Jelentések gyakoriságának vizsgálata a Magyar WordNet-ben

Kiss Márton[1], Vincze Veronika[1], Alexin Zoltán[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
H-6720 Szeged, Árpád tér 2.
{mkiss, vinczev}@inf.u-szeged.hu
[2] Szegedi Tudományegyetem, Szoftverfejlesztés Tanszék
H-6720 Szeged, Árpád tér 2.
alexin@inf.u-szeged.hu

Kivonat: A WordNet strukturális felépítését és a Google keresőprogram szolgáltatásait felhasználva olyan kísérletet hajtottunk végre, amely vizsgálja a WordNetben előforduló szavak jelentéseinek gyakoriságát. A vizsgált szó jelentéseit a hiponímia – hipernímia relációkban lévő synsetek felhasználásával különbözteti meg (kiegészíti ezekkel a szavakkal a keresőkifejezést) és tárolja a Google által visszaadott becsült előfordulási számot. A megkülönböztetés eredményeképpen megállapítható, hogy egy adott jelentés relatív gyakorisága az összes jelentés előfordulására nézve. A kísérlet eredményeit összehasonlítottuk a SZTE Informatikai Tanszékcsoport által épített WSD korpuszban található jelentésgyakoriságokkal. E munkálatok fontos szerepet töltenek be egy magyar nyelvű jelentés-egyértelműsítő szoftver készítésében.

1 Bevezetés

Adott szó jelentései előfordulási arányainak meghatározásához a jelentés hiponímia és hipernímia relációkkal hivatkozott synseteket használtuk fel. Az alapötlet az volt, hogy egy jelentést meghatároz, ha a WordNet fastruktúrájában közeli (közvetlenül alatta vagy felette) elhelyezkedő synsetekkel fordul elő együtt egy weboldalon, egy dokumentumban.

2 A szópárok, kifejezések lekérdezése

A kutatás kezdetekor megvizsgáltuk azokat a módszereket, melyekkel nagy mennyiségű (napi több ezer, esetleg több tízezer) keresési eredményt lehet lekérdezni a Googletől. Négy megoldást vizsgáltunk, mely a Google kereső által visszaadott becsült találati számokat (ERC_{kif}) kéri le: HTML protokoll feletti lekérés, Google SOAP API, Google AJAX API, Google AJAX API használata HTTP protokoll felett. Erre az összehasonlításra azért volt szükség, mert a Google nem ad pontos találati számot, csak egy becslést közöl és ráadásul ez a becslés a különböző technikai megoldások-

ban sem egyezik meg. Ezeket a lehetőségeket összehasonlítottuk és kiválasztottuk a megfelelőt.

3 A jelentések gyakoriságának meghatározása

Meghatároztuk az A_{wi} tulajdonsághalmazokat, mellyel elkülöníthetünk egy adott jelentést. A w szóhoz tartozó A_{wi} tulajdonsághalmaz i: a WordNetben is használt jelentésindex; I: adott szó összes jelentésének halmaza; $i \in I$. Az A_{wi} *az* i jelentéséhez tartozó hiperním (w_{i_hip}) és hiponím (w_{i_hyp}) szavak, szókapcsolatok halmaza. Azzal a megkötéssel, hogy azon szavak vagy kifejezések, melyek a vizsgált w szó, más n jelentéseinél ($n \in I$ és $n \neq i$) is előfordultak, nem vettük figyelembe, tehát a közös ős- vagy gyerekhivatkozásokat kihagytuk.

A tulajdonsághalmazok meghatározása után lekérdeztük w szó összes $i \in I$ jelentését a jelentésekhez meghatározott A_{wi} tulajdonsághalmazban található összes szóval. A lekérdezésben használt kifejezés felépítése, tehát:

$$kif_n = w_i + n, \text{ ahol } n \in A_{wi} \quad (2)$$

Majd adott i jelentéshez tartozó becsült előfordulási számokat összegezzük:

$$w_{i\,ERC} = \sum_{n \in A_{wi}} \mathrm{ERC}_{\mathrm{kifn}} \quad (2)$$

Ezen értékek figyelembevételével már könnyen meghatározható volt adott w_i jelentés relatív gyakorisága.

4 A WSD korpusz és a jelentésgyakoriságok összehasonlítása

A WSD korpusz 39 synset szemantikus annotációját tartalmazza és minden synsethez 300-350 előfordulás található. A 39 synset jelentéseihez tartozó, a Google segítségével kapott relatív gyakoriságokat összehasonlítottuk a WSD korpuszban található gyakoriságokkal. A WSD korpusz a WordNet jelentéseinek felhasználásával készült, ugyanakkor a jelentések nem fedték egymást egy az egyben, így az összehasonlítás nehézkes munka volt, mert minden szóalak esetében meg kellett feleltetni a WordNet és a WSD korpusz jelentéseit egymásnak.

Az így összepárosított jelentések gyakoriságát vizsgáltuk a WSD korpuszban és a Google-ben. Az eredmények időnként egybevágtak a két módszert tekintve (pl. *század, jár*), más esetekben azonban a jelentésgyakoriságok éles eltérést mutattak (pl. *kormány*, *program*). Utóbbi jelenség valószínűleg a WSD korpusz tematikai egyöntetűségének köszönhető.

1. táblázat: A WSD korpusz és a kutatási eredményeink összehasonlítása.

szó:jelentés	korpusz %	Google %	szó:jelentés	kor-pusz %	Google %
jár: 3 volt	6,5	10.3	kormány: 1 irányítóeszköz	0,3	41,3
jár: 4 tánc	0,7	8.6	kormány: 3 biciklikormány	0,0	2,0
jár: 5 valahogyan	17,7	5.8	kormány: 2 szerv	98,4	56,7
jár: 7 működik	0,3	2.9	kormány: 3 egyéb	1,4	0
jár: 8 előfizető	0	10.4	program: 1 szabadidő	7,0	64,3
jár: 9 valakinek	12,3	30.0	program: 2 célok	74,8	0,2
jár: 10_együtt	18,1	8.6	program: 3_műsor	1,6	26,0
jár: 11 tartozik	1,3	18.9	program: 4 számítógép	16,5	9,5
jár: 12 egyéb	25,8	0	század_n_1 évszázad	99.7	97,0
jár: 13 valahol	1,0	1.3	század_n_2 katonai	0.3	3,0
jár: 14 valamiért	1,0	0			
jár: 15 közeledik	3,8	0			
jár: 16 valakivel	0,3	4.2			

5 Tervek, a kutatás folytatása

Kutatásaink célja egy magyar jelentés-egyértelműsítő rendszer előkészítése. Ehhez azonban szokásos módszerekkel tanuló korpuszt készíteni, amelyben az egyes jelentések megfelelő számban fordulnak elő, nem lehetséges. Mindenképpen olyan technikai megoldásokra van szükség, melyek az elérhető legnagyobb korpuszon (interneten) megtalálható dokumentumok alapján becslik meg az előfordulások gyakoriságát.

Hivatkozások

1. Szarvas György, Hatvani Csaba, Szauter Dóra, Almási Attila, Vincze Veronika, Csirik János: Magyar jelentés-egyértelműsített korpusz, Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Magyarország (2007), 158-165
2. Agirre et al.: Personalizing Page Rank for Word Sense Disambiguation, The First KYOTO Workshop, Amsterdam, Netherlands (2009)
3. Gabrilovich, Evgeniy, Markovitch, Shaul: Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, Israel, Haifa
4. Strube, Michael, Ponzetto, Simone Paolo: Genetic WikiRelate! Computing Semantic Relatedness Using Wikipedia, Heidelberg, Germany
5. Miháltz M.: Towards A Hybrid Approach To Word-sense Disambiguation In Machine Translation. In Proceeding Modern Approaches in Translation Technlologies Workshop at RANLP-2005, Borovets, Bulgaria (2005)