# Formalization of Unstructured Content to Semantic Form

Sandi Pohorec, Milan Zorman

University of Maribor, Faculty of Electrical Engineering and Computer Science,
Smetanova ulica 17, SI-2000 Maribor, Slovenia
{Sandi.Pohorec, Milan.Zorman}@uni-mb.si

**Abstract.** This paper will present an approach for knowledge extraction from unstructured content. Unstructured content refers to information syndicated online, with little structure. The information extracted from these content sources is semantically tagged and stored in the form of semantic graphs. The approach is upgraded with the transformation of semantic graphs to a relational database. The goal of the content acquirement is to build a repository of formal knowledge, similar to DBpedia, for content in Slovene language. The formalized content will use the database as the permanent storage and would be provided in various formats suited for machine processing (RDF dump, API, SPARQL endpoint).

**Keywords:** natural language processing, knowledge formalization, semantic representation, ontology construction, relational data, databases, knowledge engineering.

## 1 Introduction

The current generation of the World Wide Web has transformed the users from passive content consumers to active content creators. Therefore user based content is a rich source of information to be used in data mining or artificial intelligence projects. In the scope of this paper the term *user based content* will be used according to the following restrictions: content is available online, content is not peer reviewed or published in literature, content is available in various formats, the content is often in the form of news items or blog posts. With regard to the content we have limited the research on the content available from web sources: documents and web pages. The content channels are limited to online news services and blog sites. The content is obtained with the use of a RSS agent and a custom, focused crawler.

The main focus of the research is the transformation from free-text (or semi-structured text) to semantically annotated graph representations, which also provide the performance for real-life application. The motivation is to create a structured data set in Slovene language. The complete collection will be offered to the community in the form of a structured data dump, typically in RDF triples. The approach can be compared to the DBpedia [1] project. Related work includes work of various Slovene researchers in the field of natural language processing [5] [6]. Currently there is still a

very limited amount of formal, annotated data in Slovene language. With the exception of the Slovene WordNet (SWN) [7] which is interlinked with the use of the Princeton WordNet index (SWN was created with the use of the expand model), other resources offer little in the form of semantic cross-linking with other formal data sets. The data set we are creating will be interlinked with the Web of Data [3] [4]. The approach we are using can be described as a bottom-up approach in the sense that we are starting with the extraction of instances (individuals) and generalize them into concepts, which are later cross-linked with existing vocabularies within Linked Data. The process is segmented in the following phases: content acquisition and aggregation, syntactic preprocessing, semantic net construction, semantic tagging, ontology construction (concepts, instances and properties) and transformation to relational data storage.

The paper is segmented in sections that follow the major process phases. The next section introduces the concept of ontology, ontology web language, the Linked Open data and DBpedia projects. Section three defines the content in more detail. In section four we present the processing of the content in natural text. This is continued in section five where experimental results are presented. The final section is the conclusion.

## 2 Formal Data and the Web

Natural language is by nature ambiguous, therefore it is not machine understandable. Scientific community has been researching the methods for machine processing of natural text for many years. The analysis of natural language texts with the goal of extracting structured information is common known as Information Extraction (IE). The process works on unstructured texts and extracts formal unambiguous data. A subarea of IE is named entity recognition (NER). NER focuses on extracting several different types of information: entities (people, places, organizations, etc.) mentions (entities indirectly referred), relationships between entities and events involving entities. Extracted information is converted and stored in a machine readable format.

Machine readable formats are usually some type of a semantic graph. Artificial Intelligence (AI) community has proposed the use of ontologies for referring to formal descriptions of individual domain. The ontology is composed of terminology and assertion components [4].The most widely accepted definition of ontology is by Thomas Grubber [9]: *Ontology is a formal specification of a conceptualization.* In the context of formal data on the web, specifically the Web Ontology Language (OWL), ontology provides the capability to define classes, properties, instances (individuals) and the relationships between them [4]. OWL instances are represented in the meta-data language RDF (Resource Description Framework). RDF is similar to conceptual modeling approaches like Entity-Relationship diagrams, common in the database world. The working logic of RDF is the expression of facts by making statements. The statements are triples connecting subjects to objects with predicates. Two types of RDF statements are possible: literal valued and resource valued statements. Literal valued statements are statements where subject (resource) is linked by the predicate (property) to object (literal value), while resource valued statements link resources

with resources. On the example statement "the paper has the color white", the RDF representation would be: subject – *paper* is linked by the predicate- *has the color* to the object – *white*. A series of RDF triplets form a directed labeled graph.
RDF is also the graph based data model that is used for the structuring and linking of data in the Linked Data initiative. Linked Data [2] is a project to create a web of machine readable data. The main difference from a user perspective is that the WWW (World Wide Web, hypertext Web) uses HTML, as the primary units of content, while Linked Data uses structured data in RDF. The hyperlink construct of WWW is replaced with data links between arbitrary things that are described by RDF [10]. Linked Data is created on the four main principles, set by Berners-Lee [10]: 1) things are named with URI's, 2) HTTP URI's are used so lookup is possible for humans, 3) on lookup useful information is provided in standard form (RDF*, SPARQL) and 4) information includes links to other URI's with additional information.

The most visible application of the Linked Data principles is the Linking Open Data (LOD) project [11]. Many large datasets form the LOD cloud, one of them being DBpedia [1]. DBpedia uses Wikipedia to extract structured data and expose it publicly.

## 3 Content Formats and Acquisition

Web 2.0 enables the end-users to become active content publishers. The content is interesting from the perspective of formalizing common knowledge therefore it was included as one of primary content sources, the other being news items, The content can be classified with regard to the format in which it is accessible: unstructured HTML content (news items, blog posts, content syndicated through RSS) or semi-structured content (microformats, based on XHTML).
HTML is a markup language designed for visual formatting; it lacks any semantic annotation capability. Therefore content in HTML format is basically plain natural language. An extension of XHTML, that provides integrated semantic markup are microformats. Microformats use the XHTML class attribute system and reuse existing elements to present semi-structured information. The microformats technology stack is composed of four layers: XML, XHTML, elemental microformats and compound microformats [2].

For the content acquisition purposes a custom aggregator was implemented. The application works as a focused crawler for the HTML content (user blogs and other social networks) and as RSS aggregator for the content in RSS format (news items). The application is comprised of three modules: 1) a focused HTML crawler, 2) an RSS aggregator and 3) the preprocessing module (transformation to plain-text and segmentation). Every module retrieves the content and stores it in the temporary storage in the database.

## 4 Natural Language Processing

One of the main focus points in this research was the verification and analysis of the ability to implement the syntactic and semantic analysis entirely within the database environment. This part of the processing is applied to the sources that contribute natural language content, as content in microformats is semi-structured and extraction is simplified (few well known formats enable easy extraction of semi-structured data).The NL texts are stored in the temporary data store. The transformation to plain-text form, the syntactic and semantic analysis is performed using in-database technologies. The reasoning behind this was based on the following factors: natural language processing (NLP) usually requires large dictionaries, databases natively support parallelization, minimal user interface is required for NLP (processing only), T-SQL, compared to general programming languages, is easier to maintain and update and there is no need to transfer large quantities of data from permanent storage (database) to the processing application.

The process of the processing of the natural language content is in several successive phases, as shown in Fig. 1. The first phase is the lexical analysis which encompasses the transformation to plain-text (PT), text segmentation (tokenization, abbreviated as SEG. in Fig. 1) and part-of-speech (POS) analysis. The second phase is the semantic analysis where named entity recognition (NER) is performed. The third step is the resolution of equality (concepts and individuals that previously exist with either the same or a different label). The transformation to plain-text is straightforward and will not be presented in greater detail.

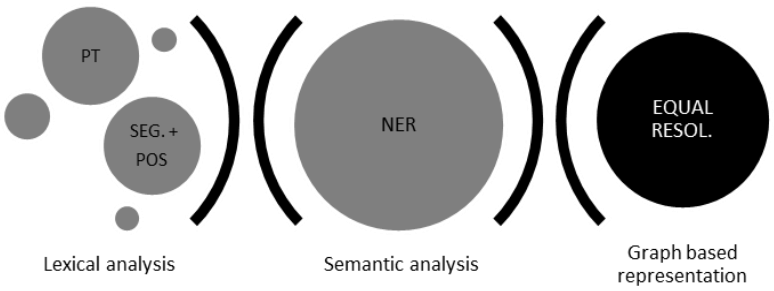


**Fig 1:** The NLP process in sequential order. The first step, the lexical analysis, encompasses transformation to plain-text (abbreviated as "PT"), segmentation (abbreviated as "SEG") to logical units, words, sentences and paragraphs and part-of-speech (abbreviated as POS) analysis. The second step, the semantic analysis is mainly concerned with named entity recognition (NER), although semantic analysis is also a part of the disambiguation phase of SEG and POS. The third step is the resolution of equality; where the same entities are linked together and their properties reanalyzed and joined.

### 4.1 Text Segmentation - Tokenization

Natural language texts are sequences of characters. The texts lack formal, explicit boundaries (words, sentences). Alphabet languages separate words with spaces. Usually NL applications require words, sentences and paragraphs, which are used as basic units of processing [12]. The most common problems of tokenization is the ambiguity of punctuations (a period is used in abbreviations, acronyms, dates etc.), the use of abbreviation, artificially separated words (line breaks in word processors) and various typing errors (missing spaces). We have approached the issue of tokenization with an algorithm heavy dependent on pre-existing knowledge. The reasoning for this is that a highly reliable and accurate tokenizer is required to resolve ambiguity at the semantic level. For instance when a person is named in the text and the first name is abbreviated (for example "S. Pohorec"), the tokenizer has no other means of knowing, that the period is not a sentence delimiter, than to understand that the multiword entity (first letter of the name, followed by period, space and surname) is a personal name. The ability to recognize multiword entities and classify them correctly required some degree of semantic understanding. The rules of syntax are ambiguous and only semantic reasoning can resolve the ambiguity. We have approached the problem of semantic reasoning only for the Slovene language while we use available tokenizers for the English language (English content is only used for entities that are the same across languages, like names). The problem was approached by the accumulation of known exceptions to the syntactic rules: personal names, dates, abbreviations etc. Also the known exceptions were cross-checked with a large corpus of Slovene natural language texts where the non-alphabet characters were classified manually. The most common types that can be used for the punctuation ambiguity include: name entities, dates, numbers, multiword entities, addresses (physical, email, web). The knowledge gained was formalized as rules. The rules have left-side conditions and right-side decisions.

### 4.3 Part-of-speech Analysis

With the tokenization (segmentation) problem, the diversity between English and Slovene language can be overlooked in majority of cases. The languages share basically the same areas where ambiguity of punctuations occurs and therefore solutions are quite similar. This similarity however does not exist when we consider the area of part-of-speech analysis. The Slovene language distinguishes the following word "types": Noun, Verb, Adjective, Pronoun, Determiner, Article, Adverb, Adposition, Conjunction, Numeral, Interjection and Particle. The complete list of properties (such as number, sex, case, etc.) shows 144 possible property values. The richness of the language greatly enhances the problems of ambiguity. Our approach for POS analysis includes the approach introduced in [13], where a custom pattern based tagger is used along with the well-known TnT [14]. This approach is further enhanced with linguistically tagged corpus of 1 million Slovene words; Jos 1M [15]. The corpus was partially manually checked and verified to contain correct lemma and POS tags. The results of our POS analysis are texts tagged with POS tags according to the MULTEXT specification [16]. Each tag is a string where each character

represents the value of the properties associated with the part-of-speech. For example tag »Ncmsa« is used to describe a noun (N) whose Type is common (c), Gender is masculine (m), Number is singular (s) and the Case is accusative (a).

### 4.4 NER and Semantic Graph Construction

Named entity recognition focuses on extracting entities from unstructured texts. It is the goal of this research to use NER to extract individual entities and use the implicit description in the surrounding text to map the individuals to the underlying concepts. The named entities are extracted with a combination of the POS analysis and the word frequency occurrence. The process is backed by lookups to structured data sources. We will explain the process of NER and the mapping of individuals to the ontology concepts on a news item, published about the meeting between Presidents Obama and Karzai. The research is focused on the Slovene language, but an English example will be used for easier understanding. The content of the example news item in natural language is:

[Title] "*Obama will host Karzai at White House*."
[SHORT DESCRIPTION] "*President Obama will welcome Afghan President Hamid Karzai to the White House on Wednesday for a second day of partnership talks*."

From the title we can extract three known entities: Obama, Karzai and the White House. All three were tagged to be nouns. The POS analysis suggests that "White House" is a geographical location, because "at" is a preposition (suggesting location or time, resolved as location, since the words do not represent any known time reference) and it is followed by a noun. The link between "Obama" and "Karzai" are the auxiliary verb and verb "will host". Therefore the meaning of the sentence is clear. Individual "Obama" will perform the activity of *hosting* to individual Karzai. Additional information is then obtained from the short description. "Obama" and "Karzai" are defined as both being "Presidents". They share the activity of "partnership talks." A semantic graph of the meaning of this item is then constructed. The concepts and instances are verified against known structured data sources for links of the "sameAs" type. Presidents are verified to be "sameAs" the concept of *nationalLeader*. The complete semantic graph is presented in Fig 2.

### 4.5 Transformation to Relational Storage

The data store was designed on the basis of the RDF data model with three top level tables (subjects, properties, objects). Although it is important to note that statements that would translate to RDF literal valued statements are transformed in the sense that literal values are stored in the table of objects (with added flag that indicates object type). This enables less redundancy in the data store and simplifies the querying of the data store. The basic tables (subjects, properties, objects) were also enhanced with additional tables that describe the source of the statements (in order to establish

source trustworthiness) and tables that store content classification tags that were added to the content by the publisher (for instance news items are prefixed by categories in the form of *#Politics*). Especially the content classification tags are used for the disambiguation of homonyms. If we consider the meaning of the word *AJAX*, it is clear it is a homonym. AJAX is a name of a cleaner (for washing dishes), AJAX is also shorthand for Asynchronous JavaScript and XML, AJAX is also a football club and so forth. Explicit content tags are of great use when disambiguation is required since they obviously state what domain the content is about.

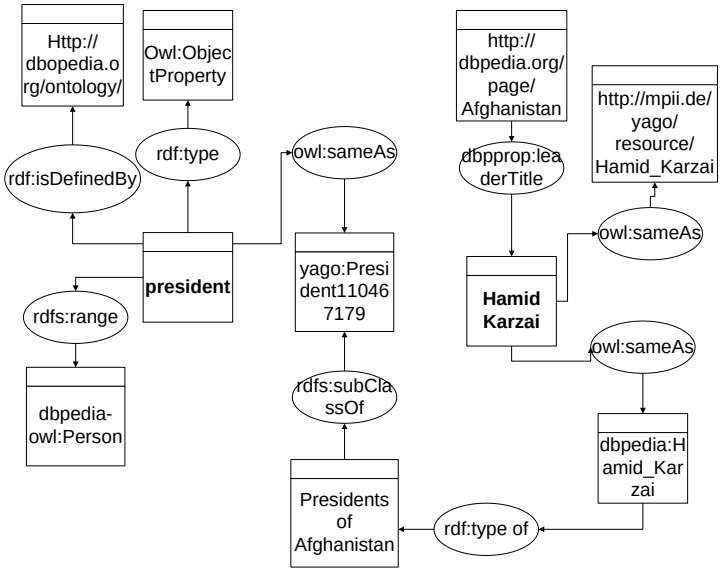


**Fig 2:** The semantic graph representing the concept of "president" linked through structured data sources to the individual in the example "Hamid Karzai". The diagram also represents the "sameAs" properties that link individual in our data store to the same individual in the LOD cloud. This enable cross-referencing and additional data discovery with link propagation.

## 5 Experimental Results

This section will focus on the results obtained from the automated processing of natural language content in the form of Slovene language news items. The news items were obtained from RSS sources in the timespan from March 2010 until September 2010. The news items numbered 55,018. Table 1 lists some statistics on the news items in regard to the size of the processed content.

The process of named entity recognition (without manual interventions) has extracted 4095 distinct entities which were distributed across various areas. The top 7 areas according to the classified number of entities are listed in table 2. The detailed classification that was done by the context in the news content and structured data sources has detected that the maximum level of ambiguity in the processed content (the number of meanings of an entity) to be three. Generally the news source

classification was used to resolve the ambiguity for a particular news item. The text content of the news items was analyzed to obtain the mapping of entities to concepts. Table 3 shows the final statistics of the experiment with regard to the number of concepts, individuals and properties that were obtained in the experiment.

The final numbers seem quite low when considering the amount of text that was processed. News titles alone contained over 370,000 words. When comparing the total number of words (14,236,506) with the total number of concepts, properties and individuals (24,701) we see that the relation between the natural language content and the structured content is 576.35 words for each structured element. We should stress that this relation is the consequence of the automated process with no human intervention. Therefore only structured elements that had the confidence level above 90% were extracted. The use of complex statements combined with a highly inflectional language (Slovene) obviously restricts the automated extraction of structured content. The system fails to classify content when the entity is being referenced through multiple sentences, only being named in the first sentence. Also as the source used was popular media many informal names (abbreviations, nicknames) are used in the text. As these are a part of the broader knowledge on the topic and are often not known (not found in any formal data source) they are ignored by the system.

**Table 1:** Statistics on the size of processed content

| **Property** | **Value** |
|---|---|
| Number of processed news items | 55,018 |
| Number of words | 14,236,506 |
| Number of all words in news titles | 373,062 |
| Maximum number of words in a news title | 30 |
| Average number of words in the news title | 6,78 |
| Average number of words in the news content | 237 |

**Table 2:** Distribution of top entity types

| **Area** | **Number of entities** |
|---|---|
| Persons | 1507 |
| Geographic location | 220 |
| Sports | 201 |
| Business | 160 |
| Government | 87 |
| Population | 55 |
| Entertainment industry | 21 |

**Table 3:** Final results of the experiment

| Property | Value |
|---|---|
| Distinct concepts | 1708 |
| Distinct entities | 4095 |
| Average number of properties per entity | 4,38 |
| Slovene entities | 94% |
| Foreign entities | 6% |

## 6 Conclusion

The main focus of the research was the construction of a formal repository of structured content. The content was required to be machine readable and would support querying, discovery of additional facts with link propagation and basic reasoning. The focus was on content in Slovene language because the many structured data sources that exist (and are freely available) are in foreign languages (foremost English). This research uses natural language content published online as its primary source of statements. Natural language content requires significant resources for the extraction of formal facts (statements). At the same time it is true that the majority of the processing requires little or none at all human interaction. The automation of the NL analysis and the experimental results give the research high value in the context of creating structured content.

Currently the acquirement of content is still an ongoing process. The experiment, conducted on a part of acquired content, has shown that very large quantities of content are required for the extraction of a significant amount of data. Several factors contribute to this: the use of highly complex language constructs that prove too difficult to analyze (referencing entities across multiple sentences, use of informal names etc.), the large domain area of the content (for instance news items cover domains from computer science, politics, economy, health, etc.) and the repetition of known facts.

Future work will focus on greater integration with structured sources in foreign languages, which will assist the discovery of entities (especially entities that are already described in LOD cloud) and research into the area of linking entities across different languages. The latter is especially important for the content in Slovene language as it would ensure cross linking with the LOD on a large scale. We will focus also on improving the grammatical capabilities of the system in order to improve the extraction rate when processing complex sentences.

## References


1. The DBpedia project, http://dbpedia.org/About.
2. Khare, R. and Celik, T.: Microformats: a pragmatic path to the semantic web. In Proceedings of the 15th international conference on World Wide Web (WWW '06). ACM, New York, USA, pp. 865-866, (2006).
3. Bizer, C., Heath, T. and Berners-Lee, T.: Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, Special Issue on Linked Data (2009).
4. Hausenblas, M., Karnstedt, M.: Understanding Linked Open Data as a Web-Scale Database. In: Proc. of the Second International Conference on Advances in Databases Knowledge and Data Applications (DBKDA), pp. 56-61(2010).
5. Erjavec, T., Fišer, D.: Building Slovene WordNet. In: Proc. of LREC 2006, Genoa, Italy (2006).
6. Fiser, D., Pollak S., Vintar S.: Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources. In: Proc. of Language Resources and Evaluation, (2010).
7. Erjavec, T., Fišer, D.: Building the Slovene Wordnet: first steps, first problems. In: Proc. of the Third International WordNet Conference, Jeju Island, Korea (2006).
8. Lacy, LW.: Owl: Representing Information Using the Web Ontology Language, Trafford Publishing, 2005, ISBN1-4120-3448-5.
9. Gruber TR.: A translation approach to portable ontology specifications. *Knowl. Acquis.*Vol. 5, Nr. 2, pp. 199-220, 1993.
10. Berners-Lee, T.: Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html [last accessed 03.04.2011]
11. Linking Open Data, an W3C SWEO community project, http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData [last accessed 03.04.2011]
12. Grefenstette, G., Tapanainen, P.: What is a word, what is a sentence? Problems of tokenization. In: Proc. of the 3rd International Conference on Computer Lexicography, pp. 79-87 (1994).
13. Pohorec S., Verlič M. and Zorman M.: Information extraction from concise passages of natural language sources. In: Proc. of the 14th east European conference on Advances in databases and information systems (ADBIS'10), Barbara Catania, Mirjana Ivanović and Bernhard Thalheim (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 463-474 (2010).
14. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the sixth conference on Applied Natural Language Processing, pp. 224-231 (2000).
15. Erjavec T., Fišer D., Krek S., Ledinek, N.: The JOS Linguistically Tagged Corpus of Slovene. In: Proc. of the Seventh International Conference on Language Resources and Evaluation (2010).
16. Erjavec T.: MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation (2004).