EXTRAKCE INFORMACÍ Z TEXTOVĚ ORIENTOVANÝCH ZDROJŮ WEBU

Jan Dědek, Peter Vojtáš

Katedra softwarového inženýrství, Matematicko-fyzikální fakulta, Univerzita Karlova v Praze

1 Abstrakt

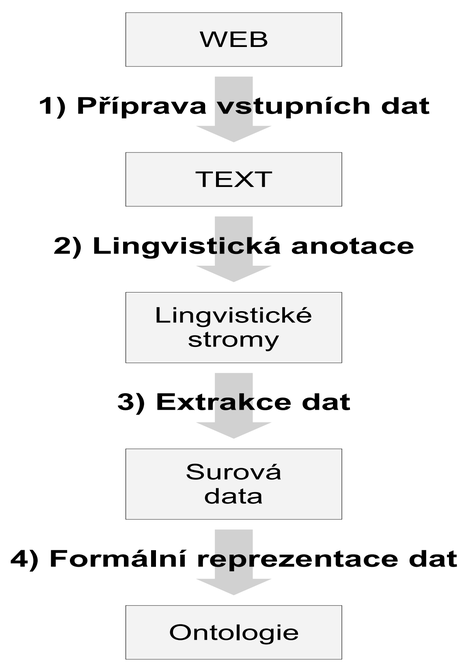
V tomto příspěvku se zabýváme extrakcí informací z webových zdrojů převážně textového charakteru. K tomuto účelu jsme se pokusili využít několik lingvistických nástrojů pro zpracování přirozeného textu v češtině. Jmenovitě se jedná o nástroje pražského projektu PDT a český WordNet. Cílem příspěvku je 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.

2 Sémantická datová extrakce

Pokusili jsme se extrahovat některé informace z textů aktuálního zpravodajství HZS z různých regionů ČR, které MVČR poskytuje na svých stránkách
`http://www.mvcr.cz/rss/regionhzs.html`.

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 následující sekce.
4. **Formální reprezentace dat**
   V této fázi sémanticky interpretujeme extrahovaná data pomocí konceptů vhodné ontologie.

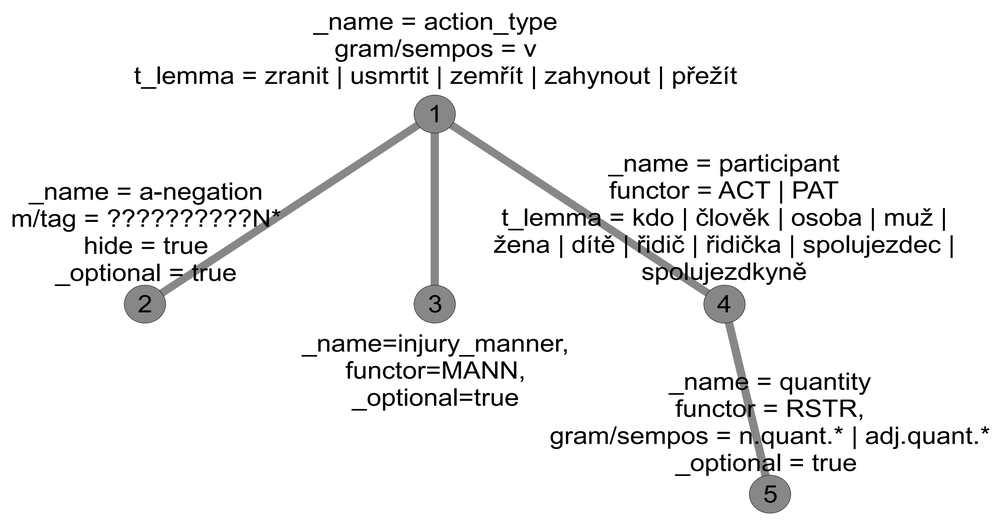
Proces sémantické extrakce:



3 Extrakce informací nad lingvistickými stromy

Vyzkoušeli jsme jednoduchou extrakční metodou založenou na deterministických pravidlech pro extrakci. Tato pravidla odpovídají dotazům aplikace Netgraph [4].

Extrakční pravidlo – dotaz aplikace Netgraph:



Ukázka z výstupu extrakce:

```
<injured_result>
      <action type="zranit">
            <sentece>
                  Při požáru byla jedna osoba lehce zraněna -- jednalo se
                  o majitele domu, který si vykloubil rameno.
            </sentece>
            <sentece_id>T-vysocina63466.txt-001-p1s4</sentece_id>
            <negation>false</negation>
            <manner>lehký</manner>
            <participant type="osoba">
                  <quantity>1</quantity>
                  <full_string>jedna osoba</full_string>
            </participant>
      </action>
      <action type="zemřít">
            <sentece>
                  Ve zdemolovaném trabantu na místě zemřeli dva muži -- 82letý
                  senior a další muž, jehož totožnost zjišťují policisté.
            </sentece>
            <sentece_id>T-jihomoravsky49640.txt-001-p1s4</sentece_id>
            <negation>false</negation>
            <participant type="muž">
                  <quantity>2</quantity>
                  <full_string>dva muži</full_string>
            </participant>
      </action>
      <action type="zranit">
            <sentece>Čtyřiatřicetiletý řidič nebyl zraněn.</sentece>
            <sentece_id>T-jihomoravsky49736.txt-001-p4s3</sentece_id>
            <negation>true</negation>
            <participant type="řidič">
                  <full_string>Čtyřiatřicetiletý řidič</full_string>
            </participant>
      </action>
</injured_result>
```

4 Tools for machine annotation – PDT 2.0 [2]

Tools for machine annotation jsme použili pro lingvistickou anotaci textů. Jedná se o skupinu nástrojů, které provádějí plně automatickou lingvistickou analýzu českého textu. Ze surových českých vět vytvářejí lingvistické závislostní stromy.

Název nástroje	Autory udávaná úspěšnost
Segmentation and tokenization	precision: 98,0%, recall: 91,4%
Morphological analysis	2,5% nerozpoznaných slov
Morphological tagging	93,0% značek správně
Parsing (Collins)	81,6% závislostí správně
Analytical function assignment	precision: 92%
Tectogrammatical analysis [3]	precision, recall závislostí: 90,2%, 87,9%
	precision, recall f-značek: 86,5%, 84,3%

5 Český WordNet – doména dopravních prostředků

Pomocí českého WordNet-u [5] jsme chtěli zobecnit extrakční pravidla. Tento krok jsme však zatím nerealizovali. Zdá se, že přímé nasazení WordNet-u by nebylo příliš přínosné, protože WordNet nepokrývá doménu dopravních prostředků přesvědčivě (viz vzdálenost slov *autobus* a *kamion* ve stromu lexikální dědičnosti.) WordNet však může posloužit jako dobrý základ pro vytvoření vhodné doménové lexikální sítě.

Ukázky ze stromu lexikální dědičnosti:

- entita:1
  - objekt:1
    - celek:1
      - artefakt:1, výtvor:2, výrobek:2
        - vybavení:2
          - přepravní prostředek:1, transportní prostředek:1
            - ▷ **veřejná doprava:1**
              ⇒ *autobus:1, autokar:1*
            - ▷ **dopravní prostředek:1**
              - kolové vozidlo:1
                - samohybné vozidlo:1, vozidlo s vlastním pohonem:1
                  - motorové vozidlo:1
                    - nákladní automobil:1
                      ⇒ *kamion:1*

- motorové vozidlo:1
  - motocykl:1
  - nákladní automobil:1
  - obojživelné vozidlo:1
  - auto:1, vůz:2
  - pohřební vůz:1
  - sněžný pluh:1, pluh:2
  - golfový vozík:1

- nákladní automobil:1
  - dodávka:3
  - sklápěč:1, vyklápěcí nákladní automobil:1
  - tahač:1
  - pick-up:1, malý nákladní automobil:1
  - hasící vůz:1, požární stříkačka:1
  - rozhlasový vůz:1
  - kamion:1
  - nákladní automobil s přívěsem:1
  - popelářský vůz:1, popelářské auto:1, bobr:3

- auto:1, vůz:2
  - limuzína:1
  - elektrický vozík:1
  - závodní vůz:1, závodní automobil:1
  - sportovní vůz:1
  - kabriolet:1, sporťák:1
  - vrak:3
  - limuzína:2
  - hlídkový vůz:1, policejní vůz:1
  - sériový automobil:1
  - cestovní vůz:1
  - džíp:1
  - kupé:1
  - kabriolet:3
  - kombi:1
  - taxi:1
  - ambulance:1, sanitka:1, pohotovost:6, záchranka:1, sanita:1

6 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 LDC2006T01*, LDC, Philadelphia 2006.

[3] Klimeš V. Transformation-Based Tectogrammatical Analysis of Czech. *Proceedings of Text, Speech and Dialogue 2006* Lecture Notes in Computer Science Volume 4188, Springer, Berlin Heidelberg 2006. 135–142.

[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. 211–222.

[5] Pala K., Ševeček P. *The Czech WordNet, final report.* Technical report, Masarykova univerzita, Brno, 1999.