Building a Model of Disease Symptoms Using Text Processing and Learning from Examples

Marek Jaszuk*†, Grażyna Szostek†, Andrzej Walczak†* and Leszek Puzio*†
*University of Information Technology and Management Rzeszów, Poland
†Military University of Technology Warsaw, Poland
Email: marek.jaszuk@gmail.com, grazyna.szostek@gmail.com
awalczak@wat.edu.pl, lpuzio@wsiz.rzeszow.pl

***Abstract*—The paper describes a methodology of building a semantic model of disease symptoms. The fundamental techniques used for creating the model are text analysis and learning from examples. The text analyser is used for extracting a set of symptom descriptions. The descriptions are a foundation for delivering a user interface, necessary for collecting patient cases. Given the cases a semantic model is built, which is achieved through clusterisation and statistical analysis of cases. The approach to creating the model eliminates the need of direct model manipulation, because the meaning is retrieved from association to diseases instead of purely linquistic interpretation of symptom descriptions. Detection of synonyms is also completely automatized.**

I. Introduction

BUILDING models of knowledge is a very important topic in the domain of artificial intelligence and knowledge management systems. The most common approach is associated with the intensively developed Semantic Web technology. This technology requires representing knowledge in the form of an ontology. Such models are typically built for some specific domains such as biomedical sciences, or various branches of business and industry. The fundamental element of every ontology is a set of concepts from the particular domain. The concepts are combined using a set of semantic relations defined within the ontology. As a result we get a hierarchc structure in the form of a directed graph with nodes being the ontology concepts (semantic classes), and the edges being the semantic relations.

Unfortunately the task of ontology building requires a lot of effort, and engagement of experts from the given field. One of the most important obstacles, that the ontology builders have to overcome, is the proper identification of concepts which should be used as the ontology nodes. It is assumed, that the nodes have to represent particular meanings instead of their possible verbal descriptions. Thus the ontology building process requires identifying synonyms among the verbal descriptions. To represent the meaning, usually one of the possible descriptions is chosen. The problem is, however, that particular verbal expressions can represent multiple meanings depending on the context of their use. As a consequence they can be classified as representatives of completely different meanings. Another difficulty are the subtle differences in meaning between the particular expressions. In consequence it is frequently very difficult to decide, whether the particular expressions should be considered representatives of the same concept or their meaning is distinctive enough to separate them. All such decisions are left to the person constructing the ontology, and are the reason of frequent hesitations, which slow down the whole process.

The high labour consumption is not the only consequence of the difficulties mentioned above. The drawback of most ontologies constructed today is their highly subjective character. All individual decisions about defining particular classes and the possible relations between them influence the final shape of the constructed ontology. As a result a given domain can be described by many different models. This is obviously not what is desired. The domain knowledge is only one, and the properly constructed model should be independent of the particular persons building it.

Another obstacle against efficient ontology building is the large size of the models that need to be developed for real world problems. The size is a simple consequence of the huge number of concepts and the possible relations between them. Considering that the domain knowledge is usually contained in resources like books, technical articles, or the Internet, the ontology 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 that have already been created. Some of the most well known examples are: OntoLearn [1], Text-to-Onto [2], OntoGen [3], ASIUM [4], TextStorm/Clouds [5], SYNDIKATE [6], ISOLDE (Information System for Ontology Learning and Domain Exploration) [7]. There is number of approaches based on utilising clustering algorithms [8]–[10] for building ontologies. A good overview of the current state of the art in the field of ontology learning can be found in [11], [12]. To assess the ontology learning methodologies, several surveys have been made [13]–[15]. According to their findings most of the systems are semi-automated tools for supporting domain experts in creating ontologies. Complete automation and elimination of user involvement is hard and can be applied only in cases where high quality of the knowledge model is not obligatory.

978-83-60810-22-4/$25.00 © 2011 IEEE

In this paper we demonstrate an approach to building a model of medical symptoms in association with a set of diseases. The obstacles encountered during building this type of a model belong to the categories already mentioned. The first of them is the reach vocabulary used for describing symptoms, which takes effect in a huge size of the model to be created. The particular symptoms can be described by different combinations of words. The additional difficulty results from the fact that the descriptions which could be considered synonymic, do not always represent exactly the same meaning. There are frequently subtle differences in meaning between the possible alternatives. This results in additional difficulties in indentifying the important concepts. In consequence using standard methodologies requires a lot of effort and time to complete the task. Moreover the final result is always influenced by the personal habits, and knowledge of the model creator.

In our approach a different methodology is employed. It allows for avoiding the most important problems. One of the main assumptions is using NLP methods for text analysis and extracting information which could be important for the assumed task. This delivers a huge database of verbal constructions which are potential descriptions of symptoms. The second stage of the work is based on utilizing the database of verbal constructions. This approach is however completely different than the standard methodologies. It does not assume any direct manipulation of the model by human experts. This stage is replaced by collecting patient cases, and completing the model construction by training the system on these cases. The identification of concepts is achieved by applying clustering algorithm in the space of possible descriptions. In this way the clusters represent the particular meanings, which are the building blocks of the model. The advantage of this approach is that no human is responsible for identifying the meaning standing behind the verbal descriptions. The only expectation from the experts entering the cases is that they should describe the symptoms according to their best knowledge. To do that they can use the set of descriptions delivered by the text analyzer, but they are not restricted to it. They do not need to obey any special restrictions on the verbal constructions to be used. It is even advantageous if the cases are entered by different specialists having different habits. In this way the model structure is resistant to the subjectivity of experts taking part in its creation. Also the human effort during construction of the model is lesser than required during direct manipulation of the model. This is a consequence of the fact that conscious analysis of a complex model is replaced with a relatively simple task of describing cases.

It should also be mentioned that the described system is build for the Polish language. To be more precise, the language specific features are implemented in the module used for text processing. This module is responsible for identifying associations between words in text. The main features of the language, which influence the module structure is extensive inflection and free order of words in sentence. Using linguistic rules and sentence schemas, we are able to identify sets of associated words forming tree structures. These structures are potential symptom descriptions. Of course a tree of words is not a natural knowledge representation for a human user. To increase the readability of the symptom representation, the tree structures are reduced to flat sequences of words, which resemble the original representation of knowledge extracted from sentences. Such descriptions form the initial database of symptoms which is further purified during the process of collecting cases and learning from examples. Except of the method of creating trees of associated words, the remaining part of the system is universal and free of language specific features. Thus it can easily be moved to another language, if an appropriate method of identifying associations between words would be developed.

The paper is organized as follows. Sec. II discusses the methodology of text processing used for extracting symptom descriptions. In Sec. III the process of collecting patient cases is presented. Sec. IV presents how the semantic model of diagnostic knowledge is built through clusterization and statistical analysis of cases. In Sec. V the results of experiments with text processing are presented.

II. Extraction of symptom descriptions from text

The text searching mechanism is founded on the observation, that from the perspective of verbal construction, every symptom descriptions have a common structure. This structure has a form of a tree of words with root being a noun in the nominative case. The case of course can be determined for inflective language like Polish. Every symptom description contains at least one noun in nominative. The branches of the symptom description tree are formed of the words associated to the root noun.

The discussed text analysis methodology has some common elements with other known algorithms. First of all it includes gramatical tagging of words. This task can be solved using several different approaches. Some of the examples are the Stanford POS tagger [16] or the MXPOST [17]. For Polish the most well known tagger is the TaKIPI [18]. Another issue is the analysis leading to finding the relations between words in a sentence. An example of a system realizing this task is the Multparser [19]. Our approach is not directly based on any of the existing solutions, however, it contains some of their elements. The discussed system is strictly task specific, and developing our own solution allowed for introducing the necessary optimizations. Although this does not mean that the solution is not applicable to other domains after some minor modifications. We do not do any comparisons to other algorithms here. This is because the paper is devoted to presenting the general idea of the methodology designed to build the model of medical diagnostic knowledge. Although we realize that the comparison of the text analysis algorithms from the computational linguistics perspective is a very important issue and this will be the subject of separate study.

The process of extracting symptom descriptions from text consists of the following steps:

1) decomposition of text into sentences;

2) reading individual sentences, and morphological analysis of words;
3) disambiguation of morphological tags;
4) discovery of morphologically related words;
5) discovering relations using sentence schemas;
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 details of every step will be described in subsequent sections.

A. Specifics of the Polish Language

As already mentioned the text analysis strongly depends on the specific features of the language for which the system is developed. There is a number of characteristic elements of the Polish language which were taken into account while building the module. Below are listed the most important of them for the defined task [20]:

- inflection
  This is a language feature meaning, that words are inflected by case, number, person, etc. The Polish language belongs to the group of inflectional languages. Inflectional properties of words influence parsing sentences. The inflection allows for determining roles of words in sentences.
- discontinuity of phrases
  Elements of noun and verb phrases do not have to occur directly next to each other in a sentence.
- free order of words in sentences
  Words of a sentence may appear in different order without affecting the meaning of the whole sentence.
- lexical polysemy
  Lexical polysemy occurs when two or more words have the same form. For example, the form *drogi* (eng. *roads*) has two lexemes:
  a) *droga* (eng. *the road*) - noun, plural, feminine gender,
  b) *drogi* (eng. *dear*) - adjective, singular, masculine gender.
- syntactic polysemy
  Syntactic polysemy occurs when several forms of the same lexeme are identical. For example, morphological analysis of the form *okna* (eng. *windows*) will give a number of interpretations, including: *noun:singular:genitive*, *noun:plural:nominative*, *noun:plural:locative*.

B. Morphological Analyzer

The morphological analyzer is a very important resource used during text processing. Our system uses the Morfeusz software package [21]. 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 [22], [23]. 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 a several other pieces of information. The analyzer data are represented in the form of finite state machines, which makes new word forms analysis impossible. Also analysis of the word context is not done (the program is not a tagger). So the problem of lexical and syntactic polysemy remains to be resolved.

C. Disambiguation of Words

Polysemy is an important factor influencing the effectiveness of discovering verbal associations. The morphological analyser generates multiple morphological tags for many of the words found in text, while only one of them is the correct one. Taking the incorrect tag leads to erroneously constructed tree of verbal associations, and as a result incorrectly extracted description. Disambiguation is thus important for reducing the number of errors in the results of text analysis.

Some of the tags can be eliminated *a priori* taking into account the character of the domain to which the text corpus refers. In this way we are able to eliminate all the tags including the vocative case, as this case is not used in medical texts. An example of a lexeme form eliminated in this way is *szybko* (eng. *fast*), which is an adverb, but could also be interpreted as the noun *szybka* (eng. *glass*) in the vocative case. Other examples of tags eliminated in this way include verbs in the imperative mood. An example is the lexeme form *dym* (eng. *smoke*) which can be interpreted as a noun in the nominative case, as well as a verb in the imperative mood. Of course for the medical texts only the first interpretation makes sense. Yet another example are depreciative forms of nouns, which tend to diminish in value the described entity. An example is *inne* (eng. *different*), which in medical texts is used only as an adjective, so the noun interpretation can be eliminated.

The second method of word disambiguation allows for reducing lexical polysemy. It is based on the observation, that the lexeme form interpretation which is inappropriate for the given context is not a subject of declination (by case, by number, by person, or by gender). As a result the number of forms in which a given lexeme appears in the text corpus is very limited (usually to one of the possible forms). As a result we are able to create a disambiguation dictionary consisting of lexemes inappropriate for the given domain. The lexemes are ignored when appropriate word interpretation is searched for. The examples of such lexemes are: *lewy* - the interpretation as an adjective is correct (eng. *left*), while the noun interpretation is ignored (bulgarian currency), *małe* - the adjective interpretation is correct (eng. *small*), the noun interpretation is ignored (eng. *young*), *normalna* - the adjective interpretation is correct (eng. *normal*), the noun interpretation is incorrect (eng. *normal (line)*).

The two presented methods of disambiguation do not guarantee removing the polysemy completely. The remaining ambiguities we try to resolve using linguistic rules, which is discussed in the subsequent section.

D. Identification of Word Associations

The objective of this stage is to eliminate the lexical and syntactic polysemy and identify relations between words using linguistic rules. There is 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. Of course this method does not allow for elimination of ambiguities completely. Sometimes there is more than one alternative of word association, which is allowed by the linguistic rules. In such a case all the possible variants of word associations are built, assuming that one of them is the one we are searching for. 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.

Except the tasks mentioned above using the linguistic rules allows 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 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.

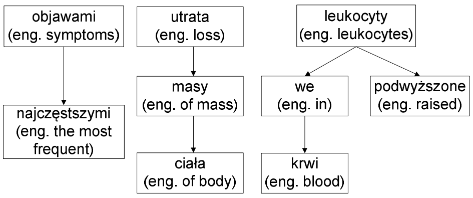

Fig. 1. Word associations extracted from a sample sentence

Let us analyze a sample sentence:
Podwyższone leukocyty we krwi i utrata masy ciała są najczęstszymi objawami. (eng. *Increased leukocytes in blood and body weight loss are the most frequent symptoms.*)

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

- noun - adjective: *objawami najczęstszymi* (eng. *frequent symptoms*);
- noun - noun in genitive: *utrata masy* (eng. *weight loss*);
- noun - noun in genitive: *masy ciała* (eng. *body weight*);
- noun - adjective: *leukocyty podwyższone* (eng. *increased leukocytes*);
- preposition - noun in locative *we krwi* (eng. *in blood*);
- noun - noun in locative: *? we krwi* (eng. *? in blood*).

In the last relation the preposition and the noun in locative were 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: *leukocyty we krwi*. 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.

The obtained word associations are presented in Fig. 1. As we can see the mechanism delivers three separate graph structures. To identify the graphs which are interesting for our purposes, we need to remember, that every symptom description contains at least one noun in the nominative case. This noun is the main element of the description verbal construction. In terms of the graph structures this means that it is the root node of the tree of words. When looking at the structures from Fig. 1 we can see that only two of them contain nouns in nominative. The nouns are: *utrata* and *leukocyty*. As a result only these two trees are considered to be the desired descriptions. The selected trees are then reduced to the flat sequences of words, which originally appeared in the text. As a result two descriptions are extracted from the sentence: *utrata masy ciala* and *podwyższone leukocyty we krwi*.

One element from the example sentence has not been discussed yet. This is the verb *są*. It cannot be associated to the other sentence elements using linguistic rules. It can, however, be associated using sentence schemas, which allow to identify the sentence subject, predicate and object. We defined a set of the most typical sentence schemas to associate verbs to the rest of the sentences. The discussed sentence contains two subjects, which are the nouns *leukocyty* and *utrata*. The sentence schema detects only the first one, and associates it to the verb. As a result we get the association *leukocyty są*. Unfortunately the schema assumes a noun in accusative to be the object of the activity expressed by the verb. No such noun could be found in the example sentence. As a result the construction retrieved by the schema is incomplete, and thus ignored.

Another thing which has not been considered yet are the ambiguities resulting from imprecision of the linguistic rules. The ambiguities lead to generating some additional word associations, which for conscious reader are obvious mistakes. Such mistakenly created associations could the following: *objawami krwi*, *utrata krwi*, and *masy krwi*. This delivers some alternatives to descriptions, which could be extracted from text. If the ambiguities are not possible to be resolved, all the possible variants are generated, assuming that one of them is the correct one.

III. Collecting Cases

The collection of descriptions extracted from text is of course far from perfect. It strongly depends on the actual contents of text corpus. It is obvious, that medical text contains not only symptom descriptions, but a lot of other information, including descriptions of medical procedures, patient treatment, etiology, or pathogenesis of diseases. All that information is unimportant for the diagnostic purposes. Unfortunately, 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 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 search 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 search 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 front of the search list. Using the described tool the experts create a database of patient cases, which can be considered 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.

The training phase of the system is necessary for identifying the descriptions correctly describing symptoms. This is easily learned from the cases. The descriptions which were frequently used are considered to be correct. The descriptions which were not used, or used occasionally, are considered to be incorrect and removed from the system. It is assumed that some level of human errors is possible, and thus the rarely used descriptions are removed, as considered to be erroneous. In this way the system is resistible to occasional human mistakes. Of course we should distinguish the erroneous descriptions, from description of rare symptoms. The descriptions of rare, but important symptoms, are identified given the correlations with diseases. If a rare description is strongly correlated with some disease, it should not be eliminated.

IV. Building the Semantic Model of Symptoms

A. Identification of Concepts

For building any semantic model it is necessary to identify the set of concepts. In our case the concepts of the model are the symptoms and the diseases. The descriptions remaining in the system after collecting a reasonable number of cases, although correct, are not symptoms yet. The reason is that some of them have the same or similar meaning. The meaning of a given description is considered to be a symptom. To discover the symptoms among the set of collected descriptions, it should be noted, that the descriptions with close meaning, have similar statistical distribution with respect to the set of diseases. This distribution is easily retrievable from the set of training cases. For practical reasons the set of diseases, which can be diagnosed is fixed, and *a priori* defined. The considered system has a modular structure, and a single module contains diseases from one domain. Currently we are experimenting with two domains: allerglogy and pulmonology. As a consequence we collect two sets of cases with diagnosed diseases from one of the two domains. The a priori defined set of diseases allows for ignoring the problem of synonymic names of diseases, and makes them immediately the set of the model concepts.

The problem remaining to be resolved is identification of symptoms. As already noted the synonymic descriptions have similar distributions of their occurrence in particular diseases registered in training cases. Identification of sets of such descriptions is possible through clusterisation. As a result of this process we get a set of symptoms, represented by identified clusters. It is of course possible that some descriptions, with different meaning, are closely correlated by occurring in the same diseases. Such descriptions are easily distinguished from synonyms. This is possible by observing, that synonymic descriptions are not used in the same cases, as no one describes the same symptom twice.

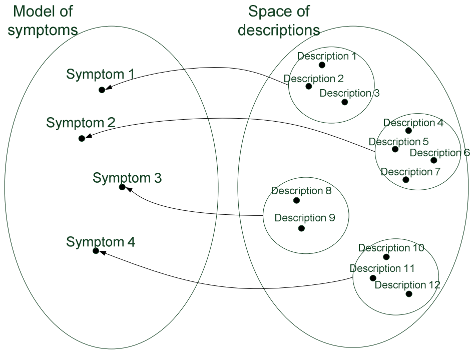

Fig. 2. Schematic mapping between clusters in the space of natural language descriptions, and the model of symptoms

After clusterisation, every description is easily assignable to its respective symptom. It is thus easy to determine the statistical distribution of symptoms with respect to diseases. This is just a result of simple summation of the distributions obtained for the synonymic descriptions. Such distribution can be utilised for construction of a Bayesian network, which can further be used as a diagnostic decision support tool.

B. Identification of Vertical Relations

The simple clusterisation leads to a model of flat list of symptoms, where no relations between them are taken into account. Such a model is good for building a Bayesian network, but is not the most accurate for more complicated purposes, like semantic reasoning. The basic element of every semantic model is a vertical hierarchy of concepts. Such a hierarchy arranges concepts in the form of a tree, where the concepts with wider meaning are parent nodes of the concepts with more narrow meaning. This type of relation seems to exist also between symptoms. For example the symptom *allergic reaction to the animal fur* could be considered a superclass of the more narrow symptom *allergic reacion to the dog's fur*. This indicates that hierarchic model of symptoms is more accurate than the flat one. The hierarchy of symptoms can be built by applying hierarchic clusterisation algorithm. The clusters obtained in this way are directly transformed into hierarchy of semantic classes.

One should be careful, however, when trying to interpret the meaning of the classes from different levels of the model hierarchy. It should be underlined that the meaning of particular descriptions is resolved not on the foundation of the linguistic interpretations, but on the foundation of statistical association to diseases. It might seem strange at first, as most of the approaches to building semantic models are founded on purely linguistic interpretation of meaning. However, the presented way of meaning determination if much better, if we have in mind, that the model is build for diagnostic purposes. The linguistic interpretation can sometimes be misleading, when one would try to associate it with possible diagnoses. There are many symptoms which could be interpreted as subsymptoms of other symptoms from the linguistic point of view, while at the same time they are associated to significantly differing diseases. In other words the diagnostic hierarchy of classes does not need to agree with the linguistic hierarchy of classes. In the diagnostic hierarchy of classes the wider meaning refers to occurrence of a symptom in a wider set of diseases, while the subclasses are more specific in the sense that they are associated with a more narrow set of diseases. When considering the mentioned example of the allergic reaction symptoms, we actually do not know if the two symptoms are actually related diagnostically. This is suggested by the linguistic interpretation, because the *dog's fur* is a subclass of the *animal fur*. But cannot be sure if the *allergic reaction to the dog's fur* indicates a subset of the diseases indicated by the *allergic reaction to the animal fur*. As a result, we cannot determine the vertical relations from the purely linguistic interpretations.

The identification of symptoms (i.e. the model concepts) on the foundation of association to diseases also influences the contents of particular synsets (i.e. the sets of synonymic descriptions). The descriptions belonging to individual clusters do not need to be actual synonyms from the linguistic point of view. It is enough if they are used as synonyms from the diagnostic perspective, i.e. their occurrence indicates the same diseases.

To underline the importance of diagnostic meaning determination on the foundation of association to diseases, it should be noted that the structure of a semantic model based purely on a linguistic perspective interpretation is always arguable. This is a result of the fact, that in the standard approach such models are built by choosing one of the possible verbal descriptions of the world. No matter how objective the model constructor tries to be, the final result is always subjective, because he is forced to choose between one of alternatives. The model built according to our approach is objective, in the sense that no individual can directly influence its structure. The structure is an outcome of the cumulated knowledge of all the experts taking part in collecting patient cases. The experts also mutually verify themselves by using or ignoring particular descriptions.

This of course does not mean that models based on linguistic relations are wrong. Everything depends on the purpose of the model. If a system is aimed at doing some kind of linguistic analyses it should be constructed in this way. In our case the aim is doing patient diagnoses, and the system should be constructed in a way which maximizes the accuracy of results.

C. Adding Restrictions to Semantic Relations for Semantic Reasoning

The foundation for knowledge representation in semantic networks is description logics. This allows for expressing all the dependencies between semantic classes in terms of logical expressions. The logical expressions are then used by semantic reasoning engines. Semantic knowledge representation is thus a powerful tool in decision support systems. Such a tool can also be used for diagnosing patients on the foundation of their symptoms.

The way of constructing the model structure has already been described. What is missing is the set of restrictions of particular semantic relations. Such restrictions are expressed with the aid of a set of logical quantifiers. What is necessary to identify the restrictions, is to find mutual dependencies between symptoms which can easily be transformed onto a set of logical expressions. Such expressions are then associated with particular relations, and in this way the model structure becomes complete and ready for performing semantic reasoning. The input for this process is the set of symptoms observed in a patient, while as a result we get the suggested diseases. Appropriately constructed reasoning could also indicate the missing symptoms, which would significantly improve the diagnosis. In this way the system is able to suggest the medical tests necessary to perform for improving diagnosis.

The construction of the logical restrictions can be determined on the foundation of statistical distribution of symptoms with respect to particular diseases. Again the patient cases are the key resource to identify the model construction elements. Given a set of cases with a particular disease diagnosed it is possible analyse mutual occurrence of the symptoms in particular cases. This data is transformable into a set of logical expressions, such as logical sum, product, or any other statistically relevant dependency. These are the restrictions we are searching for. Such an analysis should be performed for all the diseases from the domain of interest. As a result we get a powerful tool of semantic reasoning.

V. Results of Experiments

The experiments that we are able to describe at this stage of the work refer to extraction of descriptions from text corpus. The results of clustering and building the semantic model will be described later. Currently we are working on collecting appropriately large collection of patient cases.

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. The main text resources were [24] for allergology and [25] 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 foundation of 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 described in the text. 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. The average number of words in every description was 5.4 for allergology, and 4.8 for pulmonology. The number of retrieved descriptions is not the only factor which is important. As already mentioned the analyser is not able to distinguish, whether the extracted descriptions refer symptoms or to anything else. It is thus important to assess the rate of the number of symptom descriptions to the number of other descriptions. This rate strongly depends on the specifics of the analysed text. In some texts the symptoms appear rather sparsely, while other are almost entirely devoted to describe symptoms. Thus the observed rate ranges from 10-20% up to 80-90%. The assessed overall rate is about 50%. This amount is huge enough to cover significant part of the possible verbal descriptions of symptoms from the given domain, and be the basis for describing patient cases. The missing element will be completed during collecting cases.

VI. Conclusion

The paper describes a methodology of building a model of diagnostic knowledge. The idea of the system assumes two stages in the model creation process. The first of them is text analysis in order to extract verbal constructions describing symptoms. The second stage aims to collecti patient cases and build the model of symptoms on the foundation of the patient cases. As a result of the whole we get a semantic model built of symptoms and diseases. What distinguishes this approach from other solutions typically applied in building semantic models, is that no direct manipulation of the model is required. The system structure is learned from examples. The key tool for extracting the model structure is clusterisation and statistical analysis of particular symptoms occurence in the cases.

The model is designed for diagnostic purposes. In the simplest case the diagnostic process can be supported by the Bayesian network constructed on the foundation of the data collected during the model construction. The more advanced algorithms lead to construction of a semantic network with hierarchic structure of symptoms, and description logics rules. This allows for performing reasoning based on a semantic inference engine. It should be underlined, that the meaning of the particular concepts forming the model, is not determined on the foundation of their linguistic interpretation. The meaning is a result of associations between the symptoms and the diseases. Such solution is the required when diagnosing a patient is the task. The natural language descriptions are used for human communication only, while the computational model is subordinated the diagnostic purposes. Trying to interpret the model structure in terms of natural language associations could be even misleading, so no one is supposed to do it.

The method of meaning determination not only rises the quality of the model. It also simplifies the model construction process. This is due to eliminating the task of the direct model manipulation by experts. In this way no human needs to care about choosing the most accurate world description method. Carrying about synonyms is also not required. These tasks are completely automatized. The model constructed in the described way is also resistant to the subjectivity. This is a problem which appears when a model is constructed by a human expert which has to choose among one of possible model structures. In our approach the knowledge is extracted from cases, which are entered by more than one person. In this way the model is a resultant of all the individual habits and knowledge represented by the human users.

The experiments described in the paper refer to extraction of textual descriptions from the text corpus. The lexical analyser is able to extract the required information from text. The criterion for assessing the quality of the set of extracted descriptions is the rate of descriptions actually describing symptoms, to the other descriptions. This parameter does not depend only on the analyser, but also on the quality of the text. By quality of the text we mean the density of symptom descriptions which appear in the text. This factor depends on the authors' writing style. The density of symptom descriptions is important, because the analyzer works only on the foundation of grammatical rules of sentence construction. It is not able to interpret the meaning of the extracted descriptions. As the grammatical construction of a symptom description is no different than description of any other entity, we are not able to filter the undesired descriptions. Such a mechanism would be very helpful, but currently we are not able to deliver it.

The undesired descriptions are, however, not an obstacle which would make impossible delivering the user interface with a collection of symptom descriptions. This interface is necessary for describing the patient cases. The extracted set of descriptions is huge enough, to cover significant range of possible symptom descriptions. The key which allows for efficient entering of cases is an efficient search mechanism and system of weights. Of course we are not able to guarantee that all the required descriptions are extracted from text, thus during collecting the training cases the descriptions are open for edition. In this way it is possible to correct existing descriptions, or create new descriptions from scratch. Such descriptions are immediately available to other users entering the cases. This allows for mutual verification within the team responsible for collecting the cases.

The work on collecting a reasonable number of cases is in progress, so the effects of the second stage of building the model were not described here. This will be a field of experiments with clusterisation, and statistical analysis of cases in order to build the final model, which further will be used as a part of a decision support system.

Acknowledgment

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] P. Velardi, R. Navigli, A. Cucchiarelli, F. Neri. "Evaluation of ontolearn, a methodology for automatic learning of ontologies," in *Ontology Learning from Text: Methods, Evaluation and Applications*, P. Buitelaar, P. Cimmiano, B. Magnini, Eds. IOS Press, 2005, pp. 92-106.

[2] A. Maedche, S. Staab. "Ontology learning for the Semantic Web." *Intelligent Systems*, vol. 16, pp. 72-79, Mar. 2001.

[3] "The OntoGen system web site." Internet: http://ontology-learning.net/wiki/OntoGen, [Jun. 28, 2011].

[4] D. Faure, T. Poibeau. "First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX," in *Proc. Ontology Learning ECAI-2000 Workshop*, 2000, pp. 7-12.

[5] F. Pereira, A. Cardoso. "Clouds: A Module for Automatic Learning of Concept Maps." in *Lecture Notes in Computer Science*, vol. 1889/2000, pp. 468-470, 2000.

[6] U. Hahn, M. Romacker. "The syndikate text knowledge base generator," in *Proc. of the 1st International Conference on Human Language Technology Research*, San Diego, 2001, pp. 328-333

[7] N. Weber, P. Buitelaar. "Web-based ontology learning with ISOLDE," in *Proc. of the ISWC Workshop on Web Content Mining with Human Language Technologies*, Athens, 2006.

[8] L. Karoui, M.A. Aufaure, N. Bennacer. "Context-based Hierarchical Clustering for the Ontology Learning," in *Proc. of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, Hong Kong, 2006, pp. 420-427.

[9] S. Sung, S. Chung, and D. McLeod. "Efficient concept clustering for ontology learning using an event life cycle on the web," in *Proc. of the 2008 ACM symposium on Applied computing*, 2008, New York, pp. 2310-2314.

[10] S. Kok, P. Domingos. "Extracting Semantic Networks from Text Via Relational Clustering," in *Proc. of the 2008 European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer-Verlag Berlin, Heidelberg, 2008, pp. 624-639.

[11] P. Buitelaar, P. Cimiano. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, Amsterdam: IOS Press, 2008.

[12] W. Wong. "Learning Lightweight Ontologies from Text across Different Domains using the Web as Background Knowledge." Doctor of Philosophy thesis, University of Western Australia, Crawley, 2009.

[13] A. Gomez-Perez, D. Manzano-Macho. "Deliverable 1.5: A survey of ontology learning methods and techniques." OntoWeb Consortium, Internet: http://www.csd.uoc.gr/ hy566/A survey of ontology learning methods and techniques.pdf, [Jul. 30, 2011]

[14] M. Shamsfard, A. Barforoush. "The state of the art in ontology learning: A framework for comparison." *Knowledge Engineering Review*, vol. 18, pp. 293-316, Dec. 2003.

[15] Y. Ding, S. Foo. "Ontology research and development: Part 1 - a review of ontology generation." *Journal of Information Science*, vol. 28, pp. 123-136, Apr. 2002.

[16] K. Toutanova and C.D. Manning. "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger." in *Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 63-70.

[17] A. Ratnaparkhi. "A Maximum Entropy Part-of-Speech Tagger." in *Proc. of the First Empirical Methods in Natural Language Processing Conference*, 1996, pp. 250-255.

[18] M. Piasecki. "Polish Tagger TaKIPI: Rule Based Construction and Optimisation." *Task Quarterly*, vol. 11(1-2), pp. 151-167, 2007.

[19] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov and E. Marsi. "MaltParser: A language-independent system for data-driven dependency parsing." *Natural Language Engineering*, vol. 13(2), pp. 95-135. Jun. 2007.

[20] S. Szpakowicz. "Formalny opis składniowy zdań polskich." Warsaw: Warsaw University Publishing House, 1983. (in Polish)

[21] M. Woliński. "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, Jun. 2006.

[22] A. Przepiórkowski. "The IPI PAN Corpus. Preliminary Version." Warsaw: Institute of Computer Science PAS, 2004.

[23] M. Woliński. "System znaczników syntaktycznych w korpusie IPI PAN." *Polonica*, vol. XXII/XXIII, pp. 39-55, 2003.(in Polish)

[24] W.H.C. Burgdorf, G. Plewig, H.H. Wolff, M. Landthaler , "DERMATOLOGIA Braun-Falco," Lublin: Czelej, 2010. (in Polish)

[25] A. Szczeklik. Choroby wewnętrzne. Kraków: Medycyna praktyczna, 2006. (in Polish)