Automatické generovanie slov patriacich do jednej syntakticko-sémantickej triedy

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 spôsob automatického generovania syntakticky a sémanticky podobných slov.

Keywords: ontológia, koncept, sémantická podobnosť, asociatívne učenie, synonymá, synsety

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 automatické generovanie slov patriacich do jednej syntakticko-sémantickej triedy.

Náš prístup je založený na neuro-fyziologickom modeli spracovania talamo-kortikálnej informácie od R. Hech-Nielsena [3]. 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. 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. 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 nezávislé 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 násobok vzájomnej pravdepodobnosti symbolov (diskrétnych premenných) *i, j* a pravdepodobností ich výskytu, kde *i* a *j* sú nezávislé.

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

Ak sa dva symboly vyskytujú v spoločnom kontexte náhodne, bude reprezentácia takáto *S(i,j)* < 1. Ak *S(i,j)* > 1, tak budú tokeny považované za asociované. To bude naším hlavným parametrom počas procesu učenia. Tento parameter – prah (*treshold*) môže byť využitý pri filtrovaní viac nenáhodných výskytov. Váha asociácie medzi dvoma symbolmi *i* a *j* je daná vzorcom:

$$w(i, j) = \frac{p(i, j)}{p(j)} \quad (3)$$

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. 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. Váhy v tomto prípade iba znázorňujú, že asociácia bola vytvorená.

2.2 Konsenzus

Konsenzus je naším základným nástrojom na rozpoznávanie fráz. V našej práci je fráza definovaná ako „syntakticky“ validný vstup rozpoznaný systémom. Výpočet sa vykonáva nad oknom (niekoľkých tokenov) a je vlastne hľadaním odpovede, ktorá najviac potvrdzuje najslabšiu hypotézu.
Je dané okno *n* tokenov $t_1, t_2, \ldots, t_n$. Maximálna kontextová vzdialenosť v takom okne bude *n-1*. Posledný token v okne, t_n, nazvime dotazovacím tokenom. Všetky ostatné tokeny potom hlasujú cez relevantné fascikle (F_1, F_2, …, $F_{(n-1)}$) určené ich kontextovou vzdialenosťou a vytvárajú množiny odpovedí M_1, M_2, …, $M_{(n-1)}$. Prienik týchto množín potom tvorí množinu možných odpovedí v danom okne a nazývame ho „plná množina konsezov**“:**

$$t_n \in M; M = \{M_1 \cap M_2 \cap \ldots \cap M_{(n-1)}\} \quad (4)$$

Množina *M* obsahuje *m* tokenov m_1, m_2,…, m_m, ktoré sú asociované s každým tokenom v okne cez príslušnú kontextovú vzdialenosť. Tieto asociácie ale majú obvykle rozdielne váhy. Aby sme podporili najslabšiu hypotézu musíme vyjadriť silu konsenzu $s(m_j)$:

$$s(m_j) = \min(F_1(t_1, m_j), F_2(t_2, m_j), \ldots, F_{(n-1)}(t_{(n-1)}, m_j)); j = 1,2,\ldots,m \quad (5)$$

Odpoveď, ktorá potom najviac podporuje najslabšiu hypotézu je potom t_{answer}:

$$t_{answer} = m_j ; s(m_j) = \min(s(m_1), s(m_2), \ldots, s(m_j)) \quad (6)$$

Pri učení sme ale použili okrem dopredných fasciklov aj fascikle. S inverznými fasciklami je možné vypočítať konsenzus pre dotazovaný token umiestnený na akomkoľvek políčku okna.

2.3 Výtváranie synoným

V tejto práci sú synonymá definované ako tokeny prislúchajúce do nejakej konkrétnej triedy. Názov „synonymum“ bol podľa nás najbližšou aproximáciou medzi existujúcimi názvami. Vstupom do algoritmu na výpočet synoným je nami zvolený známy token. Výstupom algoritmu je množina tokenov, ktoré sú k tokenu zadanému syntakticky blízke, ale rozšírené aj o tokeny z jedného sémantického poľa, čo implicitne definuje samotnú triedu.

Algoritmus na výpočet množiny synoným:
1. Pre zadaný token T, vyber všetky asociácie z fasciklov F_{-2}, F_{-1}, F_1, F_2 a vlož ich do množín M_{-2}, M_{-1}, M_1, M_2.
2. Nájdi úplnú množinu konsezov za dotázaný token T pre všetky kombinácie M-2 x M-1 x T x M1 x M2. Výsledok pridaj do množiny M.
3. V množine M sa nachádza množina všetkých synoným k dotázanému tokenu.

Príklady synoným:

Systém sme učili nad množinou anglických textov pozostávajúcich z náhodne vybraných titulov beletrie z elektronickej knižnice guternberg.org. Korpus pozostával z 1 GB textu v anglickom jazyku. Veľkosť lexikónu bola 3000 slov a prah pre jednotlivé príklady bol použitý v rozsahu 6.0-10.0 vzhľadom na mohutnosť výslednej množiny odpovedí s váhou konsenzu väčšou ako 0.001.

moon → star, stars, moon, lights, bright
white → brown, grey, black, white, red, yellow, blue, silk, gray, green, purple
william → joseph, james, william, edward, smith, thomas
helen → polly, laura
tree → trees, plants, tree, cloud, plant, pine, branches, green, grass
song → song, chorus, voices, sounds
two → two, fifteen, eight, twelve, seven, three, forty, twenty, six, four, five, thirty

3. Záver

Naša metóda dosiahla sľubné výsledky pri generovaní slov patriacich do jednej syntakticko-sémantickej triedy. Pri použití kvalitnejšieho korpusu, zväčšeniu lexikónu a ďalšom doladení algoritmu by mohol poslúžiť na generovanie množín konceptov pre jednotlivé pojmy. Vytvorením množín synoným pre každé slovo lexikónu by sme mohli získať vstup do ďalších zhlukovacích algoritmov. 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