Angol szavak szinonimáinak automatikus keresése

Dobó András

I. Programtervező informatikus MSc

Témavezetők: Dr. Alexin Zoltán egyetemi adjunktus, Dr. Csirik János egyetemi tanár

SZTE TTIK Informatikai Tanszékcsoport

Nagyméretű dokumentumhalmazból valamilyen információ szempontjából releváns dokumentumok kiválasztásának a leghatékonyabb módja egy keresőrendszer használata. Az elmúlt időszakban számos kutatás irányult a keresőrendszerek hatékonyságának növelésére. Az egyik gyakran alkalmazott hatékonyság növelő módszer a lekérdezés kiterjesztése, ami az adott lekérdezésben található szavakat további, a lekérdezésre releváns szavakkal, például az eredeti szavak szinonimáival, egészíti ki.

E dolgozat célja olyan statisztikai alapú módszerek bemutatása, melyek alkalmasak angol szavak szinonimáinak automatikus megtalálására. Az így meghatározott szinonimák felhasználhatók keresőrendszerek hatékonyságának növelésére a lekérdezés kiterjesztésének segítségével. Az algoritmusok alapötlete az, hogy a szinonim szavak szintaktikailag is hasonlóan viselkednek. A dolgozatban leírt módszerek minden szóhoz egy tulajdonsághalmazt vagy tulajdonság-vektort rendelnek a szó szintaktikai kapcsolatainak egy nagy nyelvi adatbázisban történő előfordulásai alapján. A szavak összehasonlítása e tulajdonságvektorok alapján, különféle hasonlósági mértékek alkalmazásával történik.

Az ismertetett módszereket először a Modell-Alapú Szemantikus Kereső Rendszer (MASZEKER) NTP projekt keresőrendszerében kívánjuk felhasználni. A tervek szerint ez egy szigetontológiára épülő keresőrendszer lesz, amely strukturált szemantikai előzetes tudást használ fel elsősorban technikai szabadalmakban való keresésre. A teszteléshez elsőként 1000 gyógyszerészeti témakörből vett angol nyelvű szabadalom szövegét használtam fel angol szakszavak és általános köznyelvi szavak szinonimáinak keresésére. Ezután a módszereket kipróbáltam nagyjából 9000, tíz különböző témakörből vett szabadalom, valamint a British National Corpus adatbázisán is. Az eredmények egy nagyméretű lexikai adatbázis segítségével kerültek kiértékelésre.

A keresőrendszerek fejlesztése mellett az automatikusan megtalált szinonimák hasznosak lehetnek különféle nyelvészeti kutatások szempontjából, illetve alkalmasak fogalomhálók készítésének támogatására is.