

Extraction of Semantic Information From web Resources

Jan Dědek

Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. The paper addresses a problem of extraction of semantic information from Czech texts from the Web. The method described in this paper exploits existing linguistic tools created originally for a syntactically annotated corpus, Prague Dependency Treebank (PDT 2.0). We are working on development of a system which captures text of web-pages, annotates it linguistically by linguistic tools, extracts data and interprets the extracted data semantically in terms of web ontologies. The proposed extraction method is based on extraction rules – tree queries, which are adopted from the Netgraph application. Semantic interpretation of these rules provides semantics of the extracted data. We present some initial experiments in the domain of reports of traffic accidents.

Introduction

The idea of the Semantic Web [Berners-Lee et al., 2001] is to create an universal medium for sharing of information not only among humans but also among software agents. The main step is to put machine-understandable information to the Web. First we need standard formalism for expressing semantics of information or data. This is well in progress in web ontological modeling (see W3C 2004). Further we need to create semantic data, i.e. data annotated by ontological concepts sufficient for machine processing. Semantic annotation of Web resource (mostly Web page) can be done either during the creation of the resource (by author or data generator) or after it (third party annotation). The third party annotation assumes we can extract (recognize) information from Web resources.

In this paper we describe initial experiments with information extraction from traffic accident reports of fire departments in several regions of the Czech Republic. We would like to demonstrate the prospects of using linguistic tools developed in the Institute of Formal and Applied Linguistics in Prague.

Motivation

The Ministry of Interior of the Czech Republic presents on its Web pages¹ also reports from fire departments of several regions of the Czech Republic. These departments are responsible for rescue and recovery after traffic accidents. These reports are rich in information, e.g. where and when an traffic accident occurred, which units helped, how much time it took them to show up on the place of accident, how many people were injured, killed etc. An example of such report can be seen in the Figure 1. We are trying to extract some of these information and provide it to possible software agents.

Semantic extraction

We propose a relatively straightforward process for the extraction of semantic data from text-based web-resources. This process consists of four steps. The Figure 2 describes it.

1. Extraction of text

The linguistic annotating tools process plain text only. In this phase we have to extract the text from the structure of a given web-resource. In this first phase we have used RSS

¹<http://www.mvcr.cz/rss/regionhzs.html>

feed of the fire department web-page. From this we have obtained URLs of particular articles and we have downloaded them. Finally we have extracted the desired text by means of a regular expression. The extracted text is highlighted in the Figure 1.

2. Linguistic annotation

In this phase the linguistic annotators process the extracted text and produce corresponding set of dependency trees representing the deep syntactic structure of individual sentences. The exploited linguistic tools will be described later.

3. Data extraction

We use the structure of tectogrammatical (i.e. deep syntactic) dependency trees to extract relevant data. We will describe the extraction method in more detail later.

4. Semantic representation

This phase consists of data transformation or conversion to the desired ontology format. It is quite important to choose suitable ontology that will properly represent semantics of the data. Semantic interpretation of data comes from semantic interpretation of extraction rules and it will be also demonstrated later.

Ministerstvo vnitra
home navigace vyhledávání změna vzhledu

Zpravodajství
Informace z resortu o tom, co se stalo, co se děje i co se připravuje

HZS Jihomoravského kraje

Zubatého 1, 614 00 Brno, telefon 950 630 111,
<http://www.firebrno.cz>
Zpravodajství v roce 2006

15.05.2007

V trabantu zemřeli dva lidé
K tragické nehodě dnes odpoledne hasiči vyjžděli na silnici z obce Česká do Kuřimi na Brněnsku.

Nehoda byla operačním středisku HZS ohlášena ve 13.13 hodin a na místě zasahovala jednotka profesionálních hasičů ze stanice v Trávnově. Jednalo se o čelní srážku autobusu Karosa s vozidlem Trabant 601. Podle dostupných informací trabant jedoucí ve z Brna do Kuřimi zřejmě vyjel do protisměru, kde narazil do linkového autobusu dopravní společnosti ze Žďáru nad Sázavou. Ve zdemolovaném trabantu na místě zemřeli dva muži – 82letý senior a další muž, jehož totožnost zjišťují policisté.

Hasiči udělali na vozidle protipožární opatření a po vyšetření a zadokumentování nehody dopravní policií vrak trabantu zaklesnutý pod autobusem pomocí lana odtrhli. Po odstranění střechy trabantu pak z kabiny vyprostili těla obou mužů. Obě vozidla – trabant i autobus, pak postupně odstranili na kraj vozovky a uvolnili tak jeden jízdní pruh. Únik provozních kapalin nebyl zjištěn. Po 16. hodině pomohli vrak trabantu naložit k odtahu a asistovali při odtahu autobusu. Po úklidu vozovky krátce před 16.30 hod. místo nehody předali policistům a ukončili zásah.

Odkazy

Hasiči

- Generální ředitelství
- hl. m. Praha
- Jihočeský kraj
- Jihomoravský kraj
- Karlovarský kraj
- Královéhradecký kraj
- Liberecký kraj
- Moravskoslezský kraj
- Olomoucký kraj
- Pardubický kraj
- Plzeňský kraj
- Středočeský kraj
- Ústecký kraj
- kraj Vysočina
- Zlínský kraj

Euroskop.cz

V této rubrice Zpravodajství

- Aktualizace stránek
- Archiv zpravodajství
- Bleskové zpravodajství
- RSS
- Boj proti korupci
- Digitální televize
- Hasiči
- Hlavní zprávy
- Ministerstvo
- Od dopisovatelů (neoficiální)
- Policie
- Regiony
- Servis nejen pro novináře
- Schengenská spolupráce
- WebEditorial

Na našem serveru v jiných rubrikách

- Aktuality Národního archivu

Figure 1. Example of the web-page with a report of a fire department

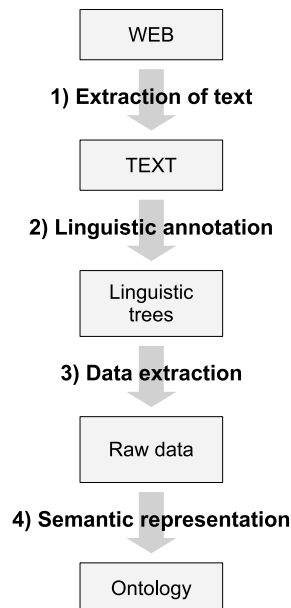


Figure 2. Schema of the extraction process

Linguistic tools for automatic annotation of texts

In this section we will describe the linguistic tools that we have used to produce linguistic annotation of texts. These tools are being developed in the Institute of Formal and Applied Linguistics² in Prague (ÚFAL), Czech Republic. They are publicly available – they have been published on a CD-ROM under the title PDT 2.0 [Hajič et al., 2006] (first five tools) and in [Klimeš, 2006] (Tectogrammatical analysis). These tools are used as a processing chain and at the end of the chain they produce tectogrammatical [Mikulová et al., 2006] dependency trees. The table in Figure 3 shows some details about these tools. Short descriptions of particular tools follows.

1. **Segmentation and tokenization** consists of tokenization (dividing the input text into words and punctuation) and segmentation (dividing a sequences of tokens into sentences).
2. **Morphological analysis** [Hajič, 2000] assigns all possible lemmas and morphological tags to particular word forms (word occurrences) in the text.
3. **Morphological tagging** [Hajič, 2000] consists in selecting a single pair lemma-tag from all possible alternatives assigned by the morphological analyzer.
4. **Collins' parser – Czech adaptation** [Collins et al., 1999] Unlike the usual approaches to the description of English syntax, the Czech syntactic descriptions are dependency-based, which means that every edge of a syntactic tree captures the relation of dependency between a governor and its dependent node. Collins' parser gives the most probable parse of a given input sentence.
5. **Analytical function assignment** [Sgall et al., 2002] assigns a description (*analytical function* – in linguistic sense) to every edge in the syntactic (dependency) tree.
6. **Tectogrammatical analysis** [Klimeš, 2006] produces linguistic annotation at the tectogrammatical level.

²<http://ufal.mff.cuni.cz>

Name of the tool	Evaluation results (proclaimed by authors)
Segmentation and tokenization	precision(p): 98,0%, recall(r): 91,4%
Morphological analysis	2,5% unrecognized words
Morphological tagging	93,0% of tags assigned correctly
Collins' parser – Czech adaptation	81,6% dependencies assigned correctly
Analytical function assignment	precision: 92%
Tectogrammatical analysis	dependencies p: 90,2%, r: 87,9%
[Klimeš, 2006]	assignment of f-tags p: 86,5%, r: 84,3%

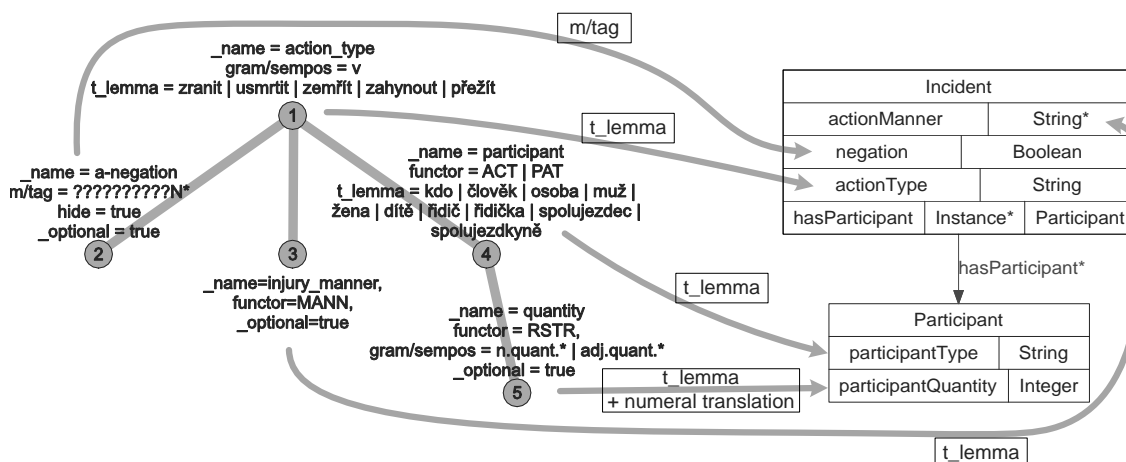
Figure 3. Linguistic tools for machine annotation

The data extraction and semantic interpretation

The extraction method we have used is based on extraction rules. These rules correspond to query requests of Netgraph application. The Netgraph application [Mírovský, 2006] is a linguistic tool used for searching through a syntactically annotated corpus of a natural language. It was originally developed for searching the analytical and tectogrammatical levels of the Prague Dependency Treebank, a richly syntactically annotated corpus of Czech [Hajič et al., 2006].

Netgraph queries are written in a special query language. An example of such Netgraph query can be found in the Figure 4 – left side. The Netgraph is a general tool for searching trees, it is not limited only to the trees in the PDT format. In our application we use it for searching the tectogrammatical trees provided by a set of language processing tools described in the previous chapter. The tectogrammatical trees have a very convenient property of containing just the type of information we need for our purpose, namely the information about inner participants of verbs - actor, patient, addressee etc.

The tectogrammatical (deep syntactic) level of representation is more suitable for our purpose than the analytical (surface syntactic) level of representation of the structure of each sentence. The inner participants (actor, patient, addressee etc.) provide much more reliable information about the actual meaning of the sentence than the syntactic roles (subject, object etc.). The roles of a subject or object are misleading especially in the case of passive sentences, where usually the subject of the sentence corresponds to the patient or addressee while the actor is expressed by an object of the passive sentence.

**Figure 4.** Netgraph query – extract rule and its semantic interpretation.

```

<QueryMatches>
  <Match root_id="T-vysocina63466.txt-001-pls4" match_string="2:0,7:3,8:4,11:2">
    <Sentence>
      Při požáru byla jedna osoba lehce zraněna - jednalo se
      o majitele domu, který si vykloubil rameno.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="injury_manner" attribute_name="t_lemma">lehký</Value>
      <Value variable_name="participant" attribute_name="t_lemma">osoba</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">jeden</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49640.txt-001-pls4" match_string="1:0,13:3,14:4">
    <Sentence>
      Ve zdemolovaném trabantu na místě zemřeli dva muži - 82letý senior
      a další muž, jehož totožnost zjišťují policisté.
    </Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zemřít</Value>
      <Value variable_name="participant" attribute_name="t_lemma">muž</Value>
      <Value variable_name="quantity" attribute_name="t_lemma">dva</Value>
    </Data>
  </Match>
  <Match root_id="T-jihomoravsky49736.txt-001-p4s3" match_string="1:0,3:3,7:1">
    <Sentence>Čtyřiatřicetiletý řidič nebyl zraněn.</Sentence>
    <Data>
      <Value variable_name="action_type" attribute_name="t_lemma">zranit</Value>
      <Value variable_name="a-negation" attribute_name="m/tag">VpYS---XR-NA---
    </Value>
      <Value variable_name="participant" attribute_name="t_lemma">řidič</Value>
    </Data>
  </Match>
</QueryMatches>

```




Figure 5. Example of the result of the extraction procedure.

Extraction

The extraction works as follows: the extraction rule is in the first step evaluated by searching through a set of syntactic trees. Matching trees are returned and the desired information is taken from particular tree nodes.

We have evaluated the extraction rule shown in the Figure 4 by using a set of 800 texts of news of several Czech fire departments. There were about 470 sentences matching the rule and we found about 200 numeric values contained in the node number 5. Small part of the result of the extraction is shown in the Figure 5. This result contains three pieces of information extracted from three articles. This extracted data should be interpreted in terms of some ontology. This could be done for example by XSLT transformation but we have not implemented this step yet. The extraction rule (Figure 4) is a result of a learning procedure of a human designer. We are going to support and automatize the procedure of learning extraction rules in our future work.

Semantic Interpretation

The Figure 4 shows connection between the extraction rule on the left and an ontology instance on the right. Each piece of information taken by the extraction rule can be interpreted as instance of given ontology. Problematic part of the interpretation is the translation of the linguistic information to the semantic one. This time our proposal of the extraction system counts with several supported translation like translation of numerals to numbers, translation of lexical content (*t_lemma*) of a query node or detection of negation present in a query node.

Gathering similar words

The Figure 4 shows that it would be useful to gather words with similar meanings in our extraction rules. For example, the rule in the Figure 4 contains long disjunctions of similar words (nodes with numbers 1 and 4). These disjunctions could be replaced with some kind of expression telling that we are looking for any word from some semantic category (e.g. human beings). For this purpose we wanted to use the Czech WordNet [Pala and Smrž, 2004].

After we have explored the records of the Czech WordNet (CzWN) related to the domain of our interest (car accidents, etc.) we have decided not to involve CzWN in the extraction process. The reason is that the coverage of the vocabulary of our domain is rather poor and the semantic connections of words are sometimes unfortunately missing. But we can supply the missing information to CzWN or we can build up a new domain-specific word-net based on the ground of CzWN.

Conclusion

We have presented a proposal of a system for semantic extraction of information from Czech text on Web pages. Our system relies on linguistic annotating tools from ÚFAL and the tree querying tool Netgraph. Our contributions are in fact initial experiments in text extraction from downloaded pages, formulation of a rule-based extraction method and demonstration of semantics of the extracted data.

More details can be found in Dědek 2007. In the future we would like to extend this method by domain oriented lexical net and semiautomatic search for interesting extraction rules.

Acknowledgments. This work was partially supported by the Ministry of Education of the Czech Republic (grant MSM0021620838).

References

- Berners-Lee, T., Hendler, J., and Lassila, O., The semantic web, a new form of web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, 284, 34–43, 2001.
- Collins, M., Hajič, J., Brill, E., Ramshaw, L., and Tillmann, C., A Statistical Parser of Czech, in *Proceedings of 37th ACL Conference*, pp. 505–512, University of Maryland, College Park, USA, 1999.
- Dědek, J., *Semantic annotation of data from web resources (in Czech)*, Master’s thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech rep., 2007.
- Hajič, J., Morphological Tagging: Data vs. Dictionaries, in *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, pp. 94–101, Seattle, Washington, 2000.
- Hajič, J., Hajičová, E., Hlaváčová, J., Klimeš, V., Mírovský, J., Pajas, P., Štěpánek, J., Vidová-Hladká, B., and Žabokrtský, Z., Prague dependency treebank 2.0 cd-rom, Linguistic Data Consortium LDC2006T01, Philadelphia 2006, 2006.
- Klimeš, V., Transformation-based tectogrammatical analysis of czech, in *Proceedings of the 9th International Conference, TSD 2006*, no. 4188 in Lecture Notes In Computer Science, pp. 135–142, Springer-Verlag Berlin Heidelberg, 2006.
- Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Uřešová, Z., Veselá, K., and Žabokrtský, Z., Annotation on the tectogrammatical level in the prague dependency treebank. annotation manual, Tech. Rep. 30, ÚFAL MFF UK, Prague, Czech Rep., 2006.
- Mírovský, J., Netgraph: A tool for searching in prague dependency treebank 2.0, in *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT)*, edited by J. Hajič and J. Nivre, 5, pp. 211–222, Prague, Czech rep., 2006.
- Pala, K. and Smrž, P., Building czech wordnet, *Romanian Journal of Information Science and Technology*, 2004, 79–88, URL http://www.fit.vutbr.cz/research/view_pub.php?id=7682, 2004.
- Sgall, P., Žabokrtský, Z., and Džeroski, S., A Machine Learning Approach to Automatic Functor Assignment in the Prague Dependency Treebank, in *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, edited by R. M. Rodríguez and C. P. S. Araujo, vol. 5, pp. 1513–1520, European Language Resources Association, 2002.
- W3C, Owl web ontology language guide, URL <http://www.w3.org/TR/owl-guide/>, 2004.