Dolovanie konceptov z textov v prirodzenom jazyku

Viliam Ročkai, Marián Mach

Katedra kybernetiky a umelej inteligencie
Technická univerzita, Letná 9/B, 042 00 Košice, Slovenská republika
viliam.rockai@gmail.com, marian.mach@tuke.sk

Abstrakt Slovníky wordnetového typu už dlhé roky nachádzajú široké uplatnenie nielen v doméne spracovania textov v prirodzenom jazyku. Ich využitiu v NLP je venované nepočetné množstvo prác, no oblasť ich automatického vytvárania stále naráža na mnoho problémov. Tak sa najbežnejšími postupmi stávajú automatické preklady podobných slovníkov v iných jazykoch, poprípade prístupy so silnými znalosťami o jazyku, nad ktorým pracujú. Pojmy sú v slovníku zoradené hierarchicky podľa sémantických relácií (hypernymá, hyponymá, príbuznosť...). Keďže s využitím teórie asociatívneho učenia pojmov je možné podobné stromy pojmov vytvárať, existuje predpoklad na jej využitie pri budovaní podobných slovníkov. Táto práca sa zaoberá možnosťou využitia teórie asociatívneho učenia pojmov v procese budovania slovníku wordnetového typu, na základe správneho zaradenia nových pojmov do podstromu hierarchicky zoradených pojmov v slovníku.

Keywords: asociatívne učenie pojmov, Wordnet, spracovanie prirodzeného jazyka

1 Úvod

Asociatívne učenie pojmov (AUP) je prístup založený na neuro-fyziologickom modeli spracovania talamo-kortikálnej informácie od R. Hecht Nielsena [4]. Model predpokladá existenciu fixného lexikónu symbolov v ľudskom talame, ktorý sa vytvorí v skorom vývinovom štádiu jedinca. Učenie potom pozostáva z vytvárania asociácií medzi neurónmi, odrážajúc tak neurónové spojenia medzi kortikálnymi regiónmi. Najnovšie znalosti o teórii talamu sa nachádzajú v [5]. Dynamika učenia je namodelovaná podľa hebbovskej metódy. Asociatívne učenie pojmov je teda nekontrolované učenie nad prúdom symbolov. Cieľom je naučenie sa znalostnej reprezentácie štruktúry do podoby sémantických sietí, prípadne hierarchií konceptov. Metóda je založená na indukcii asociácií medzi symbolmi vzhľadom na ich spoločné výskyty v kontextovom okne, cez ktoré samotné učenie prebieha. Detailný popis prístupu vrátane procesu učenia sa nachádza v [6].

2 Asociatívne učenie pojmov

Nutná podmienka pre vytvorenie asociácie medzi párom symbolov je daná ich štatisticky nenáhodným spoločným výskytom. Implementuje sa pomocou vzťahu na výpočet signifikancie:

$$S_i(a,b) = \frac{p(a,b)}{p(a)p(b)} \tag{1}$$

kde a a b sú diskrétne náhodné premenné a i je ich kontextová vzdialenosť. Vzájomná signifikancia $S_i(a,b)$ je potom definovaná ako podiel pravdepodobnosti spoločného výskytu symbolov (diskrétnych premenných) a, b k súčinu apriórnych pravdepodobností ich výskytu. Ak hodnota $S_i(a,b)$ prekročí vopred určenú prahovú hodnotu, dané symboly sa považujú za asociované. Je mierou toho, ako veľmi sa líši pravdepodobnosť spoločného výskytu dvoch javov $p(a,b)$ od hodnoty, ktorú by sme očakávali ak by boli dané javy na sebe nezávislé (teda $p(a)p(b)$) [1]. V kontexte tejto práce sa symbol chápe ako term, čiže je reprezentáciou (konkrétne indexom) jedného konkrétneho slova (termu). Kontextová vzdialenosť je pre potreby tejto práce definovaná ako vzdialenosť dvoch slov vo vete, a to ako počet slov, ktoré sa medzi nimi nachádzajú.

2.1 Výpočet sémantickej príbuznosti

Prístup je založený na Jaccardovom indexe [3] počítania podobnosti alebo diverzity dvoch množín:

$$J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|} \tag{2}$$

Vzťah (2) popisuje výpočet podobnosti dvoch množín A a B. Obor hodnôt Jaccardovho indexu je $< 0, 1 >$, takže nie je potrebné ho ďalej normalizovať. V AUP obsahuje každý symbol množinu asociovaných symbolov, teda svoje asociované okolie v rôznych kontextových vzdialenostiach. Okolie symbolu x v kontextovej vzdialenosti i definujeme ako $O_i(a)$:

$$O_i(a) = \{t \in L | S_i(a,t) > prah\} \tag{3}$$

$O_i(a)$ je teda množinou všetkých takých symbolov z množiny známych symbolov, lexikónu L, pre ktoré platí, že sú so symbolom a asociované v kontextovej vzdialenosti i. Podobnosť dvoch symbolov a a b vzhľadom na kontextovú vzdialenosť i môžeme definovať ako:

$$r_i(a,b) = \frac{|O_i(a) \bigcap O_i(b)|}{|O_i(a) \bigcup O_i(b)|} \tag{4}$$

V prípade, že je každej kontextovej vzdialenosti priradená váha w_i, dostávame vzťah na výpočet podobnosti dvoch symbolov $Rel_w(a,b)$:

$$R_w(a,b) = \sum_i w_i r_i(a,b) \tag{5}$$

Oborom hodnôt $Rel_w(a, b)$ je množina $< 0, \sum_i w_i >$. Ak sú váhy pre jednotlivé kontextové vzdialenosti volené tak, že ich súčet dáva hodnotu 1, nie je ďalej nutné výsledné hodnoty normalizovať pretože sú z intervalu $< 0, 1 >$. Výpočtom podobností vybraného symbolu so všetkými symbolmi z lexikónu L dostávame zoradený zoznam podobností.

3 Vytváranie zhlukov konceptov bez znalostí o jazyku

Výpočtom podobností dotazovaného symbolu k všetkým ostatným symbolom z lexikónu sa dá vytvoriť zoznam symbolov zoradený podľa hodnoty podobnosti k dotazovanému symbolu. Základným problémom pri hľadaní podobných symbolov k danému symbolu je určenie hranice v zozname zoradenom podľa podobnosti k vstupnému symbolu. Takýto zoznam síce ilustruje schopnosť AUP identifikovať podobné slová k vstupnému symbolu, ale čelí viacerým problémom. Základné problémy takéhoto zoznamu sú:

- Dĺžka zoznamu - dĺžka zoznamu je vo väčšine prípadov len o málo menšia ako je celková veľkosť použitého lexikónu.
- Chyby v usporiadaní symbolov - často sa v zozname medzi relevantnými symbolmi objavia symboly, ktoré sú z hľadiska empirického poznania jazyka menej relevantné.
- Nejednoznačnosť významu symbolov - slovo „May“ môže mať význam „mesiac“, ale aj „snáď“. Usporiadanie ostatných slov z lexikónu podľa podobnosti je potom silne závislé na korpuse, nad ktorým AUP akumuloval asociácie.

Tieto problémy sa pokúša riešiť prístup zhlukovania symbolov na základe ich podobnosti popísaný v tejto kapitole. Pri zhlukovaní pojmov je potrebné hľadaný zhluk nejako definovať. Pri výpočte jednoduchého zoznamu zoradeného podľa sémantickej podobnosti bol použitý ako vstup práve jeden symbol. Tento symbol môže patriť do viacerých významových množín. Ako príklad významovej množiny symbolov sa dá chápať napríklad množina sesterských synsetov Wordnetu [1]. Kým symbol reprezentovaný slovom „January“ patrí podľa Wordnetu len do jedného synsetu, symbol reprezentovaný slovom „March“ už patrí do štrnástich synsetov (okrem významu „mesiac“ má aj významy ako „pochod“ a iné). Vo Wordnete sú synsety vymenované množiny termov, ktoré sú definované rovnako vymenovaním termov, ako aj krátkou všeobecnou definíciou ich významu. Vzhľadom na fakt, že AUP nemá žiadne znalosti o jazyku nad ktorým sa učil asociácie odpadá možnosť popisnej definície zhluku. Kým slovo „March“ sa vyskytuje v mnohých synsetoch, dvojica slov „March“ a „January“ existujú v dvoch sesterských synsetoch a to v podstrome označujúcom mesiace gregoriánskeho kalendára. Tieto dva termy teda v rámci Wordnetu môžu poslúžiť ako postačujúce na explicitnú definíciu jedného konkrétneho synsetu. Existuje teda predpoklad, že kombináciou zoznamov kvantifikovaných podobností k dvom alebo viacerým

[1] Dostupné online: http://wordnet.princeton.edu/

termom by mohlo byť možné automaticky vytvárať množiny konceptov podobných synsetom vo Wordnete. Predpokladáme, že pre všetky symboly v rámci jedného synsetu obsahujúceho viac ako dva symboly, bude platiť, že si budú približne rovnako podobné. Pre lepšie pochopenie a vizualizáciu problému zavedieme nový pojem sémantickej vzdialenosti dvoch symbolov:

$$D_w(a,b) = 1 - R_w(a,b) \tag{6}$$

kde $D_w(a,b)$ označuje hodnotu vzdialenosti dvoch symbolov a a b a $R_w(a,b)$ označuje ich podobnosť. Každý zoznam termov zoradených podľa podobnosti k dotazovanému termu vieme prepísať pomocou (6) ako zoznam termov zoradených podľa vzdialenosti k tomuto termu. Samotná vzdialenosť od jediného symbolu ale nie je postačujúcou znalosťou pre vytvorenie zhluku, ktorého prvky by patrili do jednej jasne ohraničenej sémantickej oblasti. Predpokladáme, že všetky termy z jednej ohraničenej sémantickej oblasti budú od seba navzájom približne rovnako vzdialené a táto vzdialenosť by mohla poslúžiť ako fiktívna hranica sémantickej oblasti. Sémantická oblasť môže byť definovaná vymenovaním prvkov:

$$M_c = \{s_1, s_2, \cdots, s_n\} \tag{7}$$

kde M_c je množina označujúca samotnú sémantickú oblasť, symboly $s_1, s_2, \cdots, s_n$ sú jej prvkami a n je počet jej prvkov, teda mohutnosť množiny M_c. Príslušnosť nového symbolu t, nezaradeného v M_c, k tejto sémantickej oblasti je definovaná vzťahom:

$$\frac{1}{n}\sum_{i}^{n} D_w(i,t) \leq \frac{1}{C^2(n)}\sum_{i}^{n}\sum_{j=i+1}^{n} D_w(i,j) \tag{8}$$

kde i a j sú prvkami sémantickej oblasti M_c, n je mohutnosťou sémantickej oblasti M_c a $C^2(n)$ označuje počet dvojprvkových kombinácií všetkých symbolov zo sématickej oblasti M_c. Vzťah (8) sa dá interpretovať tak, že symbol t patrí do blízkosti sémantickej oblasti M_c vtedy, ak priemerná vzdialenosť termu t od všetkých prvkov sémantickej oblasti M_c je menšia alebo rovná priemernej vzdialenosti medzi všetkými dvojicami z danej sémantickej oblasti.

4 Ukážka prístupu na sémanticky silne ohraničených množinách

Pre názornú ukážku tohto prístupu je vhodné vybrať množiny, u ktorých je príslušnosť do danej skupiny jednoznačná a zároveň sa jedná o slová, ktoré by mali možnosť dostať sa do ohraničeného lexikónu najpoužívanejších slov. Trojica bola zvolená preto, lebo v nej počet všetkých vzdialeností, ktoré majú symboly medzi sebou, je zhodný s počtom vzdialeností, ktoré majú voči symbolu mimo trojice. Ako príklad môže poslúžiť dvanásť mesiacov v roku, sedem dní v týždni alebo číslovky. Systému zadáme náhodnú trojicu vybranú z týchto množín a na výstupe očakávame správne doplnenie o ďalšie (v ideálnom prípade všetky) prvky z danej množiny. Na akumuláciu asociácií bol použitý korpus textov v prirodzenom jazyku získaný z náhodnej podmnožiny článkov projektu Wikipedia

Tabuľka 1. Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných mesiacov

vstup	january, june, november
výstup	april, march, august, february, september, may, june, november, december, july, january, october

Tabuľka 2. Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných dní

vstup	tuesday, wednesday, sunday
výstup	saturday, thursday, monday, sunday, tuesday, wednesday

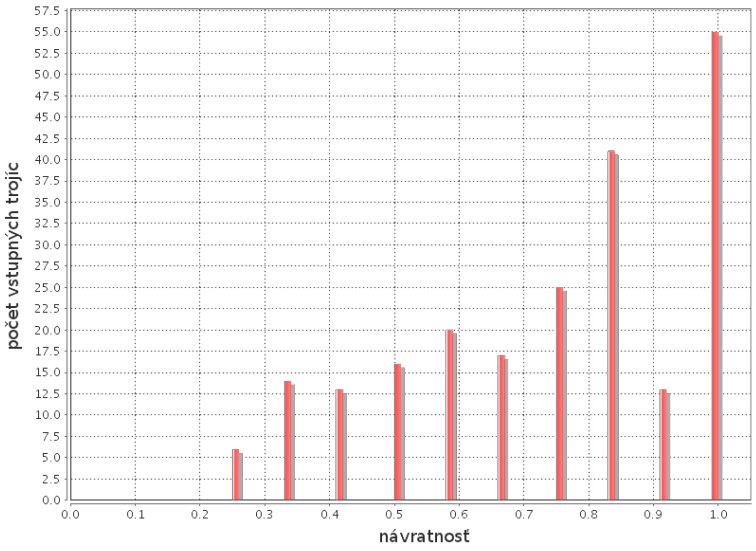
Tabuľka 3. Výstup zhlukovania symbolov na základe vstupnej trojice troch náhodných čísloviek

vstup	three, fifty, twenty
výstup	twelve, forty, four, three, twenty, ten, two, seven, five, fifty, fifteen, six, thirty

o celkovej veľkosti $3GB$. Ako lexikón bolo použitých 6000 najpoužívanejších slov z rovnakého korpusu. Asociácie boli získavané v štyroch kontextových vzialenostiach (s hodnotami 1, 2, 3, 4) s hodnotou prahu 2 a pri výpočte podobnosti mali všetky váhy hodnotu $^1/_4$. Výsledok týchto vstupov možno vidieť v tabuľkách 1,2 a 3. V ďalšom prípade bolo zo slov reprezentujúcich dvanásť mesiacov v roku vytvorených všetkých 220 kombinácií trojíc, ktoré poslúžili ako iniciálna podoba sémantickej oblasti. Pre každú z týchto trojíc bola znova vypočítaná príslušnosť ostatných symbolov z lexikónu do blízkosti sémantickej oblasti, ktorú definovali. Ani pri jednej z 220 trojíc nebol žiaden výsledný symbol slovo, ktoré by neoznačovalo mesiac v roku. Úplná množina 12 mesiacov bola nájdená v 25% prípadov. V ostatných prípadoch sa výstup pohyboval od troch (zhodných so vstupnou trojicou) do jedenástich mesiacov. Histogram hodnôt návratnosti prístupu je znázornený na grafe 1, kde na osi y je znázornený počet vstupných trojíc a na osi x je znázornená im prislúchajúca hodnota návratnosti (úplnosti hľadanej oblasti). Priemerná návratnosť pre všetkých 220 trojíc mala hodnotu 0.738.

5 Záver

Táto práca sa zaoberala možnosťou využitia teórie asociatívneho zhlukovania pojmov na základe ich sémantickej blízkosti. Obsahuje popis postupu, ktorý túto úlohu dokáže v nejakom rozsahu riešiť. Ako príklad poslúžilo generovanie jasne sémanticky ohraničených množín. Vstupom algoritmu boli tri slová reprezentujúce sesterské synsety Wordnetu (napr. mesiace, dni, číslovky) a systém sa snažil nájsť ďalšie slová bez akejkoľvek znalosti jazyka. To sa mu podarilo s úplnou presnosťou a veľmi uspokojivou návratnosťou. Systém bol prezentovaný na veľmi úzkej množine príkladov a bol obmedzený na anglický jazyk. Ďalší



Obr. 1. Histogram hodnôt návratnosti pre všetky kombinácie vstupných trojíc symbolov.

výskum by mal byť sústredený na skúmanie jeho funkčnosti pri širšej množine vstupov a iných jazykoch.

Poďakovanie

Tento príspevok vznikol s podporou VEGA grantu MŠ SR č. 1/0042/10 "Metódy identifikácie, anotovania, vyhľadávania, sprístupňovania a kompozície služieb s využitím sémantických metadát pre podporu vybraných typov procesov."

Literatúra

1. Bouma G., Normalized (Pointwise) Mutual Information in Collocation Extraction, In Proceedings of the Conference of the German Society for Computational Linguistics 2009, GSCL-2009, Potsdam, Germany, 2009
2. Fellbaum C.: *WordNet: An Electronic Lexical Database*, The MIT Press, Cambridge, MA, 1998
3. Jaccard, Paul (1901), "Étude comparative de la distribution florale dans une portion des Alpes et des Jura", Bulletin de la Société Vaudoise des Sciences Naturelles 37: 547–579.
4. Nielsen R. H.: *A Theory of the Cerebral Cortex*, ICONIP, 1998
5. Nielsen R. H.: *Confabulation Theory: The Mechanism of Thought*, Springer, 2007
6. Ročkai V., 2005. *Mining of Concepts and Semantic Relations from Texts in Natural Language* Diplomová práca, Katedra kybernetiky a umelej inteligencie, Fakulta elektrotechniky a informatiky, Technická univerzita, Košice, 2005