Asociatívne učenie konceptov

Viliam Ročkai

Department of Cybernetics and Artificial Intelligence, Technical University of Košice,
Letná 9, 041 20 Košice, Slovak Republic, viliam.rockai@gmail.com

Abstrakt. Pre ľudí nepredstavuje určenie toho, či dve slová spolu súvisia, problém. Dokonca dokážu bez problémov určiť, či dané slová spolu súvisia viac alebo menej. Napríklad ak by sme sa stretli s pojmami "pes" a "mačka" automaticky by sme vedeli, že sa v oboch prípadoch jedná o domácich miláčikov. Takisto by pre nás nebol problém určiť, že slovo "mačka" bude viac súvisieť s pojmom "pes" ako s pojmom "klávesa". V tomto článku popíšeme prístup na výpočet sémantickej príbuznosti konceptov založenom na teórii asociatívneho učenia pojmov.

Keywords: ontológia, koncept, sémantická podobnosť, asociatívne učenie

1 Úvod

Výstavba ontológií sa v poslednom čase stáva veľmi populárnou oblasťou výskumu. Súčasné metódy a výskum sa ale zaoberá hlavne poloautomatickou výstavbou ontológií založenej na procesoch spracovania prirodzeného jazyka. Zvyčajne sa jedná o napĺňanie už existujúcej ontologickej štruktúry inštanciami nájdenými v samotnom texte. Bežne potrebujú silne predspracovaný textový korpus a nesnažia sa "naučiť" ontológiu od základov. V tomto článku predstavujeme metódu na indukovanie hierarchií konceptov z korpusu textov v prirodzenom jazyku založenej na učení štatistickými metódami. Experimenty ukazujú, že táto metóda je schopná efektívne merať sémantické podobnosti párov tokenov, čo poskytuje prvý krok k budovaniu hierarchie konceptov (jednoduchej ontologickej štruktúry). Výstavba hierarchie konceptov v našom prípade znamená zhlukovanie slovných tokenov do morfologicky alebo sémanticky blízkych skupín, čo umožňuje automatické vytváranie slovníkov z textov v prirodzenom jazyku.

Náš prístup je založený na neuro-fyziologickom modeli spracovania talamo-kortikálnej informácie od R. Hech-Nielsena [3], čo do neho vnáša istú dávku biologickej hodnovernosti. Model predpokladá existenciu fixného lexikónu "symbolov" v talame ľudského mozgu, ktorý sa vytvorí v skorom vývinovom štádiu ľudského jedinca. Učenie potom pozostáva z vytvárania asociácií medzi neurónmi, odrážajúc tak neurónové spojenia medzi kortikálnymi regiónmi. Dynamika učenia siete je namodelovaná podľa Hebbovskej metódy [2]. Asociatívne učenie pojmov je nekontrolované učenie nad prúdom symbolov. Cieľom je naučiť sa znalostnú reprezentáciu 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. Nutná podmienka pre vytvorenie asociácie medzi párom tokenov je daná ich štatisticky nenáhodným spoločným výskytom. Potom ako je asociácia vytvorená sa v priebehu učenia mení iba jej váha [6].

2 Asociatívne učenie konceptov

2.1 Reprezentácia

Vzhľadom na teóriu talamo-kortikálneho učenia vysvetlenom v [3], naša implementácia [6] pracuje s tokenmi, ktoré sú dané ako množina excitovaných neurónov v nejakom kortikálnom regióne. Token je tak reprezentovaný ako invariantný senzorický vstup a dá sa vnímať ako atribút pozorovaného objektu. V teórii sú reprezentáciou týchto atribútov symboly. Koncept je definovaný ako symbol so svojím asociovaným okolím. Koncept môže byť teda aktivovaný rôznymi množinami symbolov. Vytvorenie asociácie medzi symbolmi závisí od toho, či je ich spoločný výskyt náhodný, alebo nenáhodný. Signifikancia môže byť vypočítaná podľa nasledujúceho vzorca:

$$S(i, j) = \frac{p(i, j)}{p(i).p(j)} \quad (1)$$

kde *i* a *j* sú diskrétne náhodné premenné. Vzájomná signifikancia *S(i,j)* je založená na vzájomnej informácii z teórie informácií. V našom prípade, s použitím korpusu prirodzeného jazyka, chápeme *p(i)* a *p(j)* ako pravdepodobnosti, že sa slovo *i* alebo *j* vyskytlo v texte. *p(i,j)* označuje pravdepodobnosť, že sa tieto slová v texte vyskytli spolu. Vzájomná signifikancia je definovaná ako podiel vzájomnej pravdepodobnosti symbolov (diskrétnych premenných) *i, j* a pravdepodobností ich výskytu. Ak by *i* a *j* boli nezávislé platilo by:

$$p(i, j) = p(i).p(j) \quad (2)$$

Teda ak sa dva symboly vyskytujú v spoločnom kontexte náhodne (teda *i* a *j* sú nezávislé), bude reprezentácia takáto:

$$S(i, j) = \frac{p(i, j)}{p(i).p(j)} = 1 \quad (3)$$

Ak *S(i,j)* > *tresh*, kde *tresh* je parameter hodnoty prahu z intervalu $<1,\infty)$, tak tokeny budú v našom prípade považované za asociované. To bude naším hlavným parametrom počas procesu učenia. Nazveme tento parameter prahom a môže byť využitý pri filtrovaní viac nenáhodných výskytov. Váha asociácie medzi dvoma symbolmi *i* a *j* sa počíta iba v prípade, že sú dané symboly asociované a je daná vzorcom:

$$w(i, j) = \frac{p(i, j)}{p(j)} \quad (4)$$

Váha asociácií je jediná znalosť, ktorá sa učí. Tieto váhy sú reprezentované ako fascikle - naše znalostné bázy $F_x = w(i,j)$ matice, kde x označuje kontextovú vzdialenosť. Učíme niekoľko znalostných báz F_x pre rôzne kontextové vzdialenosti (vlastnosti) daných dvojíc symbolov [6].
V prirodzenom jazyku chápeme kontextovú vzdialenosť ako počet slov medzi danou dvojicou tokenov. Napríklad vo vete "Peter ľahko vyšplhá aj na najvyššiu skalu" vezmime za referenčný token slovo "aj". Kontextové vzdialenosti pre ostatné slová vo vete potom budú (vyčíslené v zátvorkách):
Peter (-3) ľahko (-2) vyšplhá (-1) aj (0) na (1) najvyššiu (2) skalu (3).
V našom prípade je proces učenia, ukladania váh, prevedený na štyroch priamych (kladných) a štyroch (záporných) kontextových vzdialenostiach. Takže sme ukladali váhové informácie pre fascikle F_{-4}, F_{-3}, F_{-2}, F_{-1} a fascikle F_1, F_2, F_3, F_4.
Výpočet váh $w(i,j)$ je daný pomocou nasledujúcich vzorcov:

$$S(i, j) = \frac{\frac{C(i, j)}{T}}{\frac{C(i)}{T} \cdot \frac{C(j)}{T}} \quad (5)$$

$$w(i, j) = \frac{\frac{C(i, j)}{T}}{\frac{C(j)}{T}} \quad (6)$$

Kde $C(i,j)$ je spoločný výskyt tokenov i a j, a T je počet tokenov, s ktorými sa doteraz v procese učenia model stretol, a ktorý sa v priebehu učenia inkrementuje. Podiel týchto premenných definuje pravdepodobnosti vo vzorcoch 5 a 6. V tomto článku nevyužívame hodnotu váh kvôli zjednodušenému výkladu. Váhy v tomto prípade iba znázorňujú, že asociácia bola vytvorená, pretože signifikancia bola vyššia ako prahová hodnota potrebná na vytvorenie asociačného spojenia.

2.2. Sémantická príbuznosť

Ľuďom nerobí najmenší problém povedať či dve slová spolu súvisia [3]. Napríklad ak sa stretneme s dvojicou slov - „auto“ a „bicykel“, tak s istotou vieme povedať, že tieto slová spolu súvisia, obe totiž označujú dopravné prostriedky. Taktiež nám v bežnom živote nerobí najmenší problém posúdiť, či dvojica slov spolu súvisí viac alebo menej. Napríklad bez pochýb určíme, že slovo „bicykel“ má bližšie k „autu“ ako k „žehličke“. Ale je možné nejakým spôsobom priradiť tejto príbuznosti dvoch slov nejakú konkrétnu hodnotu? S použitím znalostí nahromadených počas procesu učenia sme vytvorili vzorec na výpočet sémantickej príbuznosti $R(x,y)$.

$$R(x, y) = \frac{|O(x) \bigcap O(y)|}{|O(x) \bigcup O(y)|} \quad (7)$$

kde *O(x)* označuje sémantické okolie symbolu *x* z lexikónu L. *O(x)* je taká množina symbolov, kde pre každý token *t* z lexikónu *L* existuje asociácia medzi tokenom *t* a symbolom *x*. Je teda množinou všetkých tokenov asociovaných s tokenom *x.* Ako je možné vidieť na vzorci 9, asociácie sa vytvárajú až potom, ako signifikancia prekročí prahovú hodnotu *tresh.*

$$L = \{t_1, t_2, t_3, ..., t_n\} \quad (8)$$

$$O(x) = \{ t \in L \,|\, S(x, t) > tresh \} \quad (9)$$

Lexikón *L* je množina všetkých známych tokenov. Vzorec 7 je podielom mohutnosti množiny spoločných vlastností zdieľanými oboma symbolmi *x* a *y* a mohutnosti množiny zjednotenia vlastností oboch symbolov. Keďže tieto vlastnosti sa viažu na fascikle, ktorých sme použili pri učení viac, je potrebné na to prihliadať aj pri výpočte koncovej hodnoty:

$$R(x, y)_{int} = \sum v_i \cdot \frac{|O_i(x) \bigcap O_i(y)|}{|O_i(x) \bigcup O_i(y)|} \quad (10)$$

$R(x,y)_{int}$ reprezentuje váhovanú hodnotu príbuznosti symbolov *x* a *y*. Váha priradená i.-temu fasciklu označujeme v_i.

3. Výpočet podobnosti

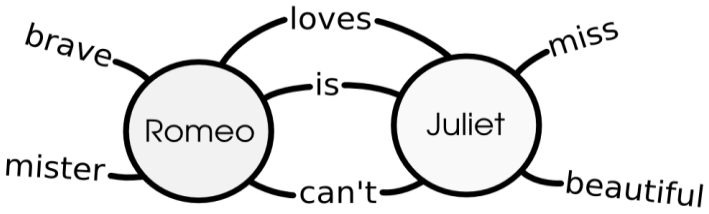
Výpočet podobnosti (príbuznosti) je podrobnejšie popísaný v stati 2.2. Vzhľadom na to, že sa jedná o veľmi dôležitú časť našich experimentov, je vhodné si celý proces výpočtu sémantickej podobnosti ukázať na príklade.

Príklad:
Dané sú tri tokeny „Romeo“, „Juliet“ a „Poison“, ktorých asociatívne okolie poznáme. Chceme vypočítať sémantickú podobnosť tokenu „Romeo“ k tokenu „Juliet“ a k tokenu „Poison“. Asociácie sú definované vo fasciklooch F_{-1} a F_1 takto:

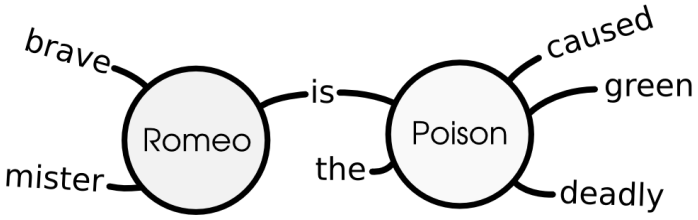
Romeo → brave (-1), mister (-1), loves (1), is (1), can't (1)
Juliet → beautiful (-1), miss (-1), loves (1), is (1), can't (1)
Posion → the (-1), is (1), caused (1), green (1), deadly (-1)

Pre zjednodušenie výpočtu sme do výpočtu zahrnuli len 2 fascikle F_1 a F_{-1}. Týmto fasciklom sme priradili váhy o veľkosti 1. Počet asociácií medzi „Romeo“ a „Juliet“

je 3, ale počet všetkých asociácií týchto tokenov je 7. Použitím vzorca 10 dostaneme číselné vyjadrenie príbuznosti 3/7 = 0.42.



Obr. 1. Podobnosť pojmov „Romeo“ a „Juliet“



Obr. 2. Podobnosť slov „Romeo“ a „Poison“

Číselné vyjadrenie príbuznosti medzi "Romeo" a "Poison" je 1/7 = 0.14.
S prihliadnutím na teóriu asociatívneho učenia pojmov je zrejmé, že koncept „Romeo“ je podobnejší konceptu „Juliet“ ako konceptu „Poison“.

4. Experimenty

4.1. Korpus nad umelou gramatikou

Na otestovanie našej metódy sme vygenerovali 10 000 viet nad umelou gramatikou. Na vytvorenie korpusu sme použili program The Simple Language Generator (SLG), ktorý po definícii danej umelej gramatiky poskytuje možnosť tvorby aj relatívne zložitejších vetných konštrukcií. V definícii gramatiky, ktorá poslúžila ako vstup do SLG, sme použili 2 členy (a, the), 7 prídavných mien, 10 podstatných mien a 18 slovies. Urývok z vygenerovaného korpusu vyzerá napríklad takto:

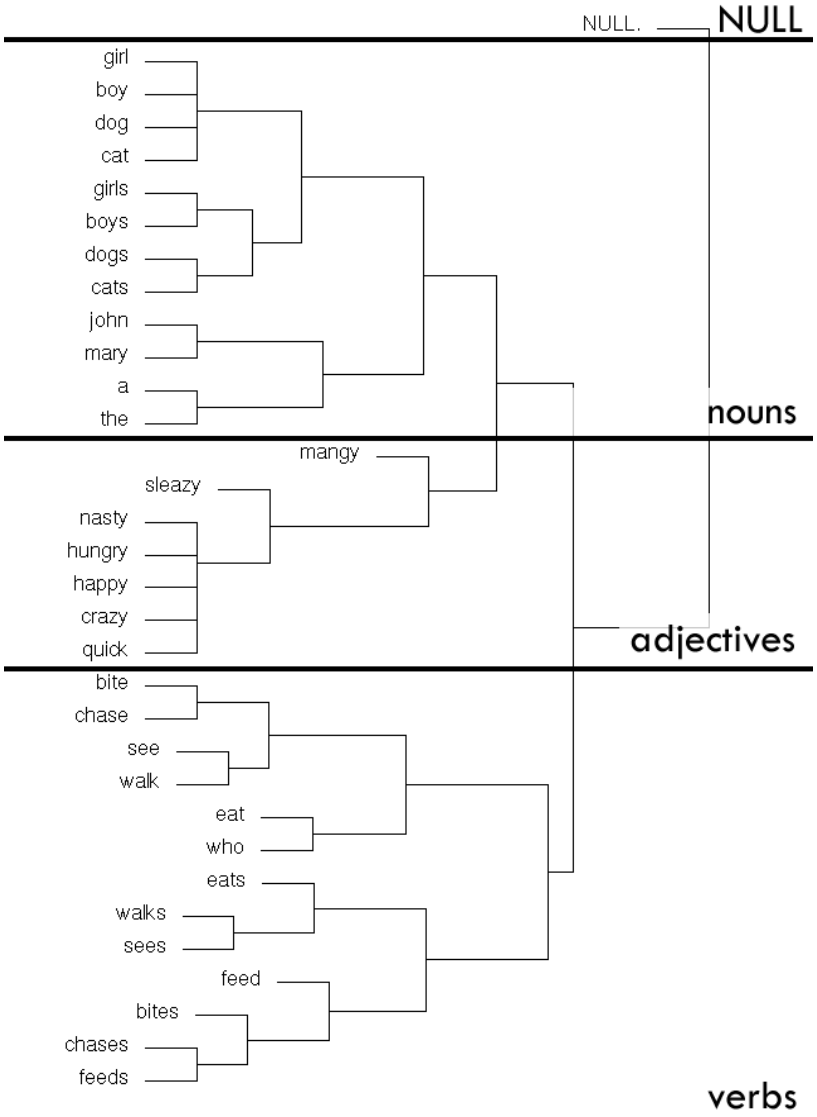
…John walks. mangy dogs bite. a mangy cat walks. the nasty dog bites the girls who walk. the boy who eats walks. the boy who walks hungry dogs who chase John walks a cat who eats. the dogs who Mary walks bark. a crazy cat chases John. girls walk

hungry dogs who chase a cat. John who feeds Mary who cats who the cats see bite feeds dogs…

Na tomto korpuse sme spustili učenie a na naučených znalostných bázach (pri prahu *tresh* = 5) sme následne previedli zhlukovanie pre všetky symboly z Lexikónu. Potom sme vypočítali maticu podobnosti.
Matica podobnosti *S* je taká matica, kde každý prvok $S_{i,j}$ je podobnosť tokenov *i* a *j*. Počíta sa pomocou vzorca (11). Váhy sú uložené v štyroch fascikloch $F_1,\ldots, F_4$ a štyroch inverzných fascikloch $F_{-1},\ldots,F_{-4}$, pričom indexy označujú kontextovú vzdialenosť. Váhy použité vo vzorci (10) boli $v_1 = v_{-1} = 0.4$, $v_2 = v_{-2} = 0.3$, $v_3 = v_{-3} = 0.2$ a $v_4 = v_{-4} = 0.1$. Tieto hodnoty boli použité pri oboch experimentoch. Presný vzorec je odvodený od vzorca (10) :

$$S_{i,j} = \sum_{\substack{k=-4 \\ k\neq 0}}^{4} v_k . \frac{|O_k(i) \bigcap O_k(j)|}{|O_k(i) \bigcup O_k(j)|} \qquad (11)$$

Táto matica poslúžila ako vstup do hierarchického zhlukovacieho algoritmu *hclust* vo výpočtovom programovacom jazyku R. Táto funkcia prevedie hierarchickú zhlukovaciu analýzu pre *n* objektov využívajúc pri tom maticu rozdielností. Na začiatku je každý objekt priradený do svojho vlastného zhluku. Algoritmus pracuje iteratívne a v každej iterácii spojí dva najpodobnejšie zhluky, až sa dopracuje k jediného zhluku. V každej iterácii sú vzdialenosti medzi zhlukmi prepočítavané pomocou Lance-Williamsového vzorca aktualizácie rozdielnosti [5] s prihliadnutím na použitú metódu zhlukovania. My sme v našich experimentoch použili metódu úplnú.



Obr. 3. Dendrogram vypočítaný nad korpusom z umelej gramatiky

Výsledný dendrogram je znázornený na obrázku 3.

Z obrázku je zrejmé, že algoritmus pozhlukoval tokeny do morfologických skupín s 99% úspešnosťou. Jedinú výnimku tvoria členy - tokeny „a“ a „the“, ktoré sú ale v texte veľmi frekventované, čiže si vytvorili mnoho asociácií. To by mohla vyriešiť úprava funkcie na výpočet vzájomnej signifikancie pre veľmi frekventované tokeny.

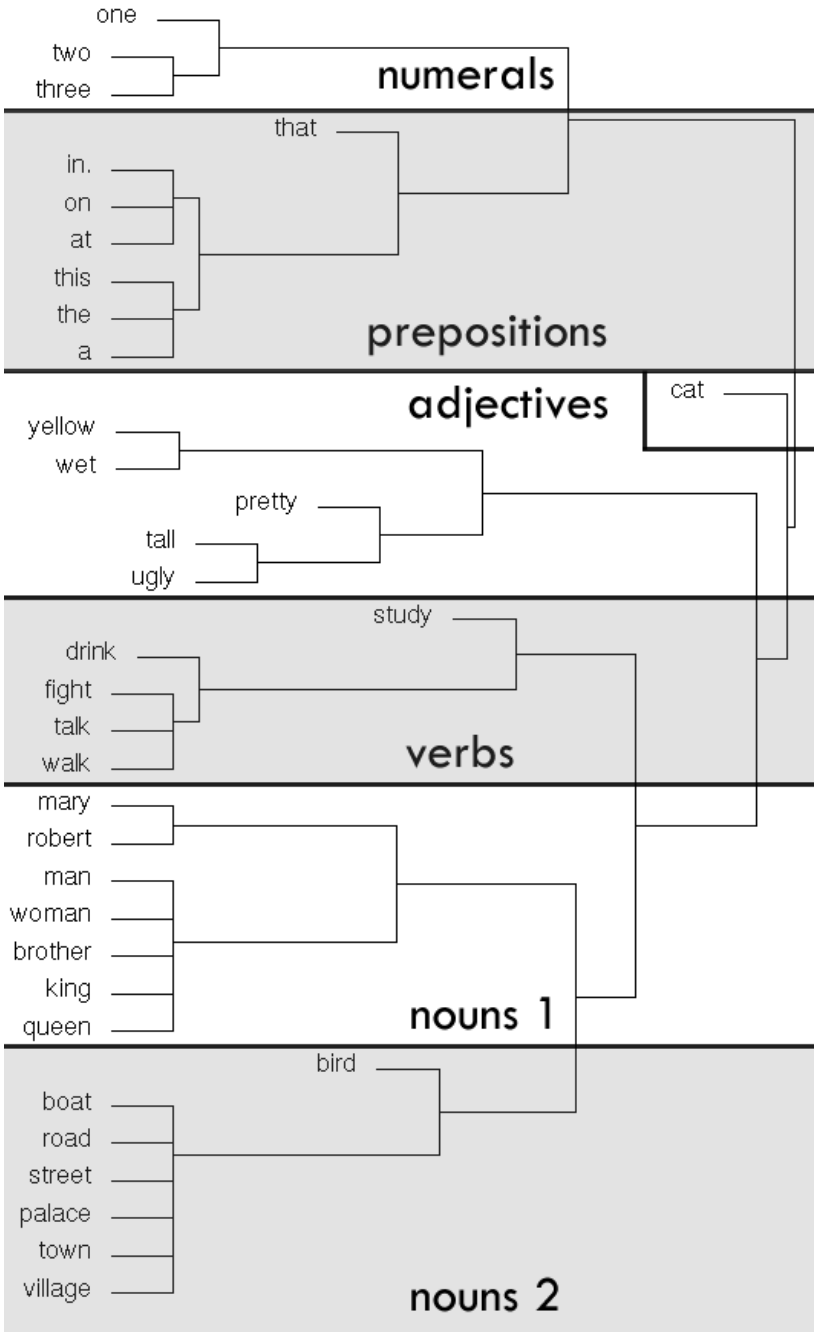
4.2. Korpus nad prirodzeným jazykom

Korpus nad prirodzeným jazykom bol získaný z projektu Gutenberg (www.gutenberg.com) a pozostával z náhodne vybraných publikácií v anglickom jazyku. Korpus mal veľkosť 200MB a skladal sa z čisto textových súborov.

Ukážka z korpusu nad prirodzeným jazykom:

…But at the foot of this cliff grew a tree, gnarled and stunted, the which, as Beltane watched, Black Roger began to climb, until, being some ten feet from the ground, he, reaching out and seizing a thick vine that grew upon the rock, stepped from the tree and vanished into the face of the cliff. But in a moment the leaves were parted and Roger looked forth, beckoning Beltane to follow. So, having climbed the tree, Beltane in turn seized hold upon the vine, and stumbling amid the leaves, found himself on his knees within a small cave, where Roger's hand met his…

Proces učenia prebiehal nad lexikónom o veľkosti 3500 tokenov. Táto veľkosť bola určená experimentálne, keďže 3500 tokenov pokrývalo 90% najfrekventovanejších tokenov zo všetkých tokenov, ktoré sa v texte vyskytli. Ďalšie zväčšovanie lexikónu už neprinášalo významnú zmenu v rozpoznávaní nových slov. Počas procesu učenia sme taktiež zaznamenávali nárast počtu vytvorených asociácií. Približne po tom, ako bolo spracovaného 70% korpusu, sa nárast asociácií skoro zastavil. Toto je pre nás dôležitý merateľ kvality naučenia korpusu, a v tejto fáze je bezpečné proces učenia zastaviť.
Na demonštráciu úspešného procesu učenia sme opäť zostavili maticu podobností (opäť pri prahu *tresh* = 5) nad malou skupinou (30) ručne vybraných tokenov. Tento počet síce neposkytuje úplne objektívny pohľad na úspešnosť a kvality našej metódy, ale slúži skôr ako malá ukážka práce nad reálnym korpusom a motivácia do budúcej výskumnej činnosti. Rovnakým spôsobom ako v experimente s umelou gramatikou sme vytvorili dendrogram zobrazený na obr. 4.



Obr. 4. Dendrogram vypočítaný nad korpusom z prirodzeného jazyka

5. Záver

Naša metóda dosiahla veľmi dobré výsledky pri zhlukovaní tokenov do morfologických skupín nad umelou gramatikou. Zlé priradenie členov anglického jazyka by nás malo nasmerovať k úprave vzorcov, aby sa vedeli lepšie vysporiadať s veľmi frekventovanými tokenmi. Na výslednú hierarchiu tokenov má ale veľký dopad aj správne nastavenie základného parametra – prahu. Experimenty nad korpusom z prirodzeného jazyka priniesli veľmi sľubné výsledky. Naším ďalším cieľom pri budovaní ontológií bude priraďovanie konceptov k tokenom. Keďže sme schopní vytvárať zhluky sémanticky blízkych tokenov, ďalším krokom by mohlo byť pomenovanie týchto zhlukov konceptami. Tieto koncepty by mohli byť vybrané priamo z daných zhlukov a mohli by byť reprezentované ako množiny synoným podobné tým z WordNetu [1].

Poďakovanie
Práca prezentovaná v tomto príspevku vznikla za podpory Vedeckej grantovej agentúry Ministerstva školstva SR a Slovenskej akadémie vied (VEGA) v rámci projektu č.1/4074/07 s názvom "Metódy anotovania, vyhľadávania, tvorby a sprístupňovania znalostí s využitím metadát pre sémantický popis znalostí".

Referencie

1. Fellbaum C., WordNet: An Electronic Lexical Database, The MIT Press, Cambridge, MA, 1998.
2. Hebb D.O. ,*The Organization of Behavior*. ,John Wiley, New York, USA, 1949
3. Hecht-Nielsen R., A Theory of Cerebral Cortex, Proceedings of the International Conference on Neural Information Processing (ICONIP98), 1998
4. Kende R.: Ontology Enabled Information Retrieval, Dissertation Thesis, University of Technology in Kosice, Slovakia, 2006
5. Lance G.N., Williams W.T. : A general theory of classificatory sorting strategies 1. Hierarchical systems., Computer Journal, Vol. 9, pp. 373-380, 1967
6. Rockai V.: Mining of Concepts and Semantic Relations from Texts in Natural Language, Diploma Thesis, University of Technology in Kosice, Slovakia, 2005

Annotation:

Associative learning of concepts

For humans it is such an easy task to determine whether two words are related to each other. It‘s even easy for us to determine the quantity of their relatedness. For example if we took two words „dog“ and „cat“ we should see, that they both are pointing to common pets. And sure we know, that the word „cat“ is more related to the word „dog“ than to the word „radioactivity“. In this article, we are introducing a method for computing the relatedness of word pairs.