Original scientific paper

Encyclopedic Knowledge as a Semantic Resource

Marko Orešković
National and University Library in Zagreb
Hrvatske bratske zajednice 4, Zagreb, Croatia
moreskovic@nsk.hr

Ivana Kurtović Budja
Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
ikurtov@ihjj.hr

Mario Essert
Faculty of Mechanical Engineering and Naval Architecture
Ivana Lučića 5, Zagreb, Croatia
messert@fsb.hr

Summary

Extraction of semantic information from unstructured natural language texts is currently a hot topic in the field of computational linguistics. The majority of researchers agree that the core of the problem is in determining "semantic domains", which would give the computer the real meaning of the text. The aim of this research is creation of semantic domains from online encyclopedic texts. This article explains two different approaches to morpho-syntactic analysis that could be used to achieve that.

Key words: knowledge extraction, semantic domains, ontologies, encyclopedia

Introduction

Categorization as seen by Aristotle is the first method of setting the order in a well-organized set of data (information), i.e. creating knowledge about the world around us. Today we can say that Aristotle's categories are the first attempt to establish formal ontology (Bosančić 2016) and the emergence of science of classification – the taxonomic/hierarchical organization of those categories, i.e. the knowledge that is immanent. Over the past three centuries, encyclopedias have kept and maintained that knowledge as an organized set of information. The methods of categorization and classification have been improved for decades, but the core content stayed the same. However, a shift from the printed (paper) media first to the digital one and subsequently by incorporating a network structure in encyclopedic work, essentially changes not only the preparation and processing of encyclopedias (Jecić 2013) but also their content. It is not just about the ease of access to information (by clicking hypertext

links – surfing to the new 'hints' of a virtual encyclopedia) but the possibility of deeper entrance into meaningful aspects of information, both for human and the machine. The machine has the ability to learn and acquire knowledge. That is how an artificial intelligence is developed. As an example, we can use the definition of knowledge from a network encyclopedia (Wikipedia):

Knowledge is a familiarity, awareness, or understanding of someone or something, such as facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning. (Wikipedia; https://en.wikipedia.org/wiki/Knowledge)

Such definition allows a user to understand the meaning of the main term (knowledge) through its parts, hypertext links (facts, information, description,...), and that new knowledge can further be expanded by opening the links from the definition, so it is just a matter of time and effort which user has to take to acquire knowledge (transfer it to his/her biological data store/brain). The machine already has that information stored in itself, and it does not need any additional collection and processing time. The data connections (in the databases) are exact and strong which enables fast information retrieval.

This kind of 'computer knowledge' can be used (by user) in one of the following ways:

1. In some domains of human activity to solve a problem in steps based on the stored information. An algorithm selects one of the predefined options in each step (Alberico and Micco 1990). These are so called expert systems.
2. By collecting facts that are not specially arranged, categorized or classified, but hidden in the information itself. These are so called inherent ontologies.

This paper will present such research – retrieving semantic information which is neither apparently visible nor explicitly mentioned. An already published, existing knowledge stored in public repositories from the Internet will be maximally used. The aim of the paper is to show how by changing the structure of the language lexicon (Orešković, Brajnović, and Essert 2017), very difficult natural language tasks can be performed (e.g. metaphorical expressions detection, metonymy detection, or detection of any other tropes that people use in everyday writing or speech). In the second chapter, the basic ideas in similar research studies will be presented in order to compare them with a proposed one (in the third chapter). The fourth chapter of the paper will show the available resources and their constraints regarding our main goal. Finally, we will show the realization of the proposed idea – a segment of a network framework that, based on the encyclopedic knowledge, will create the semantic domains which are the basis for the discovery of stylistic figures of the natural language.

Related work

There are numerous studies in a field of an explicit knowledge extraction from online resources, that relied on globally available knowledge databases (most often Wikipedia and WordNet) in order to build a machine readable ontologies. They all aimed to provide a method for information extraction that would have a higher or at least the same quality like those that are manually made. One such ontology is YAGO (Suchanek, Kasneci, and Weikum 2007) which is created by an automated process of information extraction from both Wikipedia and Wordnet. The aim was to create a larger (in terms of the number of facts) and better (by adding additional knowledge) ontology which can later be used as a valuable resource in computational linguistic studies.

Another research project in this filed is Java-based WikiPedia Library (JWPL) and Java-based WiKTionary Library (JWKTL) information extractor developed by (Zesch, Müller, and Gurevych 2008). These libraries use Wikipedia's and Wiktionary's API's to extract explicit knowledge and serve it over API for usage in other projects.

Project Wikify! (Mihalcea and Csomai 2007) also deals with automatic information extraction from textual documents. It parses the text in search of relevant keywords which are then linked to Wikipedia articles.

For Croatian language there were very few researches that deals with semantic. Most significant one is by (Šnajder and Almić 2015).

A new approach to the word tagging

Binary information and associated computer data types (integer, real, string, ...) allow only binary representation of words, but do not provide any grammatical or meaningful features related to them. In order for a computer to 'learn' a natural language, it is necessary that each stored word is tagged with a more appropriate character. The process is similar to coding (at a higher level) of any terms in an encyclopedic dictionary; without the definition, the word itself has no meaning (what is 'Knowledge', from the example above, becomes clear only when the description is read: '… is a familiarity, awareness, or understanding of someone or something, …'). At the level of grammar, mark-up/tagging means that each word must be associated with the part-of-speech category it belongs to (e.g. noun, verb, adjective) and/or other relevant grammatical features (e.g. gender, number, etc…). Several sets of tags were designed for grammatical mark-up, called the tagsets, which are then manually or semi-automaticaly associated to alphabetically ordered lists of words. The manual mark-up process is of course slower, but more accurate than the automated one, especially for languages characterized by rich and complex morpho-syntactic structures (such as a Croatian). Regardless of the way of mark-up, there is another key improvement to that process. As demonstrated at the EURALEX conference (Orešković, Čubrilo, and Essert 2016), instead of a classical vector grammar-semantic features (e.g. MULTEXT-East), it is possible to introduce a new, so-

called T-structure that brings significant possibilities. In the T-structure, there is no difference in the definition and usage of grammatical and semantic features, although they are still being considered separately (as grammatical: WOS - *word of speech* or as semantic: SOW - *semantics of word*). The MULTEXT-East tagset, designed by a unification of different languages in this area, is replaced by a structure that allows the diversity of language features, and highlights the differences that people tend to use in their language. Compared to the linear tagging, the T-structure introduces taxonomic (ontological) characteristics of the word. Instead of making an ontology for a domain (an area of interest) from a dictionary or a word list, each word represents a unique ontology within itself in this case, and its features form branches of a tree to any depth. The information that is stored in the branches of this tree, (except values – string or numeric literals) may contain a link to other branches, a new ontology or a repository of a network information. For the purpose of creating a general linguistic Syntactic and Semantic Framework (SSF), automated links with open network repositories (the Croatian Language Portal (HJP), The Miroslav Krleža Institute of Lexicography's online encyclopedia and Wortverbindungen http://www.lingua-hr.de by Dr. Stefan Rittgasser) are created. In the same way, the "ontology of words" is expanded with other online resources (CrowN, Wikify, ...) or specialized dialects/thesauri in digitized form (Šarić, Google translate, etc.).

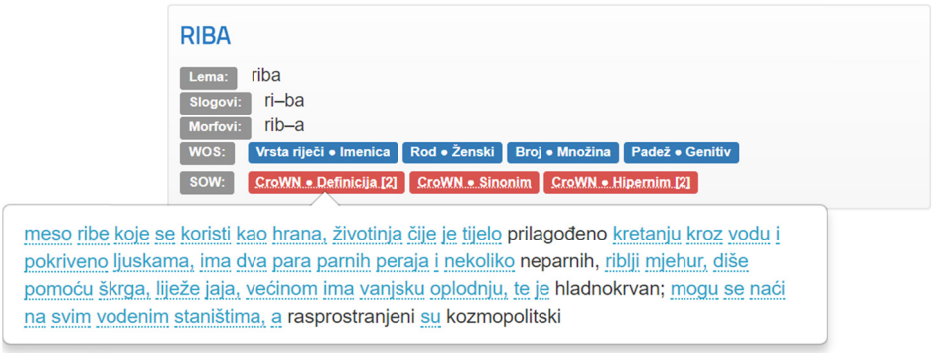

Figure 1. Lexical entry definitions in SSF

A user (or the machine if called via API function) along with grammatical information (WOS; blue tags) also sees a rich semantic information (SOW; red tags) collected locally or from other network resources (respecting the copyrights of their owners) as shown in Figure 1. It is very important for the SSF user to have all the information in the same place, and even more important to have that information interconnected and linked (notice a blue links inside the definition that are direct links to other lexical entries). For automatic information gathering, the existence of already embedded (human) knowledge about

lexical entry is highly important (e.g. different meanings of the same word (homonyms), or the same meaning of different words (synonyms), which is suitable for creating semantic domains, for both literal and metaphoric meaning). This possibility is used concisely in the research presented in Chapter 5 of this paper. Finally, it is worth mentioning that the T-structure can be extended in both (WOS or SOW) directions. Each registered user can make their own tree tags, but of course, they have to make sure that new tags are applied to the words in the dictionary. Such user operations will not interfere with the work and the results obtained by other users.

Open network repositories and dictionaries

By open network repositories, we imply permanent projects that are updated on a daily basis, and are in open (free) access and often allow the user to collaborate in the form of contributions or comments. Certainly the most famous such repository is Wikipedia, which is envisioned as an organized set of encyclopedias written in different languages. Wikipedia content today is provided in about 300 languages, most commonly associated with official and/or national languages. Ten years ago, the famous Miroslav Krleža Institute of Lexicography and its Knowledge Portal http://enciklopedija.lzmk.hr, which encompasses several digitized editions, in line with the Croatian Family Lexicon, began to embark on Wikipedia prestigious footsteps. The SSF uses this information (available online) in agreement with the Institute's disclaimer: "It is permitted to use or quote individual articles in parts or as a whole with the indication of the source". This gives an SSF a new dimension - indexing each word from the definition and creating links to other words from selected network repositories. For CroWN (Croatian version of Wordnet) such indexing means increasing semantic links in the repository itself for at least the order of magnitude. In a similar way, grammatical information from the Croatian Language Portal (http://hjp.znanje.hr) is retrieved, and then automatically compared with those obtained from the morphological generator (Markučič and Govedić 2013) over the "Hrvatska riječ" dictionary (Pinjatela 2001) which is a part of the SSF. The comparison and verification of multiword expressions is achieved with the help of http://www.lingua-hr.de and the algorithm for more accurate decomposition of words into morphs and syllables. In that way the SSF becomes an integrator (hub) of network frameworks that give its dictionary/thesaurus characteristics that have only been theoretically discussed in the models of Generative Lexicon (Pustejovsky 1991) and Meaning-Text Theory (Melcuk 1981). The Figure 2 shows an example of one lexical entry with all the accompanying features (Syllable, Morph, WOS, SOW, MWE etc.)

Figure 2. Lexical entry with all accompanying features in SSF

Formal or computer ontology, defined by (Gruber 1993) as an explicit specification of the conceptualization of a particular domain or a shorter "conceptualization specification", provides the highest level of machine-readability of a particular area. The T-structure has all the features of ontology, at the word level, that is, the lowest level of natural language, which makes it especially strong. In formal and machine-readable (higher) ontologies, developed with the Protégé (or similar) tools, the knowledge is defined by classes and subclasses, their instances, objects, data properties and their interconnections, thus allowing an extremely high degree of formalization. It should be emphasized that the creation of ontology is not the greatest step in knowledge creation, but a process or idea that enables the development of new knowledge from an existing ontology (Antoniou and Harmelen 2004). That is achieved by this kind of lexicon organization. It is easy to notice the two-way activity - the computer in its handling of stored knowledge in an intuitive, understandable, simple and visible way, represents the knowledge to the user and the user controls that knowledge and extends it to new information and/or links to other repositories. In that way the user and the machine create a kind of symbiosis with their complementary work, both of them expanding the basic components of knowledge – the data and their relationships (information). The data and information together makes a knowledge that in the end, philosophically speaking, by understanding and evaluation forms a wisdom (that is known as DIKW hierarchy - which was long represented as a fundamental model: (Ackoff 1989) and (Liew 2007), but also challenged by (Frické 2009). After showing some limitations of word ontology, primarily because of dialectal variation, which is particularly present in Croa-

tian, algorithms that extract the core information from a knowledge written in the sentences of the definition for a given word from the online encyclopedia will be shown. That knowledge is then transcribed in an ordered or unordered sequence of information which is included as a domain of the word in the T-structure and serves for recognition and extraction of the knowledge from other sentences. That leads into one recursive learning and checking cycle, which then reaches a growing precision and accuracy of new vocabulary/thesauri in multiple iterations.

Dialectological extensions of the dictionary base

The Croatian language has three main dialects: Čakavian, Kajkavian and Štokavian, which began as separate language systems. Each of these three is divided into minor dialects, among which, within the same dialect, there are also considerable differences. The Čakavian and Kajkavian dialects differ most from Standard Croatian language. Moreover, the Northern Čakavian dialect is lexically closer to the Slovenian and the Southern Čakavian is lexically closer to the Štokavian dialect. A complete picture of Croatian language can be attained only by combining diachronical and synchronical study. Large amounts of linguistic data can only be processed electronically. Since there is still no single historical corpus of Croatian, one that would cover all three dialects, and since many of the research points (villages) for the Croatian language atlas have not yet been studied, the Croatian language and its history are known to us more or less fragmentarily.

The aim of the project Dialects of Makarska coast – diachrony and synchrony is to study the Štokavian dialect of Makarska coast from a dialectological, textological and sociolinguistic point of view. The data obtained through these efforts will be digitalized and eventually integrated into the framework of the Croatian language. The data obtained through field work will be coordinated with the data obtained from historical texts. For example, in the old texts a noun form *šćeta* has been regularly attested, which form is also spoken today in the Makarska coast (Brela, Podgora), as confirmed by our field data. A computer program will link this form with the standard-language form *šteta*. In order to achieve this, it is necessary to precisely define linguistic characteristics of the Makarska coast dialect and correlate them with the corresponding forms of the standard Croatian. For example, once we define the idiom of Makarska coast as šćakavian (i. e. that psl.*stj and *skj gave šć) and standard Croatian as štakavian (i. e. that psl.*stj and *skj gave št), the computer program will easily link the dialectal forms of gušćerica, šćucat and šćap with standard-language forms gušterica, štucati and štap. By linking these lexems on the phonological and morphological level we are linking them on the semantic level. This is the most used procedure by authors of dialectal dictionaries: in the glossary they usually attach meanings verified in the standard Croatian language to dialectal words. Another way of obtaining the meaning of a particular word is by inferring from

the totality of verified texts of certain area, in this case the Makarska coast: based on the context in which a particular word appears, its meaning can be determined and recorded. As such it completely corresponds to the meaning of the word in the standard language, narrows it, or alters it altogether.
The idea is that a research of the same kind is to be carried out on the whole of the Croatian language. Machine-generated data should present a true picture of the Croatian language, connect the history of the Croatian with its present and show all the changes that have taken place in the language.
In the process, the primary meaning of a word will be determined, the semantic field of each word will be identified, as the overlapping of these fields, their semantical convergence and divergence, all of which would create a respectable encyclopedic knowledge base.

The creation of a semantic domains and their usage

In the context of the SSF a semantic domains represent a set of words that are meaningfully and closely related to a specific word. Building of such domains can be done either manually or automatically. The manual creation of domains is certainly a more difficult process, because it involves a huge human effort. In this article we propose a new way of creating semantic domains based on the definitions of words derived from publicly available online sources. All definitions of words from the available online resources such as the Miroslav Krleža Institute of Lexicography's online encyclopedia or Croatian language portal, are part of the SSF. Beside the fact that all elements of definitions are interconnected, and form a new semantic network, they are also a part of a SSF lexicon, which enables expansion of such domains so long as words with definitions exist in the database. There are two main approaches to a semantic domains creation within the SSF, and both are performed in four steps. In the first step, the definition from an online resource is obtained for a given word, which is then processed with extraction algorithm. The first approach (like shown in Figure 3) extracts the subject, predicate and the object. The result is then lemmatized. The final result is a set of words that makes the semantic domain for a given word. Since every word has a number of properties that have an ontological (T-structure) attached to them, so the name/word itself is also a property, it was easy to make a difference. For example, it is enough to know the accent of the word (which is also one of the properties) or WOS/SOW of words neighbors together with a suitable morphosyntactic pattern to determine what the term is it about.

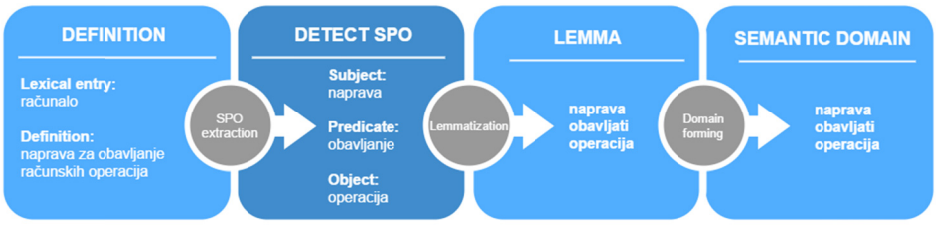

Figure 3. Creation of a semantic domain by extracting subject, predicate and object

The second approach differs only in the second step, where instead of the subject, predicate and the object extraction, definition elements are filtered using WOS marks (Figure 4). A user can define the tags the filter he will use, thereby automatically affecting the size and scope of the newly created semantic domain.

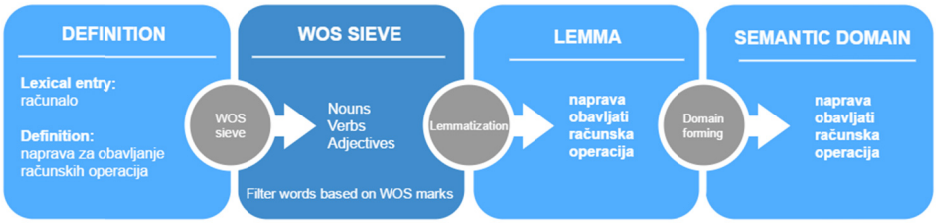

Figure 4. Creation of a semantic domain by WOS mark filtering

Such domains can be further recursively expanded with new elements resulting from the repeated process of extraction.
The main usage of such domains in the SSF is in the process of tropes detection (e.g. metaphors extraction). The algorithm that extracts metaphors from the text consists of three steps (Orešković et al. 2017). The first step is also the simplest because the metaphor is found based on its match within the repository. Stored and tagged multiword expressions are lemmatized and matched with the text. The second step is performed by using a syntactical and semantic patterns inside a defined virtual (semantic) domains, whose formation is described above. The last step is by detailed WOS/SOW analysis of each word in a sentence.

Conclusion

This research showed that creation of "semantic domains" can be done by using online encyclopedic articles. The Miroslav Krleža Institute of Lexicography's online encyclopedia (http://enciklopedija.lzmk.hr/) was used as an example. Also, complex programming support was created to achieve a domain creation. Two approaches are developed: through the open-class words (nouns, verbs, adjectives) and through the word function in the sentence (subject-object-predi-

cate) – dependency grammar. Because of the dialect richness of the Croatian language, program support for standardizing dialect forms to standard language forms was added. The created domains can be used to extract semantic information from unstructured text even at the most semantic level (stylistic figures: metaphor, metonymy, etc.)

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which were helpful in improving the article.

Funding

This work has been supported in part by the Croatian Science Foundation under the project IP-06-2016.

References

Ackoff, RL. 1989. "From Data to Wisdom." *Journal of Applied Systems Analysis*.

Alberico, R. and M. Micco. 1990. "Expert Systems: For Reference and Information Retrieval."

Antoniou, G. and F. Van Harmelen. 2004. "A Semantic Web Primer."

Bosančić, B. 2016. "Proces Stjecanja Znanja Kao Problem Informacijskih Znanosti." *Libellarium: Journal for the Research of Writing,*.

Frické, M. 2009. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy." *Journal of Information Science*.

Gruber, TR. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition*.

Jecić, Zdenko. 2013. *Studia Lexicographica.* [s.n.]. Retrieved May 31, 2017 (http://hrcak.srce.hr/114403).

Liew, A. 2007. "Understanding Data, Information, Knowledge and Their Inter-Relationships." *Journal of Knowledge Management Practice*.

Markučič, Joško and Klemen Govedić. 2013. "Morphological Generator of Croatian Language."

Melcuk, I. A. 1981. "Meaning-Text Models: A Recent Trend in Soviet Linguistics." *Annual Review of Anthropology* 10(1):27–62.

Mihalcea, Rada and Andras Csomai. 2007. "Wikify!: Linking Documents to Encyclopedic Knowledge." *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* 233–42.

Orešković, M., M. Čubrilo, and M. Essert. 2016. "The Development of a Network Thesaurus with Morpho-Semantic Word Markups." *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* 273–79.

Orešković, Marko, Marta Brajnović, and Mario Essert. 2017. "A Step towards Machine Recognition of Tropes." P. 71 in *Third International Symposium on Figurative Thought and Language*.

Pinjatela, Krešimir. 2001. "Hrvatska Riječ."

Pustejovsky, J. 1991. "The Generative Lexicon." *Computational Linguistics*.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. "Yago." *Proceedings of the 16th International Conference on World Wide Web - WWW '07* 697.

Šnajder, Jan and Petra Almić. 2015. "Modeling Semantic Compositionality of Croatian Multiword Expressions." *Informatica* 39(3):301.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary." *Linguistics* 1646–52.