Melléknevek szűk szemantikai osztályainak detekciója a Magyar Nemzeti Szövegtárban jelentés-egyértelműsítés céljából

Héja Enikő[1], Takács Dávid[1]

[1] MTA Nyelvtudományi Intézet
{eheja, takdavid}@nytud.hu

A jelentés-egyértelműsítés a hazai és nemzetközi nyelvtechnológia egyik központi problémája. Számos alkalmazás (pl. információkinyerés, gépi fordítás) számára kiemelkedő jelentőségű. A jelentés-egyértelműsítés két részfeladatra bontható: (1) megfelelő jelentéstár kiválasztása, amelynek elemei hozzárendelhetőek a szövegekben szereplő tokenekhez, (2) megfelelő algoritmus kiválasztása, amely ezt a hozzárendelést elvégzi. Az annotátorok közötti egyetértést mérő vizsgálatok bizonyítják (l. [3, 4]), hogy a már létező, nem kifejezetten jelentés-egyértelműsítés céljából kialakított egynyelvű adatbázisok (pl. *Petit Larousse*, *Magyar Értelmező Kéziszótár*) alapján a jelentés-egyértelműsítés még az emberek számára is nehéz vagy megoldhatatlan feladat. Tehát az intuíción alapuló, nem jelentés-egyértelműsítés céljából létrehozott jelentéstárak alkalmatlanok a gépi jelentés-egyértelműsítésre.

Mivel kontextuális információ alapján tudjuk csak automatikusan lehorgonyozni a megfelelő jelentést, a jelentéstár kialakításánál is célszerű kizárólag a kontextuális információra támaszkodni. Ez a megközelítés egybecseng a jelentés disztribúciós felfogásával, amit Firth [2] így fogalmazott meg: "*You shall know a word by the company it keeps*". Az általunk javasolt módszer fontos tulajdonsága, hogy kizárólag disztribúciós információt vesz figyelembe, a jelentéstár kialakításánál az emberi intuíciót figyelmen kívül hagyja, és a jelentéstár szerkezetére vonatkozóan semmilyen előzetes megkötést nem teszünk.

Az általunk végzett kutatás célja, hogy felügyelet nélküli tanulással egy melléknevet tartalmazó jelentéstárat készítsünk a Magyar Nemzeti Szövegtár adatai alapján. [1] klikk-klaszterezési (*clique-based clustering*) eljárását alkalmazva az MNSZ-ben annotált főnév-melléknév kapcsolatokra azt várjuk, hogy a létrejövő klaszterek a melléknevek szűk szemantikai osztályaival (pl. színnevek) esnek egybe. Ezek a kontextus alapján létrejövő klaszterek képezhetik a jelentés-egyértelműsítő rendszer jelentéstár-komponensét.

Az eljárás az alábbi lépésekből áll: (1) az annotált korpusz alapján felépítjük a melléknevek disztribúciós mátrixát, ahol minden melléknevet a módosított főnevek halmazával jellemzünk. (2) Ebből a mátrixból egy távolsági mérték alkalmazásával meghatározzuk az egyes melléknevek közötti kontextuális távolságot. (3) Egy megfelelő vágási paraméter alkalmazása után a Bron-Kerbosch algoritmussal meghatározzuk a létrejövő gráf teljes részgráfjait, vagyis klikkjeit. (4) Az így létrejött, jellemzően

kis elemszámú klikkeket egy klaszterezési eljárással összeolvasztjuk, aminek eredményeképpen az egyértelműsítés számára megfelelő finomságú felosztást kapunk.

A klikk-klaszterezés járulékos előnye, hogy a lépések során az egyes melléknevek a különböző jelentéseik szerint egyszerre több klaszterben is szerepelni fognak. Emellett mindvégig megőrizzük a kontextuális információkat, így fölépíthetünk egy, a melléknév-kontextus párok halmazát a jelentések halmazára leképező függvényt, amelyet közvetlenül használhatunk a jelentés-egyértelműsítés során.

Jelen kutatás célja annak vizsgálata, hogy felügyelet nélküli tanulással a fent javasolt módszerrel létrehozható-e egy olyan jelentéstár, amely a magyar melléknevek jelentés-egyértelműsítésének alapjául szolgálhat.

Bibliográfia

1. Ah-Pine, J., Jacquet, G.: Clique-Based Clustering for improving Named Entity Recognition systems. In: EACL 2009, 12th Conference of the European Chapter of the Association for Computational Linguistics. Athens, Greece (2009)
2. Firth, J. R.: Papers in Linguistics. Oxford University Press, London (1957) 1934–1951
3. Kuti, J., Héja, E., Sass, B.: Sense Disambiguation — "Ambiguous Sensation"? Evaluating Sense Inventories for verbal WSD in Hungarian. In: Proceedings of the LREC W22. Malta (2010)
4. Véronis, J.: Sense tagging: does it make sense? In: Wilson, A., Rayson, P. McEnery, T. (szerk.) Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech. Peter Lang, Frankfurt (2003)