Learning Medical Diagnostic Knowledge from Patient Cases

Marek Jaszuk[1], Grażyna Szostek[2], Andrzej Walczak[2,1], and Leszek Puzio[1]

[1] University of Information Technology and Management, ul. Sucharskiego 2, Rzeszów, Poland,
[2] Military University of Technology, Information Systems Institute, ul. Kaliskiego 2, 00-908 Warsaw, Poland
{marek.jaszuk,grazyna.szostek}@gmail.com,awalczak@wat.edu.pl

Abstract. The paper describes a methodology designed for building medical diagnostic knowledgebase. The purpose of the knowledgebase is collecting information about diagnostic technologies, symptoms and diseases. Its key feature is the distinction between textual descriptions of symptoms and the symptoms themselves. The collection of symptom descriptions is initially built from text using natural language processing tools, and is further refined by medical experts while entering patient cases. The patient cases are the training data necessary for identifying the meaning standing behind the textual descriptions. In other words the system identifies the sets of synonymic descriptions, and the sets become the symptoms stored in the knowledgebase. The task is achieved by clusterisation of descriptions with respect to their distribution in the space of diagnosed diseases.
Keywords: medical knowledgebase, natural language processing, semantic model.

1 Introduction

Building models of knowledge is a very important topic in the domain of artificial intelligence and knowledge management systems. Such models are required for processing huge amounts of data contained in various database systems or stored in less formalized formats, like textual resources. The key problem, the creators of such models have to deal with, is the distinction between the variety of natural language expressions used to describe real word entities, and the actual meaning of the expressions. To build the model of knowledge it is necessary to identify the meanings standing behind the verbal expressions, and all the possible associations between the meanings. This is especially important for the intensively developed Semantic Web technologies and the ontologic knowledge representation [1].

There are a number of obstacles preventing from efficient building of knowledge models. The most important of them is the difficulty of making the mapping between the natural language expressions and the meanings standing behind the expressions. For a human, identifying the meaning standing behind a particular

CONCURRENCY, SPECIFICATION AND PROGRAMMING
M. Szczuka et al. (eds.): Proceedings of the international workshop CS&P 2011
September 28-30, Pułtusk, Poland, pp. 238-249

expression is not a problem. However, identifying all the possible forms of expressing identical meaning becomes a significant challenge. Also building a model which would incorporate all the possible associations between meanings is not easy. The situation becomes even more difficult when we realize the number of concepts used in some domains of knowledge. The biomedical sciences which lay within the scope of our interest use thousands of terms which are the building blocks of the model to be created.

Considering that the domain knowledge is usually contained in resources like books, technical articles, or web pages, the model building process can be supported by extracting the important information from text. This approach is founded on a number of techniques coming from the natural language processing field (NLP). The purpose of using such methodologies is identification of concepts important for the domain and the possible relations between the terms. This approach resulted in a number of ontology learning systems such as OntoLearn [2], Text-to-Onto [3] or OntoGen [4]. A good overview of the current state of the art in the field of ontology learning can be found in [5, 6]. According to the conclusions which can be found there, the contemporary ontology building systems are semi-automatic tools for doing text processing and extracting potentially relevant information. If high precision of the model is required there is no way to avoid manual verification of the model by domain experts.

In this paper we demonstrate an approach to building a knowledgebase aimed at collecting the data about symptoms, and associate them to respective diagnostic technologies and diseases. The data aggregated in the knowledgebase is the foundation for building the model of knowledge, which can be further used as the input for computational tools supporting patient diagnosis. The set of diagnostic technologies and the diseases, are not the subject of the data acquisition, because they are defined *a priori*. The data that needs to be collected refers only symptoms which are described by various natural language expressions.

The tools for supporting the diagnostic process, which we currently take into consideration are bayesian networks, and semantic networks. Both of them will not work properly if the symptoms are described using synonymic expressions. So the problem that needs to be solved to use these tools, is the proper identification of concepts standing behind the expressions. Our system is able to learn the concepts from examples through clusterisation. The data to be clusterised are the patient cases. They need to be collected not only to identify the symptoms, but also to build the computational models. All the data needed to construct the bayesian network or the semantic model are contained in the set of cases. As a result, the process of building the model of symptoms and accompanying computational tools becomes completely automatized, and no direct manipulation to the models is required.

An important element which supports the model creation process is text processing used for extracting the expressions describing symptoms from text. These expressions form the initial contents of the set of descriptions, which are further used for describing patient cases. It should be underlined, that the described system is built for the Polish language. The method of text processing

is based mainly on the language inflective character, and thus moving it to other languages requires significant changes. This especially refers to non-inflective languages like English.

The paper is organized as follows. Sec. 2 discusses the general structure of the knowledgebase. In Sec. 3 the methodology used for extracting symptom descriptions from text is described. Sec. 4 presents the method used for refining the set of symptom descriptions, and collecting patient cases as the training data for the system. Sec. 5 presents how the model of symptoms is built through clusterization and statistical analysis of cases. The general assumptions about constructing the computational models are also presented there.

2 The Structure of the Knowledgebase

When considering the medical diagnostic knowledge, three things come to mind: diagnostic technologies, symptoms, and diseases. The diagnostic technologies are a tool for collecting information about a patient. The information has a form of symptoms. All of the symptoms are expressible as verbal expressions. For a physician this is a daily routine to describe symptoms using natural language. The problem is, however, that the verbal expressions are not a good input for a computational system aimed at diagnosing a patient. The reason is that many of the symptoms are describable using multiple verbal expressions. In other words, the descriptions have their synonyms. Moreover the meaning of particular expressions can have a different range.

The input required for a diagnostic system is the actual meaning, i.e. we want the system to interpret the synonymic descriptions in the same manner. For a physician, identification of meanings standing behind a description is not a problem. Unfortunately, for a computer system this is a serious problem. There are a number of approaches to automatic synonym identification using text processing. Most of them are based on the Harris distributional hypothesis [7]. Such approaches are not applicable to our purposes. The reason is that the distributional hypothesis is applicable to single words or to simple phrases. The meaning we are searching for refers to more complicated verbal constructions.

The symptom descriptions take very different forms. The simplest ones are single nouns, like *gorączka* (eng. fever). But in general case a symptom description takes a form of a more complicated verbal expression including almost every part of speech, like nouns, adjectives, verbs, adverbs, prepositions, etc. As an example let us serve a sentence taken from a computer tomography record *Zmiany włókniste w płacie dolnym płuca prawego* (eng. fibrous changes in the bottom lobe of the right lung). There is no NLP method which would allow to identify the synonymic meaning among such a variety of expressions. Thus in our approach we assume to learn the meanings not by text processing, but by learning from examples. The details of this process are described in Sec. 5.

The distinction between the natural language symptom descriptions and their meaning is reflected in the structure of the knowledgebase. The descriptions and the symptoms are separate entities (see Fig. 1). The remaining components are

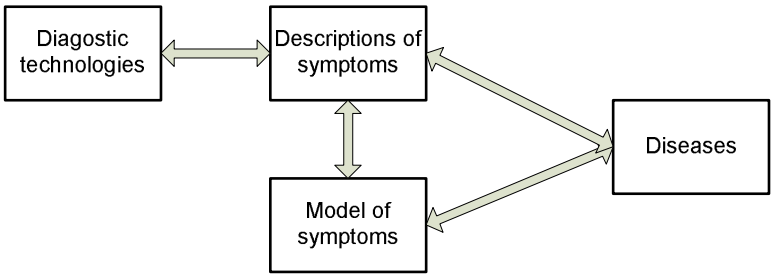

Fig. 1. The structure of the medical diagnostic knowledgebase

the technologies and the diseases. The technologies and the descriptions are responsible for human communication. Before describing any symptom the user needs to decide what kind of diagnostic procedure the symptom is resultant from. The diagnostic technologies introduce modularity within the set of all possible descriptions. Every technology has a set of associated descriptions and every description belongs to some technology. Thus by choosing one of the technologies the user restricts the set of possible descriptions he can choose from. This allows for more efficient searching among the descriptions.

The symptoms and diseases are responsible for the computational process, which aims at diagnosing a patient. The diseases also take part in the human communication process. In the phase of gathering the training data for the system, the expert users enter the diseases diagnosed for patient cases. For the end user the diseases will be presented as the result of the system diagnosis. The symptoms represent the meanings hidden behind the set of descriptions. The model of symptoms is determined automatically from the patient cases and the users have no direct control over this model. The method used for symptom identification is discussed in Sec. 5.

The additional element which accompanies the knowledgebase is the collection of patient cases. The cases are used only for extracting the knowledge model, and thus are not a permanent component of the knowledgebase. The knowledgebase is also accompanied by computational tools created to support the diagnostic process. The tools are strictly related to the model of symptoms. In fact the set of symptoms can be different depending of the computational tool to be used. The symptoms required for the Bayes network can be created by any clusterisation method, because no special structure within the set is required. The semantic network assumes hierarchic representation of knowledge, and thus the set of symptoms should be organized in this way. This is achieved by more sophisticated clusterisation algorithms.

3 Extraction of symptom descriptions from text

The foundation for training the system is patient cases. Every case consists of textual descriptions of the patient symptoms and the disease diagnosed in a

patient. The set of diseases is fixed, because it is taken from a catalogue. So the only element we need to determine, are symptom descriptions. These could be created from scratch by humans, but they could also be extracted from medical texts using NLP methods. This simplifies and speeds up the work of expert users responsible for collecting the cases. Thus an efficient text processing mechanism is required, which will allow for extracting valuable verbal constructions from text.

The text processing mechanism is founded on the observation, that from the perspective of verbal construction, every symptom description has some common structure. This structure has a form of a tree of words with root being a noun in the nominative case. The branches of the symptom description tree are formed of the words associated to the root noun. The case, of course, can be determined only for inflective language like Polish. The described methodology is thus language specific. There are some symptom descriptions, which do not contain the noun in nominative, but using them is rare, and thus they are not taken into account. The process of extracting symptom descriptions from text consists of the following steps:

1. decomposition of text into sentences;
2. morphological analysis of words within individual sentences (tagging);
3. disambiguation of morphological tags;
4. discovery of morphologically related words;
5. discovering relations using sentence patterns;
6. identification of nouns in the nominative case and building trees of words associated to every such noun;
7. reduction of every tree to a flat sequence of words;

The key resource which allows for performing the text analysis is the morphological analyzer. Our system uses the Morfeusz software package [8]. It assigns one or several tags expressing potential morphological interpretations to the analyzed word (lexeme form). The analyzer is based on a system of tags developed for the IPI PAN Corpus [9, 10]. The contents of the tags includes the basic form of the lexeme, information about the part of speech (lexeme class - noun, adjective, verb, etc.), number (singular or plural), case (nominative, genitive, etc.), gender (feminine, masculine, etc.), and several other pieces of information.

3.1 Identification of Word Associations

The most important step in extracting the symptom descriptions from text is identification of word associations. The purpose of this process is to eliminate the lexical and syntactic polysemy and identify relations between words using linguistic rules. There are a number of such rules which are characteristic for the Polish language, and their use in sentence construction indicates related words. Below are some of the most important rules used for disambiguation:

– linking preposition
  A compound consisting of preposition and a noun is expressed by inflectional noun ending, which is specific for the case acceptable in this link.

- links between nouns and nouns in genitive
  As it can be observed, when two nouns are directly next to each other in a sentence, the last of them is usually in the genitive case. This feature allows to disambiguate the category of the case for the second noun.
- links between nouns and adjectives
  The dependency between a noun and an adjective is expressed by the characteristic inflectional endings. These endings are characteristic for the number, case and gender, which are common for both of the words.

When the linguistic rules are applied we are able to identify the lexeme forms which are related and eliminate the lexemes which do not create relations. After applying the linguistic rules also the knowledge about the subject and the predicate of a sentence is collected. This knowledge will be used when applying sentence patterns. The linguistic rules allow also for establishing relations between words. Some of the most important relation types, resulting directly from the rules are listed below:

- noun - adjective
  It is a relation which occurs between a noun and the corresponding adjective, e.g. *płuco prawe* (eng. *right lung*), *wydzielina ropna* (eng. *purulent discharge*), *ciśnienie niskie* (eng. *low pressure*), etc.
- noun - noun in locative
  It is a relation between two nouns, where the second noun is in the locative case. Morphological analysis discovers only the argument in the locative case, which in case of symptoms specifies the place of occurrence, e.g. *w płucach* (eng. *in lungs*), *na powierzchni* (eng. *on the surface*), *we krwi* (eng. *in blood*), etc. The argument specifying what occurred in the specified place remains to be found in the sentence. As it could be observed the noun in the locative case has also an associated preposition, which is a result of a separate rule.
- noun - noun in genitive
  This type of relation associates two nouns occurring in the text immediately next to each other, where the second noun is in the genitive case. For example: *skóra głowy* (eng. *skin of the head*), *masa ciała* (eng. *body weight*), *grzybica stóp* (eng. *mycosis of feet*), etc.

Let us analyze the already mentioned sample sentence:
Zmiany włókniste w płacie dolnym płuca prawego (eng. Fibrous changes in the bottom lobe of the right lung).

The linguistic rules allow for generating the following set of relations from the sentence:

- noun - adjective: *zmiany włókniste* (eng. *fibrous changes*);
- noun - adjective: *płacie dolnym* (eng. *bottom lobe*);
- noun - noun in genitive: *płacie płuca* (eng. *lobe of the lung*);
- noun - noun in genitive: *płuca prawego* (eng. *of the right lung*);
- preposition - noun in locative: *w płacie* (eng. *in lobe*);
- noun - noun in locative: *? w płacie* (eng. *? in lobe*).

In the last relation the preposition and the noun in locative are treated as one entity. This is because the relation refers to them as a whole. It can also be observed that the last relation has an unidentified element which is not indicated by any linguistic rule. Assuming that the noun fitting the relation is the closest noun before the noun in locative, we get the missing argument of the relation. The resulting relation is thus: *zmiany w płacie* (eng. *changes in the lobe*). It should be remembered, however, that in general case resolving the missing argument of the rule is not so simple, because of free word ordering in the Polish language.

After assembling all the relations into single structure we get a tree of words. The root of the tree is the noun in nominative *zmiany* (eng. *changes*). The branches of the tree are formed of the remaining words. As it has already been mentioned every symptom description contains at least one noun in the nominative case. This is what distinguishes the potentially interesting verbal constructions from all the other. So the extracted tree is for our system a candidate for a symptom description.

After reducing the extracted tree back to the flat sequence of words, we get its version which is readable for the human user. The above example is quite idealized, because all the words from the sentence were associated into one tree, and thus become an element of a single symptom description. In many cases, however, we get multiple verbal constructions. Some of these constructions are eliminated completely, because of lacking noun in nominative. The use of sentence patterns also has not been demonstrated. The patterns are applicable only in situations where a verb is used. They allow for making associations between the subject, the verb and the remaining elements of the sentence.

3.2 Results of Text Analysis

The text corpus used for the experiments came from two domains of medicine: allergology and pulmonology. To be more precise the experiments were carried out separately on texts from the two domains. The size of the corpuses is rather small. For allergology it is 95kB, and for pulmonology it is 265kB. Unfortunately there are not too many texts in Polish which could be used for the analysis, and thus the small size of the corpus. The main text resources were [11] for allergology and [12] for pulmonology. We selected only the book chapters and paragraphs, which actually describe symptoms. Including any other fragments of texts would deteriorate the results. This results from the fact that the analyser is based only on the grammatical construction of the sentences. It is not able to interpret the meaning of the analysed text. The grammatical structure of symptom descriptions is no different than grammatical structure of any other entity. As a result any text processed by the analyser delivers a set of descriptions, no matter if it refers to symptoms or not. The careful selection of texts is thus important, if we want to avoid getting too many useless descriptions. As a result of text processing we got 1080 descriptions for allergology and 2810 descriptions for pullmonology. The difference in numbers is the obvious consequence of the

corpora sizes. Such a collection of descriptions seems to be sufficiently large to be the starting point for describing the patient cases.

4 Gathering the Training Data for the System

The collection of descriptions extracted from text is of course far from perfect. It strongly depends on the actual contents of text corpus. As already mentioned, the mechanism extracting information from text is based only on morpho-syntactic rules and is not able to interpret the meaning of extracted information. As a result the collected descriptions include except symptoms, also a lot of other unwanted information. Also some part of the descriptions is incorrect due to grammatical ambiguities which we were not able to resolve.

Fortunately, the unwanted information is not so huge problem, as it could initially seem. The condition is an efficient searching mechanism, which allows for quick finding of the desired description in the database. Given such mechanism, medical experts can quickly describe symptoms observed in patients. The most efficient searching mechanism that we are able to deliver is based on suggestions to a typed sequence of characters. This mechanism is well known from the Google search web site. Using this mechanism the user is always able to find the desired description after typing an adequate number of characters. The search mechanism is additionally supported by weights assigned to the descriptions. The weights indicate the descriptions, which are frequently used, and should be moved to the top of the search list. Using the described tool the experts create a database of patient cases, being the training patterns for the system.

Of course we are not able to guarantee that any possible symptom description that an expert could ever think of is available in the collection extracted from text. Thus the description chosen by the user should be open for edition. In this way it is always possible to complete or correct the missing parts of the expression, or even build it from scratch. Every new description is then registered in the system and available for other users. Collecting patient cases leads to refining the whole set of symptom descriptions, and leaving only the descriptions which are actually useful.

5 Building the Model of Symptoms

As mentioned earlier the symptoms are the meanings hidden behind the set of textual descriptions. Some of the descriptions represent identical or close meaning. The purpose of this phase is thus identification of sets of descriptions with synonymic meaning. For a human this is a difficult task, because, he needs to deal with a large vocabulary, and the differences in meaning are sometimes very subtle. Fortunately this task can be automatized given the patient cases.

Before introducing the method used for identifying particular meanings, first we have to realize the actual sense of the term “meaning”. It comes out from the purpose of the designed system. The meaning should indicate the possible diseases given the set of symptom descriptions entered by the user. Thus the

meaning results from the associations between the descriptions and the diseases. The associations are delivered within the cases collected as the training data for the system. To identify the meaning of a particular description, it is enough to analyse its statistical distribution with respect to particular diseases in which it appeared. The descriptions which have the same meaning also have the same distribution. If it is not true, this means, that users by choosing one or the other description associate them with different diseases. Thus their actual diagnostic meaning is different.

The above considerations are valid for the descriptions which were used regularly during collecting cases. Only then their distribution can be determined with sufficient precision. Thus the descriptions used occasionally are eliminated from the system. Their rare use indicates, that they might be not well formulated, and are not what most of the users are searching for. Of course the rare use can result from rare occurrence of some symptoms. The importance of a given description can be checked by analysing its correlation with particular diseases. If no such correlation could be found the description introduces no value to the system and thus should be omitted.

If the similarity of distributions determines the meaning of particular descriptions, the task of identifying the symptoms can be realized by clusterisation. The clusters will group the descriptions indicating the same diseases, and thus having the same diagnostic meaning. It should be underlined, that the diagnostic meaning is not the same as the linguistic meaning. The natural language descriptions can introduce many expressions, which from the linguistic perspective have distinctive meanings. The meanings can be, however, not distinctive enough from the diagnostic perspective. Let us take for example two expressions: *ból brzucha* (eng. *belly ache*) and *silny ból w okolicy brzucha* (eng. *strong ache in the belly region*). From the linguistic perspective they are not the same. This is mainly because the additional adjective *silny* (eng. *strong*) appearing in the second phrase. But from the diagnostic perspective the difference is very subtle. Both of the descriptions could appear in the same diseases. If the cases will indicate that the two phrases have similar distributions with respect to diseases, they will be classified as representatives of the same symptom.

It should be reminded that the set of diseases, which is the foundation for determining the diagnostic meaning is fixed and defined *a priori* by experts. This makes the analysis much easier. Otherwise we would be forced to lead the analysis in the space spanned by the synonymic names of diseases and take into account the possible relations between the disease names.

The described methodology leads to transforming the natural language descriptions into a set of symptoms. Given the symptoms we can construct computational model for supporting diagnosis. The approach which seems the most obvious and straightforward is constructing the Bayes network. Building it requires probabilities of particular symptoms and their combinations appearing in diagnosed diseases. The probabilities are easily determined from the probabilities of particular descriptions appearing in the cases. The probability of a symptom is the sum of probabilities of its particular descriptions.

But the Bayes network is not the only model which can be built given the cases. In practice any computational model can be constructed, which can be learned from examples. We are considering building a semantic network model from the cases. Such a model is more sophisticated, because it is built as a network of relations between a number of concepts. The concepts in our case are symptoms and diseases. There are in general two types of relations in every semantic network: vertical and horizontal. The vertical ones are the type of association which relates more general concepts to more specific ones through the *isSubclassOf* relation. In this way the more general concepts become superclasses and the more specific ones their subclasses. Associating the classes through the described relation leads to a hierarchic structure of all concepts in the domain. The horizontal relations are all the other associations between the concepts, where no particular hierarchy is assumed.

The purpose of vertical relations in our case is to create the hierarchy of symptoms. The diseases are assumed to be unrelated, so we do not need to look for any relations between them. The vertical relations between symptoms can be easily seen when analysing samples of descriptions. Let us take for example the two descriptions: *ból głowy* (eng. *headache*), and *napadowy ból głowy* (eng. *paroxysmal headache*). The first of them is the superclass of the second one. The additional adjective makes the description more specific, but the *napadowy ból głowy* is still *ból głowy*. Implication in the opposite direction is not true. The example illustrates quite a typical situation, where adding any adjective to a more general symptom description makes the description more specific. Thus the description can be considered a representative of a subclass of some more general class in the model.

The presented example is based only on the linguistic interpretation of descriptions. It should be remembered, however, that the discussed model is based on the notion of diagnostic meaning. Thus we cannot be sure if the two presented descriptions are representatives of two distinct classes. This will be true if the added adjective is significant enough, that the modified phrase indicates different diseases. In many cases the additional words modifying the original phrase will not introduce relevant diagnostic meaning, and thus a new class will not be created.

The presented considerations show the relevance of the vertical hierarchy in the model. To build the structure a hierarchic clusterisation algorithm can be applied. In such an approach the descriptions representing more general diagnostic meaning will be identified as clusters containing subclusters of descriptions with more specific meaning.

The remaining element of the semantic model structure is the horizontal relations. The most important of them associate symptoms and diseases. Such relations come directly from the patient cases after clusterisation of descriptions. Of course we are interested in relations with significant statistical meaning. Thus the statistical analysis of cases allows for extracting the relevant relations among all of the possible. Also some relations between symptoms are possible. Such relations can be both important for supporting diagnosis and for indicating symp-

toms desirable for improving the quality of diagnosis. The second case could be used for suggesting a physician additional examinations to be done. Practically all the statistically relevant relations between symptoms can be detected through analysis of cases. The most important of the relations seem to be the mutual co-occurrence of particular symptoms. Such a relation can indicate the typical symptom configurations for particular diseases.

6 Conclusion

In the paper we described a methodology designed for building a model of medical diagnostic knowledge. The model consists of four main components which are: diagnostic technologies, verbal descriptions of symptoms, model of symptoms, and diseases. The key point in building the model is collecting diagnosed patient cases. As the whole diagnostic knowledge in medicine is based on diagnosed cases, this seems the best of possible approaches to building a computer system able to diagnose patients. The knowledge acquired in this way is not disturbed by any human interpretations. The important problem that we had to solve is the human-computer communication factor. This requires identifying the set of verbal expressions valuable for describing symptoms. The expressions are initially extracted from text and further refined by a team work of medical experts. For computational purposes extracting the meanings standing behind the expressions is obligatory. The meanings are extracted from the set of patient cases by means of clusterisation.

The data extracted from the cases leads to constructing a model of symptoms, which by further analysis can be transformed into a computational model. Such a model can be a powerful tool for supporting patient diagnosis by indicating the diseases that the patient possibly can suffer from, as well as by suggesting other medical examinations to be done in order to improve the diagnosis. Our considerations are focused on constructing a bayesian network and a semantic network. In practice, however, any computational model which is possible to construct by learning from examples can be taken into account.

It should be also underlined, that the model is easily extendible by adding appropriate cases. The symptom model is constructed automatically, so no manual manipulation in its structure is required. Adding a new technology requires adding patient cases containing results of the technology. Adding a new disease is done in a similar way. It requires collecting a set of cases with the disease diagnosed. This allows for determining appropriate distributions of symptoms associated to the newly added disease.

The paper is focused on explaining the structure of the knowledgebase, and justifying the method used for identification of meanings. We realize that many of the details need further explanation. This especially refers to the NLP methods used for extracting the symptom descriptions from text and the algorithms used for constructing the computational models. As there was not enough space to discuss all the details they will be described separately.

Acknowledgement

This work was financially supported by the European Union from the European Regional Development Fund under the Operational Programme Innovative Economy (Project no. POIG.02.03.03-00-013/08).

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web, Scientific American Magazine, (2001)
2. Velardi, P., Navigli, R., Cucchiarelli, A., Neri, F.: Evaluation of ontolearn, a methodology for automatic learning of ontologies. In: Buitelaar, P., Cimmiano, P., Magnini, B., (eds.) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press (2005)
3. Maedche, A., Staab, S.: Ontology learning for the Semantic Web, Intelligent Systems, IEEE 16, pp. 72–79 (2001)
4. The OntoGen system, `http://ontology-learning.net/wiki/OntoGen`
5. Buitelaar, P., Cimiano, P.: Ontology Learning and Population: Bridging the Gap between Text and Knowledge. Series information for Frontiers in Artificial Intelligence and Applications, IOS Press (2008)
6. Wong, W.: Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge. Doctor of Philosophy thesis, University of Western Australia (2009)
7. Harris, Z.: Distributional Structure. Jerrold J. Katz (ed.) The Philosophy of Linguistics. Oxford University Press, pp. 26–47 (1985)
8. Woliński, M.: Morfeusz - a Practical Tool for the Morphological Analysis of Polish, Intelligent Information Processing and Web Mining, Advances in Soft Computing, vol. 35, pp. 511-520. Springer, Berlin / Heidelberg (2006)
9. Przepiórkowski, A.: The IPI PAN Corpus. Preliminary Version. Institute of Computer Science PAS. Warsaw (2004)
10. Woliński, M.: System znaczników syntaktycznych w korpusie IPI PAN. XXII/XXIII. pp. 39-55, Poloniki (2003) (in Polish)
11. Burgdorf, W.H.C., Plewig, G., Wolff, H.H., Landthaler, M.: DERMATOLOGY Braun-Falco, Czelej (2010) (in Polish)
12. Szczeklik, A.: Internal diseases. Medycyna praktyczna, Kraków (2006)(in Polish)