View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by University of Szeged

Angol nyelvű összetett főnevek értelmezése parafrázisok segítségével

Dobó András[1], Stephen G. Pulman[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
H-6720 Szeged, Árpád tér 2.
dobo@inf.u-szeged.hu
[2] University of Oxford, Department of Computer Science,
Wolfson Building, Parks Road, Oxford, OX1 3QD, Egyesült Királyság
stephen.pulman@cs.ox.ac.uk

Kivonat: Az angol nyelvben gyakran használnak összetett főneveket, melyek jelentésének meghatározása számos számítógépes nyelvészeti probléma megoldásának fontos eleme. Egy olyan módszert mutatunk be cikkünkben, mely alkalmas két szóból álló angol nyelvű összetett főnevek értelmezésére parafrázisok segítségével, ahol parafrázisok alatt igéket és elöljárószavakat értünk. Ez a módszer először megfelelő parafrázisokat keres statikus korpuszokban, majd webes kereséseket alkalmaz a helytelen parafrázisok kiszűrésére. A módszerünk által visszaadott parafrázisokat angol anyanyelvű személyekkel értékeltettük ki. Az első, második, illetve harmadik helyen visszaadott parafrázisokra rendre átlagosan 3,1842, 2,7687, illetve 2,5583 pontot adtak az értékelők megfelelősségük alapján (1-től 5-ig terjedő skálán), ami véleményünk szerint biztató eredmény a feladat nehézségét figyelembe véve.

1 Bevezetés

Mind az írott, mind a beszélt angolban bőségesen előfordulnak összetett főnevek (noun compound), melyek Downing [1] definíciója alapján főnevek olyan sorozatai, melyek egy főnévként viselkednek (az angol nyelvben az összetett főneveket külön kell írni). Értelmezésük, különösen gyakori használatuk miatt, nélkülözhetetlen számos számítógépes nyelvészeti probléma megoldásához, mint például a gépi fordításhoz és információ-visszakereséshez. Például amikor egy információ-visszakereső rendszer a *plastic bottles* (műanyag palackok) kifejezéshez keres információkat, akkor szükséges tudnia, hogy a *bottles that are made of plastic* (műanyagból készült palackok) kifejezésről talált információ releváns-e.

Első gondolatra statikus szótárak használata megfelelőnek tűnik e feladat megoldására, azonban még a gyakran használt összetett főnevekre is kis lefedettséget adnak e szótárak [2], és az összetett főnevek gyakorisági spektruma Zipf-eloszlást mutat [3], vagyis a legtöbb összetett főnévnek nagyon ritka az előfordulása.

E kutatás célja a két szóból álló angol nyelvű összetett főnevek automatikus értelmezése statikus korpuszok segítségével. Wright [4] és Nakov és Hearst [5] nyomán úgy gondoljuk, hogy az összetett főnevek parafrázisokkal (paraphrase - igék és elölj-

rószavak) történő értelmezése célravezetőbb, mint korlátozott számú absztrakt kategória alkalmazása, mivel lényegében végtelen különböző összetett főnév létezik és finom jelentésbeli különbségek kifejezésére is képesek. Továbbá úgy gondoljuk, hogy parafrázisok egy sorrendbe állított listája alkalmasabb e szószerkezetek értelmezésére mint egyetlen parafrázis, mivel egy gyakran nem elég egy összetett főnév teljes jelentéskörének megadására. Például, a *malaria mosquito* (malária moszkitó) egy lehetséges értelmezése a következő sorrendbe állított parafrázis lista lehetne:

1. carry (hordoz)
2. spread (terjeszt)
3. be infected with (által fertőzött)

, mivel *a malaria mosquito is a mosquito that carries / spreads / is infected with malaria* (a malária moszkitó egy olyan moszkitó, ami maláriát hordoz / maláriát terjeszt / malária által fertőzött).

A kidolgozott módszer olyan parafrázisokat keres a felhasznált statikus korpuszban, melyek alkalmasak az input összetett főnév értelmezésére. A módszer alapja az, hogy megkeresi azokat a mondatokat a korpuszban, amelyek egy parafrázis segítségével mondatba foglalják az adott összetett főnevet, megszámlálja, hogy az egyes parafrázisok hányszor fordultak elő a szókapcsolattal, majd e gyakoriságok alapján létrehoz egy rendezett listát. Ezt az alapötletet később több módon kibővítettük. Algoritmusunkat korábban angol nyelven már bemutattuk a Dobó és Pulman [6] cikkben.

2 Kapcsolódó munkák

2.1 Kategóriaalapú módszerek

Vannak olyan nyelvészeti elméletek, mint például Levié [7], melyek szerint az összetett főnevek mindegyike besorolható kis számú kategóriák valamelyikébe a főnevek között fennálló szemantikai kapcsolat alapján. Sok korábbi összetett főnév értelmezési módszer ezeken az elméleteken alapszik, és ennek megfelelően az összetett főneveket absztrakt kategóriákba sorolással próbálja meg értelmezni.

Rosario és Hearst [8] például 18 absztrakt osztály használatát indítványozza és egy olyan általános gépi tanulási módszert alkalmaz biomedikai összetett szavak osztályozására, mely doménspecifikus lexikai hierarchiával rendelkezik.

Nastase és Szpakowicz [9] szintén gépi tanulási módszereket alkalmazó algoritmust publikált összetett szavak klaszterezésére. Ehhez a WordNetből és a Roget's Thesaurusból kinyert tulajdonságokat használtak, és 30 klasztert definiáltak, melyek 5 szuperklaszterbe tartoztak.

Azonban az ebbe a csoportba tartozó módszereket számos kritika érte. Habár megvan az az előnyük, hogy megragadják az összetett főnevekben megtalálható általános kapcsolatokat, az általuk felhasznált kis számú kategória korlátozza is őket [2]. Downing [1] az egyike azoknak, akik leginkább kritizálják ezeket a módszereket. Szerinte olyan sokféle összetett főnévi kapcsolat létezik, hogy azt felsorolni lehetetlen, és nagyon sok olyan kapcsolat van ezek között, mely egyetlen általánosan használt kapcsolati kategóriába sem illeszkedik bele. Véleménye szerint az is problémát okoz, hogy mivel a használt kategóriák száma limitált, ezért a kategóriák homályosak, többértel-

műek lehetnek, és így különböző belső kapcsolattal rendelkező összetett főnevek is azonos kategóriákba kerülhetnek. Továbbá azt is nehéz lenne megállapítani, hogy a kategóriáknak mely halmaza lenne a legmegfelelőbb az összetett szavakban megtalálható kapcsolatok osztályozására, mivel a kimondottan összetett szavakkal foglalkozó nyelvészek sem értenek egyet a még fő kategóriákban sem [10].

2.2 Parafrázisalapú módszerek

Az előző alfejezetben említett problémák egy lehetséges megoldása az, ha parafrázisokat, vagyis igéket és elöljárószavakat, használunk az összetett szavak értelmezésére előre definiált absztrakt kategóriák helyett. Parafrázisok használata esetén a lehetséges kapcsolati kategóriák számát csak az adott nyelv szókincse korlátozza, továbbá még nagyon finom jelentésbeli különbségeket is ki lehet velük fejezni, valamint nincs egyetlen olyan összetett főnév sem, amely egyetlen kategóriába sem illik bele [2]. Ezért a parafrázis alapú módszerek az elmúlt években egyre népszerűbbek lettek.

Az egyik korai parafrázis alapú összetett szavakat értelmező módszert Laurer [10] fejlesztette ki. Ugyan parafrázisokkal dolgozik, mégis csak nyolc elöljárószót alkalmaz parafrázisként, ezért ez a módszer még inkább a kategóriaalapú módszerek családjába tartozik, és rendelkezik azok hátrányaival.

Ezzel szemben Nakov és Hearst [5], valamint Nakov [11] módszere már ténylegesen parafrázisalapú, az összetett szavak értelmezéséhez webes keresések által visszaadott szövegtöredékekből nyeri ki a parafrázisok listáját azok gyakoriságával együtt.

A SemEval-2 Workshop 9. feladatának [2] megoldására is született számos módszer. A feladatban adott összetett szavak egy listája és minden összetett szóhoz adott lehetséges parafrázisok egy halmaza. A cél olyan algoritmus írása volt, mely minden összetett szóhoz visszaadja a parafrázisok rendezett sorozatát, ahol a rendezés alapja az, hogy a parafrázisok mennyire megfelelőek az összetett szóhoz.

Erre a feladatra Nulty és Costello [12] egy olyan módszert dolgoztak ki, mely a tanító halmazból kinyert parafrázis gyakoriságokat használja fel úgy, hogy az általánosan használt parafrázisokat előnyben részesíti a kevésbé általánosakkal szemben.

A feladat megoldásához Wubbennek [13] teljesen más volt a stratégiája: egy osztályozó algoritmust hozott létre a WordNetből, a tanító halmazból és a Web 1T 5-gram Corpusból kinyert tulajdonságok alapján.

3 Módszerünk bemutatása

Célunk egy olyan módszer létrehozása volt, mely alkalmas tetszőleges két szóból álló angol nyelvű összetett főnév értelmezésére úgy, hogy ha bemenetként megkapja összetett főnevek egy listáját, akkor mindegyikhez visszatérjen parafrázisok egy rendezett listájával, igéket és elöljárószavakat használva parafrázisként.

Majdnem minden összetett szóban a második szó a fej (alaptag), míg az első az alárendelt tag, ami a fej egy tulajdonságát határozza meg. A két szó által alkotott összetett szó szintaktikailag úgy viselkedik, mint ahogy a feje [5], [10]. Munkánk során feltettük, hogy ez a tulajdonság az értelmezendő összetett szavakra fennáll, ezért módszereinkkel csak olyan parafrázisokat kerestünk, melyeknek alanya az összetett szó második főneve és tárgya az összetett szó első főneve.

3.1 A két alapmódszer

Az összetett szavakhoz megfelelő parafrázisok keresésére és kinyerésére két alapmódszert dolgoztunk ki.

Az alany-parafrázis-tárgy hármasokat alkalmazó módszer. Alapötletünk az volt, hogy oly módon tudunk megfelelő parafrázisokat találni egy összetett szóhoz, hogy ha egy statikus korpuszban keresünk olyan mondatokat, melyek egy parafrázis segítségével mondatba foglalják az adott összetett szót. Ehhez az algoritmus végigolvassa az alkalmazott korpuszt és megkeresi az összes olyan előforduló *(a, p, t)* hármast, melyben:

- *p* egy ige, melynek *a* az alanya és *t* a közvetlen tárgya
- *p* egy elöljárószavas ige, melynek *a* az alanya, az elöljáró az igével szorosan egybe tartozik (particle) és *t* az elöljárószavas ige közvetlen tárgya
- *p* egy elöljáró, ami *a*-nak egy módosítószava, és *t* a közvetlen tárgya az elöljárószónak

Ez a kinyerési módszer nagyon hasonló Nakov [11] módszerének ahhoz a részéhez, mely során a webes kereső által visszaadott, nyelvtanilag elemzett szövegtöredékekből kinyeri a tulajdonságokat az összetett szavakhoz.

Ez után a parafráziskinyerési módszer után módszerünk minden egyes bemeneti összetett főnévhez megkeresi azokat az *(a, p, t)* hármasokat, ahol *t* az összetett szó első, *a* pedig a második főneve. Ennek eredményeképpen megkapjuk parafrázisok egy listáját minden összetett főnévhez, az összetett főnév és a parafrázis együttes előfordulási gyakoriságával együtt. Ez az együttes előfordulási gyakoriság lesz a parafrázis pontszáma az adott összetett szóhoz. Például, ha 50 darab *(a=story, p=be about, t=adventure)* hármast talál az algoritmus, akkor az *adventure story* összetett főnév *be about* parafrázisához 50-es pontszámot rendel.

Ugyan az e módszerünk által megtalált parafrázisok általában megfelelőek voltak, nagyon kevés parafrázist talált az algoritmus még gyakori összetett főnevek esetén is, mivel az összetett szavak ritkán voltak ilyen módon mondatba foglalva. Így kipróbáltunk egy másik módszert is, mely a precision rovására magasabb recallal rendelkezik.

Az alany-parafrázis és parafrázis-tárgy párokat használó módszer. Ennek a módszernek az alapötlete az, hogy ha létezik olyan parafrázis, melynek a vizsgált összetett szó második főneve gyakran az alanya és első főneve gyakran a tárgya, akkor nagy esély van arra, hogy ez a parafrázis alkalmas az összetett szó értelmezésére. Ezért ez a módszer a korpusz végigolvasása közben azokat az *(a, p)* párokat keresi meg, melyekben:

- *p* egy ige, melynek *a* az alanya
- *p* egy elöljárószavas ige, melynek *a* az alanya és az elöljáró az igével szorosan egybe tartozik (particle)
- *p* egy elöljáró, ami *a*-nak egy módosítószava

Továbbá megkeresi azokat a *(p, t)* előfordulásokat is, melyekben:

- *p* egy ige, melynek *t* a közvetlen tárgya
- *p* egy elöljárószavas ige, melyben az elöljárószó az igével szorosan egybe tartozik (particle) és *t* az elöljárószavas ige közvetlen tárgya
- *p* egy elöljárószó, aminek *t* a közvetlen tárgya

E párok kinyerése után az algoritmus olyan *(a, p)* és *(p, t)* párokat keres egy összetett főnévhez, melynek második szava *a* és első szava *t*. Ez két parafrázislistát eredményez, egyet a második főnévhez (alanyhoz), egyet pedig az első főnévhez (tárgyhoz). Ebből a két listából egy olyan *(a, p, t)* listát kell létrehoznia, mely rangsorolja a parafrázisokat az összetett szó értelmezésére való alkalmasságuk szerint. Ehhez megkeresi azokat a parafrázisokat, melyek mindkét listában szerepelnek, és ezeket belerakja a közös listába, egy, a két listában talált gyakoriságból számolt pontszámmal.

Azonban szimplán gyakoriságok használata itt nagyon nagy problémát jelent: attól függetlenül, hogy az összetett szó első (tárgy) vagy második (alany) főnevét tekintjük, a hozzá megtalált leggyakoribb parafrázisok olyan nagyon gyakori igék, mint a *be*, a *do* vagy a *make*. Ezért a kombinált listában is ezek az igék szerepelnének legmagasabb pontszámmal, és ezek egyike sem jellemzi jól az összetett szavakat. Azért, hogy ezt elkerüljük, mind az *(a, p)* és *(p, t)* párok esetén pontonkénti kölcsönös információt [14] használtunk a gyakoriságok helyett. Az *(a, p)* és *(p, t)* párok pontonkénti kölcsönös információját ezután az algoritmus összeszorozza, és a parafrázisok ezzel a pontszámmal kerülnek be a közös *(a, p, t)* listába.

Például, ha az *(a=bottle, p=be for)* párnak és a *(p=be for, t=water)* párnak rendre 40 és 50 a gyakorisága, a *bottle* szó 500-szor és a *be for* kifejezés 2000-szer fordul elő *(a, p)* párban, valamint a *water* szó 800-szor és a *be for* kifejezés 1500-szor fordul elő *(p, t)* párban, továbbá az algoritmus összesen 2000000 *(a, p)* párt illetve 1500000 *(p, t)* párt talál, akkor a *be for* parafrázis *water bottle* szóhoz vett pontszáma 37,7153 lesz ezzel a módszerrel.

Mivel a 0 értéknél kisebb pontonkénti kölcsönös információ negatív asszociációt (disszociációt) jelent, ezért csak azokat a parafrázisokat vettük figyelembe, melyek esetén az *(a, p)* és a *(p, t)* pár is pozitív pontonkénti kölcsönös információval rendelkezik. Továbbá, mivel a pontonkénti kölcsönös információ instabil kis gyakoriságok esetén [14], ezért az 5-nél kisebb *(a, p)* vagy *(p, t)* gyakorisággal rendelkező parafrázisokat nem vettük figyelembe.

Azért, hogy módszereink hatékonyabban működjenek, mindkét módszer esetén az összes szót lemmatizáltuk, és a keresést is az összetett főnevek szavainak lemmájával végeztük. A szavak lemmáját a WordNet segítségével határoztuk meg.

3.2 A felhasznált korpuszok és azok előfeldolgozása

A parafrázisok kereséséhez a British National Corpust és a Web 1T 5-gram Corpust használtuk fel. Azért, hogy a megfelelő *(a, p)* és *(p, t)* párokat, illetve *(a, p, t)* hármasokat az algoritmusok ki tudják nyerni, szükséges a korpusz szavai között fennálló nyelvtani kapcsolatok azonosítása. Ehhez a British National Corpusnak egy a C&C CCG automatikus nyelvtani elemzővel [15] feldolgozott példányát használtuk fel, melyben így a nyelvtani kapcsolatok már explicit módon adottak voltak.

A rendelkezésünkre álló Web 1T 5-gram Corpus azonban nem volt még nyelvtanilag elemezve. Az automatikus nyelvtani elemzéshez szükséges idő hiányában egy alternatív megoldást választottunk. A korpuszt szófajilag elemeztük a C&C CCG automatikus szófaji elemzővel, majd szófaji minták alapján próbáltunk a szavak között fennálló nyelvtani kapcsolatokra következtetni. Például, ha egy 4-gram a *főnév ige névelő főnév* szófaji mintával rendelkezik, akkor nagy annak az esélye, hogy az első *főnév* az *ige* alanya, míg a második *főnéve* az *ige* tárgya. Ezt és ehhez hasonló mintákat használtunk fel a nyelvtani kapcsolatok kinyerésére a Web 1T 5-gram Corpus esetén. Mivel a rövid szövegtöredékek automatikus szófaji elemzése nagy hibával jár, ezért csak a 4- és 5-gramokat használtuk fel.

3.3 Elöljárószavak

Az elöljárószóval rendelkező parafrázisokat különlegesen kezeltük az alany-parafrázis és parafrázis-tárgy párokat használó modell esetében: ha a modellünk egy ilyen parafrázist talál, akkor két *(a, p)* párt nyer ki a szövegből. Egy olyat, amelyben a parafrázis tartalmazza az elöljárószót, és egy olyat is, amelyben nem. Az elöljárószó nélkülit azért, mert egy olyan mondatból, mint a *"The professor teaches at a university"* logikusnak látszik az *(a=professor, p=teach)* pár kinyerése. Így ha például van egy *(p=teach, t=anatomy)* párunk is, akkor a két párt összekapcsolva megkaphatjuk a *teach* parafrázist az *anatomy professor* összetett szóhoz. Az is szükséges, hogy módszerünk kinyerjen egy *(a, p)* párt az elöljárószóval együtt is, mivel egyébként nem lenne képes elöljárószót tartalmazó parafrázisok megtalálására egyetlen összetett főnév esetében sem. A *(p, t)* párok és *(a, p, t)* hármasok esetén nincs szükség speciális bánásmódra.

3.4 Passzív parafrázisok

A passzív parafrázisok abban különböznek a többi parafrázistól, hogy látszólagos alanyuk valójában a cselekvés tárgya. Ezért egy olyan (a, p_1) párnak, melyben p_1 egy elöljárószó nélküli passzív parafrázis, lényegében ugyanaz a jelentése (legalábbis a mi szempontunkból), mint egy olyan (p_2, t) párnak, melyben $a=t$ és p_2 a p_1 parafrázis aktív alakja. Ezért logikus lenne az ilyen, lényegében azonos jelentésű párokat együtt kezelni, gyakoriságukat közösen számolni. Ennek érdekében ha algoritmusunk egy olyan (a, p_1) párt talál, melyben p_1 parafrázis elöljárószó nélküli és passzív, akkor ezt egy olyan (p_2, t) párként menti el, melyben $a=t$ és p_2 a p_1 parafrázis aktív alakja. Például a *"The pizza was eaten"* mondatból az alany-parafrázis és parafrázis-tárgy párokat használó modellünk a *(p=eat, t=pizza)* párt nyeri ki. Mivel a passzív parafrázisoknak nem lehetnek közvetlen tárgyai, ezért nem létezhetnek olyan *(p, t)* párok és *(a, p, t)* hármasok, melyekben *p* egy elöljárószó nélküli passzív parafrázis.

Azoknál a passzív parafrázisoknál pedig, melyek tartalmaznak egy olyan *by* elöljárószót, melynek van közvetlen tárgya, ez a tárgy valójában a cselekvés alanya. Ezért egy olyan (a_1, p_1, t_1) hármas, melyben a p_1 parafrázis passzív és tartalmazza a *by* elöljárószót, lényegében ugyanolyan jelentéssel bír, mint egy olyan (a_2, p_2, t_2) hármas, ahol $a_2=t_1$, $t_2=a_1$ és p_2 a p_1 parafrázis aktív alakja elöljárószó nélkül. Tehát az ilyen, lényegében azonos jelentésű hármasokat is érdemes együtt kezelni, gyakoriságukat közösen számolni. Így például a *"The house was built by an architect"* mondatból az

alany-parafrázis-tárgy hármasokat használó módszerünk az *(a=architect, p=build, t=house)* hármast nyeri ki. Az olyan *(a, p)* és *(p, t)* párokat, melyekben *p* szintén egy passzív parafrázis a *by* elöljárószóval, az alany-parafrázis és parafrázis-tárgy párokat alkalmazó modellünk ehhez nagyon hasonlóan kezeli. Az olyan passzív parafrázisokat, melyek a *by*-tól eltérő elöljárószót tartalmaznak, nem kell speciálisan kezelni.

A fent leírt átalakítások miatt azoknak az *(a, p, t)* hármasoknak, valamint *(a, p)* és *(p, t)* pároknak a gyakorisága, melyekben *p* egy passzív parafrázis a *by* elöljárószóval, az átalakított verzióikhoz lettek elmentve. Ezért, annak érdekében, hogy algoritmusunk ehhez hasonló parafrázisokat is megtalálhasson összetett főneveinkhez, mindkét alapmódszerünk keres aktív, elöljárószó nélküli parafrázisokat a megfordított összetett szóhoz is (melyben a főnevek sorrendje fel lett cserélve; lehet, hogy így nem egy tényleges főnevet kapunk, de ez számunkra most lényegtelen). Ha talál ilyen parafrázist, akkor annak a passzív, *by* elöljárószóval kiegészített változatát használja fel, a megtalált parafrázis gyakoriságával.

Vagyis, ha például a *band concert* összetett szóhoz keres az algoritmus passzív, *by* elöljárószót tartalmazó parafrázist, akkor az alany-parafrázis-tárgy hármasokat használó módszerünk a szövegből kinyert *(a=band, p, t=concert)* alakú hármasokat keres. Például az *a=band, p=give, t=concert)* hármas esetén az algoritmus elmenti a *be given by* parafrázist a *band concert* összetett szóhoz, a talált hármas pontszámát felhasználva. Ez a másik alapmódszerünk esetén is nagyon hasonlóan működik.

3.5 Ambitranzitív igék

Angolban az igék lehetnek szigorúan tárgyasak, szigorúan tárgyatlanok, illetve ambitranzitívak [16], ahol az utolsó kategóriába tartozó igék tárgyas és tárgyatlan igeként is funkcionálhatnak. Jó példa szigorúan tárgyas igére a *like* és a *recognise*, szigorúan tárgyatlanra az *arrive* és a *run*, és ambitranzitívre a *break* és a *read*. Perlmutter [17] Unaccusative Hypothesise szerint a tárgyatlan igék két csoportra bonthatók: az unakkuzatív igék azok, melyek látszólagos alanya valójában a cselekvés tárgya (például *arrive*), és az unergatív igék azok, melyek látszólagos alanya ténylegesen a cselekvés alanya (például *run*). Ehhez nagyon hasonlóan az ambitranzitív igéket is két csoportra oszthatjuk: a páciens alanyú ambitranzitív igék azok, melyek unakkuzatív módon viselkednek intranzitív esetben és az ágens alanyú ambitranzitív igék azok, melyek unergatív tulajdonságúak intranzitív esetben [18]. Egy tipikus páciens alanyú ambitranzitív ige a *break*: a *"the window broke"* kifejezés valójában azt jelenti, hogy *"someone or something broke the window"*. Egy gyakori ágens alanyú ambitransitive ige pedig a *read*, mivel a *"she reads"* kifejezésben *she* ténylegesen a cselekvés alanya.

Tehát páciens alanyú ambitranzitív igék intranzitív használatakor módszerünk a cselekvés tényleges tárgyát (ami a látszólagos alany) helytelenül a cselekvés alanyaként nyerné ki. Ez hibákat eredményezne az összetett szavak értelmezésében. Azonban megfigyelhetjük, hogy az intranzitív esetben használt páciens alanyú ambitranzitív igék pontosan úgy viselkednek, mint a passzív igék: látszólagos alanyuk valójában a cselekvés tárgya. Ezért ezeket az igéket ugyanolyan módon kezeljük algoritmusunkban, mint a passzív igéket, és ezzel a fent leírt problémát kiküszöböljük. A páciens alanyú ambitranzitív igék felismeréséhez a Levin [19] által megadott átfogó listát használtuk fel.

3.6 Szinonimák, hipernimák, testvér szavak és szemantikailag hasonló szavak használata a magasabb recall elérése érdekében

Ugyan az általunk felhasznált korpuszok viszonylag nagyok, alapalgoritmusaink még így sem találnak bennük sok összetett főnévhez parafrázist. Kim és Baldwin [20] hipotézisét követve mi is úgy véljük, hogy hasonló jelentéssel bírnak azon összetett főnevek, melyek egymáshoz szemantikailag hasonló szavakból állnak. Így annak érdekében, hogy az összetett szavak értelmezésénél magasabb recallt tudjuk elérni, nemcsak az eredeti összetett szavakhoz kerestünk parafrázisokat, hanem azok olyan módosított változataihoz is, melyekben valamelyik (esetleg mindkettő) szót helyettesítettük az eredeti szó egy szinonimájával, hipernimájával, testvér szavával vagy pedig egy hozzá szemantikailag hasonló szóval. A szavak szinonimáit, hipernimáit és testvér szavait a WordNetből nyertük ki, míg a szavakhoz szemantikailag hasonló szavakat Lin [21] pusztán statikus korpuszokat felhasználó módszerével határoztuk meg.

3.7 A helytelen parafrázisok kiszűrése webes keresések segítségével

Az összetett szavak értelmezésére a korpuszból kigyűjtött parafrázisok sajnos sokszor nem helyesek, különösen az alany-parafrázis és parafrázis-tárgy párokat használó módszerünk esetén, illetve akkor, ha az összetett szó szavait a módszer helyettesítheti a szavak szinonimáival, hipernimáival, testvér szavaival vagy a szóhoz szemantikailag hasonló szavakkal. Ezért algoritmusunkat kibővítettük egy második lépéssel is, mely segít annak eldöntésében, hogy a megtalált parafrázisok közül melyek helyes értelmezései az összetett főneveknek, így növelve az algoritmus által elért precisiont.

Ehhez a lépéshez úgy döntöttünk, hogy webes kereséseket alkalmazunk a Google és a Yahoo! keresőrendszerek segítségével. Feltettük, hogy ha egy parafrázis alkalmas egy adott összetett szó értelmezésére, akkor léteznie kell legalább néhány olyan weblapnak, mely mondatba foglalja az összetett szót a parafrázis segítségével. Ezért minden (összetett szó, parafrázis) párhoz webes kereséseket indítottunk, és a parafrázisokat a keresésekre visszaadott lapok számának segítségével újrarendeztük.

Először egyszerű kereséseket próbáltunk ki, hasonlókat a Nakov és Hearst [5] és Nakov [11] által használtakhoz: egy n_1 n_2 összetett szó és p parafrázis esetén az összes lehetséges *"n2Infl THAT p n1Infl"* alakú lekérdezéssel kerestünk a keresőrendszerben, ahol *n1Infl* és *n2Infl* rendre az n_1 és n_2 főnevek lehetséges ragozott, illetve ragozatlan alakjai lehetnek, a *THAT* pedig vagy egy üres szó vagy az egyike a következő három vonatkozó névmásnak: *that*, *which* és *who*. Egy adott (összetett szó, parafrázis) párhoz tartozó összes ilyen alakú lekérdezésre visszaadott lapok számát összegezve definiáltuk az (összetett szó, parafrázis) pár webes pontszámát.

Azonban még ezek a keresések sem adtak vissza minden helyes (összetett szó, parafrázis) párhoz találatot. Ezért ezeket a kereséseket kibővítettük. Egyrészt úgy, hogy az igei parafrázisok esetén nemcsak a jelen idejű alakjukat használtuk fel, hanem egyéb igeidejű alakjaival is keresést indítottunk. Továbbá olyan kereséseket is használtunk, melyek joker karaktereket (*), 0 és 9 közötti számút, is tartalmaztak. Ezeket a joker karaktereket a parafrázis (*p*) és az első főnév (*n1Infl*) közé raktuk.

Miután egy adott (összetett szó, parafrázis) párhoz elvégeztük a fent leírt webes kereséseket és azok segítségével meghatároztuk a pár webes pontszámát, a pár végle-

ges pontszámát az eredeti pontszámának és a webes pontszámának segítségével számoltuk ki a következőképpen:

$$pontszám_{végső} = \ln(pontszám_{eredeti} + 1) * \ln(pontszám_{web} + 1) \quad (1)$$

ahol $pontszám_{eredeti}$ a pár eredeti és $pontszám_{web}$ a pár webes pontszáma. Az algoritmus ezután a parafrázisokat végső pontszámuk segítségével rendezi sorba.

4 Eredmények

A módszerek kiértékeléséhez a SemEval-2 Workshop 9. feladatának tesztadathalmazát használtuk fel. Ennek a feladatnak a célja olyan algoritmusok írása volt, melyek képesek az összetett főnevekhez már előre megadott lehetséges parafrázisokat megfelelősségük szerinti sorrendbe rakni. A mi algoritmusunk e feladat megoldásánál többre képes, ugyanis nincs szüksége bemenetként a lehetséges parafrázisok egy listájára, hanem a lehetséges parafrázisokat automatikusan nyeri ki a felhasznált korpuszból. Mivel módszerünk nem használja fel bemenetként az összetett főnevekhez adott lehetséges parafrázisok listáját, így olyan parafrázisokat is visszaad, melyek nincsenek ezen a listán. Ez okból kifolyólag a feladathoz biztosított kiértékelőt nem tudtuk módszereink teljesítményének mérésére felhasználni.

Helyette megkértünk 5 angol anyanyelvű személyt, hogy segítsenek módszerünk kiértékelésében. Mindegyiküknek odaadtuk a módszerünk által a bemeneti összetett szavakra visszaadott (összetett szó, parafrázis) párosok listáját, és ők minden párhoz egy 1 és 5 közé eső pontszámot rendeltek, ami a parafrázis minőségét adta meg (1: egyáltalán nem megfelelő, 5: teljesen megfelelő).

A limitált emberi erőforrás miatt nem tudtuk módszerünk összes változatát a felkért személyekkel kiértékeltetni, ezért a módszereink különböző változatait először mi magunk értékeltük ki, és csak az általunk legjobbnak vélt eredményeket adtuk oda a felkért személyeknek. Továbbá, szintén a kiértékelést gyorsítandó okból csak a tesztadatbázis első 50 összetett szavát használtuk fel. Mivel úgy véljük, hogy néhány parafrázis teljesen elegendő egy összetett szó teljes jelentéskörének a leírásához, ezért minden összetett szóhoz a módszerünk által visszaadott parafrázisok közül a három legmagasabb pontszámmal rendelkezőt vettük figyelembe.

Saját teszteléseink során arra az eredményre jutottunk, hogy a legjobban egy kombinált módszer teljesített. Ez két módszer kombinációjával jött létre: az egyik nem használ helyettesítő szavakat a parafrázisok kereséséhez, míg a másik felhasználja a WordNetből kinyert testvér szavakat az összetett szó eredeti szavainak helyettesítésére. A kombinált módszer a két módszer által visszaadott parafrázisok listáját egyesíteti, miután a testvér szavakat is alkalmazó módszer által visszaadott parafrázisokat újrapontozza a következőképpen:

$$pontszám_{új} = \frac{pontszám_{eredeti} * pontszám_{legalacsonyabb,nincsHelyettesítés}}{pontszám_{legmagasabb,helyettesítésTestvérSzavakkal}} \quad (2)$$

ahol $pontszám_{eredeti}$ az (összetett szó, parafrázis) pár eredeti pontszáma, $pontszám_{legalacsonyabb,nincsHelyettesítés}$ a helyettesítő szavakat nem használó módszer által visszaadott parafrázisok közül legkisebb pontszámmal rendelkezőnek a

pontszáma és $pontszám_{legmagasabb,helyettesítésTestvérSzavakkal}$ a helyettesítésként testvér szavakat alkalmazó módszer által visszaadott parafrázisok közül a legmagasabb pontszámmal rendelkezőnek a pontszáma. Ez által az újrapontozás által a második módszer által visszaadott legjobb parafrázis pontszáma meg fog egyezni az első módszer által visszaadott legrosszabb parafrázis pontszámával. Az ugyanazon módszer által visszaadott parafrázisok pontszáma közti arány így nem változik meg, viszont a kombinálás e módja előtérbe helyezi az első, lényegesen magasabb precisionnel rendelkező módszer által visszaadott parafrázisokat. Ahol pedig az első módszer nem ad vissza a kiértékeléshez elegendő (legalább 3) parafrázist, ott a lista kiegészül a második módszer által visszaadott parafrázisokkal. A kombinált módszerek közül mindkettő alany-parafrázis-tárgy hármasokat alkalmazott és a Web 1T 5-gram Corpust használta fel parafrázisok keresésére.

Az egyesített lista létrehozása után a listában szereplő parafrázisok mindegyikét újrapontozza webes keresések segítségével, a 3.7. alfejezetben leírt módon. A különböző webes pontozási módszereket a SemEval-2 Workshop 9. feladatának teszthalmazán automatikusan kiértékeltük a feladathoz adott kiértékelő segítségével. Ez alapján az a webes keresési módszer érte el a legjobb eredményt, amelyik a Google keresőrendszert, az igéknek csak a jelen idejű alakját és 0 és 1 közötti darabszámú joker karaktert használ, továbbá a keresésekben nem alkalmaz vonatkozó névmásokat.

Mielőtt a felkért személyek által visszaadott értékelésekből következtetéseket vontunk le, szükséges volt a személyek értékelésben való egyetértésének az igazolása. Amennyiben az értékelő személyek közt jelentős az egyet nem értés, akkor az általuk adott értékelés nem megbízható, és abból következtetéseket nem lehet levonni. Az adatok megbízhatóságának vizsgálatára Krippendorff [22] alfa metrikáját alkalmaztuk. A megbízott személyek által visszaadott értékelésre 0,435-ös alfa értéket kaptunk, vagyis jelentős volt közöttük az egyet nem értés. Ezért azt a 39 (összetett főnév, parafrázist) párt, melynek szórása legalább 1,5 volt, elvetettük. A maradék 111 párra kapott alfa érték 0,696 lett, amit már elfogadhatónak találtunk a feladatra.

A megbízott személyek értékelését úgy használtuk fel, hogy megnéztük azt, hogy átlagosan milyen pontszámot adtak a módszerünk által első, második és harmadik helyen visszaadott parafrázisokra: ezek rendre 3,1842, 2,7687 és 2,5583 voltak. Ez az eredmény azt mutatja, hogy a módszereink által visszaadott parafrázisok átlagban közepesen megfelelőek, és a visszaadott parafrázislistákban előrébb szereplő parafrázisok átlagban jobbak, mint a sorban később szereplő társaik. A feladat nehézségeit figyelembe véve úgy gondoljuk, hogy ezek az eredmények biztatóak, különösen annak fényében, hogy még az angol anyanyelvű értékelők között is nagy az egyet nem értés sok összetett szó értelmezésének tekintetében.

Azt az 5 összetett szót, melyen az algoritmus a legjobb, illetve a legrosszabb eredményt érte el a visszaadott (és nem elvetett) parafrázisok tekintetében, az 1. és 2. táblázatban foglaltuk össze.

5 Konklúzió

Cikkünkben egy olyan módszert mutattunk be, mely alkalmas két főnévből álló angol nyelvű összetett szavak automatikus értelmezésére. Módszerünk először statikus korpuszokban keres az összetett szó értelmezésére alkalmas parafrázisokat, majd webes

kereséseket alkalmazva újrapontozza őket. A módszerünk által első, második és harmadik helyen visszaadott parafrázisokra az anyanyelvi értékelők átlagosan 3,1842, 2,7687 és 2,5583 pontot adtak megfelelőségük alapján (1-től 5-ig terjedő skálán), amit a feladat nehézségeit figyelembe véve biztató eredménynek tartunk.

Mint ahogy azt a 3.2 alfejezetben említettük. idő hiányában nem tudtuk a Web 1T 5-gram Corpust nyelvtanilag elemezni, és a nyelvtani kapcsolatok kinyeréséhez szófaji mintákat használtunk fel. Ez a módszer azonban lényegesen nagyobb hibával jár, mint az automatikus nyelvtani elemzés, ezért a jövőben mindenképpen szeretnénk a már nyelvtanilag elemzett Web 1T 5-gram Corpuson is lefuttatni algoritmusainkat, mely módosítással reményeink szerint eredményeink tovább javulnának. Ezen felül szeretnénk algoritmusainkat további, még nagyobb korpuszok alkalmazásával is kipróbálni, melyek használata szintén kedvezően hathatna az eredményekre.

1. táblázat: Az az 5 összetett szó, melyen az algoritmus a legjobb eredményt érte el.

Összetett főnév, zárójelben a visszaadott parafrázisok	**Átlagos pontszám**
broadway youngster (be in)	4,7500
cell membrane (surround)	4,6000
cattle population (be of)	4,4000
arts museum (be of, be devoted to, be for)	4,3333
business sector (be of)	4,2000

2. táblázat: Az az 5 összetett szó, melyen az algoritmus a legrosszabb eredményt érte el.

Összetett főnév, zárójelben a visszaadott parafrázisok	**Átlagos pontszám**
anode loss (be at, be)	1.5000
bird droppings (be in, be for, be)	1.2667
bow scrape (be)	1.2500
activity spectrum (be in)	1.0000
altitude reconnaissance (-)	1.0000

Hivatkozások

1. Downing, P.: On the creation and use of English compound nouns. Language, Vol. 53 (1977) 810–842
2. Butnariu, C., Kim, S.N., Nakov, P., Seaghdha, D.O., Szpakowicz, S., Veale, T.: Semeval-2010 Task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In: 5th International Workshop on Semantic Evaluation. Taberg Media Group AB, Talberg, Sweden (2009) 100–105
3. Séaghdha, D.O.: Learning compound noun semantics. University of Cambridge, Cambridge, UK (2008)
4. Wright, D.G.S.: Noun-verb associations for Noun-Noun Compound Interpretation. Oxford University Working Papers in Linguistics, Philology & Phonetics, Vol. 8 (2003) 175–190
5. Nakov, P., Hearst, M.: Using Verbs to Characterize Noun-Noun Relations. In: Euzenat, J., Domingue, J. (eds.): Artificial Intelligence: Methodology, Systems, and Applications. Springer, Berlin / Heidelberg, Germany (2006) 233–244

6. Dobó, A., Pulman, S.G.: Interpreting noun compounds using paraphrases. Procesamiento del Lenguaje Natural, Vol. 46 (2011) 59–66
7. Levi, J.N.: The syntax and semantics of complex nominals. Academic Press, New York, USA (1978)
8. Rosario, B., Hearst, M.: Classifying the semantic relations in noun compounds via a domain-specific lexical hierarchy. In: 2001 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Stroudsburg (2001) 82–90
9. Nastase, V., Szpakowicz, S.: Exploring noun-modifier semantic relations. In: 5th International Workshop on Computational Semantics. Association for Computational Linguistics, Stroudsburg (2003) 285–301
10. Lauer, M.: Designing statistical language learners: Experiments on noun compounds. Macquarie University, Sydney, Australia (1995)
11. Nakov, P.: Using the Web as an Implicit Training Set: Application to Noun Compound Syntax and Semantics. University of California at Berkeley, Berkeley, USA (2007)
12. Nulty, P., Costello, F.: UCD-PN: Selecting General Paraphrases Using Conditional Probability. In: 5th International Workshop on Semantic Evaluation. Taberg Media Group AB, Talberg, Sweden (2010) 234–237
13. Wubben, S.: UvT: Memory-based pairwise ranking of paraphrasing verbs. In: 5th International Workshop on Semantic Evaluation. Taberg Media Group AB, Talberg, Sweden (2010) 260–263
14. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics, Vol. 16 (1989) 22–29
15. Clark, S., Curran, J.R.: Parsing the WSJ using CCG and log-linear models. In: 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, Stroudsburg (2004) 103–110
16. Dixon, R.M.W., Aikhenvald, A.U.: Introduction. In: Dixon, R.M.W., Aikhenvald, A.U. (eds.): Changing valency: Case studies in transitivity. Cambridge University Press, Cambridge (2000) 1–29
17. Perlmutter, D.: Impersonal passives and the unaccusative hypothesis. In: 4th Annual Meeting of the Berkeley Linguistics Society. BLS, Berkeley, USA (1978) 157–189
18. Mithun, M.: Valency-changing derivation in Central Alaskan Yup'ik. In: Dixon, R.M.W., Aikhenvald, A.U. (eds.): Changing valency: case studies in transitivity. Cambridge University Press, Cambridge (2000) 84–114
19. Levin, B.: English verb classes and alternations: A preliminary investigation. The University of Chicago Press, Chicago, Il (1993)
20. Kim, S.N., Baldwin, T.: Interpreting noun compounds using bootstrapping and sense collocation. In: 10th Conference of the Pacific Association for Computational Linguistics. Pacific Association for Computational Linguistics, Melbourne, Australia (2007) 129–136
21. Lin, D.: An information-theoretic definition of similarity. In: 15th International Conference on Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco (1998) 296–304
22. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage Publications, Thousand Oaks, CA, USA (2004)