Sémantický web a jeho technologie (3)

Pavel Smrž, Tomáš Pitner, FI MU

Závěrečná část volného seriálu o sémantickém webu se vrací „k tomu podstatnému“. Prezentuje pojmy, technologie a prostředky možná ne tak bezprostředně využitelné jako bylo RSS, avšak perspektivně velmi významné: ontologie, wordnety a nástroje jejich zpracování. Informace zde uvedené jsou velmi stručným přiblížením rozsáhlých několikaletých zkušeností *Laboratoře zpracování přirozeného jazyka* (NLP) na FI v této oblasti.

1 Co jsou ontologie

Vize sémantického webu předpokládá organizaci a provázanost informací umožňující jejich zpracování a „pochopení“ nejen člověkem, ale i počítači. Prostředkem k dosažení tohoto (dlouhodobého) cíle mají být mj. ontologie.

Ontologie v tradičním filosofickém pojetí označuje *nauku o bytí* (jsoucnu). Vzhledem k tomu, že dnešní stroje nemají dostatek inteligence, aby mohly být rovnocenným partnerem člověka, je jejich „jsoucno“ modelováno obvykle pomocí sady pojmů, které zadává člověk. V oblasti informatiky jsou dnes ontologie chápány jako *explicitní, formální specifikace pojmů a vztahů mezi nimi*.

Cílem ontologie je *definovat společné, jednotné chápání určité třídy pojmů*. Je zřejmé, že „jednotnosti“ se snáze dosáhne v určité uzavřené oblasti – doméně. Proto dnes existuje poměrně velké množství tzv. *doménových ontologií*. Aktuální otázkou výzkumu je, zda pro úspěšnou realizaci informačních systémů budoucnosti postačí rozšiřování počtu těchto „kamínků mozaiky“ sémantického webu, či zda je nutné zapojit i nějakou obecnou, široce pojatou, „všeobjímající“ ontologii.

Problém je samozřejmě zejména s přívlastkem „všeobjímající“. Vzhledem k tomu, že názory lidí na různé skutečnosti se často rozcházejí, je těžké si představit, jak by obecná, všeobecně přijímaná ontologie mohla vzniknout. I kdybychom se však spokojili s tím, že široce zaměřenou ontologii pro potřeby sémantického webu zpracuje jen nějaká skupina lidí podle svého nejlepšího vědomí a svědomí, zůstává stále mnoho nevyřešených těžkostí. Nejvýznamnější je patrně časová náročnost tvorby takového zdroje.

K překonání tohoto slabého místa může vést tradiční cesta – pokusit se použít to, co je již hotovo. Cílem takové přístupu k budování ontologií je integrace existujících lexikálních databází (viz dále), rozsáhlých bází znalostí budovaných pro jiné účely (např. *Cyc*[1]) a dalších zdrojů, jejich „vyčištění“, zpřesnění uložených informací a nakonec integrace do jednotné ontologické struktury.

2 Wordnet

Nejpopulárnějším výchozím zdrojem pro budování ontologií je *wordnet*[2]. Jak název napovídá, jedná se o síť slov, spojených sémantickými vztahy. Základními stavebními kameny wordnetu jsou synonymické řady – *synsety* – tvořené slovy, která mají v určitém kontextu totožný význam. Jednotlivé prvky synsetů se nazývají *literály*. Wordnet dále obsahuje celou řadu sémantických vazeb mezi literály a zejména mezi synsety. Nejvýznamnějším typem jsou hyper-/hyponymické vztahy spojující synsety s obecnějším/konkrétnějším významem, např. *flanel* je druhem *textilie*. Důležité jsou dále vztahy meronymie, zachycující vztahy mezi částí a celkem, např. *nos* je částí *obličeje*.

V předchozím odstavci jsme mluvili zjednodušeně o slovech. Synsety však obsahují i *slovní spojení* (sousloví), např. „vysoká škola“. Obecně tedy mluvíme o *jazykových výrazech*. Literály dále obsahují *číselný identifikátor významu*, např. anglické podstatné jméno *bank:1* označuje finanční instituci, *bank:2* – břeh.

První a současně největší (americký) wordnet vznikl a je dále vyvíjen a rozšiřován na Princentonské univerzitě pod vedením George Millera a Christinne Felbaum [1]. V aktuální verzi 2.0 obsahuje více než 100 000 synsetů složených přibližně z dvojnásobku literálů. V roce 1996

[1] `http://www.cyc.com`

[2] `http://www.cogsci.princeton.edu/wn`

odstartoval evropský projekt *EuroWordNet* s cílem vytvořit wordnety pro další jazyky (holandštinu, španělštinu, italštinu, francouzštinu, němčinu, estonštinu a češtinu) a provázat všech je do multilingválního databáze.

3 Český wordnet

Budováním české části byl pověřen tým Laboratoře zpracování přirozeného jazyka na Fakultě informatiky MU, vedený doc. Palou. Vzhledem k velmi dobrým výsledkům nám byla dále nabídnuta spolupráce na navazujícím projektu BalkaNet, jehož cílem je vybudovat srovnatelné jazykové zdroje pro dalších pět balkánských jazyků (bulharštinu, rumunštinu, řečtinu, srbštinu a turečtinu) a dále rozšiřovat český wordnet.

Aktuální velikost českého wordnetu je přibližně 30 000 synsetů [2]. Avšak hlavním cílem, který si v projektu BalkaNet český tým vytkl, je vytvoření kvalitního editoru wordnetových databází. Autoři originálního wordnetu sice nabízejí volně dostupný prohlížeč wordnetu, avšak veškerou editaci provádějí v jednoduchých textových souborech. Ani wordnetový editor vyvíjený v rámci projektu EuroWordNet firmou Lernout & Hauspie nesplňoval požadavky nového projektu. Byl poměrně drahý pro mnoho zájemců o tvorbu wordnetů, byl omezen na platformu MS Windows a navíc přestal být vyvíjen s koncem projektu a krachem firmy L & H. Proto byla na FI vytvořena zcela nová aplikace, která dostala název *VisDic* [3].

4 Editor VisDic

Jako základní formát, s nímž VisDic pracuje, byl zvolen jazyk XML. Bylo navrženo vhodné schéma pro uložení informací z wordnetových databází a veškerá data předchozích projektů byla převedena tak, aby mohla být zobrazována a editována v novém nástroji. VisDic umožňuje definovat různé pohledy na data. Nejpoužívanějším je zobrazení obsahu synsetu a všech jeho vazeb a dále hypero-hyponymický strom (viz obr.1).

5 Kritika wordnetu

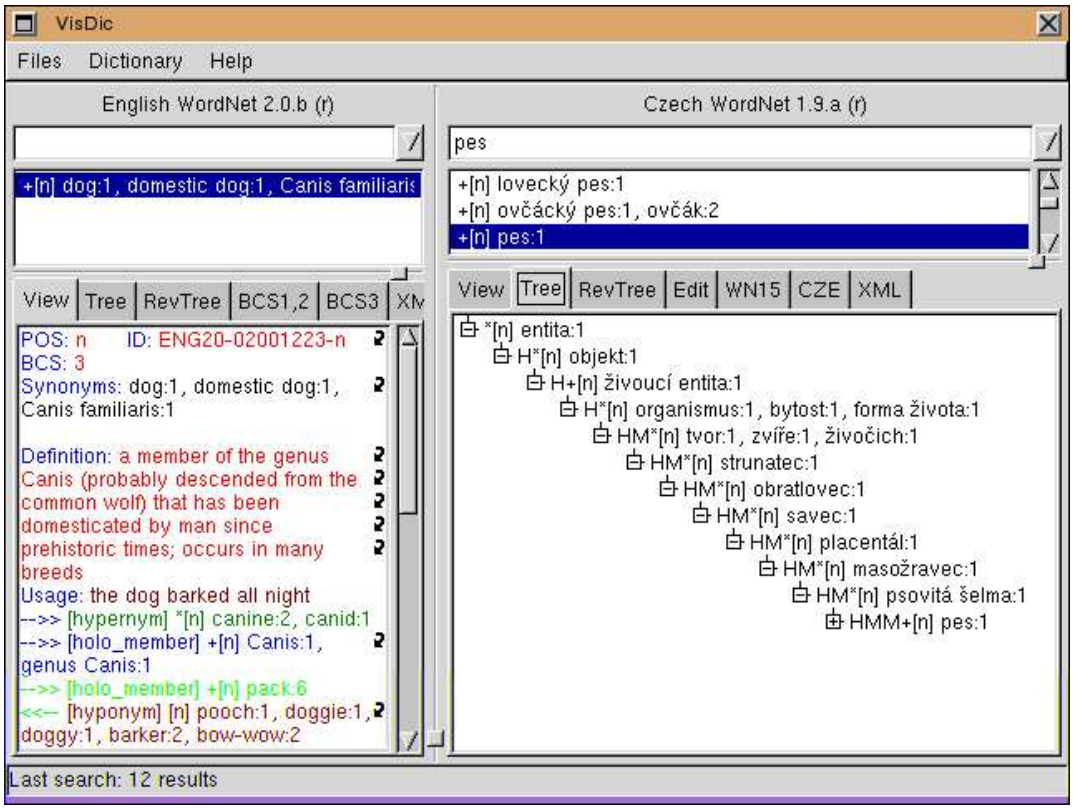
Wordnet je často kritizován z mnoha různých pozic. Zejména tradiční lexikografové vidí mnoho problémů v nejasně definované (a v čase se měnící) koncepci tvorby hesel (synsetů). Nicméně wordnet a celá řada jeho „klonů“ získal obrovskou popularitu. Přestože byl původně zamýšlen jako model mentálního slovníku člověka, největší uplatnění dnes nachází jako významný lexikografický zdroj. Až neuvěřitelná popularita wordnetu je dána vedle jeho velkého rozsahu i faktem, že od samého počátku byl poskytován princentonským týmem volně. Stejnou filosofii jsme přijali pro poskytování editoru VisDic, který je možné získat na stránkách `http://nlp.fi.muni.cz/projects/visdic`. Zdá se, že je tento přístup opět úspěšný, neboť VisDic je dnes vedle všech členů konsorcia BalkaNet využíván kolegy z Německa, Slovenska, Mad'arska, Ruska a dalších zemí.

6 Wordnet a ontologie

Pokud přijmeme jisté zjednodušení, můžeme literály v rámci jednoho synsetu považovat za výrazy označující jeden pojem. To je důležité pro spojení wordnetu s ontologiemi. Wordnet je dnes nejen výchozím zdrojem pro budování nových ontologií, ale také pro „zakotvení“ existujících i nově vznikajících ontologií k výrazům přirozeného jazyka (angličtiny a, pomocí mezijazykových vazeb, dalších jazyků). Příkladem může být nově vytvořená ontologie *SUMO* (*Suggested Upper Merged Ontology*) vytvořená IEEE. Na stránkách věnovaných této iniciativě (`http://ontology.teknowledge.com`) je možné získat kompletní mapování z wordnetu na SUMO a zpět. Ontologie SUMO je zajímavá také tím, že na výše zmíněných stránkách je k dispozici v různých formátech, které se dnes pro ukládání ontologií používají – KIF, OWL, LOOM a Protege.

7 Závěrem

Volný seriál tří článků k sémantickému webu a jeho technologiím končí. První díl se věnoval především základním pojmům a technologiím, jako jsou metadata a prostředky jejich zpracování. Ve středu pozornosti druhé části následně byly zcela praktické, okamžitě použitelné standardy a nástroje zpřístupňující jednoduchá metadata o webových zdrojích (RSS). Konečně tato, poslední,



Obrázek 1: Wordnet v editoru VisDic

část se znovu věnuje ontologiím a souvisejícím praktickým prostředkům (wordnet).

Literatura

[1] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. Five papers on wordnet. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.

[2] Karel Pala and Pavel Smrz. Building Czech wordnet. *Romanian Journal of Information Science and Technology*, 2004. (to be published).

[3] Tomáš Pavelek and Karel Pala. Visdic – a new tool for wordnet editing. In *Proceedings of the First International Global WordNet Conference*, Mysore, India, 2002. Central Institute of Indian Languages. □