Agnieszka KONYS

West Pomeranian University of Technology, Szczecin

doi:10.15199/48.2015.11.49

An Approach for Ontology-Based Information Extraction System Selection and Evaluation

Abstract. *The process of Information Extraction (IE) allows us to retrieve different types of information from natural language text by processing them automatically. Ontology-based information extraction (OBIE) is a subfield of information extraction. An increasing number of existing OBIE system may cause a problem with selection the most suitable solution. The general aim of this paper is to provide an approach for OBIE system selection and evaluation. It should ensure knowledge systematization and help users to find a proper solution that meets their needs.*

Streszczenie. *Zastosowanie ekstrakcji informacji pozwala na pozyskiwanie różnych typów informacji w języku naturalnym, jednocześnie umożliwiając automatyczne jego przetwarzanie. Systemy ekstrakcji informacji oparte na ontologiach są poddziedziną ekstrakcji informacji. Rosnąca ich liczba oraz różnorodność podkreślają ważność problemu, jednocześnie wskazując na możliwość występowania problemu związanego z ich doborem. Celem artykułu jest prezentacja podejścia wspierającego proces doboru i oceny systemów ekstrakcji informacji opartego na ontologiach.* ***Dobór i ocena systemu ekstrakcji informacji opartego na ontologiach.***

Keywords: OBIE systems, Semantic Web, ontology, Information Extraction, evaluation measures for OBIE.
Słowa kluczowe: systemy OBIE, Semantic Web, ekstrakcja informacji, miary oceny dla OBIE.

Introduction

Nowadays, the World Wide Web (WWW) is the most popular way to distribute information. Despite this, the problem of finding or extracting relevant information out of a large amount of data from the Web exists. An enormous amount of data is spread across the Web and an increasingly a huge number of users is making advanced use of it. The users expect search engines to understand natural language and perceive the intent behind the words they type in , but most of the time they get lots of irrelevant and unimportant web documents. Hence, the process of searching information in the Web might be a little bit time consuming to find relevant content in the Web resources. Very often information gathered in Web resources is not easily processable by machines. A huge number of data available on the Web and their heterogeneity may cause a lot of problems. Due to dynamic development of new technologies and an increasing amount of data available on the Web, the greater role is assigned to efficient data collection, analysis and processing.

Another problem concerns the heterogeneity of documents published on the Web. The existence of different input sources both structured (e.g. HTML, XML/XML Schema, RDF/OWL, relational data) and unstructured (e.g. text, documents, images) poses new challenges of WWW. The process of retrieving information from unstructured or semi-structured sources may cause some problems. Another problem concerns the linking between different types of documents available on the Web.

Search has changed dramatically over the last years and it is seemed that it is still early days for the rapidly changing environment. In 2001, Berners-Lee published the idea of Semantic Web, which extends the WWW by a formal and machine interpretable layer. In recent years the Semantic Web community has been very active and productive in this research field [13]. One of its main purposes published by Berners-Lee is to provide a meaningful representation of machine readable data over the Web. It means that machines are capable of rightly interpreting the data [14]. In this research area a great amount of work is dedicated to improving ontology engineering. This includes techniques to discover correspondences and to match similar concepts automatically [15].

In last decade the terms related to Semantic Web (e.g. Information Extraction (IE), Knowledge Extraction, ontologies etc.) become the significant elements in the efficient way of information retrieval, processing and supporting availability of machine readable data. This success, widespread usage and commercialization emphasizes their role of WWW community. A close relation between ontology-based information extraction (OBIE) and the Semantic Web is noticeable [12]. OBIE systems generate semantic content which is known as Semantic Annotation for the Web pages. Semantic agents can directly process semantic content for Information Retrieval.

A number of existing OBIE systems emphasizes the role of this element in Semantic Web community. An increasing number of existing OBIE system may cause a problem with selection the most suitable solution. The general aim of this paper is to provide an approach for OBIE system selection and evaluation. It should ensure knowledge systematization and help users to find a proper solution that meets their needs. The main advantage is that a user does not need to check many sub-sites to find relevant information of OBIE systems and their functionalities. Furthermore, it enables information classification, what has a great impact in the future work.

Information extraction

The process of Information Extraction (IE) is based on automatic retrieving certain types of information from natural language text [1]. The general aim of this is to process natural language text and to retrieve occurrences of a particular class of objects or events and occurrences of relationships among them [2]. Another definition describes information extraction as a form of natural language processing, in which certain types of information must be recognized and extracted from text [3]. It can be defined also as a task of identifying, collecting and normalizing relevant information from NL text and skipping irrelevant text passages. Information Extraction has employed various algorithms and methods for information retrieval.

In literature [1, 12], IE system is described as a system that processes a set of web pages and extracts information from a given domain. IE systems do not attempt an exhaustive deep NL analysis of all aspects of a text. Rather, they are built in order to analyse or “understand” only those text passages that contain information relevant for the task at hand [4].

The IE system functionality is described in the fig. 1. On the input IE system may have the possible values: IE model that includes: specification of lexical knowledge, extraction rules, and ontology, and a set of NL texts, which

may encompass technical reports press releases, online-documents, or emails. On the output the following values are obtained: target knowledge structure, i.e. a set of instantiated and related concepts and attributes [1, 12].

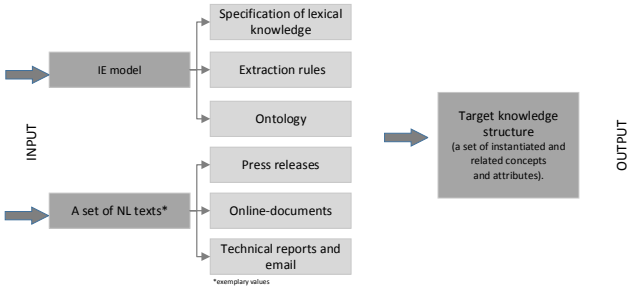

Fig.1. IE system functionality

Ontology-based information extraction (OBIE)

The term ontology-based information extraction (OBIE) has recently emerged as a subfield of information extraction [5]. OBIE is different from traditional IE because it finds type of extracted entity by linking it to its semantic description in the formal ontology [6]. Moreover, ontologies are used by the information extraction process, and the output is generally presented through an ontology. The definition presented by [26] describes OBIE as a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information, and to present the output using ontologies.

Wimalasuriya and Dou [1] presented three key characteristics of OBIE systems that make OBIE systems different from general IE systems. The first of them is the process of natural language text documents. The inputs of OBIE systems are limited to unstructured or semi-structured documents. The next thing encompasses information extraction process, which is guided by an ontology to extract things such as classes, properties and instances. The third characteristics presented by authors is a possibility to present the output using ontologies: OBIE systems must use an ontology to represent the output. It is worth to notice that Information Extraction has employed various algorithms and methods for information retrieval. OBIE does not require a specific algorithm or a method. It is based on ontology, which is used to support and guide algorithms for efficient and relevant information extraction [7].

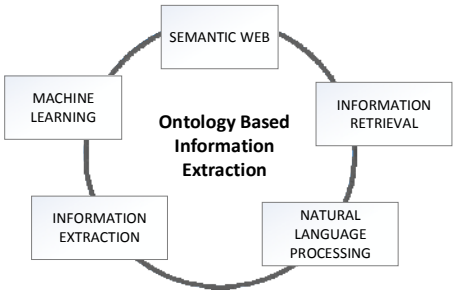

Fig.2. OBIE system influence areas

OBIE system is under the influence of the following areas: machine learning, Semantic Web, Information Extraction and retrieval, and Natural Language Processing (fig. 2). OBIE systems [8, 9, 10, 11] allow to automatically process the information contained in natural language text. Moreover, they create semantic contents for the Semantic Web, and additionally, they improve the quality of ontologies. One of the most important potentials of OBIE is its ability to automatically generate semantic contents for the Semantic Web [1, 5].

A general architecture of OBIE system

In literature, the most frequently cited general architecture of OBIE system is proposed by [1]. In many cases OBIE systems do not include or exploit all of these components. Some elements from this general architecture were used to construct the procedure. On base of this, the proposal of a general architecture of OBIE system is presented. It is still developed and modified. It contains the additional modules: QAS selection support (presented in [29]), a tool supporting automatic ontology construction processes (presented in [28]), and an approach for OBIE system selection and evaluation (fig. 3). The Protégé software is used as an ontology editor. It is worth to notice that OBIE system can be a part of a larger system.

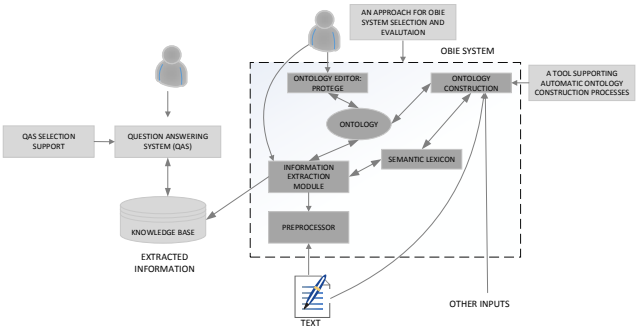

Fig.3. Architecture of OBIE system with additional modules

On input, a preprocessor component converts the text to a format that can be handled by the information extraction module. Information Extractor populates the Knowledge base and is queried by users. Output of the OBIE system contains the information extracted from the text. In addition, the output might also include links to text documents from which the information was extracted [1].

An analysis of selected approaches for OBIE

An analysis of literature provides a description of different OBIE systems, methods, approaches and projects supporting them. A number of available solutions emphasizes their role and importance. Moreover, it seems that this field of interest is still developed. The analyzed solutions differentiate between each other in terms of information extraction processes, available types of sources, the extraction of ontology components, the update processes and more.

A number of tools and approaches to support information extraction processes exists. They offer a wide range of functionalities. The analysis of literature allows to indicate the selected solutions: GATE, UIMA, sProUT, SOBA, Text-To-Onto, and OntoX [1, 12, 16, 17, 18, 19, 20, 21]. A number of existing solutions is quite big, but due to the lack of development and the time of publishing the proposed analysis is limited to these chosen approaches. It is worth to emphasize that OBIE method only guides the system how to pull out efficient and relevant information using the Information Extraction methods. A tool offers a wide range of capabilities. Most of them is available for free with the full support. The table 1 presents the selected solutions within a short description of each of them.

Table 1. The characteristics of selected OBIE systems

Name	Short description
GATE (General Architecture for Text Engineering) [16]	It is an open-source, web-based, collaborative text annotation framework. It offers users to carry out complex corpus annotation projects, involving distributed annotator teams. Moreover, it supports a user providing an infrastructure for developing and deploying software components that process human language. It provides a set of NLP tools including a tokeniser, a gazetteer, a POS tagger, a chunker, and parsers, etc. The process of relation extraction is supported by the following components: English Tokeniser, Sentence Splitter, POS Tagger, NP Chunker, VP Chunker, BuChart Parser, MiniPar Parser, WordNet.
UIMA (Unstructured Information Management	The aim of UIMA framework is to provide a platform for natural language processing (NLP) applications. It enables analyzing large volumes of unstructured information in order to discover knowledge that is relevant to an end user. Moreover, it allows to decompose applications into components.
sProUT (Shallow Processing with Unification and Typed featured structures	It combines finite-state techniques and unification-based algorithms. It is a platform for development of multilingual STP (shallow text processing) systems. It consists of the following components: finite-state machine toolkit, a regular compiler, a finite-state machine interpreter, a typed feature structure package, and linguistic processing resources. It provides basic grammars for the annotation of persons, locations, numerals and date and time expressions.
SOBA (SmartWeb Ontology-based Annotation) [19]	It is a component for OBIE system from soccer web pages for automatic population of a knowledge base that can be used for domain-specific question answering. It offers the following information extraction methods: Linguistic rules, Gazetteer lists, Analyzing tags. It is possible to provide only html files from a given domain. This system uses a corpus of web pages about soccer games as its source. SOBA retrieves the web pages that it processes using its own web crawler. It extracts the information from heterogeneous sources like tabular structures, text, image caption in a semantically integrated way. The SOBA OBIE system extracts information from HTML tables into a knowledge base that uses F-Logic. This system uses a corpus of web pages about soccer games as its source.
Text-To-Onto [20]	This system uses the SMES (Saarbücker Message Extracting System) system to construct an ontology. It is composed of several components. It produces an under-specified dependency structure as the output, which is basically a partial parse tree. This structure is used for information extraction. It uses documents from a domain as a source. The information is extracted using partial parse trees. It leverages data mining and natural language processing techniques to assist the user in the development and maintenance task by analyzing web data and suggests modelling decisions. Text-To-Onto is used with KAON comprehensive server architecture.
OntoX [21]	It is a method for ontology-driven information extraction. It is based on linguistic rules that use regular expressions with the elements of ontologies such as classes and properties, resulting in “extraction ontologies. It allows to extract information from natural language text without using any knowledge resource except an input ontology.

A comparative analysis of selected OBIE systems

The presented characteristics of selected approaches were a basis for a comparative analysis. The comparative analysis includes the results of work presented by [1, 5, 12]. The criteria proposed by authors were used to create classes and sub-classes. A table 2 presents the comparative analysis of selected OBIE systems, including the set of 5 main criteria and the set of 22 sub-criteria: Information extraction method (Linguistic rules, Gazetteer lists, Analyzing tags, Partial parse trees), Ontology construction and update (Off-the-shelf, not updated, "Constructed by process", Manually defined, N/A), Components of the ontology extracted (Instances, Property values, Classes, Taxonomy, Datatype property values, Other relationships), Types of sources (HTML files from a domain, XML files from a domain, Documents from a domain), Offered support (Tool, Other supporting solutions, Components, Providing an infrastructure for developing components).

Table 2. A comparative analysis of selected OBIE systems

Criteria	Sub-criteria	GATE	UIMA	sProUT	SOBA	Text-To-Onto	OntoX
Information Extraction Method	Linguistic rules	+	+	+	+		+
	Gazetteer lists	+	+		+		
	Analyzing tags	+	+		+		
	Partial parse trees					+	
Ontology Construction and Update	Off-the-shelf	+	+		+		
	not updated			+	+		+
	Constructed by process					+	
	Manually defined	+	+				+
	N/A					+	
Components of the Ontology Extracted	Instances	+	+		+		+
	Property values	+	+		+		
	Classes	+		+		+	
	Taxonomy	+		+		+	
	Datatype property values	+	+				+
	Other relationships	+	+			+	
Types of Sources	HTML files from a domain	+	+		+		
	XML files from a domain	+	+				
	Documents from a domain	+	+	+		+	+
Offered support	Tool	+	+	+	+	+	+
	Other supporting solutions	+	+				+
	Components	+	+	+	+	+	
	Providing an infrastructure for developing components	+	+			+	

The measures for OBIE evaluation

The evaluation of the quality of ontological classification is an important element of semantic web technology. Traditionally, OBIE systems are evaluated using Precision, Recall and F-Measure. There are the most popular measures. They provide a binary decision of correctness for each entity in the text, i.e. by comparing the key (gold standard) and system responses, they classify the result as either right or wrong in each case [24, 25]. Precision shows the number of correctly identified items as a proportion of the total number of items identified. Then recall shows the number of correctly identified items as a proportion of the total number of correct items available. F-measure is a weighted average of the two metrics. The specification of these methods are detailed described in many works, e.g. in [25, 27].

Apart from this, the new measures are developed and proposed by [11, 24, 25]. The authors criticized the traditional metrics (Precision, Recall). They claim that using precision and recall with OBIE systems can be problematic because these metrics are binary in nature. They state that OBIE systems should be evaluated in a scalar manner, allowing different degrees of correctness [24].

Another existing metric proposed by [22] is called Learning Accuracy (LA). It is classified as cost-based and distance-based metric. It allows to measure how well a concept had been added in the right level of the ontology,

but it can be equally applied to measure how well the instance has been added in the right place.
This metric was used by [23] to evaluate how well an ontology has been populated. Learning Accuracy (LA) essentially measures ”the degree to which the system correctly predicts the concept class which subsumes the target concept to be learned”.

Maynard et al. have defined two metrics called Augmented Precision (AP) and Augmented Recall (AR) that can be used for OBIE systems [24]. These measures combine the concepts behind precision and recall with cost-based metrics. Augmented Precision and Recall measure that takes into account the ontological distance of the response to the position of the key concepts in the hierarchy.

Another considered solution is called Balanced Distance Metric (BDM) [24, 25]. The general aim of this approach is to take into account the more flexible nature of ontologically-based applications. It is concentrated on ontological similarity. The BDM computes semantic similarity between two annotations of the same token in a document.

The process of selection a proper metric for OBIE evaluation depends on the user. It is worth to notice that this filed is still developed and, as a consequence, it requires improvement and standardisation.

An approach for OBIE system selection and evaluation

Different types of available OBIE systems may cause the problem of a selection of a proper solution. Information of OBIE systems is scattered, and the whole process of information gathering about OBIE systems is time-consuming. The general aim of presented approach for ontology-based information extraction system selection is to provide a possibility to choice a proper OBIE system to a given decision situation. Moreover, it should provide the knowledge systematization of OBIE systems domain, and enables a time reduction for OBIE system selection. The aim of proposed approach is to support a user in a selection process and to reduce a time necessary for gathering information of a given OBIE system and offered tools.

The process of retrieving information of OBIE systems and offered functionalities by them took some time. It is worth to notice that existing solutions differentiate between each other, and unnecessarily user does not have to have a specified knowledge of available OBIE systems. The proposed approach helps to find a solution that suits the best to user preferences. An application of reasoning mechanisms supports the selection process and provides a set of results with regards to the user requirements. The short description of available metrics for OBIE system evaluation is included. Finally, the approach will be enhanced by a adding an evaluation criteria and metrics for OBIE system evaluation. It will ensure a complex solution for OBIE system selection and evaluation processes (fig. 4).

A proposed approach is based on ontology application. It is implemented in OWL language. On base of analysis of selected OBIE systems the set of criteria (5) and sub-criteria (22) was defined. It was a basis for the taxonomy and ontology construction as a next step. A general procedure of an ontology construction consists of the following phases: (1) defining a set of criteria, (2) taxonomy construction, (3) ontology construction, (4) formal description, (5) defined classes creation, (6) reasoning process, (7) consistency verification, (8) a set of results.

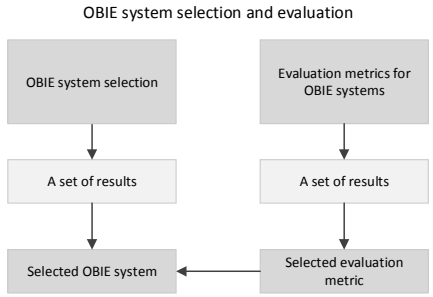

Fig.4. An approach for OBIE system selection and evaluation

A figure 5 presents a part of classification criteria (including class hierarchy) of selected OBIE system and graphical visualisation of the ontology.

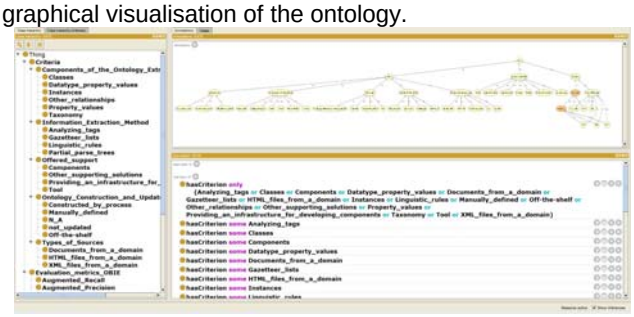

Fig.5. A part of classification criteria of selected OBIE system

A figure 6 depicts the relations between selected the OBIE systems.

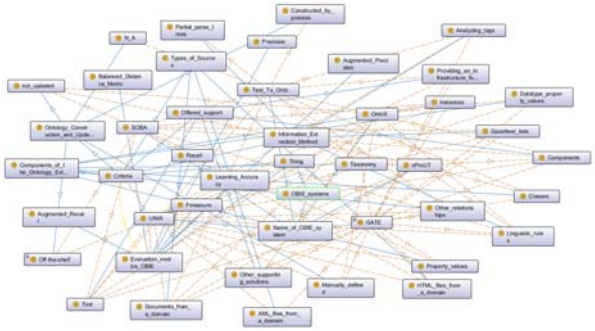

Fig.6. Existing relations between selected the OBIE systems

Case study: an approach for OBIE system selection and evaluation

The case study presents practical examples of the approach for OBIE system selection. It is worth to emphasize that a decision-maker may specify a non-limited set of queries. It is supposed that a decision-maker is looking for the OBIE system that fulfills a set of pre-defined requirements. It is assumed that preferable solution should have all of the following requirements: Offered support: Components, Offered support: Tool, Information extraction method: Gazetteer lists, Information extraction method: Linguistic rules, Components of the ontology extracted: Instances. The application of the reasoning mechanism provided a set of results with regard to the pre-defined requirements. In this case 3 OBIE systems (GATE, UIMA, and SOBA) fulfil this defined set of criteria (fig. 7).

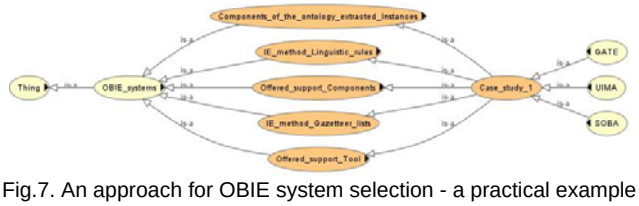

Fig.7. An approach for OBIE system selection - a practical example

It is possible to define a non limited set of queries and definitions in a given ontology. The process of OBIE system evaluation can be done by using one of available metrics for OBIE system evaluation. The next step will encompass the complex support in this process, providing a solution which recommends a metric for a given problem.

Conclusion

This paper presents the an approach for OBIE system selection and evaluation. The general aim of proposed approach was to provide knowledge systematization in this research area, and to support a decision-maker in a selection process of OBIE system.

In this paper, the analysis of selected OBIE systems was presented. The short description of evaluation measures for OBIE systems was provided. The comparative analysis allowed to construct the set of criteria and sub-criteria of available OBIE systems. The results from the comparative analysis were used to create the ontology, which was the element of the proposed approach. It ensured the support in OBIE system selection. Next, the process of OBIE system evaluation might be supported by using one of available metrics for OBIE system evaluation.

The future researches will encompass the support in this process, providing a solution which helps in recommendation of a metric for a given problem. It is worth to emphasize that the proposed approach for OBIE system selection and evaluation is a part of the more complex work, including support in the following areas: Question Answering Systems selection [29], automatic ontology construction processes [28], and application of Natural Language Interface to query ontologies.

REFERENCES

[1] Wimalasuriya D.C., Dou D., Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36, (2010), No. 3, 306

[2] Russell S., Norvig P., Artificial Intelligence: A Modern Approach, 2nd edition, Prentice-Hall, Englewood Cliffs, NJ (2003), 848-850

[3] Riloff E., Information extraction as a stepping stone toward story understanding, In: Ram A., Moorman K., (eds), Understanding Language Understanding: Computational Models of Reading, MIT Press, Cambridge, MA (1999)

[4] Maedche A., Neumann G., Staab S., Bootstrapping an Ontology-Based Information Extraction System, In: Studies in Fuzziness and Soft Computing, Intelligent Exploration of the Web, Springer, (2002)

[5] Wimalasuriya D.C., Dou D., Using multiple ontologies in information extraction. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, New York (2009)

[6] Nebhi K., Ontology-Based Information Extraction from Twitter, Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, Mumbai (2012), 17-22

[7] Shah R., Jain S., Ontology-based Information Extraction: An Overview and a Study of different Approaches, International Journal of Computer Applications, 87, (2014), No. 4

[8] Wu F., Weld D.S., Autonomously semantifying wikipedia. In: Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management, New York (2007)

[9] Kietz J., Maedche A., Volz R., A method for semi-automatic ontology acquisition from a corporate intranet. In: Proceedings of the EKAW'00 Workshop on Ontologies and Text, Springer, Berlin (2000)

[10] Cimiano P., Handschuh S., Staab S., Towards the self-annotating web. In: Proceedings of the 13th International Conference on World Wide Web, New York (2004)

[11] Maynard D., Peters W. Li Y., Metrics for evaluation of ontology-based information extraction. In: Proceedings of the WWW 2006 Workshop on Evaluation of Ontologies for the Web, New York (2006)

[12] Wimalasuriya D.C., Dou D., Components for Information Extraction: Ontology-Based Information Extractors and Generic Platforms, CIKM'10, Canada, (2010)

[13] Berners-Lee T. et. al., The Semantic Web, Scientific American (2001)

[14] Ramprakash et. al., Role of Search Engines in Intelligent Information Retrieval on Web, Proceedings of the 2nd National Conference, India (2008)

[15] Bedini I. et. al., Janus: Automatic Ontology Builder from XSD files, Proceedings of the World Wide Web Conference, Beijing, China (2010)

[16] Bontcheva K., Tablan V., Maynard D., Cunningham H., Evolving GATE to meet new challenges in language engineering, Natural Language Engineering, 10, (2004), 349-373

[17] Ferrucci D, Lally A (2004) UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. Natural Language Engineering, 10 (3-4), 327–348

[18] Drozdzynski W., Becker M., Krieger H.-U., Piskorski J., Schäfer U., Xu F., SProUT (Shallow Processing with Unification and Typed Feature Structures) (2002), available at: http://sprout.dfki.de

[19] Buitelaar P., Siegel M., Ontology-based Information Extraction with SOBA. In: Proceedings of the 5th International Conference on Language Resources and Evaluation, Italy, (2006)

[20] Maedche A., Staab S., The Text-To-Onto Ontology Learning Environment. In: Software Demonstration at the Eighth International Conference on Conceptual Structures, Springer-Verlag, Berlin, (2000)

[21] Yildiz B., Miksch S., OntoX - a method for ontology-driven information extraction. In: Proceedings of the 2007 International Conference on Computational Science and Its Applications, Springer, Berlin, (2007)

[22] Hahn U., Schnattinger K., Towards text knowledge engineering. In Proc. of 15th National Conference on Artificial Intelligence (AAAI-98), MIT Press, (1998), 524–531

[23] Cimiano P., Staab S., Tane J., Automatic Acquisition of Taxonomies from Text: FCA meets NLP, In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, (2003), 10-17

[24] Maynard D., Benchmarking ontology-based annotation tools for the semantic web, In UK e-Science Programme, All Hands Meeting (AHM2005) Workshop "Text Mining, e-Research and Grid-enabled Language Technology", (2005)

[25] Maynard D., Peters W., Li Y., Metrics for evaluation of ontology-based information extraction, In WWW 2006 Workshop on "Evaluation of Ontologies for the Web" (EON), Scotland, (2006)

[26] Adrian, B.: Incorporating ontological background knowledge into Information Extraction. In: Maynard, D. (ed.), ISWC 2009: Doctoral Consortium, Washington, US (2009)

[27] Li Y., Bontcheva K., Cunningham H., Perceptron-like learning for ontology based information extraction. Technical report, University of Sheffield, UK, (2006)

[28] Konys A., A Tool Supporting Mining Based Approach Selection to Automatic Ontology Construction, In Proceedings of: th Multi Conference on Computer Science and Information Systems 2015 9th European Conference on Data Mining, Spain, (2015)

[29] Konys A., Knowledge-based Approach to Question Answering System Selection, 7th International Conference on Computational Collective Intelligence Technologies and Applications, Madrid, Spain, (2015)

***Author**: dr inż. Agnieszka Konys, Zachodniopomorski Uniwersytet Technologiczny w Szczecinie, Wydział Informatyki, Katedra Inżynierii Oprogramowania, Zakład Symulacji Komputerowej i Ochrony Informacji, ul. Żołnierska 52, 71-210 Szczecin, E-mail: Agnieszka.Konys@zut.edu.pl*