Část I

Teoretické pojednání

Kapitola 1

Wordnet

1.1 Motivace vzniku

Od počátků snah o zpracování přirozeného jazyka (NLP, natural language processing) bylo nutné poskytnout programu data o lexiku ve zpracovávaném textu, ať už ona data byla jakákoliv. Kupříkladu pro překlad se mělo za to, že stačí ekvivalentní dvojice ve zdrojovém a cílovém jazyce, později se přidal kontext v případě statistického strojového překladu spolu s dalšími informacemi, jako je například slovní druh. Tradičně se lexikální materiál ukládá způsobem nikoliv diametrálně odlišným od papírových slovníků určených pro lidské uživatele. Ty obvykle obsahují abecenedně (či podle jiného indexu) seřazené jednotlivé záznamy s potřebnými informacemi o slovech, z nichž pak program může čerpat při zpracování textu.

cit?

Jak uvádí Pala; Ševeček, (2013), uspořádání lexikálního materiálu v takovéto formě je sice vhodné pro člověka, ale nikoliv pro strojové zpracování, a to z několika důvodů. Kromě toho, že vyhledávání v abecedním seznamu je relativně pomalé , struktura tradičního slovníku kvůli onomu abecednímu řazení inherentně vzdaluje slova, jež člověk chápe jako nějakým způsobem blízká. Tato blízkost může vyplývat ze vztahu volné synonymie, antonymie, podřazenosti, nadřazenosti, etc. Pokud si tedy například uživatel výkladového slovníku nepříliš obeznámený s daným jazykem vyhledá určité heslo, dozví se sice pravděpodobně jeho význam, ale nebude schopen své znalosti prohlubovat dále zjištěním, kupříkladu jaké je slovo odpovídá opačnému významu.

nejaka citace?, dohledat neco, jak takovy slovniky byly ulozeny...

Dalším všeobecným problémem při využití tradičních slovníků k počítačovému zpracování jazyka je fakt, že lexikografové předpokládají u uživatele slovníku značné encyklopedické znalosti. Zařazují tak do slovníku jen informace dle jejich názoru důležité pro rozlišení (differentia specifica) a zařazující do kontextu či přiřazující k určité nadřazené třídě objektů (genus proximum). Vyhledá-li si tedy člověk ve Slovníku spisovného jazyka českého heslo vlk, zjistí následující:

citace

vlk: psovitá šelma šedě (n. šedožlutě) zbarvená, žijící v Evropě, Asii a v Sev. Americe

Definice a priori předpokládá, že uživatel je obeznámen s tím, co je šelma a co je pes. Pokud takovou znalostí neslyne (což je vcelku představitelné například u cizince), je nucen si tato slova ve slovníku najít a podívat se na jejich definice (pomiňme nyní netriviální úkol převést slovo psovitá na základní tvar pes). Pokud nerozumí definicím ani nadřazených slov, musí pokračovat v hierarchii dále a dále.

Z uvedeného případu plyne, že jakkoliv je možné správným vyhledáváním hyperonym[1] dospět k tomu, že vlk je (Nevěřilová; Ulipová, 2014) konkrétní entita našeho vesmíru, živá bytost o čtyřech končetinách, savec nějakým způsobem příbuzný se psovi, má šedou srst etc., je takový proces dosti komplikovaný. Případ s cizincem se sice nemusí zdát zcela relevantní, protože se dá předpokladat, že daný člověk má, byť v jiném jazyce, stejné základní znalosti předpokládané lexikografy jako člověk český. Situace je však dramaticky jiná u počítače. Na rozdíl od člověka totiž počítač nemá žádné předchozí znalosti, tudíž musí projít celým procesem popsaným výše, aby byl schopen kupříkladu určit, že vlk může umřít (ježto je živá bytost). Protože však tradiční slovníky typu SSJČ byly vytvářené pro papírové médium, neobsahují žádné propojení ve stylu *toto je odkaz na hyperonymum*, a počítač tudíž jen těžko může zjišťovat, na které vlastně slovo se to má podívat, aby se dobral podstaty pojmu vlk.

V zájmu automatizace vyhledávání ve slovníku vznikaly tzv. strojově čitelné slovníky[2], což je pojem souhrnně označující lexikální databáze. Podle množství informací, které taková databáze obsahuje, pak lze tyto dělit na slovníky, taxonomie a ontologie. Je evidentní, že obyčejný slovník neobsahuje oproti tradičnímu papírovému slovníku navíc žádné metainformace, takže je počítač při jeho užívání v podstatě omezen na elektronický listovač (Miller et al., 1990).

Naznačeny tedy byla jedna klíčová vlastnost, kterou databáze významů musí oproti tradičnímu slovníku mít, aby byla použitelná pro počítačové zpracování přirozeného jazyka. Jsou jí pojmenované sémantické vztahy mezi slovy. Díky nim je totiž možno jednoznačně určit, které slovo (či slova) je v takové databázi konkrétnímu slovu nadřazené, které je jeho specifikací, označením jeho částí, etc. S touto myšlenkou tedy vznikl Wordnet - lexikální síť provázaná sémantickými vztahy obsahující vedle vztahů ještě definice slov a jejich metainformace (například slovní druh). (Pala; Ševeček, 2013)

1.2 Historie Wordnetu

[1] nadřazené slovo

[2] machine readable dictionary

Seznam literatury

Miller, George A et al., 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4): s. 235–244.

Nevěřilová, Zuzana; Ulipová, Barbora, 2014. A System for Predictive Writing. In: Horák, Aleš; Rychlý, Pavel (ed.). *8th Workshop on Recent Advances in Slavonic Natural Language Processing* [online]. Brno: NLP Consulting, s. 11–18, [cit. 2015-10-22]. Dostupné z: ⟨`http://nlp.fi.muni.cz/raslan/raslan14.pdf`⟩.

Pala, Karel; Ševeček, Pavel, 2013. Česká lexikální databáze typu WordNet. *Feedback*, 48(A47).