Magi

edík

26. března 2017

Obsah

Část I

Teoretické pojednání

Kapitola 1

Princetonský WordNet

Princetonský WordNet je prvním wordnet vůbec. Vznikal na univerzite v Princetonu pod G. A. Millerem od poloviny 80. let 20. století. Vzhledem k tomu, že byl prvním wordnetem, bylo k němu referováno jako k WordNetu, bez přívlastku. Ačkoliv tento stav v podstatě přetrvává dodnes, oproti době jeho vzniku se situace změnila, vzniklo několik dalších wordnetů a nastala tudíž potřeba je rozlišit. V anglickém prostředí se obvykle pojmem WordNet míní ten princetonský a všechny ostatní wordnety mají přívlastek. Příkladem nechť je Balkanet či Eurowordnet. Ačkoliv v mezinárodním prostředí je obvyklé přívlastek „princetonský“ používat, bude tato práce pracovat s následujícím rozlišením:

- *WordNet* (ve významu princetonský WordNet)
- *wordnet* (ve významu obecné sémantické sítě založené ideou, obsahově či strukturálně na WordNetu)
- konkrétní wordnety, např. *Balkanet*

a samozrejme bych to actually mohl dodrzovat, tohle jsem si vymyslel az po napsani tehle kapitoly, lol

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 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.

[1] nadřazené slovo

1.1.1 Strojově čitelné slovníky

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).

Míra, jakou se strojově čitelný slovník odliší od pouhé zdigitalizované formu papírového slovníku a přiblíží se k pokročilé lexikální databázi, lze vyjádřit v několika stupních. V případě, že slovník má jednotlivé významy uspořádány v hierarchii dle nadřazenosti–podřazenosti, lze jej označit za taxonomii, tedy systém s hlubší strukturou než pouze abecedním řazením hesel.

nejakej link, kde budou významy/senses vysvetleny

Dalším stupněm už je skutečná lexikální databáze, která má jednotlivé významy propojeny rozličnými vztahy, počínaje onou základní hyperonymií a hyponymií a pokračuje kupříkladu vztahy meronymie[3] či antonymie[4]. Kromě vztahů mezi významy bude taková lexikální databáze obsahovat zřejmě i další informace, tedy nějaké kategorie slov, jejich popis, etc. Databáze tak popsaných významů propojených sémantickými vztahy může být nazývána ontologií.(GarshoI, n.d.)

1.1.2 Od slovníků k WordNetu

Výše uvedená opozice papírového slovníku a ontologie ilustruje rozdíly tradičního slovníku a počítačově zpracovatelné lexikální databáze. Už ze samotného významu takové databáze je evidentní jeden klíčový rozdíl – tradiční slovníky, jsouce řazené abecedně, od sebe oddalují některá hesla, jež by bylo vhodné mít pohromadě (Pala; Ševeček, 2013). Příkladem budiž kostra a její části, např. lebka. V SSČ[5] i SSJČ se u lebky uvádí, že jde o kostru hlavy. Lze tedy s jistou rezervou tvrdit, že heslo obsahuje své holonymum[6], opačně to však již nefunguje. Z celkem evidentních důvodů nejsou u hesla kostra uvedeny všechny její části. Tento příklad příhodně ukazuje i jistou nesystémovost tradičních slovníků, která je pro počítačové zpracování fatální.

Naznačeny tedy byly vlastnosti, jež by databáze významů měla oproti tradičnímu slovníku mít, aby byla použitelná pro počítačové zpracování přirozeného jazyka. Především jde o systémovost vztahů. Hypero/hyponymie je vztah oboustranný, tudíž by mělo být možné se stejnou cestou dostat od

[2] machine readable dictionary
[3] vztah *je částí*, tedy např. dveře je meronymem trolejbusu
[4] protikladu
[5] Slovník spisovné češtiny
[6] vztah opačný k meronymii; tedy např. dům je holonymem pro okno, dveře, práh etc.

nadřazeného slova k podřazenému a naopak. Dále jsou podstatné 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, která by dle poznatků psycholingvistiky odrážela uspořádání lexikálního materiálu v lidském mozku (o tom v dalších kapitolách). (Pala; Ševeček, 2013) Zde by bylo na místě poznamenat, že ačkoliv se tak z odstavců výše může čtenáři jevit a i všeobecně je to často tvrzeno, WordNet není ontologií v pravém slova smyslu, protože něco něco.. `https://en.wikipedia.org/wiki/WordNet#WordNet_as_a_lexical_ontology`

ref

a tady tomu vubec nerozumim, ale prijde mi to relevantni

1.1.3 Psycholingvistické hledisko

G. A. Miller, který je tvůrcem WordNetu, se po spolupráci s Chomskym na základních kapitolách jeho *Handbook of Mathematical Psychology* , která se věnuje spíše syntaktickému popisu jazyka, se zaměřil společně s Johnsonem-Lairdem na výzkum, jakým způsobem je lexikální materiál uložen v lidském mozku. Tento přístup je označován jako psycholingvistika a jeho počátky jsou spojeny s průzkumem asociací a budování modelu mentálního slovníku člověka. Výchozí myšlenka, jež se odráží i ve způsobu organizace WordNetu, spočívá v tom, že slovní zásoba není organizována abecedně, jako tomu je v tradičních slovnících, ale spíše konceptuálně.

cit

Jednou z otázek tohoto směru bylo, jakým způsobem je organizována paměť. Aby člověk byl schopen určit pravdivostní hodnotu výroku Kanárek může létat, musí použít svou dlouhodobou pamět. Její organizace je pak možná (minimálně) dvěma způsoby. První, redundantní, by vypadal tak, že by u každého ptáka bylo uloženo, že je schopen létat. Druhý, již na první pohled výrazně méně náročný na úložný prostor, by příznak schopnosti létat měl uložený pouze u kategorie pták. V případě, že by bylo třeba zjistit, zda kanárek léta, by bylo nutno pak zapojit inferenční proces ve stylu *kanárek je pták, tudíž může létat.* (Collins; Quillian, 1969)

Jak Collins; Quillian, (1969) dále uvádí, lze předpokládat, že v případě prvního způsobu organizace paměti by člověk mohl kteroukoliv informaci o příznacích (vlastnostech) z paměti vyvolat za konstantní čas. Naproti tomu v případě způsobu druhého by extrakce příznaku z významu v hierarchii položeného výše měla trvat delší čas než extrakce příznaku přítomného přímo u významu, jenž je objektem věty. Důvodem by měla být nutnost zapojení inferenčního procesu.

Pokus, kterým podpořili Collins; Quillian, (1969) druhý, neredundantní, způsob organizace paměti, spočíval v tom, že testovací subjekty, dobrovolníci z řad zaměstnanců společnosti Bolt Beranek and Newman, měly určovat, zda je výrok pravdivý, či nepravdivý. Měli tak činit co nejpřesněji a v co nejkratším čase, přičemž byla měřena rychlost jejich reakce. Ukázalo se,

reakční doba při určování pravdivosti výroku Kanárek umí létat[7] je delší než při určování pravdivosti výroku Kanárek umí zpívat[8] a ještě delší při určování výroku Kanárek má kůži[9]. Důvodem pro tyto progresivní prodlevy podle nich právě byla zvětšující se vzdálenost od významu kanárka ke významu, u něhož byl uložen příslušný příznak, tedy umí zpívat, umí létat, resp. má kůži. Příznak umí zpívat totiž je pravděpodobně uložen přímo u kanárka, jelikož jej odlišuje od ostatních ptáků, zatímco příznak umí létat je obecným znakem ptáků, tudíž je uložen u významu pták. V poslední řadě pak příznak má kůži bude patrně uložen u významu zvíře, který je oněch tří v hierarchii nejvýše, a ze všech tudíž od kanárka nejdále.

Jelikož WordNet G. A. Millera je založen na podobném principu , lze cit říci jistým způsobem odráží organizaci lexika v lidském mozku.

Zde je na místě uvést poznámku o dalším pricnipu, o nějž se WordNet opírá, a to organizace slovní zásoby podle základních slovních druhů. Testováním anaforických a komparativních výrazů se ukázalo, že lidé dokážou vcelku jednoduše určit slovní druh určitého výrazu. Je tedy nasnadě, že taková informace musí být přítomna u každého významu, což vede, oproti předchozímu principu, k jisté redundantnosti systému. Existuje totiž mnoho slov (zvláště např. v angličtině), která zastupují jak substantivum, tak verbum, a zdá se, že tyto významy, byť označovány stejným výrazem (angl. show, popř. česky stát), jsou uloženy zvlášť a mají svou vlastní množinu příznaků. Tento koncept podporuje i fakt, že se evidentně chovají zcela odlišně v rovině syntaktické.Miller et al., (1990)

1.2 Sémantické vztahy WordNetu

Jak bylo naznačeno výše, koncept WordNetu je založen na lexikální sémantice, tedy představě, že slovo je kombinací slovní formy a významu, nebo slovního významu. Slovní forma je projevem „fyzickým“, tedy je to vyřčená či napsaná instance významu. Jak je zjevné z přirozeného jazyka, nelze počítat s tím, že by zobrazení významu na formu bylo bijektivní, tedy každý význam byl namapován jedna ku jedné na slovní formu. Mapování je v přirozeném jazyce tzv. více ku více, tedy jedna forma může zastupovat více významů a jeden význam může být vyjádřen více formami. Je velmi časté, že jedna slovní forma zastupuje více významů. Kupříkladu slovní forma koruna může zastupovat význam měny, vrcholku stromu, vladařského odznaku, etc. Toto zobrazení se nazývá homonymií[10]. Možný je samozřejmě i opačný případ, kdy více významům slouží k vyjádření jedna slovní forma, což je nazýváno polysémií[11]. Příkladem může být forma kolej, již lze interpretovat

[7] angl. A canary can sing
[8] angl. A canary can sing
[9] angl. A canary has skin
[10] totožnost formy
[11] víceznačností

jako referenci k stopě po voze, případně dvojici kolejnic jako vodící dráze pro dopravní prostředky a zároveň jako zařízení vysoké školy pro ubytování studentů.

cit. SSJC

cit SSJC

Miller et al., (1990) popisují výše naznačené vztahy pomocí takzvané lexikální matice. Ta názorně zobrazuje formy synonymní (F_1 a F_2) a formy polysémní (F_2):

`tabulka z miller1990introduction pg. 4: http://i.imgur.com/sohtwe5.png`

V dalších podkapitolách budou rozebrány podrobněji, nikoliv však vyčerpávajícím způsobem, sémentické vztahy konstituující Wordnet. Jelikož se sémantické vztahy pro jednotlivé slovní druhy liší, bude tato kapitola strukturována primárně právě podle slovních druhů a až sekundárně podle sémantických vztahů.

tady by jeste slo pokracovat opisovanim dalsi casti toho clanku (miller1990introduction pg 5 »>)

1.2.1 Frekvenční distribuce sémantických vztahů ve WordNetu

1.2.2 Substantiva

Synonymie

Synonymie je centrálním organizačním vztahem pro substantiva ve WordNetu. Na praktických aplikacích je tento jev nejlépe pozorovatelný, jelikož při vyhledání konkrétní formy je uživateli obvykle nabídnut výběr z jednotlivých významů dané formy. Aby byly od sebe významy oné formy odlišitelné, nabídka běžně sestává z výpisu skupin forem (tzv. synsetů, o tom později), přičemž každá skupina náleží k jednomu významu a obsahuje formy danému významu přiřazené. Kupříkladu při vyhledání slova kolo tak je uživatel konfrontován s několika skupinami, které obsahují zhruba následující:

- kolo (1),
- jízdní kolo (1), bicykl (1), kolo (2),
- kružnice (1), cívka (1), kolo (3),

přičemž čísla (zde) v závorce značí index významu dané formy v daném synsetu. Reprezentace v různých aplikacích a různých wordnetech se liší (standardem bývá číslo významu psát za dvojtečku), koncept však zůstává neměnný.

Navzdory zdánilivé jednoduchosti výše uvedeného konceptu je všeobecnou otázkou, jak synonymii pojímat. Striktní teorie (obvykle připisovaná Leibnizovi) praví, že dvě slova jsou synonymní, pokud se jejich záměnou nikdy nezmění pravdivostní hodnota výroku. Lingvistickou interpretací tohoto poněkud matematicko-logického výroku může pak být, že synonymní dvě slova jsou v případě, že se jejich záměnou nikdy neporuší význam (zhruba

ona pravdivostní hodnota) a gramatičnost výroku. Je nasnadě, že takto striktně synonymní slova budou pospolu v jazyce těžko přežívat, jelikož je dokázáno, že jazyk tíhne k ekonomičnosti, která by takovým soužitím dvou slov byla hrubě porušena. Pravděpodobně jedinými obecně uznávanými synonymy jsou obvykle dvojice cizího slova a domácího slova, například internacionální a mezinárodní. S relativně vysokou jistotou lze tvrdit, že jejich záměnou se nikdy pravdivostní hodnota výroku nezmění, stejně tak jako jeho gramatičnost. Stále však zůstává ve hře stylistika, která může být podobnou náhradou narušena (např. z důvodu cílové skupiny čtenářů či stylistické příznakovosti jednoho ze slov (srov. zajímavý a interesantní)). Co se tendence k ekonomičnosti jazyka týče, lze předpokládat, že v těchto případech převládá potřeba synonym k eliminaci opakování určitých slov v textu a tím zajištění jeho stylistické uhlazenosti.

Volnější interpretace synonymie také počítá s kontextem. Dvě slova jsou synonymní, jsou-li bez způsobení škod nahraditelná alespoň ve stejném kontextu. Jako příklad mohou posloužit formy board a plank . V kontextu dřevařství mohou tyto dvě formy pravděpodobně bez problému být nahrazeny jedna za druhou, ovšem v případě, že je forma board použita ve významu comittee, těžko ji lze nahradit za formu plank, neboť by se věta obsahující takové nahrazení stala zcela nesmyslnou.

najit ceske priklady

Bylo by nanejvýš logické považovat synonymii za vztah diskrétní, tedy že dvě formy buďto synonymní jsou, či nejsou. Z logického hlediska to nepochybně z již uvedeného vyplývá, ovšem lingvisticko-filosofický náhled vycházející z poznatků reálného jazyka se na tuto problematiku poněkud liší. Jak bylo dokázáno, synonymie v striktním slova smyslu je velice vzácná. Její volnější interpretace je značně častější, ale také výrazně vágnější – kontext, v němž dvě formy synonymní jsou může být velmi široký, či naopak velice úzký. Záměna některých dvojic (obecně n-tic, lze ale předpokládat, že mnoho forem nebude mít dvě další synonymní) může měnit stylistiku a význam výpovědi více či méně, přičemž ony dvě formy stále dle daných kritérií lze považovat za synonymní. Nelze tedy než vyvodit, že synonymie, minimálně z pohledu přirozeného jazyka, je jevem graduálním, a některé formy jsou tak *synonymnější* než jiné.

Zaměnitelnost forem podporuje ještě jeden koncept, na němž je WordNet postaven, a to fakt, že jednotlivé významy jsou seskupovány podle slovních druhů. Jak již bylo řečeno, tento systém vede k jisté redundantnosti, jelikož zvláště v syntetických jazycích, jako je kupříkladu angličtina, lze nalézt mnoho případů, kdy identická slovní forma zastupuje více slovních druhů. Významy, které taková slovní forma zastupuje (napříč slovními druhy), mohou být velice blízké, nikdy však nebudou stejné (nelze říci, že význam slovesa run[12] a substantiva run[13] je identický). Jejich záměnou by se sice nestalo

check, nekecam?

[12] běžet
[13] běh

vůbec nic, jelikož čtenář či posluchač textu, v němž taková záměna nastala, by automaticky formu interpretoval ve prospěch správného slovního druhu, avšak pokud by slovní druh byl nějakým způsobem „vynucen“ (nechť nyní čtenář pomine úvahy, jakým způsobem lze „vynutit“ slovní druh formy), stala by se výpověď zcela negramatickou a nesmyslnou.

Seskupování významů podle slovních druhů a seskupování forem dle vztahu synonymie se tedy zdá v případě lexikální databáze určené pro strojové zpracování jako vhodným konceptem. Oproti tradičním slovníkům se totiž počítačově zpracovávaná lexikální databáze nemusí potýkat s problémem lidského faktoru – jednotlivé synonymické řady je stroj schopen prohledávat, na rozdíl od člověka, velice účině, a nahradí tak v případě, že WordNet používá člověk, neúčinné lidské procházení restříkového obsahu.

psano v chvatu, mohlo by se to mozna trochu uhladit...

1.2.3 Antonymie

Kapitola 2

Další wordnety

Podle vzoru princetonského WordNetu začaly postupně vznikat i další sémantické sítě založené na stejném konceptu. Tyto sémantické sítě se samozřejmě svou strukturou do menší či větší míry liší , hlavním kritériem pro to, aby mohly být považovány za wordnet je to, aby obsahovaly synsety a hyponyma . Jelikož se tato práce bude primárně zabývat wordnetem českým, bude pro srovnání uvádět dva hlavní evropské vícejazyčené wordnety, a to Eurowordnet a Balkanet.

nakou kurva citaci

http://globalwordnet.org/wordnets-in-the-world/

Seznam literatury

Collins, Allan M; Quillian, M Ross, 1969. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2): s. 240–247.

GarshoI, Lars Marius. Metadata? Thesauri? Taxonomies? Topic Maps! *2004—10—26)[2010-6—19] http://www. ontopia. net/topicmaps/materials/tm-VS—thesauri. htm1.* Dostupné také z: ⟨`http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html#N773`⟩.

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

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