Automatic Term Recognition Needs Multiple Evidence

Natalia Loukachevitch
Research Computing Center of
Lomonosov Moscow State University
Russia, Moscow, Leninskie Gory, 1/4
E-mail: louk_nat@mail.ru

Abstract

In this paper we argue that the automatic term extraction procedure is an inherently multifactor process and the term extraction models needs to be based on multiple features including a specific type of a terminological resource under development. We proposed to use three types of features for extraction of two-word terms and showed that all these types of features are useful for term extraction. The set of features includes new features such as features extracted from an existing domain-specific thesaurus and features based on Internet search results. We studied the set of features for term extraction in two different domains and showed that the combination of several types of features considerably enhances the quality of the term extraction procedure. We found that for developing term extraction models in a specific domain, it is important to take into account some properties of the domain.

Keywords: term extraction, domain-specific thesaurus, machine learning

1. Introduction

Automatic extraction of domain terms from texts is a subject of constant interest in automatic document processing (Zhang et al. 2008; Wong et al. 2008). Contemporary information systems usually contain documents related to broad domains, which requires development of large terminological resources. Term extraction to develop such resources should be based on processing of large amount of documents. In addition, existing terminological resources need periodic updates.

It is known that manual term selection is an enough complicated, often subjective procedure (Nazarenko & Zargayouna, 2009). Besides the real practice of term selection may be quite different depending on a specific type of a terminological resource under development.

At first, terminologists collect the terminology of a domain to present it in terminological dictionaries. The main principle of term selection in this case is based on the necessity to provide term definitions (Wuster, 1979; Shelov, 2002).

Second, developers of traditional information-retrieval thesauri, maintenance of which is regulated by international and national standards (ISO-2788, Z39.19), usually create a more detailed terminological system of so-called descriptors (or in other words "authorized terms"). The main purpose of the descriptors is to cover main topics of documents in the domain (LIV, 1984). So manuals and standards on information-retrieval thesaurus development provide detailed principles for multiword term selection. For example, American standard on construction of monolingual thesauri Z39.19 considers such principles as:

- frequency in domain-specific texts and importance for the domain community (literary warrant),
- splitting the parts would lead to ambiguity or loss of a meaning,
- one component of a phrase is too vague,
- meaning of the compound term as a whole is not the sum of the meanings of its parts etc.

But the traditional information-retrieval thesauri usually do not contain very specific terms, or term variants to decrease the inconsistency of manual indexing (Z39.19, Will, 2004; LIV, 1984).

At last, if terminological resources for natural language processing and automatic indexing are created, they require an even more detailed description of the domain terminology, including ambiguous expressions, specific terms, detailed lists of term synonyms and variants to provide better matching between the pre-described terms and real texts (Spasic et al., 2005; Buitellar et al., 2006).

The aforementioned issues lead to the following conclusions:

- terminologists use a range of principles to select terms including the purpose of the resource under development. Therefore term selection in a domain is an inherently multifactor process and automatic term recognition should be based on many different factors. For example, a phrase with the relatively low frequency, mutual information and other statistical measures obtained from a domain text collection can be a term variant for a well-known term, what makes it noteworthy at least for computational terminology applications;

- multiple principles for term selection should be modelled with various linguistic and statistical features combined contemporary machine learning tools;

- it is necessary to train multifactor machine-learning models on those resources whose purpose coincides with the purpose of a terminological resource under construction;

- besides, it is important to study the possibility of the transfer of multifactor models developed for the same-type resources among different domains.

In this paper we consider an experiment on development of a multifactor model of term extraction for a specific type of terminological resources – a domain ontology for automatic text processing in information retrieval applications. We have developed a series of such resources including Socio-Political thesaurus (Loukachevitch & Dobrov, 2004), Ontology on Natural Sciences and Technologies (Dobrov & Loukachevitch, 2006), Avia-Ontology (Dobrov et al., 2003), Banking thesaurus for the Central Bank of the Russian Federation and others. To support our terminological work we now try to develop and test term-extraction models adjusted to the type of resources we create.

Extracting terms we utilize a combination of three types of features:

- features based on a domain-specific text collection,
- features obtained from an Internet search engine,
- features obtained from a domain-specific thesaurus.

Including thesaurus-based features, we simulate the situation when the thesaurus partially exists. We want to study its potential to recognise new terms. Besides, this type of features should adjust the term recognition process to a specific type of a terminological resource. An important point of our research is also to study the stability of the term extraction model among different domains.

2. Description of Experiment: Data and Evaluation

We conduct our study in two domains. The first domain is the very broad domain of natural sciences and technologies. The second one is domain of banking and bank regulation. For both domains we have Russian thesauri, developed manually, which we use as a basis for evaluation of term extraction methods (see section 2.1).

Besides, there are Russian domain-specific text collections used for development of these thesauri. From the text collections, we have extracted single words and multiword expressions. Two-word expressions belong to two types of noun groups: *Adjective+Noun* and *Noun+Noun_in_Genitive*.

The extracted expressions were initially ordered in descending order of their frequencies. Terminologists usually work with these term candidate lists paying more attention to expressions with high frequencies. However it was noted that the important terms could have medium or low frequencies because of the unbalance of text collections. So the aim of our new term extraction method is to reorder the extracted expressions to get more approved terms in the top of the candidate list. We experimented with five thousands of the most frequent two-word expressions from these lists.

2.1. Terminological Resources Used for Evaluation

Ontology on Natural Sciences and Technologies comprises Russian terminology in a very broad domain of natural sciences including mathematics, physics, chemistry, geology and elementary biology. It was created for automatic text processing of scientific documents such as automatic conceptual indexing, search results visualization, search query expansion, automatic text categorization, text summarization etc. The wide scope of the ontology is intended to support an interdisciplinary research, to serve as a general source of terminology described in a formalized way. The current volume of Ontology on Natural Sciences is more than 150 thousand terms (Dobrov & Loukachevitch, 2006).

Banking thesaurus was created during a state contract with the Central Bank of the Russian Federation. It comprises the terminology related to activity of the Central Bank, including such issues as banking activity, banking regulation, monetary politics, macroeconomics. Now it includes about 15 thousand terms.

In structure, both terminological resources are similar to classical information-retrieval thesauri (ISO 2788), having descriptors, corresponding to concepts of the domain; synonyms and term variants attached to the descriptors; relations between the descriptors.

At the same time, the resources are intended to be used in automatic text processing (in contrast to classical information-retrieval thesauri for manual indexing) and therefore they have considerable coverage of their domains, in particular, including a lot of term variants, occurred in real texts of the domain. For example, synonyms and term variants of descriptor *CURRENCY DEPRECIATION* are presented as follows:

currency devaluation,
depreciate the currency,
depreciate the money,
depreciation of currency,
depreciation of money,
devaluate,
devaluation,
devaluation of currency,
devaluation of money

This feature of our resources facilitates evaluation of term extraction methods (Nazarenko & Zargayouna, 2009). So we suppose that all term variants have been already described in our gold standards.

2.2. Measure for Evaluation of Term Extraction Performance

The evaluation of term candidates extracted from texts is a complicated procedure, because of, for example, subjectivity of domain experts, variability of terms (Nazarenko & Zargayouna, 2009).

We suppose that term extraction is needed for a broad domain with thousands of terms and term variants. A term extraction procedure is based on processing of large domain-specific text collections consisting of hundreds and thousands megabytes of texts. From these texts a ranked list of term candidates is generated. The real domain terms should be situated mainly in the top of the list to facilitate expert work or automatic exploitation of such a list. So we want to evaluate reordering performance of various methods of term recognition.

To evaluate the reordering performance of methods we use the measure of average precision adopted from information retrieval (Manning et al., 2009). Average precision AvP in the task of term extraction is calculated as follows.

Suppose that in an ordered list of expressions there are k terms, and pos (i) – the position of the i-th term from the beginning of the list. Then the precision on the level of the i-th terminological expression PrecTerm$_i$ in an ordered list is PrecTerm (pos (i)), that is the value of precision PrecTerm$_i$ is calculated at the time of inclusion to the list of i-th term and is equal to the percentage of terms in the list from 1 to pos (i) positions. Average precision for the given ordered list is equal to the average value of PrecTerm$_i$:

$$\text{AvP}=\frac{1}{k}\sum_{1}^{k}\Pr ecTerm_i$$

3. Features for Term Candidate Reordering

For extracted phrases we compute features of three types:

- features based on a domain-specific text collection,
- features obtained from an Internet search engine,
- features obtained from a domain-specific thesaurus.

Each type of features allows us to model different aspects of domain terms.

3.1. Features Based on Domain Specific Collection

We use several features calculated on the basis of a domain-specific text collection. The chosen features reveal different properties of domain terms.

Frequency in the collection (Freq). This feature is often used in term extraction methods because it is known that terms have to be frequent in domain-specific texts and the most frequent phrases of a domain include large share of domain terms.

Mutual information (MI). The feature is also very popular in extraction of terms and is calculated as follows:

$$MI(ab) = log\,(\frac{N \cdot freq\,(ab)}{freq\,(a) \cdot freq\,(b)})$$

where *ab* – is a two-word phrase, *freq ()* is the frequency of phrases or words in the collection, *N* – number of words in the collection. The feature indicates difference between real co-occurrences of a phrase and independent occurrences of phrase components.

Cubical Mutual Information (MI3). This feature is a modification of MI feature. In corpora research it was shown that this feature better orders low frequent phrases (Daille et. al., 1998):

$$MI_3\,(ab) = log\,(\frac{N \cdot freq^3\,(ab)}{freq\,(a) \cdot freq\,(b)})$$

Insideness. Insideness is calculated as the inverse ratio between the phrase frequency and the maximal frequency of a three-word expression comprising the given phrase.

$$Inside\,(ab) = \frac{freq\,(*ab*)}{freq\,(ab)}$$

This feature is intended to reveal truncated word sequences – parts of real terms. The similar phenomenon is modeled by C-value feature, described in (Maynard & Ananiadou, 2000).

3.2. Features Based on Internet Search

An important characteristic of a domain term is "termhood" that is relevance to the domain *(*Kageura & Umino, 1996*)*. The known way to estimate "termhood" is a comparative analysis of a given text collection with a contrast text collection. The huge collection of Internet texts can serve as such a contrast collection.

In previous research the Web was used for developing domain specific corpora (Penas et al., 2001; Baroni & Bernardini, 2004). (Turney, 2003) exploits the Web to obtain the most important domain terms using so called coherence feature, ranking higher term candidates that co-occur with other candidates in Web documents.

In our study we extract several phrase features from the Web and combine them with other types of features (collection-based and thesaurus-based). We obtain Internet-based features using xml-interface of Russian Search Engine Yandex on the basis of specially formulated queries. For our experiments we utilised so-called search snippets - short fragments of texts explaining search results.

Use of the Internet search is important for the following reasons. First, a text collection of a broad domain is often not sufficient because a lot of fairly significant terms may have relatively low frequencies in it. Involvement of the Internet helps us get additional information on such terms. Secondly, the use of information from the Internet allows us to find out if a given phrase is rigidly connected with the domain.

To calculate the Internet-based features, 100 snippets from search results were utilised. The snippets from the same query were merged into one document and processed by a morphological processor. As a result, for each set of snippets, lemmas (words in a dictionary form) were extracted and their frequencies of occurrence were calculated.

So, for every query we obtain a vector of lemmas with corresponding frequencies. The snippets were generated for the whole phrases and their constituent words. We denote S_{ab} – a vector of lemma frequencies derived from phrase snippets, S_a, S_b - vectors of lemmas from constituent word snippets. Using such vectors, the following types of features were calculated.

Scalar Features: Scalar1, Scalar2, Boolean1, Boolean2. The first group of Internet-based features are scalar products of snippet vectors: $<S_{ab}, S_a>$ **(Scalar$_1$)**, $<S_{ab}, S_b>$ **(Scalar$_2$)**. Many domain-specific terms have specificity of their meanings, which can not be deduced from their components (so-called non-compositionality). This specificity usually can be revealed using comparison of contexts of a phrase and its component words. The usual way to do this is to find scalar products between vectors of contexts. Also we calculated scalar products of Boolean variants of snippet vectors (vector elements are from {0, 1}) : $<Sb_{ab}, Sb_a>$ **(Boolean$_1$)**, $<Sb_{ab}, Sb_b>$ **(Boolean$_2$).**

Features of semantically specific context (SnipFreq0, SnipFreq1, SnipFreq2). Another way to find specificity of a phrase is to find a single lemma that is very frequent in phrase snippets and absent (or rarely mentioned) in component snippets.

Let lemma L occur f_{ab} times in phrase snippets and occur f_a, f_b times in snippets of components. Then we calculate SnipFreq$_0$ feature as follows:

$$SnipFreq_0 = \max_L \cdot \left(f_{ab-a-b} \cdot \log\left(\frac{N - dlcol}{dlcol}\right)\right)$$

where $f_{ab\text{-}a\text{-}b} = max\ (f_{ab}\text{-}\ f_a\ \text{-}\ f_b,\ 0)$, *dlcol* is the lemma frequency in documents of a contrast collection, N – is the number of documents in the contrast collection. Factor $\log\left(\frac{N - dlcol}{dlcol}\right)$ is so-called idf-factor known from information retrieval research (Manning et al., 2009); it helps to diminish influence of frequent general words. The contrast collection is the collection of Belorussian Internet documents distributed in the framework of Russian Information Retrieval Evaluation Seminar (www.romip.ru/ en/index.html).

SnipFreq1 and **SnipFreq2** features are calculated in a similar way excluding words in a window of 1 (2) words near every occurrence of phrase *ab.* These variants of SnipFreq feature are intended to remove partial fragments of longer terms from consideration. For example, for such macroeconomic terms as *negative cash flow* and *negative cash balance* lemmas *flow* and *balance* will be very frequent in snippets of phrase *negative cash* and will be situated immediately after phrase *negative cash,* but this phrase is not a real term.

Frequency of a phrase in its own snippets (FreqBySnip). We supposed that if the value of this feature is significantly greater than 100 (sometimes this feature reached 250-300 occurrences in 100 snippets), it means that there are many contexts in which this phrase is explained in detail, is the theme of the fragment, and, most likely, this phrase denotes an important concept or a specific entity, as, for example, phrase *internal debt* in the following snippet: *The first distinction to be made is between an internal debt and an external debt. An internal debt is owed by a nation.*

Number of definitional words in snippets (NearDefWords). This feature calculates overall frequency of so called definitional words in phrase snippets. These words (as *type, class*, *define* etc.) are often used in dictionary definitions. Therefore their presence in snippets can mean that a snippet contains a definition of this phrase or the phrase is used in definition of other term. **NearDefWords** feature is equal to the number of these definitional words that appeared immediately adjacent (left or right) to the original phrase in snippets.

Number of marker words in snippets (Markers). This feature denotes number of five-ten the most important words of the domain in snippets of the phrase. For the natural science domain these words were as follows: *mathematics, mathematical, physics, physical, chemistry, chemical, geology, geological, biology, biological.*

Number of Internet page titles (SnipTitle). We calculated number of Internet page titles coinciding with a given phrase, because we supposed that the use of the phrase as the title of an Internet page stresses significance of the phrase.

3.3. Features Based on Terms of Domain-Specific Thesaurus

In many domains there are well-known terms and even information-retrieval thesauri. The third type of our features is based on the assumption that the known terms can help to predict unknown terms. For the experiments in two domains, we used the relevant thesauri. If a phrase was a thesaurus term, then it was excluded from the terminological basis for feature generation. We considered the following features obtained from a domain-specific thesaurus.

Synonym to Thesaurus Term (SynTerm). Domain documents can contain a lot of variants of the same term (Nenadic et. al., 2004). Therefore we can suppose that a phrase similar to a thesaurus term is also a term. Let *a* and *b* be components of phrase *ab*. We consider phrase *cd* as a synonym of phrase *ab* if every component word of phrase *cd* is either equal to a component word of *ab* either is a synonym of a component word of *ab*. The order of components in the phrases is unimportant.

Synonym to Non-Term (SynNotTerm). We also fix a feature of similarity to a phrase not included to the thesaurus.

Completeness of Description (Completeness). It is possible that component words *a* and/or *b* of phrase *ab* have been already described in a domain thesaurus. For example, *a* is related to thesaurus descriptor D_a, and *b* is related to thesaurus descriptor D_b. Descriptor D_a has s_a synonyms and r_a relations to other descriptors. Descriptor D_b has s_b synonyms and r_b relations to other descriptors. **Completeness** feature is a sum of thesaurus relations of component terms that is:

$$Completeness = s_a + s_b + r_a + r_b$$

If a component of a phrase is not included to the thesaurus then its s_a and r_a are equal to 0.

4. Results of Experiments

We experimented in two domains: the banking domain and the domain of natural sciences. In all experiments 5 thousand most frequent two-word expressions extracted from the corresponding text collections were used. For these expressions, all above-mentioned features were calculated. To obtain the best combination of features for term extraction, we used machine learning methods implemented in programming package RapidMiner (www.rapidminer.com). The quality of reordering was evaluated with AvP measure. The training set was three-quarters of the phrase list, the testing set was a remaining part. As basic minimal levels of AvP we used the alphabet order and the decreasing frequency order.

To find the best combination of features for phrase reordering we tested various machine learning methods from RapidMiner package. Every time logistic regression achieved maximal level of AvP. Therefore we took this method as a basic machine learning method for our experiments on term extraction.

Table 1 shows AvP values for single features and their combination obtained with logistic regression. SynTerm and SynNotTerm features are Boolean and can not be evaluated with AvP. We concluded that SynTerm feature is highly informative: if *SynTerm* (*ab*) =1 then phrase *ab* is a domain term with probability more than 80%.

From the table we can see that in both cases the same set of features and using of machine learning methods lead to much higher values of average precision. The similarity between models can be revealed because the three features (Insideness, FreqBySnip, Completeness) among four best features coincide.

However there are significant distinctions in ratios between AvP of features between domains. For example, in the banking domain AvP of the frequency feature has the highest value, features with high average precision in the science domain have relatively low values in the banking domain.

Feature	AvP (Banking) %	AvP (Natural Sciences)%
Alphabet	40%	57%
Frequency	**57%**	66%
MI	43%	64%
MI3	45%	67%
Insideness	**55%**	**75%**
FreqBySnip	**53%**	**69%**
NearDefWords	49%	**73%**
Scalar$_1$	42%	61%
Scalar$_2$	45%	60%
Boolean$_1$	49%	64%
Boolean$_2$	48%	62%
SnipFreq$_0$	34%	66%
SnipFreq$_1$	38%	67%
SnipFreq$_2$	38%	67%
Markers	40%	65%
Completeness	**52%**	**69%**
SnipTitle	50%	-
Logistic Regression	**79% (+38.6% from Freq)**	**83% (+25.8% from Freq)**

Table 1. Average Precision (AvP) for single features and logistic regression. Feature SnipTitle was not extracted for phrases in science domain.

We explain this phenomenon with relative narrowness of the banking domain. Banking documents contain a lot of terminology of neighbour domains such as economy or politics. So among extracted expressions, there are many real terms having all specific qualities of "unithood", but not related to the banking activity. In the scientific text collection the share of terms from other domains is much lower.

Also we can see relative failure of SnipFreq$_i$ features in banking domain. The reason of this phenomenon, in our opinion, is as follows: the banking domain is subject to legal regulation, therefore documents of the domain contain a lot of citations from legal acts which leads to false large values of SnipFreq$_i$.

To evaluate the significance of the proposed features we fulfilled a feature selection procedure. For science domain the selected features were Boolean$_1$, **Completeness, FreqBySnip, Inside, MI, NearDefWords**, **SynTerm**

(AvP – 82%). For banking domain the selected features were **Completeness, FreqBySnip, MI, NearDefWords,** Scalar$_1$, SnipFreq$_0$**, SynTerm** (AvP – 78%). The selected features repeated for both domains are highlighted. We can see that in both cases all three types of features are represented in the short list of features.

5. Related approaches

For many years, researchers tried to find the best statistical feature for term extraction. However the term selection procedure is an inherently multifactor process. Now machine learning methods allow for the combination of many features.

In (Pecina & Schlesinger, 2008) the combination of statistical characteristics of phrases, based on the Czech text collection, is used to extract several types of collocations (such as phrasal verbs, idioms, terms). The authors used over 80 features and obtained 20% improvement compared with the best individual feature. The authors of this paper indicate that efficiency of various features is very variable and depends on a collection, types of expressions and so on.

In (Vivaldi et al., 2001) features for extraction of medical terms are combined using the boosting algorithm. The features include information from EuroWordNet, Greek and Latin word forms, statistical measures. Some of the features are rather domain-dependent. (Aze et al., 2005) apply the genetic algorithm ROGER to combine 13 statistical features for term extraction in two domains (biology and resources). (Foo & Merkel, 2010) study applicability of rule-based machine-learning algorithm Ripper for term extraction from patent texts.

We can see that the question of development of robust machine learning models for term extraction and the possibility of the model transfer among various domains remains open.

6. Conclusion

In this paper we argue that the automatic term extraction procedure is an inherently multifactor process and the term extraction models needs to be based on multiple features including a specific type of a terminological resource under development.

We have proposed to use three types of features for extraction of two-word terms and showed that all these types of features are useful for term extraction. The set of features includes new features such as features extracted from the existing domain-specific thesauri and features based on Internet search results. The use of thesaurus-based features allows adaptation of the term extraction procedure to the type of a created terminological resource.

We showed that the combination of several types of features considerably enhances the quality of the term extraction procedure. The developed system of term extraction reorders terms in a list of candidates much better than the basic-line ordering by decreasing frequency.

We studied the set of features for term extraction in two different domains. We found that for developing term extraction models in a specific domain, it is important to take into account such properties of the domain as broad scope or narrow scope (science vs. banking) and connection with the socio-political domain, which is regulated with legal acts. We suppose that it is possible to find the main types of domains for term extraction, to select the best feature sets and special machine learning models for every type of domains.

Acknowledgements

This work is partially supported by RFBR grant N11-07-00588-a.

References

Azé, J., Roche, M., Kondratoff, Y., Sebag, M. (2005). Preference Learning in Terminology Extraction: A ROC-based Approach. Applied Stochastoc Models and Analysis: 209-219.

Baroni, M., Bernardini, S. (2004). BootCaT: Bootstrapping Corpora and Terms from the Web. In *Proceedings of LREC- 2004*: 1313-1316.

Buitelaar, P., Magnini, B., Strapparava, C., Vossen, P. (2006). Domain-Specific WSD. *Word Sense Disambiguation. Text, Speech and Language Technology*, V. 33: 275-298.

Daille, B., Gaussier, E., Lang, J. (1998). An evaluation of statistics scores for word association. In *Proceedings. of Tbilisi Symposium on Logic, Language and Computation*. CSLI Publications: 177-188.

Dobrov, B., Loukachevitch, N. and Nevzorova, O. (2003). The Technology of New Domains' Ontologies Development. *In Proceedings of X-th Intern. Conf. KDS 2003"Knowledge-Dialogue-Solution*". Varna, Bulgaria: 283-290.

Dobrov, B., Loukachevitch, N. (2006). Development of Linguistic Ontology on Natural Sciences and Technology. In *Proceedings of LREC-2006.*

Foo, J, Merkel, M. (2010). Using machine learning to perform automatic term recognition. In *Proceedings of LREC2010 Aquisition Workshop.*

ISO-2788. 1986. Documentation -- Guidelines for the establishment and development of monolingual thesauri.

Kageura, K., Umino, B. (1996). Methods of automatic term recognition: a review. *Terminology*, *3*(*2*): 259–289.

LIV. (1994). *LIV (Legislative Indexing Vocabulary).* Congressional Research Service. The Library of Congress. Twenty-first Edition.

Loukachevitch, N., Dobrov, B. (2004). Sociopolitical Domain as a Bridge from General Words to Terms of Specific Domains. *In Proceedings of Second International WordNet Conference GWC*:163-168.

Manning, Ch., Raghavan, P., Shutze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Maynard, D., Ananiadou, S. (2000). Identifying Terms by their Family and Friends. In *Proceedings. of 18th International Conference on Computational Linguistics COLING-2000.*

Nazarenko, A., Zargayouna, H. (2009). Evaluation Term Extraction. In *Proceedings. of RANLP-2009*.

Nenadic, G., Ananiadou, S., McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. In *Proceedings of International Conference on Computational Linguistics COLING-2004*: 604-610.

Pecina, P., Schlesinger, P. (2006). Combining association measures for collocation extraction. In *Proc. of Annual Meeting of the Association for Computational Linguistics ACL-2006*.

Peñas, A., Verdejo, F., Gonzalo, J. (2001). Corpus-Based Terminology Extraction Applied to Information Access. In *Proc. of Corpus Linguistics-2001*, Lancaster University.

Sato, S., Sasaki, Y. (2003). Automatic Collection of Related Terms from the Web. *The Companion Volume to the Proceedings of 41st Annual Meeting of the ACL*, Sapporo, Japan, 2003: 121–124.

Shelov, S.D. (2003). On Generic Definitions of Terms: An Attempt of a Linguistic Approach to Term Definition Analysis. *Terminology Science & Research. Journal of International Institute for Terminology Research,* Vol.14: 52-58.

Spasic, I., Ananiadou, S., McNaught, J., Kumar A. (2005). Text Mining and Ontologies in Biomedicine: Making Sense of Raw Text. *Briefings in Bioinformatics*. Vol. 6., No. 3: 239-251.

Turney, P. D. (2003). Coherent Keyphrase Extraction via Web Mining. In *Proc. the 18th International Joint Conference on Artificial Intelligence IJCAI-03*: 434–439.

Vivaldi, J., Marquez, L., and Rodriguez, H. (2001). Improving Term Extraction by System Combination Using Boosting. In *Proc. of ICML* 2001, LNCS, V2167: 515-526.

Will, L. (2004). Thesaurus consultancy. *The thesaurus: review, renaissance and revision* / Sandra K. Roe and Alan R. Thomas, editors. - New York ; London : Haworth,.

Wong, W., Liu, W., and Mennamoun, M. (2008). Determination of Unithood and Termhood for Term Recognition. In M.Song and Y.Wu. *(eds) Handbook of Research on Text and Web Mining Technologies*, IGI Global.

Wűster, E. (1979). *Einfurung in die Allgemeine Terminologielehre und terminologishe Lexicographie*. - Vien; N.Y..

Zhang, Z., Iria, J., Brewster, Ch., and Ciravegna, F. (2008). A Comparative Evaluation of Term Recognition Algorithms. In *Proceedings of Language Resources and Evaluation Conference of LREC-2008.*

Z39.19. (2005). *Guidelines for the Construction, Format and Management of Monolingual Thesauri*. – NISO.