PMI > PMI_{crit}$
Notation	MI_{crit} – is a critical value of criterion MI determined on the basis of Student's test at the level of significance $\alpha = 0.01$ and the number of degrees of freedom equal to the dimensions of the subdictionary.	PMI_{crit} – is a critical value of criterion PMI determined on the basis of Student's test at the level of significance $\alpha = 0.01$ and the number of degrees of freedom equal to the dimensions of the subdictionary.

a compromise where it cannot be. Unlike the authors of these methods, we believe that a method of separate per class term capture will allow to enhance the term extraction recall though it will not solve the problem of the term extraction precision.

The problem of the term extraction precision is that among the words of both the first and the second class there can occur not only terms but also the words. These words are statistically distributed in the same way as terms but in fact they are not terms. They form a set of false positives. For example, in the target collection of texts on geology used by us the words "geyserite" and "resort" occur only in one document and not a single time – in contractive collection. Correspondingly, as we noted in the previous section, the both words are characterized by the maximum value of pointwise mutual information though one of them is a term and the second one is not.

This is caused by imperfection of target and contrastive collections: if these collections were sufficiently full and extensive, the term "geyserite" would occur more frequently in target collection, while the word "resort" would more frequently occur in contrastive collection. Thus, after formation of both classes we have to eliminate the words, which are not terms from the each class. For this, let us analyze the causes of false positives in each class under study and propose the methods for elimination of these false positives (see Table 6).

As follow from this table, to recognize false terms among specific words, it is not sufficient to have only information on indices of word occurrence. More extensive information is needed and we can gain it using, apart from the indices of occurrence of words, the indices of their co-occurrence. This idea is not new. It is noted in [28] that traditionally the importance of a term-candidate is determined on the basis of analysis of how it is related to the other words of the text or collection. A term-candidate is considered to have a large weight if it is related to either a large number of other terms-candidates or terms-candidates which themselves have a large weight.

In this work, we, like the authors of [29], are guided by this idea and consider a co-occurrence matrix as an instrument for measuring the relations between words. The contribution of our work is the method of construction and processing of this matrix. At first, we form a matrix "documents-terms" the rows of which are documents of target

Table 6. The reasons of false alarms and how to resolve them

Class	Representative words	Specific words
The cause of false alarms	The word is often found in the target collection ($A \gg 0$), but it is quite common and contrasting collection ($A \sim B$) Uncertainty type $\frac{\infty}{\infty}$	The word is rare in contrast collection ($B \sim 0$), but it is rarely found in a target collection (B ~ A). ($B \sim A$). Uncertainty type $\frac{0}{0}$
Result	The ratio A/B did not limited to the top terms and limited to a certain threshold value for not R_{crit}	The ratio B/A is equal to 0 (or very close to 0) for both terms, and not for the terms.
Recognition method of false detection	It is to verify the conditions of $A/B < R_{crit}$, where R_{crit} – is a critical value of criterion A/B determined on the basis of Student's test at the level of significance $\alpha = 0.01$ and the number of degrees of freedom equal to the dimensions of the subdictionary.	Stepping outside of the contrast analysis.

collection and the columns are extracted specific and representative words. Then, we subject this matrix to the operation of singular decomposition to get rid of noise and rarity.

And only after that we go on to formation of the matrix "terms-terms". Since in the matrix "documents-terms" each term is a vector-column, the semantic relation between any two terms can be treated as closeness or distance between vectors corresponding to these terms using a cosine measure:

$$r_{ij} = cos(\bar{T}_i, \bar{T}_j) = \frac{\bar{T}_i \cdot \bar{T}_j}{|\bar{T}_i| \cdot |\bar{T}_j|} \quad (19)$$

where $\bar{T}_i, \bar{T}_j$ are vector-column of the matrix "documents-terms" corresponding to i-th and j-th terms, respectively, (i, j run over the whole list of terms), r_{ij} is the value of closeness, an element of the matrix of semantic relations. Determination of cosine in the first quadrant of Cartesian coordinates allows to state that maximum possible value of closeness between terms is equal to 1 and minimum possible one is equal to 0. As we are only interested in the strongest and stable relations, we will not take into account the value of closeness lower than a certain threshold. For each specific word we will evaluate the number of strong relations with other words and if this number is higher than a certain threshold value, we will consider this specific word to be term.

5 Experiments

To conduct the experiments, we used the textbook on general geology [30]. We divided all chapters of the textbook to documents and executed the necessary operations on preparation of target collection (i.e. tokenization and lemmatization). Table 7 shows representative words of the domain "Geology" among which there are both terms and non-terms.

Table 7. Table of representative terms in the domain "Geology"

№	Word	Value MI	Value A/B	Term
1	**rock**	**0,552**	**6,15**	**Yes**
2	chapter	0,524	2,60	No
3	**surface**	**0,495**	**5,13**	**Yes**
4	**process**	**0,436**	**3,34**	**Yes**
5	**geological**	**0,402**	**5,78**	**Yes**
6	**mountain**	**0,398**	**4,44**	**Yes**
7	**temperature**	**0,353**	**5,64**	**Yes**
8	**terrestrial**	**0,345**	**4,32**	**Yes**
9	**education**	**0,345**	**4,32**	**Yes**
10	**dynamics**	**0,344**	**5,73**	**Yes**
11	result	0,13	2,60	No
12	name	0,136	3,04	No
13	structure	0,136	3,04	No
14	material	0,126	2,27	No
15	condition	0,125	2,41	No
16	**lithosphere**	**0,103**	**4,17**	**Yes**

The decision on inclusion of the word into the class of representative words was made on the basis value MI, it had to be higher than the critical value 0.0192, and the decision on the termhood was made on this basis of value A/B, it must be higher than the critical value 3.1447. Table 8 presents specific words among which there are also both terms and non-terms.

The decision on inclusion of the word into the class of specific words was made on the basis of value PMI, it must be higher than the critical value 0.9449, and the words with the value MI > 0.0192 did not enter the class (they were already included into the class of representative words). The decision on the termhood of words was taken on the basis of the number of relations, it had to be higher than the threshold value 30. Unfortunately, we did not find the way to estimate this value not empirically and, most likely, our further works will deal with it. It is interesting that when our method selected the word "bomb" as a specific term we were puzzled. However, later, we found out that the word "bomb" is a specific geological term meaning a lump or a piece of lava blown out during eruption in a liquid or plastic state from the crater and acquiring a specific form in the course of flying and solidification in air.

Table 8. Table of specific terms in the subject area "Geology"

№	Word	Value PMI	Value B/A	Number of strong relation	Term
1	**bomb**	**1**	**0**	**136**	**Yes**
2	bombardment	1	0	4	No
3	**breccia**	**1**	**0**	**134**	**Yes**
4	storm	1	0	10	No
5	**bay**	**1**	**0**	**85**	**Yes**
6	vector	1	0	24	No
7	**hematite**	**1**	**0**	**45**	**Yes**
8	**hydrohematite**	**1**	**0**	**54**	**Yes**
9	abundance	1	0	2	No
10	**feldspar**	**1**	**0**	**66**	**Yes**
11	squid	1	0	9	No
12	stone	1	0	29	No
13	reed	1	0	13	No
14	resort	1	0	8	No
15	resort	1	0	27	No
16	**abrasion**	**1**	**0**	**82**	**yes**

Thus, from 1033 representative words were selected 617 terms, from 6489 specific words were selected 1266 terms. This once again confirms the assertion of authors of [11] that the conceptual apparatus of the domain is formed not so much by representative terms as by specific terms.

6 Conclusion

The main problem of the existing methods of automatic term extraction on the basis of contrasting approach is the need to find a compromise between the recall and the precision. Usually the choice of heuristics for the term selection is aimed at the most powerful assumption, resulting in not fully recovered rare terms or wrongly extracted words which are not terms. The proposed approach is based on the average weighted and pointwise mutual information. It provides separate extraction of both representative and specific terms that can simultaneously improve both the recall and the precision of term extraction. The results of our experiments on the "Geology" domain demonstrate that proposed method extracts much more valuable and deep information about the domain compared to the strong, but rough heuristic filters.

Our future work will be focus on comparing our method with other best heuristic filters. Especially we plan to compare our metrics with the TF-DCF metrics proposed in [24]. We conducted some opening experiments and were agree that TF-DCF is a very promising approach balancing extraction of representative and rare terms. For example, TF-DCF as well as PMI gives good estimates for rare terms such as Paleozoic, diopside, kyanite, granulite, cordierite, prominence etc. However, it gives low estimates for rare terms such as geomorphology, Paragneiss, zeolite, pseudomorphism, radioactivity whereas PMI give them high estimates.

References

1. Medelyan, O., Manion, S., Broekstra, J., Divoli, A., Huang, A.-L., Witten, I.H.: Constructing a focused taxonomy from a document collection. In: Cimiano, P., Corcho, O., Presutti, V., Hollink, L., Rudolph, S. (eds.) ESWC 2013. LNCS, vol. 7882, pp. 367–381. Springer, Heidelberg (2013)
2. Medelyan, O. et al.: Automatic construction of lexicons, taxonomies, ontologies, and other knowledge structures. In: Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discover, vol. 3, no. 4, pp. 257–279 (2013)
3. Fan, J., et al.: Automatic knowledge extraction from documents. IBM J. Res. Dev. **56**(3.4), 5:1–5:10 (2012)
4. Aggarwal, C.C., Zhai, C.X.: Mining Text Data. Springer Science & Business Media, New York (2012)
5. Nenadi, G., Ananiadou, S., McNaught, J.: Enhancing automatic term recognition through recognition of variation. In: Proceedings of the 20th International Conference on Computational Linguistics, p. 604. Association for Computational Linguistics (2004)
6. Ahrenberg, L.: Term extraction: A Review Draft Version 091221 (2009)
7. Kageura, K., Umino, B.: Methods of automatic term recognition: a review. Terminology **3** (2), 259–289 (1996)
8. Wong, W., Liu, W., Bennamoun, M.: Determination of unithood and termhood for term recognition. In: Handbook of Research on Text and Web Mining Technologies. IGI Global (2008)
9. Polya, G.: Mathematical Discovery: On Understanding, Learning, and Teaching Problem Solving. Wiley, New York (1981)
10. Heylen, K., De Hertog, D.: Automatic term extraction. In: Handbook of Terminology, vol. 1 (2014)
11. Weeber, M., Baayen, R.H., Vos, R.: Extracting the lowest-frequency words: pitfalls and possibilities. Comput. Linguist. **26**(3), 301–317 (2000)
12. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: ICML, vol. 97, pp. 412–420 (1997)
13. Kim, S.N., Cavedon, L.: Classifying domain-specific terms using a dictionary. In: Australasian Language Technology Association Workshop 2011, p. 57 (2011)
14. da Silva Conrado, M., Pardo, T.A.S., Rezende, S.O.: A machine learning approach to automatic term extraction using a rich feature set. In: HLT-NAACL, pp. 16–23 (2013)
15. Ahmad, K., et al.: University of surrey participation in TREC8: weirdness Indexing for logical document extrapolation and retrieval (WILDER). In: TREC (1999)
16. Gillam, L., Tariq, M., Ahmad, K.: Terminology and the construction of ontology. Terminology **11**(1), 55–81 (2005)
17. Peñas, A., et al.: Corpus-based terminology extraction applied to information access. In: Proceedings of Corpus Linguistics, pp. 458–465 (2001)
18. Kim, S.N., Baldwin, T., Kan, M-Y.: An unsupervised approach to domain-specific term extraction. In: Australasian Language Technology Association Workshop 2009, pp. 94–98 (2009)
19. Basili, R.: A contrastive approach to term extraction. In: Proceedings of the 4th Terminological and Artificial Intelligence Conference (TIA 2001) (2001)
20. Wong, W., Liu, W., Bennamoun, M.: Determining termhood for learning domain ontologies using domain prevalence and tendency. In: Proceedings of the Sixth Australasian Conference on Data Mining and Analytics, vol. 70, pp. 47–54. Australian Computer Society, Inc. (2007)

21. Sclano, F., Velardi, P.: Termextractor: a web application to learn the shared terminology of emergent web communities. In: Gonçalves, R.J., Müller, J.P., Mertins, K., Zelm, M. (eds.) Enterprise Interoperability II, pp. 287–290. Springer, London (2007)
22. Astrakhantsev, N.A., Fedorenko, D.G., Turdakov, D.Y.: Methods for automatic term recognition in domain-specific text collections: a survey. Program. Comput. Softw. 41(6), 336–349 (2015)
23. Kit, C., Liu, X.: Measuring mono-word termhood by rank difference via corpus comparison. Terminology **14**(2), 204–229 (2008)
24. Lopes, L., Fernandes, P., Vieira, R.: Estimating term domain relevance through term frequency, disjoint corpora frequency-tf-dcf. Knowl.-Based Syst. (2016)
25. Wong, W., Liu, W., Bennamoun, M.: Determining termhood for learning domain ontologies in a probabilistic framework. In: Proceedings of the Sixth Australasian Conference on Data Mining and Analytics, vol. 70, pp. 55–63. Australian Computer Society, Inc. (2007)
26. Prelov, V.: Mutual information of several random variables and its estimation via variation. Prob Inf Transm. **45**(4), 295–308 (2009)
27. Manning, C.D., et al.: Introduction to Information Retrieval, vol. 1, p. 496. Cambridge University Press, Cambridge (2008)
28. Hasan, K.S., Ng, V.: Automatic keyphrase extraction: a survey of the state of the art. In: ACL, vol. 1, pp. 1262–1273 (2014)
29. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. Int. J. Artif. Intell. Tools **13**(01), 157–169 (2004)
30. Sokolovsky, A.K. (ed.): A Textbook of General geology: In 2 volumes, vol. 1, p. 448. KDU, George Town (2006)