Jelentés-egyértelműsített szabadalmi korpusz

Nagy Ágoston, Almási Attila, Vincze Veronika

Szegedi Tudományegyetem, Informatikai Tanszékcsoport
Szeged, Árpád tér 2.
vizipal@gmail.com, {vinczev,nagyagoston}@inf.u-szeged.hu

Kivonat: A tanulmány célja, hogy megállapítsuk, hogy az angol nyelvű szabadalmakban milyen arányban fordulnak elő többjelentésű szavak, valamint azt, hogy ezek a valóságban hány különböző jelentéssel fordulnak elő e szövegekben. Kutatásaink során az A23K osztályba tartozó 60 szabadalmat tartalmazó korpuszunkban található szövegekre összpontosítunk. Előfeltételezéseink szerint a szakkifejezések és terminusok nagy része a főnév osztályba sorolható, ezek pedig adott doménen belül általában egyféleképpen használatosak. Az eredmények is azt igazolják, hogy a szabadalmakban kevesebb jelentés jelenik meg a gyakorlatban, mint amennyi a szótárakban található.

1 Bevezetés

Az ALL és a Szegedi Tudományegyetem egy közös projekt keretében vállalta egy szemantikus keresőrendszer kifejlesztését, amely elsődlegesen az angol és magyar nyelvű szabadalmakban való keresést célozza meg. A keresőrendszer hatékony működéséhez a szabadalmak morfológiai és szintaktikai elemzésén túl szükséges azok szemantikai feldolgozása is, melynek előfeltétele a szavak jelentésének előzetes meghatározása, azaz a jelentés-egyértelműsítés.

A tanulmány célja, hogy megállapítsuk, hogy az angol nyelvű szabadalmakban milyen arányban fordulnak elő többjelentésű szavak, valamint azt, hogy ezek a valóságban hány különböző jelentéssel fordulnak elő e szövegekben.

Cabré [1] alapján az az előfeltételezésünk, hogy a főnevek és igék a szabadalmakban általában csak egy jelentésben fordulnak elő, mivel ezek főleg terminusok, amelyeknek alapfeltétele, hogy lehetőleg csak egy fogalmat denotáljanak. Ettől függetlenül előfordulhat, hogy egy terminus több fogalmat jelöl, de egy doménen belül csak egyet, így ideális esetben a terminusok nem lehetnek poliszémek, csak homonímek.

2 A jelentés-egyértelműsítési feladat

A jelentés-egyértelműsítés egy szöveg adott szavának egy olyan meghatározással vagy jelentéssel történő párosítását jelenti, amely az adott szóhoz társítható más lehetséges jelentésektől élesen elkülönül. Így a feladat szükségszerűen két lépésből tevődik össze: (1) a vizsgált szöveg minden releváns szavának meg kell határozni a lehet-

séges jelentéseit, illetve (2) az adott szó minden egyes előfordulásához társítani kell a megfelelő jelentést. Az első lépésben leginkább előre megadott jelentésmeghatározásokat alkalmaznak, amelyek például a következőkből állhatnak:

- hétköznapi szótárakban megadott jelentések
- különféle szemantikai jegyek, kategóriák vagy kapcsolódó szavak (pl. szinonimák)
- kétnyelvű szótárakban megadott információk (idegen nyelvű megfelelők)

A második lépésben a szóalakok és jelentések összekapcsolása két fő információforrás alapján történhet meg:

- tág értelemben vett kontextus: különféle információt tartalmaz a szó szövegkörnyezetében, a diskurzusban stb.
- külső tudásforrások: lexikális, enciklopédikus tudás

A jelentés-egyértelműsítő eljárások hatókörük alapján és a jelentésmegkülönböztetés foka szerint két-két főbb csoportra oszthatók. Hatókör tekintetében a teljes szókincsre alkalmazható (all-words WSD) és előre megadott szóalakokon működő (lexical sample WSD) módszereket különböztethetünk meg, míg a jelentésmegkülönböztetés részletessége szerint aprólékos vagy finom (fine grained), illetve durva (coarse grained) szinteket különböztethetünk meg.

A *lexical sample* alapú módszer sokkal kevesebb előzetes munkát (pl. jelentésmeghatározások megalkotása) és időráfordítást igényel, mivel nem szükséges az adott korpusz összes többjelentésű elemének előzetes definiálása. Ezzel szemben az *all-words* módszer egy jóval nagyobb mértékű vállalkozás, amely akkor lehet hasznos, ha egy általános korpuszt kívánunk létrehozni, mert ebben az esetben jobban meg lehet figyelni, hogy milyen jelentés milyen szövegkörnyezetben fordul elő.

Durva jelentésmegkülönböztetés esetén nagyobb jelentésmezők, jelentésklaszterek jelennek meg. Ezek feldolgozhatósága egyszerűbb, és az egyértelműsítés a gépi tanuló számára – és egyben az emberi annotátor számára is – könnyebb. Finom jelentésmegkülönböztetés esetén viszont sokkal aprólékosabb különbségeket lehet kódolni, ami mindenképpen hasznos lehet bizonyos alkalmazásokban, mert specifikusabb dolgokra lehet rákeresni, de a korpusz elkészítése sokkal idő- és munkaerőigényesebb feladat. A túlzott jelentésmegkülönböztetés bizonyos esetekben még az emberi annotátorok számára is indokolatlannak tűnik, gyakoriak az eltérő annotációk, hiszen minél több a jelentés, annál nagyobb a tévesztés valószínűsége. Így, mind informatikai, mind pedig nyelvészeti szempontból 3-5 egymástól pontosan elkülöníthető jelentés felvétele tűnik a legmegfelelőbbnek, mert ezt mind az emberi annotátorok, mind pedig a különféle számítógépes algoritmusok számára is ideális működési hatékonyságot tesz lehetővé (lásd [6]).

3 Korpusz és módszer

Kutatásaink során az A23K osztályba tartozó 60 gyógyszerészeti és gyógyászati segédeszközöket leíró szabadalmakat tartalmazó korpuszunkban található szövegekre [7] összpontosítunk. Annak eldöntésére, hogy mely szónak hány jelentése van, a legújabb, 3.0-s Princeton WordNetet (PWN) használtuk [8]. Ebből adódóan az egyértelműsítést csak azokra a szavakra tudjuk elvégezni, amelyek ebben az ontológiában is szerepelnek, azaz főnevekre, igékre és melléknevekre. Noha a WordNet határozószavakat is tartalmaz, ezekkel nem foglalkoztunk, mert a határozószavak előfordulási aránya igen csekély a szövegekben, továbbá a szemantikus keresés szempontjából kis jelentőséggel bírnak. Mivel a PWN finom jelentésmegkülönböztetést alkalmaz, így a lehetséges jelentések száma szóalakonként magasnak mondható.

A többértelmű kifejezések kigyűjtését 60 szabadalmi főigénypontón végeztük el. Ezeket a főigénypontokat az Apache UIMA keretrendszerében az OpenNLP modullal mondatokra bontottuk és tokenizáltuk. Ezt követően a Stanford POS-tagger segítségével minden tokenhez hozzárendeltük annak szótövét és Penn Treebank szerinti szófaji kódját (pl. NNS többes számú főnév) [5]. Eztán kigyűjtöttük a korpuszban előforduló összes főnevet, igét és melléknevet, majd megnéztük, hogy a WordNetben ezen szavak többértelműek-e vagy sem. Ehhez a Javába is beilleszthető JAWS (Java API for WordNet Searching) alkalmazást [3] használtuk. Ezután a többértelmű szavakat a szövegkörnyezetükkel együtt elmentettük a SemEval és SensEval workshopokon [2] is használatos XML formátumba.

A korpusz annotálását két független nyelvész végezte a Sensetagger program segítségével. Azokat a szavakat egyértelműsítettük, amelyek legalább háromszor előfordultak a korpuszban, a későbbiekben azonban – hasonló elvek alapján – bővíthető az annotáció. 15 szó előfordulásait mindkét annotátor bejelölte, ezáltal lehetővé vált a korpusz konzisztenciaszintjének mérése. A szavakat szófajuk szerint annotáltuk, tehát például a *form* szó igei és főnévi jelentéseit egymástól teljesen elkülönítve kezeltük, a szófaji egyértelműsítő modul elemzésének megfelelően.

4 Eredmények

Ebben a fejezetben az elkészült korpusz statisztikáit és az elért eredményeket ismertetjük.

4.1 A jelentések eloszlása

A korpuszban található többértelmű főnevek, melléknevek és igék eloszlása az 1. táblázatban látható. Hangsúlyozzuk, hogy itt a többértelműséget pusztán a wordnetbeli jelentések alapján határoztuk meg, nem pedig a valós korpuszbeli eloszlások alapján.

1. táblázat: A WordNet alapján a szabadalmakban előforduló többértelmű szavak aránya szófajonként.

	Összes	Többértelmű	
Főnév	744	284	38,17%
Melléknév	310	115	37,1%
Ige	162	135	83,33%
Összes	1216	534	43,91%

A táblázat jól mutatja, hogy elméleti szinten leginkább a szabadalmak igéire jellemző a többértelműség.

Ezen listából azon szavakat annotáltuk kézzel, amelyek legalább háromszor fordultak elő a vizsgált korpuszban. Ezek konkrét száma szófaji lebontásban és az összesre kivetítve a 2. táblázat első oszlopában olvasható. A második oszlop mutatja az annotált szavak arányát az összes előforduló többértelmű szóhoz viszonyítva. A harmadik oszlop tartalmazza azon elemek számát, amelyek az annotáltak közül legalább két jelentéssel bírnak a szabadalmakban, végül az utolsó mutatja, hogy a korpuszban többértelmű szavak aránya mekkora az annotált szavak számához képest.

2. táblázat: Az annotált szavak aránya az összes többértelmű szó függvényében.

	Annotáltak száma	Annotáltak aránya az összes előforduló többértelmű szóhoz képest	Annotált és legalább kétértelmű szavak száma	Legalább kétértelmű szavak aránya az annotáltak közül
Főnév	164	57,74%	15	9,14%
Melléknév	52	45,22%	2	3,84%
Ige	69	51,11%	12	17,39%
Összes	285	53,37%	29	10,17%

A táblázatból jól látható, hogy az annotálás során a lehetséges többértelmű szavak kicsivel több mint a felét annotáltuk kézzel. A harmadik és a negyedik oszlopból kiderül, hogy az igék azok, amelyek a legnagyobb arányban bírnak több jelentéssel a szabadalmakban: ezen igék aránya 17,4%, míg a főneveknél ez az arány 9%, a mellékneveknél pedig 4%.

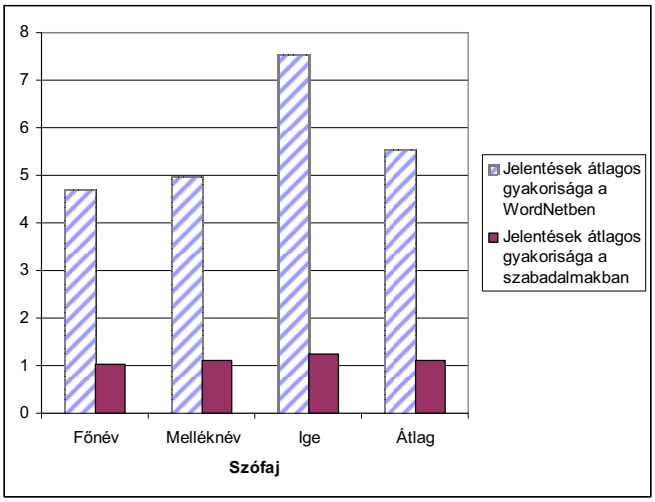
A vizsgált többértelmű szavak esetén megnéztük, hogy azok átlagosan hány jelentéssel fordultak elő mind a WordNetben, mind a szabadalmakban. A 3. táblázatban foglaljuk össze az átlagos jelentésszámot a különböző szófaji kategóriákra vonatkoztatva.

3. táblázat: Jelentések átlagos száma a WordNetben és a szabadalmakban.

	Jelentések átlagos gyakorisága a WordNetben	Jelentések átlagos gyakorisága a szabadalmakban
Főnév	4,7115	1,0385
Melléknév	4,9817	1,0976
Ige	7,5362	1,2319
Átlag	5,5509	1,1193

A 3. táblázatból jól látható, hogy a ténylegesen vizsgált és kézzel is annotált szavak esetében is az igék rendelkeznek a legtöbb jelentéssel a WordNetben, átlagban 7,5-del, míg a főnevek és a melléknevek jelentésének átlagos száma 5. A szabadalmak esetén azonban azt vehetjük észre, hogy a jelentések átlagos száma szófaji kategóriától függetlenül 1 körül van, és ez a szám az igéknél a legnagyobb, egészen pontosan 1,2319. Ez megerősíti azt a feltételezésünket, hogy a szabadalmakban nagyrészt terminusként fordulnak elő a kifejezések.

Az 1. ábra mutatja szófaji kategóriákra lebontva, hogy az adott szófaj esetén mennyi az átlagos jelentésszám a WordNetben (bal oszlop), illetve a szabadalmakban (jobb oldali oszlop).



1. ábra. Jelentések átlagos száma a WordNetben és a szabadalmakban.

Az igék között 4 darab háromértelmű (*form, reduce, make, have*) és 8 darab kétértelmű szó található. A *form* ige esetében az alábbi három jelentés figyelhető meg a WordNetben előforduló 7 jelentés közül a szabadalmakban:

4. táblázat: A *form* ige jelentései.

Jelentés száma	WordNetbeli jelentés	Példa a szabadalmakban
1	to compose or represent	
2	create (as an entity)	[…] adding to a second fluid bed dryer the fourth feed stream to **form** the granular detergent composition; […]
3	give shape or form to	[…] deforming the films to **form** a multiplicity of recesses […]
4	develop into a distinctive entity	
5	establish or impress firmly in the mind	
6	make something, usually for a specific function	A water resistant suntan gel capable of **forming** […] a water-resistant film on skin […]
7	assume a form or shape	

A wordnetbeli jelentések közül így kevesebb, mint fele használatos a szabadalmakban. Az ötös számmal ellátott jelentés például nagyon kis valószínűséggel fordulhatna elő akármilyen szabadalomban.

A szabadalmakban két jelentéssel rendelkező igék a következők: *provide, determine, combine, contain, comprise, treat, mix* és *produce*. A többi mind egy jelentéssel rendelkezik.

A melléknevek esetében kizárólag az *oral* és *lower* szó rendelkezett kettő jelentéssel a szabadalmakban, a többi mind egyjelentésű volt. Az első szó szabadalmakban előforduló két jelentését és a wordnetbéli jelentéseket az alábbi táblázat tartalmazza:

5. táblázat: Az *oral* szó jelentései.

Jelentés száma	WordNetbeli jelentés	Példa a szabadalmakban
1	of or relating to or affecting or for use in the mouth	A composition for treating diabetes to be taken in **oral** doses
2	of or involving the mouth or mouth region or the surface on which the mouth is located	tablet capable of being chewed or disintegrated in the **oral** cavity […]
3	a stage in psychosexual development when the child's interest is concentrated in the mouth; fixation at this stage is said to result in dependence, selfishness, and aggression	
4	using speech rather than writing	

A főnevek közül egyedül a *system* szónak volt kettőnél több jelentése a szabadalmakban, összesen 3 a wordnetbeli 9 helyett. Ez a három jelentés a következő volt: (1) *instrumentality that combines interrelated interacting artifacts designed to work as a coherent entity*, (2) *a group of independent but interrelated elements comprising a unified whole* és (3) *a procedure or process for obtaining an objective*. Ezen kívül 14 darab főnévnek volt legalább két jelentése a szabadalmakban.

A szabadalmakban előforduló jelentések aránya arra mutat rá, hogy noha a jelentés-egyértelműsítési feladatot finom megkülönböztetésként fogtuk fel, hiszen a WordNet alapján határoztuk meg a jelentéseket, a valóságban elégségesnek bizonyul a durva jelentésmegkülönböztetés, azaz általában 2-3 jelentéssel rendelkeznek a többértelmű szavak a szabadalmakban. Tapasztalataink azt is igazolják, hogy a gyógyszerészeti szabadalmak jelentés-egyértelműsítése nem igényli speciális gyógyszerészeti jelentéstár létrehozását, mivel egy általános célú jelentéstár (WordNet) is alkalmasnak bizonyult a feladatra.

4.2 Egyetértési ráta

A korpusz annotálását két független nyelvész végezte a Sensetagger program segítségével. Minden szófajból az öt leggyakoribb többértelmű szó előfordulásait mindkét annotátor egyértelműsítette, így mérhetővé vált az egyetértési ráta. A 6. táblázat mutatja a szófajonkénti és az összesített adatokat a mindkét annotátor által jelölt korpuszrészen.

6. táblázat: A két annotátor közötti egyetértési ráta.

	Előfordulás	Egyetértés
Főnév	211	96,68%
Ige	179	93,85%
Melléknév	62	100%
Összesen	452	96,08%

A 6. táblázat jól mutatja, hogy az annotátorok közti egyetértés igen magasfokúnak mondható. A szintén WordNet-jelentésekre épülő magyar nyelvű WSD-korpusz [6] egyetértési rátája 84,78%-os volt, amihez képest 11,4%-kal jobb teljesítményt értünk el a minta alapján. Ez arra enged következtetni, hogy szakszövegekben könnyebb feladat a jelentés-egyértelműsítés, hiszen egy adott doménen belül kisebb valószínűséggel használatosak a szavak többféle jelentésben (noha a *család* szó többértelmű, botanikai kontextusban szinte kizárólagosan a rendszertani kategóriát jelöli). Bár a magyar WSD-korpusz is homogén szövegeket tartalmaz (HVG-cikkek), azok nyelvezete és tematikája mégsem annyira kötött, mint a szabadalmaké (vö. [4]).

Különösen a melléknevek egyértelműsítése bizonyult könnyű feladatnak, noha itt számottevően kevesebb példát kellett címkézni, mint a főnevek és igék esetében. Meg kell tovább említeni, hogy a melléknevek nagy többsége egyjelentésűként fordult elő a szabadalmakban, ami tovább könnyítette az annotálást. Az egyértelműsítésre kiválasztott mintában a *form* ige bizonyult a legnehezebbnek: itt az annotátorok pusztán 52,6%-ban értettek egyet. Ennek valószínűleg az lehet az oka, hogy két jelentést (’lét-

rehoz' és 'valamilyen célra létrehoz') egymáshoz közel állónak, így nehezen megkülönböztethetőnek ítéltek az annotátorok. Az eltérően annotált esetek nagy része e két jelentést érintette.

5 Összegzés és további célok

Tanulmányunkban bemutattuk a gyógyszerészeti szabadalmakat tartalmazó jelentés-egyértelműsített korpuszunkat. A wordnetbeli és a korpuszban előforduló jelentések aránya azt tükrözi, hogy szakszövegekben, jelesül a szabadalmakban kevesebb jelentés jelenik meg a gyakorlatban is, mint ahogy azt az adatbázis alapján várhatnánk. Ez némileg megkönnyíti mind az annotátorok, mind a gépi egyértelműsítés feladatát.

Az elkészült korpuszt a jövőben szeretnénk jelentés-egyértelműsítő algoritmusok tesztelésére használni, melyek beépülnek majd a szemantikus keresőbe.

Köszönetnyilvánítás

A kutatás – részben – a MASZEKER kódnevű projekt keretében a Nemzeti Fejlesztési Ügynökség, illetve a TÁMOP-4.2.1/B-09/1/KONV-2010-0005 jelű projekt keretében az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap és az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Cabré, M. T.: Terminology. Theory, methods and applications. John Benjamins, Philadelphia PA (1998)
2. Erk, K., Strapparava, C. (eds.): Proceedings of the 5th International Workshop on Semantic Evaluation. Association for Computational Linguistics, Uppsala, Sweden, July (2010)
3. Java API for WordNet Searching (JAWS), http://lyle.smu.edu/~tspell/jaws/index.html
4. Osenga, K.: Linguistics and patent claim construction. Rutgers Law Journal Vol. 38, No. 61 (2006) 61–108
5. Stanford Log-linear Part-Of-Speech Tagger, http://nlp.stanford.edu/software/tagger.shtml
6. Vincze, V., Szarvas, Gy., Almási, A., Szauter, D., Ormándi, R., Farkas, R., Hatvani, Cs., Csirik, J.: Hungarian Word-sense Disambiguated Corpus. In: Proceedings of 6th International Conference on Language Resources and Evaluation. LREC 2008, Marrakech, Morocco (2008) 3344–3349
7. Vincze, V., Nagy Á., Klausz, Á., Almási, A., Kiss, M., 2010: Nyelvészeti problémák a szabadalmak feldolgozásában. In: Tanács, A., Vincze, V. (szerk.): VII. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2010) 168–179
8. WordNet – A lexical database for English, http://wordnet.princeton.edu/