Extrakce informací z textově orientovaných zdrojů webu

Jan Dědek, Peter Vojtáš

Katedra softwarového inženýrství, Matematicko-fyzikální fakulta,
Univerzita Karlova v Praze
Malostranské nám. 25, 118 00, Praha 1
jan.dedek@mff.cuni.cz

Abstrakt. V tomto příspěvku se zbýváme extrakcí informací z webových zdrojů převážně textového charakteru. K tomuto účelu jsme se pokusili využít několik lingvistických nástrojů pro zpracování přirozeného textu v češtině. Jmenovitě se jedná o nástroje pražského projektu PDT a český WordNet. Cílem příspěvku je přiblížit možnosti, které tyto nástroje pro extrakci informací z textu poskytují. Extrakcí informací se zde zabýváme především v kontextu sémantického webu a zkoumáme možnosti, jak tyto nástroje využít pro automatizaci sémantické anotace stránek současného webu.

Klíčová slova: Extrakce informací, lingvistika, sémantická anotace

1 Úvod

Extrakci informací z volného přirozeného textu lze využít v mnoha aplikacích. Může se jednat o extrakci kontaktních informací, časových údajů, případně informací se složitější strukturou i sémantikou.

V tomto příspěvku se zaměříme na extrakci informací z článků o akcích hasičských sborů v různých regionech ČR. Tato data zde však budou hrát spíše ilustrativní úlohu, zatím se nesnažíme data precizně zpracovat a vytěžit z nich maximum informací. Chceme pouze ukázat možnosti, které k tomuto účelu poskytují lingvistické nastroje popsané v sekci 4.

1.1 Motivace

MVČR poskytuje na svých stránkách[1] aktuální zpravodajství HZS z různých regionů ČR. Jedná se o informačně poměrně bohaté zprávy, ve kterých se například dočtete, kdy a kde se stala která dopravní nehoda, jaké hasičské sbory u akce zasáhly, za jak dlouhou dobu na místo dorazili, kolik lidí bylo při nehodě zraněno případně usmrceno atd.

[1] http://www.mvcr.cz/rss/regionhzs.html

Pokud nás ale zajímají pouze články, které se zabývají dopravními nehodami (a nikoli například hasičskými soutěžemi), případně bychom chtěli z článků spočíst nějakou statistiku, pokusit se nalézt v událostech nějaké vztahy, pak musíme potřebné informace z textu extrahovat.

Představme si ještě situaci, kdy jsou v článcích všechny relevantní informace vyznačeny – sémanticky anotovány, například pomocí RDFa [6]. V takovém případě by tyto informace mohl využít kterýkoliv softwarový agent.

2 Sémantická anotace

Na obrázku 1 je znázorněn proces extrakce sémantických informací z textově orientovaného webového zdroje. Data, která vzejdou z tohoto procesu můžeme přímo použít k sémantické anotaci zdroje.

Obr. 1. Proces sémantické extrakce.

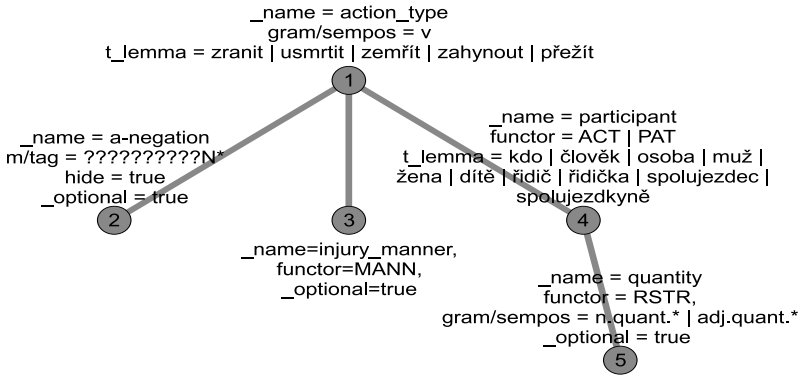
1. **Příprava vstupních dat**
   Lingvistické anotátory zpracovávají prostý text, který v této fázi musíme z webové stránky získat.
2. **Lingvistická anotace**
   Extrahovaný text předložíme lingvistickému anotátoru, který v textu rozpozná jednotlivé věty a zkonstruuje z nich lingvistické stromy.
3. **Extrakce dat**
   Pomocí lingvistické struktury jednotlivých vět extrahujeme data, která reprezentují informace vyjádřené v textu. Podrobnosti – viz sekce 3.
4. **Formální reprezentace dat**
   V této fázi sémanticky interpretujeme extrahovaná data pomocí konceptů vhodné ontologie.

3 Extrakce informací nad lingvistickými stromy

V naší práci jsme vyzkoušeli jednoduchou extrakční metodou založenou na deterministických pravidlech pro extrakci. Tato metoda je inspirovaná aplikací Netgraph [4], která umožňuje prohledávání lingvistických korpusů pomocí vlastního dotazovacího jazyka. Na obrázku 2 je vidět jeden takový Netgraph dotaz.

Při vlastní extrakci informací vezmeme všechny věty jednotlivých článků a vyhodnotíme na nich zvolený dotaz (například ten z Obr. 2). Z lingvistických anotací vyhovujících vět získáme data, která nás zajímají. Zmíněný příklad dotazu nám umožní zjistit: Kolik (uzel č. 5) jakých (4) osob bylo či nebylo (2) zraněno či usmrceno (1) při události, ke které se daná věta vztahuje.

Extrakční pravidlo z obrázku 2 jsme vyhodnotili nad osmi sty články hasičského zpravodajství zmíněného výše. Bylo nalezeno 470 vět vyhovujících pravidlu a 200 číselných hodnot v uzlu č. 5.



Obr. 2. Netgraph dotaz – extrakční pravidlo.

4 Lingvistické nástroje

Vyzkoušeli jsme několik lingvistických nástrojů, které pocházejí z Ústavu formální a aplikované lingvistiky v Praze[2] a sémantický lexikon český WordNet vyvíjený na Fakultě informatiky Masarykovy univerzity v Brně[3].

4.1 Tools for machine annotation – PDT 2.0 [2]

Jedná se o skupinu nástrojů, které provádějí plně automatickou lingvistickou analýzu českého textu. Ze surových českých vět vytvářejí lingvistické závislostní

[2] http://ufal.mff.cuni.cz
[3] http://www.fi.muni.cz

stromy. Proces anotace zahrnuje následující kroky (v závorce jsou orientační autory udávané přesnosti jednotlivých kroků, pro podrobnosti viz [2] resp. [3]).

1. Segmentation and tokenization (98%)
2. Morphological analysis (97,5%)
3. Morphological tagging (93%)
4. Parsing (81,6%)
5. Analytical function assignment (92%)
6. Tectogrammatical analysis [3] (86,5%)

4.2 Český WordNet

Pokud se podíváme na obrázek 2, napadne nás, že dlouhé disjunkce podobných slov v uzlech 1 a 4 by bylo vhodné zobecnit pomocí lexikální sítě. K tomuto účelu jsme chtěli použít český WordNet [5]. Zběžným prohledáním databáze WordNetu jsme však zjistili, že jeho přímé nasazení by naší metodě nepomohlo. Slova, která bychom v našem dotazu chtěli pomocí WordNetu vyhodnotit jako příbuzná, většinou nejsou v jeho databázi blízce propojena.

5 Závěr

Podrobnosti o této metodě je možné získat v [1]. Do budoucna bychom chtěli tuto metodu posílit o možnost použití doménové lexikální sítě a vyvinout metodu pro poloautomatické hledání zajímavých extrakčních pravidel.

Poděkování. Tato práce byla finančně podpořena projekty 1ET100300517 a 1ET100300419 AVČR.

Reference

1. Dědek J. *Sémantická anotace dat z webovských zdrojů.* Diplomová práce, KSI, Matematicko-fyzikální fakulta, Univerzita Karlova, Praha, 2007.
2. Hajič J., Hajičová E., Hlaváčová J., Klimeš V., Mírovský J., Pajas P., Štěpánek J., Vidová Hladká B., Žabokrtský Z. Prague Dependency Treebank 2.0 CDROM. *Linguistic Data Consortium.* In press, 2006.
3. Klimeš V. Transformation-Based Tectogrammatical Analysis of Czech. *Proceedings of Text, Speech and Dialogue 2006* Berlin Heidelberg 2006.
4. Mírovský J. Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0. *Proceedings of The Fifth International Treebanks and Linguistic Theories conference*, Praha 2006.
5. Pala K., Ševeček P. *The Czech WordNet, final report.* Technical report, Masarykova univerzita, Brno, 1999.
6. W3C. http://www.w3.org/TR/xhtml-rdfa-primer/, RDFa Primer.

Annotation:

Information extraction from text-based web resources

The authors present a linguistic-based method for extraction of information from text-based web resources. The paper deals with several linguistic tools for Czech, namely Tools for machine annotation – PDT 2.0 and The Czech WordNet.