Natural Language Based Concept Map Building

Marks Vilkelis, Janis Grundspenkis, Normunds Gruzitis

Abstract: *The paper represents the research on transformation of natural (Latvian) language texts into concept maps. The main point of this work is to overview and to discuss some ideas in order to describe a number of already solved tasks, linked with the topic. There are many examples of how and what to do in automated "translation" of common linguistic structures and patterns into the formal data structure. The paper gives also an overview of authors' previous research, which encouraged starting the current work.*

Key words: *Natural Language Processing, Concept Map, Concept Structure, Ontology.*

INTRODUCTION

One of the ultimate goals of early research directions in artificial intelligence is to reach a perfect man-machine interaction using natural language. Nowadays computers mainly can understand only specific machine language, programmed by specific people – programmers. The fact, that computer industry and applications are growing fast, makes humans to be involved in it more and more. Thus, the need of communication in natural language is extremely high.

The lowest level computers can operate on is mathematics. The problem is that natural language has words and sentences with no explicit mathematics in it. To operate with words (concepts) and their mutual relations, we have to use more advanced data types, which could be understandable both for a human and for a computer. There are many formal data structures, which support appropriate relationships between aspects and concepts: mind maps, frames, graphs, UML diagrams, concept maps [2] etc.

The basic aim of the research is to recognize key concepts and their relations. The next step is to transform the recognized words into concept map [2] and store it in order to create ontologies [4]. Then we could operate with this data further by answering questions, seeking relations among known concepts and even by inferring new relations. We want to make computer to understand what human says. The work described in this paper has just started, so there are more questions than answers. However, the authors would like to share some results they already have and to introduce some ideas and ways how to solve the problems faced with during the research.

The paper is organized as follows. Section 2 gives a short overview of the designed and implemented intelligent knowledge assessment system (IKAS), which usage inspired the authors of this paper to start the current research. This section also describes the concept map term. At the end of the section the main task to solve the stated problem is formulated. Already solved subtasks and successful solutions for natural language transformation into the formal data structure are described in section 3. The end of the section 3 discusses problems authors faced with during the research. This section has also some theoretical proposals. The paper ends with conclusions and future work.

PROBLEM DESCRIPTION

IKAS overview

Since year 2005 the Department of Systems Theory and Design of the Faculty of Computer Science and Information Technology of Riga Technical University (RTU) has been developing the concept map based knowledge assessment system IKAS (Intelligent Knowledge Assessment System), which usage inspired the authors of this paper to start the current research [5, 7, 8]. The system works in the following way. The teacher defines knowledge assessment area and creates concept map for it. The process of the creation of a concept map consists of the specification of relevant concepts and relationships. Thus the concept map includes all concepts and relationships among them, which are taught

and learned in the corresponding study course. Teacher's created concept map serves as a standard against which the learners' concept maps are compared. During knowledge assessment the learner solves a concept-map based task. After the learner has submitted his/her solution, the system compares the concept maps of the learner and the teacher, calculates the score of the learner's result, gathers statistical information and generates feedback which is sent back to the learner.

It is important, that there are six predefined relations that are being used in concept map creation. They are: "is a subset", "is a kind", "is an example", "is a property", "value" and "is a part". All other relations between concepts are also possible, but they are just labels defined by the teacher and are not automatically processed by the system in some advanced analyses, for instance, in recognition of hidden pattern relations [3].

Concept map terms

In the current research we will need a high level formal data structure to operate with information while transforming natural language into organized data. The choice was made to use concept maps because of some reasonable facts.

Concept maps are a kind of mental models represented by graph with labelled nodes corresponding to concepts and with directed arcs indicating relationships between pairs of concepts. A linking phrase specifies the kind of relationship between concepts. A semantic unit of a concept map is a proposition. Propositions are concept-link-concept triples which are meaningful statements about some object in the problem domain [2].

For human it is easy to understand the concrete information domain represented as a concept map, because such a data structure is close to the intuitive level of humans' perception of the world. There are quite a lot of processes in the surrounding world when one thing causes another, which could be displayed with a proposition or expression. So the main aspects for choosing concept maps are their demonstrativeness and use of visual methods of this data structure.

The important drawback, detected among many other problems researchers faced with during the development of IKAS, is that the expression of main ideas as concepts and their relationships as arcs is time-consuming operations. It is time-consuming operation to extract the main ideas as concepts and their relationships as arcs. The process of concept map input into the system is manual, so it takes some time. It would be nice if a computer would automatically perform the mentioned task or even just a part of it.

LANGUAGE ANALYSIS

In this work we are not focusing on the development of natural language processing tools. The main accent is put on concept map creation from unstructured text by using already available low level tools for morpho-syntactic analysis of natural language [1]. The work describes automatic extraction of key concepts and their relations from Latvian language texts and creation of draft concept maps. The problems, which will be mentioned regarding natural language processing, are common to many languages. Latvian belongs to the Baltic language group – it is a highly inflective synthetic language with syntactically free word order. In terms of the grammar structure Baltic languages are closely related to Slavonic languages (Eastern and Central European languages) [1].

Morphological analysis

Text analysis is a difficult, but to some extent already solved task for a number of languages. At the University of Latvia (LU) a robust morphological analyzer (this is an appropriate Java API library) has been developed [6] among other tools for Latvian language processing. For a given word form the analyzer returns its base form and a set of morphological features, like part-of-speech (e.g. noun, verb, adjective), gender, number,

case (e.g. nominative, genitive, accusative) etc. However, this is not a stochastic part-of-speech tagger – if a word form is morphologically ambiguous, all variants are returned.

The main issue is that the current lexicon of known stems and their morphological features is not very rich (ca. 50 000 lemmas): not only domain specific terms, but also a lot of commonly used words are missing. Nevertheless the API provides an interface that allows extending the lexicon, and, thus, it can be gradually adapted for both domain specific and wide coverage common sense texts. Although the single word form recognition mechanism is a helpful instrument in the text analysis, it is only the first step. The next task is to detect syntactic dependency links among the individual words.

Concept extraction

The concept extraction process from a text might look rather simple, but in practice many problems arise. Basically, we just have to take all nouns from the text (concepts could be adjectives and verbs also, but for the sake of simplicity we will take them into account later). The above mentioned morphological analyzer could provide us with such functionality. However there are several problems. Let's describe the most common ones.

Collocations. Let's take, for example, a sentence – "We have to perform Latvian language analysis with an aim to extract nouns from the text and save them in the data base" (1). The morphological analyzer can recognize word forms in this sentence and, as the result, returns the nouns illustrated in Fig.1. All nouns are correctly selected as concepts, but two pairs of nouns cannot be considered as separate concepts. One pair is "Latvian" and "language", because it is actually one concept – "Latvian language". Another pair is "data" and "base", which also compose a single concept – "data base". Thus, single word analysis provides us with just partly correct concepts taken from the text, because only individual words are taken into consideration. Collocations (multi-word units) are not recognized as a single concept.

Fig.1. Concepts extraction from the text (IKAS user interface)

There are at least two possible solutions of this problem. The first one is to create an appropriate functionality, which would be able to recognize collocations as single concept. In this case all collocations have to be known for the utility in advance and the vocabulary of such terms has to be extendable. The second way is more sophisticated. The idea is actually to split a multi-word unit into two concepts linked with one of the six mentioned predefined relations. Let's take the term "Latvian language". The concept "Latvian" is value of concept "language" (What kind of language? Latvian). Another mentioned collocation is "data base". We could introduce extra one predefined relation between concepts – "of". Thus, the concept "data base" could be separated into two concepts related with each other with "of" – "base of data". This is quite logical, because there might be many other terms linked with the same concept – base of knowledge, base of army etc.

Pronouns. Sometimes the use of pronouns makes it more complicated to understand what concept (or individual) we are speaking about even in a human-human conversation. Pronoun is some kind of anaphora to a previously introduced noun (concept/individual), but it can be ambiguous. In the above mentioned sentence (1) we use the pronoun "them".

It might be confusing what we are talking about if the sentence would be longer and syntactically more complicated.

Sometimes it is almost impossible to understand what noun has to be linked with the given pronoun in the sentence. There is no clear and general solution for this problem. Let's take a look on the next sentence – "The task is to extract the first verb from the given sentence and to save it in the data base" (2). The pronoun "it" might be linked either with the noun "verb" or with the noun "sentence", thus it is unclear even for human what we have to save in the data base verb or sentence. The possible solution to avoid such ambiguity is to ask human to substitute all detected pronouns with corresponding nouns. The text will become highly verbose, but it will be formally more understandable.

Polysemy. Homographs are the words with the same spelling but different meanings. For example noun "bass" has two meanings – it is type of fish OR low, deep voice. As the matter of fact there are two different words, which spelling is just the same. So it's impossible to distinguish them, without analyzing the whole sentence or even the whole provided text. This is extremely hard to automate such concept processing, which have the same spelling.

There are at least two possible solutions for such cases. The first one is perfunctory. The idea is to consider all concepts with the same spelling as one. To separate them, we could ask human to use collocations. Instead of just word "bass" using, "bass fish" and "bass voice" has to be used. The second solution is more fundamental. It affects concept terms in the context of the described natural language processing system. No term should be considered as concept if it's not linked with any other concepts. Thus, an isolated term could not be a concept any more, except some basic concepts and idioms.

We are one step closer to ontologies using as one elementary unit of data or knowledge. So the data structure for described system might be extended up to ontology-concept map. As ontology in its turn is also some kind of concepts and relations, so the name of structure got might be "concept maps' map" or "the map of concept maps".

Relation detecting in a pair of concepts

The problem of concept extraction is more or less defined and we can assume that it is possible to get all the nouns from a text. The most difficult task is to bind two separate concepts and to create a correct relation between them. Relation between two concepts in this paper will be designated as follows. For instance, we have concepts "C1" and "C2" linked with relation "R", so we can get an expression or proposition ("C1") – "R" –> ("C2").

Verbs as relations. Concepts typically are represented by nouns. Relations typically are represented by verbs (predicates). Let's examine the simplest example of verbs used to link two concepts: "A boy goes to a shop". There are two concepts "boy" and "shop". There is one verb in this simple sentence "goes", so we can create simple relation ("boy") – "goes to" –> ("shop"). Note that it is important to use preposition with verb, because prepositions can significantly affect expression's meaning.

Another situation in verb usage is described in the next sentence - "A dog can run". There is only one concept in the sentence, thus, at the first glance; an expression cannot be build. This is wrong, because we do have some information on hands, in spite that we cannot directly create the expression. There is the hidden second concept in this sentence; the verb "can" means the ability to do something. The sentence could be transformed in such a way – "A dog has ability to run". Thus, we've got an expression ("dog") – "has" –> ("ability"). What should we do with the rest of the sentence, i.e., with the verb's infinitive "to run"? It looks like that verb infinitives could be considered as concepts also. In this case we can create triple expression ("dog") – "has" –> ("ability") – "value" –> ("to run"). The second relation is predefined.

Adjectives. The majority of adjectives are directly linked with nouns (as attributes). Concepts can have properties, which have values expressed by adjectives. Let's examine a simple sentence – "The girl has blue eyes" (3). We have no problems with expression ("girl") – "has" –> ("eyes"). It is difficult to detect property name for adjective "blue". Possibly there must be predefined ontologies for adjectives, which would tell us what property of an object is meant by using this adjective. For example, we have to predefine the following axiom ("blue") – "is a" –> ("colour"). So, while analyzing word "blue" from the sentence (3), the system has to search predefined axioms for the adjective "blue", and if they are found in the ontology, we can say, that ("girl") – "has" –> ("eyes") <– "property" – ("colour") – "value" –> ("blue").

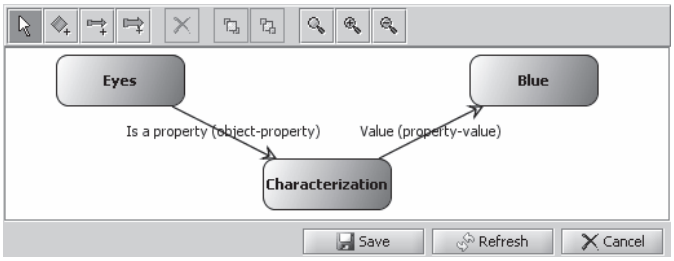

Fig.2. Detection of relations between nouns and adjectives (IKAS user interface)

If we have no any predefined ontology, surrogate concept with the name "characterization" could be inserted. The sentence "…blue eyes…" could be represented as follows: ("eyes") – "is a property" –> ("characterization") – "value" –> ("blue") (Fig. 2).

Regarding adjectives, everything is clear in simple cases, except that we have to collect some predefined information about what properties could be applied for a given adjective. This is a routine process, which has many details and nuances.

Difficulties. The question about precise detection of the relationship between different concepts is still open and possibly there is no any general solution, because natural language is highly ambiguous. At the beginning of the research we have to reduce the complexity of sentences for analysis just for task simplification. And also we could assume, that all user's sentences have the neutral word order, and there is no reverse word order used (such situations are common in Latvian). Even in this case sometimes it's unclear which noun is related to the given adjective.

In spite of mentioned assumptions, we will have also some problems with other parts-of-speech, which were not described in this paper: conjunctions and prepositions for instance. Such constructions like "not only, but also" are very hard to recognize and almost impossible to express in the concept map terms.

Detection of concept relations needs laborious further research, which has just started. It is clear that word by word analysis is not enough and full syntactic parsing of a sentence will be needed to find out the correct relations between concepts. Moreover, to solve anaphoric references, single sentence boundaries often have to be crossed.

CONCLUSIONS AND FUTURE WORK

The paper describes the basic problems of natural language processing and transforming into formal data structures such as concept maps. In the begging of this work authors gave a short history, which has led to the current research. The need to translate natural language appeared while developing intelligent knowledge assessment system, where teacher manually has to input lecture topic as a concept map. The paper gives also a short overview about the concept map terms.

The main focus of this work is put on concept extraction from unstructured text using the morphology analyzer – the Java library designed by Latvian University. This tool is

able to recognize Latvian words and to provide user with morphological information about word's type, gender, conjugation and etc. The use of this "low level" tool is the starting point of research. However, the morphology analyzer does not solve all the problems in the concept recognition process. Moreover, in future, a wide coverage and high precision part-of-speech tagger should be used instead, to avoid the morphological ambiguities.

Most fundamental problems related to the concept extraction are discussed in the paper and they are – recognition of collocations, anaphora resolution and polysemy leading to ambiguities. In spite of the mentioned problems authors believe, that the hardest task is to identify the relations between concepts in order to create simple expressions for concept map building. The work in relation seeking has started but there are more theoretical than practical results. However, some concrete problems are already identified; they are related to verbs as relations between nouns as concepts, verbs as concepts, and also detection of adjectives and maintaining them in simple expressions.

REFERENCES

[1] Barzdins G., Gruzitis N., Nespore G., and Saulite B. Dependency-Based Hybrid Model of Syntactic Analysis for the Languages with a Rather Free Word Order. Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007), Tartu, May, 2007, pp. 13–20.

[2] Canas, A. J. (2003), A Summary of Literature Pertaining to the Use of Concept Mapping Techniques and Technologies for Education and Performance Support. Technical report: Pensacola, FL.

[3] Grundspenkis, J., Strautmane, M., 2009. Usage of graph patterns for knowledge assessment based on concept maps. pp. 60 – 71.

[4] Guarino, N., 1998. Formal Ontology and Information Systems. Proceeding of FOIS'98, Trento, Italy. Amsterdam, IOS Press, pp. 3-15.

[5] Lukashenko, R., Vilkelis, M., Anohina, A. Deciding on the Architecture of the Concept Map Based Knowledge Assessment System. Proceedings of the 9th International Conference on Computer Systems and Technologies (CompSysTech'08) and Workshop for PhD Students in Computing, June 12-13,2008,Gabrovo,Bulgaria, ACM, pp.V.3-1- V.3-6

[6] Paikens P. Lexicon-Based Morphological Analysis of Latvian Language. Proceedings of the 3rd Baltic Conference on Human Language Technologies, Kaunas, October 2007), pp. 235–240.

[7] Vilkelis, M., Lukasenko, R., Anohina, A. Technical Evolution of the Concept Map Based Intelligent Knowledge Assessment System. Proceedings of the 13th East-European Conference on Advances in Databases and Information Systems, September 7, 2009, Riga, Latvia, pp. 214-221

[8] Vilkelis, M., Anohina, A., Lukashenko, R. Architecture and Working Principles of the Concept Map Based Knowledge Assessment System. Proceedings of the 3rd International Conference on Virtual Learning, October 31- November 2, 2008, Constanta, Romania, pp. 81-90

ABOUT THE AUTHORS

PhD student Marks Vilkelis, M.sc.ing., Department of Systems Theory and Design, Riga Technical University, Phone: +371 29653777, E-mail: markvilkel@inbox.lv

Prof. Janis Grundspenkis, Dr.habil.sc.ing., Department of Systems Theory and Design, Riga Technical University, Phone: +371 67089581, E-mail: janis.grundspenkis@rtu.lv

PhD student Normunds Gruzitis, M.sc.comp., Institute of Mathematics and Computer Science, University of Latvia, Phone: +371 67227486, E-mail: normundsg@ailab.lv