Összetett kifejezések automatikus azonosítása természetes nyelvű szövegekben

A DOKTORI ÉRTEKEZÉS TÉZISEI

Nagy T. István

2014. október

Témavezető: Prof. Dr. Csirik János és Dr. Farkas Richárd

Szegedi Tudományegyetem
Informatika Doktori Iskola

1. Bevezetés

A modern kommunikációs és mobil információs eszközök elterjedt használatának köszönhetően rendkívüli módon nőtt a nyilvánosan hozzáférhető információk mennyisége. Ezen információk jelentős része szöveges, természetes nyelven írt formában érhető el. E hatalmas mennyiségű adat kézi feldolgozásához óriási emberi erőfeszítés és pénzügyi befektetés szükséges, amely támogatható automatikus módszerekkel. A természetesnyelv-feldolgozás (natural language processing – NLP) a természetes nyelv számos tulajdonságát, valamint a számítógépes nyelvek széles körének fejlődését matematikailag és számítástechnológiailag modellező tudomány.

A természetes nyelveken keresztül számos módon kifejezhetünk komplex emberi gondolatokat és ötleteket. Ez többek közt a kompozicionalitás alkalmazásával érhetjük el, azaz egyszerű nyelvi elemek összetételben való használatával, aminek eredménye egy sokkal összetettebb jelentés lesz, amely kiszámolható az eredeti részek jelentéséből, illetve azok kombinációjából. A nyelvben ugyanakkor nemkompozicionális kifejezések is előfordulnak, amelyek olyan összetett kifejezések, amelyek egyedi, jelentéssel bíró egységekre bonthatók, de az egész kifejezés jelentése nem – vagy csak részben – számítható ki egységeinek jelentéséből. Az ilyen kifejezések az úgynevezett összetett kifejezések (multiword expression – MWE), amelyek lexikai, szintaktikai, szemantikai, pragmatikai és/vagy statisztikai sajátosságokkal bírnak (Sag et al., 2002; Kim, 2008; Calzolari et al., 2002). Emellett az MWE-k nem képezhetők közvetlenül az összetételek szemantikájának aggregációjával, vagyis olyan lexikai egységekből, amelyek szóközzel vannak elválasztva. Éppen ezért, az olyan természetesnyelv-feldolgozó alkalmazások használata esetén, amelyeknél szükséges a szövegek szemantikus feldolgozása, elengedhetetlen az összetett kifejezések detektálása.

Értekezésem témája az angol és magyar nyelvű összetett kifejezések automatikus detektálása. Az MWE-k mindkét nyelvben elég gyakoriak, megfelelő módon való kezelésük pedig nélkülözhetetlen számos természetesnyelv-feldolgozó alkalmazás használata – például információkinyerés, valamint -visszakeresés vagy gépi fordítás – esetén; ilyenkor fontos az összetett kifejezések kontextusban való azonosítása. Például gépi fordítás során tudnunk kell, hogy az MWE-k egy szemantikai egységet alkotnak, így annak részeit nem

külön-külön kell lefordítani. Ehhez először az összetett kifejezéseket kell azonosítanunk a fordítani kívánt szövegben.

Az alábbiakban olyan különböző összetett kifejezéseket mutatok be, amelyekre a későbbiekben részletesen kitérek majd az értekezésemben.

Az összetett főnevek (nominal compounds – NC) az összetett kifejezések egy fajtája. Az NC-k olyan lexikai egységek, amelyek két vagy több olyan elemből állnak, amelyek külön-külön is értelmesek, az egység a főnév szerepét tölti be, és az eredeti részek jelentéséhez képest extra jelentéssel bír, lásd az alábbi angol és magyar nyelvű példákat:

(a) *black sheep – fekete bárány*

(b) *stock car – marhavagon*

A félig kompozicionális szerkezetek (light verb constructions – LVC) az MWE-k egy másik típusa. Az LVC-k ige és főnév kombinációi, amelyben az ige valamennyire elveszítette jelentését, és a főnév megtartja valamely eredeti jelentését. Lásd az alábbi angol és magyar nyelvű példákat:

(a) English:
to take measure
to play a role

(b) Hungarian:
őrizetbe vesz „to take into custody”
döntést hoz „to take a decision”

A tulajdonnevek (named entities – NE) a nyelvi elemek egy további olyan csoportja, amelyek számos NLP alkalmazás során – az információ-visszekereséstől a gépi fordításig – különleges kezelést igényelnek. A tulajdonnév egy olyan kifejezés a szövegben, amely kizárólag a világ egy entitására vonatkozik, például egy szervezet vagy hely nevére. Ezek a

tulajdonnevek gyakran több mint egy szóból állnak, ezért az összetett kifejezések/összetett főnevek speciális fajtájának tekinthetők. Az összetett kifejezésekhez hasonlóan, az összetett tulajdonnevek jelentése sem vezethető vissza azok alkotórészeire. Például a *Ford Focus* egy adott típusú autó nevét jelöli, és semmi köze a *ford* vagy a *focus* szavak eredeti jelentéséhez, így indokolt az egész kifejezést egy egységként fordítani. Az NE-k az NC-khez hasonlóan a főnév szerepét töltik be. Ezen felül a hasonlóságukat jól mutatja az a tény, hogy az NC tartalmazhat NE-t (*FBI special agent*), ugyanakkor része lehet NE-nek (*Tallulah High School*), egy NE pedig másik NE-t is tartalmazhat (például *Oxford* és *Oxford University* az *Oxford University Press* kifejezésben). Másfelől, néha nem lehet egyértelműen eldönteni, hogy egy összetételi egység egy összetett főnév vagy egy tulajdonnév (pl. *Attorney General*). Bár az összetett főnevek és az összetett tulajdonnevek is több mint egy szóból állnak, egy szemantikai egységet alkotnak, így az NLP-rendszerekben egy egységként kezelendők. Mivel ezek hasonlóan viselkednek, tézisemben amellett érvelek, hogy automatikus detektálásukhoz azonos módszer használható. Értekezésem fő célja a különböző összetett kifejezések automatikus felismerése angol és magyar nyelvű, nyers szövegekben. Mivel az igei MWE-k és összetett tulajdonnevek elég gyakoriak mindkét nyelvben, azokat az angol összetett főnevekkel együtt próbálom azonosítani, ehhez pedig számos, gépi tanuláson alapuló megközelítést fogok alkalmazni.

2. Az értekezés eredményei

Az értekezésben elért főbb eredmények az alábbiakban foglalhatók össze. Felsoroljuk továbbá a kapcsolódó publikációkat is, kiemelve az értekezés szerzőjének főbb hozzájárulásait az eredményekhez.

2.1. Angol összetett főnevek azonosítása Wikipedia-alapú módszerekkel

Az összetett főnevek angol nyelvű folyó szövegekben való automatikus azonosításának érdekében szótáron, illetve gépi tanuláson alapuló megközelítéseket egyaránt vizsgáltunk kü-

lönböző korpuszokon. Ezek a megközelítések nagymértékben támaszkodtak a Wikipediára. Ismertettük, hogyan hatnak az előzetesen azonosított összetett főnevek a névelemfelismerés hatékonyságára, és fordítva: az azonosított névelemek hogyan segítik az összetett főnevek azonosítását. Úgy találtuk, hogy az összetett főnevek előzetes ismerete javítja a névelem-felismerést, míg a névelemek azonosítása segítheti az összetett kifejezések azonosítását. Ezenkívül megvizsgáltuk az automatikusan annotált tanítóhalmazon tanított gépi tanulási megközelítés hatékonyságát, és úgy találtuk, hogy ez is elfogadható eredményt képes produkálni.

Emellett megvizsgáltuk, hogyan hat az automatikusan annotált tanítókorpusz mérete a gépitanuló-megközelítés hatékonyságára. A kapott eredmények azt mutatták, hogy a nagyobb tanítóhalmazon tanított modellek jobb eredményt értek el, de a hozzáadott érték folyamatosan csökkent. (**1. tézispont**)

A Wiki50 korpuszt mutatta be Vincze et al. (2011b), valamint a korpuszon elérhető elsődleges szótárillesztési eredményeket ismertették. A szerző az összetett főnevek automatikus azonosítására implementálta a szótárillesztő megközelítését. A társszerzők a korpusz annotálásában, valamint a nyelvészeti háttér biztosításában vettek részt.

Az összetett főneveket szabályalapú megközelítéssel azonosító módszer Vincze et al. (2011a) munkájában került bemutatásra. A szerző implementálta a szabályalapú módszereket és összehasonlította a különböző jellemzők hasznosságát. A társszerzők az adatok nyelvészeti elemzéséért feleltek.

Nagy T. et al. (2011) összetett főneveket és tulajdonneveket azonosítottak folyó szövegekben, és megvizsgálták, ezek hogyan járultak hozzá a dokumentumok automatikus kulcsszavazásához. A szerző implementálta a gépi tanuló alapú összetettfőnév-azonosító megközelítést, és tesztelte azt angol nyelvű szövegeken. A társszerzők az összetett főnevek, valamint tulajdonnevek nyelvészeti elemzéséért, továbbá a kulcsszókinyerő eredményekért feleltek.

Nagy és Vincze (2013) Wikipedia-alapú megközelítéseket mutattak be összetett főnevek automatikus azonosítására. A szerző megvizsgálta, hogyan hat az automatikusan generált tanítóhalmaz mérete a gépi tanuló megközelítés hatékonyságára, valamint a Wikipedia bő-

vülése a szótárillesztő módszerre. A társszerző a kutatás nyelvészeti hátteréért felelt.

2.2. Webbányászat alapú névelem-azonosítási problémák

Mivel a névelemek is egy szemantikai egységet jelölnek, és többnyire főnévként funkcionálnak, valamint több szóból is állhatnak, az összetett főnevekhez hasonlóan kezelhettük őket. Ezért a névelemek automatikus azonosítására az összetett főnevekhez hasonló megközelítéseket alkalmazhatunk. Számos névelem-felismerési problémát ismertettek már, mi itt alapvetően a webbányászathoz köthetőekre fókuszáltunk, mint például kutatók affiliációjának kinyerése, személyes információk kinyerése, és vállalkozások elérhetőségeinek kinyerése, amelyek mind névelem-felismerési problémák.

A weboldalak általában sok zajt is tartalmazhatnak (például menüelemeket vagy hirdetéseket), amelyek jelentősen gátolhatják a különböző számítógépes nyelvészeti eszközök megfelelő működését. Ezért különböző megközelítéseket alkalmaztunk a weboldalak szöveges tartalmának egységesítésére, hogy kinyerhessük azokból a névelemeket. Első lépésben a honlapok folyószöveges részeire koncentráltunk, mivel úgy találtuk, hogy a hasznos információk legjelentősebb része itt fordul elő leggyakrabban. Ezért automatikusan azonosítottuk a releváns részeit az egyes honlapoknak. Ezután a névelemeket gépitanuló-megközelítéssel automatikusan azonosítottuk a honlapok releváns tartalmaiból. Végül feladatspecifikus szabályalapú megközelítések segítségével validáltuk a kinyert névelemeket. (**2. tézispont**)

Nagy et al. (2009) kutatók affiliációs információit azok weboldalairól automatikusan kinyerő módszert ismertetett. Személyes információk weboldalakról való automatikus kinyerését Nagy T. (2012) mutatja be. A szerző részt vett a harmadik WePS versenyen (Artiles et al., 2010), ahol rendszerével a legjobb résztvevők közt szerepelt a személyesinformáció-kinyerő részfeladaton. A vállalkozások címeit kinyerő rendszert Nagy T. (2009) ismertette.

2.3. Angol és magyar nyelvű félig kompozicionális szerkezetek automatikus azonosítása szekvenciajelölő megközelítéssel

Az igei félig kompozicionális szerkezetek folyószövegekben való azonosítására szekvenciajelölésen alapuló megközelítést implementáltunk. Eredményeinket angol és magyar, két tipológiailag különböző nyelven is ismertettük, ezzel demonstrálva megközelítésünk rugalmasságát.

Mivel a különböző típusú szövegek különböző félig kompozicionális szerkezeteket tartalmazhatnak, valamint ezek előfordulási gyakorisága is eltérő lehet a különböző doméneken, ezért az eltérő korpuszokon tanult modellek hordozhatóságát is megvizsgáltuk. A továbbiakban megvizsgáltuk, hogyan tudják egyszerű doménadaptációs módszerek a különböző domének közti különbségeket áthidalni.

A doménsajátosságok ellenére az eredményeink azt mutatják, hogy a doménen kívüli adat képes segíteni a félig kompozicionális szerkezetek eltérő doméneken való automatikus azonosításában. (**3. tézispont**)

Az igei félig kompozionális szerkezetek automatikus azonosítására szolgáló gépi tanuló megközelítést Vincze et al. (2013b) mutatja be. A szerző implementálta a gépi tanuló alapú megközelítéseket angol és magyar nyelvre, továbbá doménadaptációs módszereket alkalmazott. Továbbá vizsgálta az egyszerű doménadaptációs technikák hatékonyságát a domének közti különbségek redukálására. A társszerzők a kutatás nyelvészeti hátteréért, valamint az eredmények statisztikai elemzéséért feleltek.

2.4. Angol és magyar nyelvű félig kompozicionális szerkezetek teljes halmazának automatikus azonosítása

Ugyan a szekvenciajelölő megközelítés képes automatikusan azonosítani az igei félig kompozicionális szerkezeteket angol és magyar nyelvű folyó szövegekben, ugyanakkor nem képes kezelni az egyéb típusú szerkezeteket, úgymint a nem folytonos (SPLIT) és igeneves (PART) szerkezeteket.

Ezért a félig kompozicionális szerkezetek teljes halmazának azonosítására fókuszál-

			Tézispont			
			1	2	3	4
RANLP	2011	(Nagy T. et al., 2011)	•			
RANLP	2011	(Vincze et al., 2011b)	•			
MWEWS	2011	(Vincze et al., 2011a)	•			
TSD	2013	(Nagy és Vincze, 2013)	•			
ACTA	2012	(Nagy T., 2012)		•		
NLPIR4DL	2009	(Nagy et al., 2009)		•		
OTDK	2009	(Nagy T., 2009)		•		
ACM	2013	(Vincze et al., 2013b)			•	
ACL	2013	(Vincze et al., 2013a)				•
IJCNLP	2013	(Nagy T. et al., 2013)				•

1. táblázat. Tézispontok és a kapcsolódó publikációk közti kapcsolat.

tunk.

Az általunk bemutatott módszer először minden mondatot szintaktikailag elemzett, majd különböző jelöltkiválasztó módszerek segítségével kinyerte a lehetséges félig kompozicionális szerkezeteket. Továbbá, megvizsgáltuk ezen jelöltkiválasztó megközelítések hatékonyságát angol és magyar nyelvű félig kompozicionális szerkezetek esetén is. Ezt követően gazdag jellemzőkészleten tanított gépitanuló-modellek segítségével azonosítottuk a félig kompozicionális szerkezeteket. (**4. tézispont**)

Az angol nyelvű félig kompozícionális szerkezetek automatikus azonosítására ismertettük módszerünket (Nagy T. et al., 2013). A szerző implementálta a gépi tanuló alapú módszert és új jellemzőket definiált, valamint kifejlesztette a szintaxisalapú jellemzőkinyerő módszert, ám a kísérleti eredmények az összes szerző közös hozzájárulásának tekintendők. A társszerzők a kutatás nyelvészeti hátteréért feleltek.

Angol és magyar nyelvű félig kompozícionális szerkezeteket automatikusan azonosító gépi tanuló modelt ismertetett Vincze et al. (2013a). A szerző összehasonlította a különböző módszereket, valamint nyelvspecifikus jellemzőket implementált ezen a két tipológiailag jelentősen eltérő nyelven. A társszerzők a kutatás nyelvészeti hátteréért, valamint a nyelvek közti összehasonlításokért feleltek.

A publikációk és a fentiekben ismertett tézispontok közti kapcsolatot az 1. táblázat szemlélteti.

2.5. Összegzés és jövőbeli tervek

Az értekezésben összetett kifejezések folyó szövegekben való automatikus azonosításával foglalkoztunk. A legfontosabb eredményeink a következő módon összegezhetők:

- különböző típusú összetett kifejezések automatikus azonosítására sikeresen alkalmaztunk felügyelt gépi tanuláson alapuló megközelítéseket;
- sikeresen alkalmaztunk gépitanuló-megközelítéseket összetett kifejezések automatikus azonosítására angol és magyar nyelven;
- összetett főnevek angol nyelvű folyószövegekben való automatikus azonosításához alkalmazhatók felügyelt gépitanuló-megközelítések és Wikipedián alapuló szabályalapú módszerek;
- a névelemek előzetes ismerete segíti az összetett főnevek automatikus azonosítását, valamint a névelem-felismerést támogatják az előzetesen azonosított összetett főnevek;
- összetett főnevek automatikus azonosítása automatikusan annotált tanítóhalmazon tanított gépi modell segítségével is lehetséges;
- a névelemek automatikus azonosítása az összetett főnevek azonosításához hasonló megközelítéseket kíván, mivel azok hasonló tulajdonságokkal bírnak: a névelemek az összetett főnevekhez hasonlóan egy szemantikai egységet jelölnek, több szóból állhatnak, valamint főnévként funkcionálnak;
- igei félig kompozicionális szerkezetek automatikus angol és magyar nyelvű azonosítása feltételes valószínűségi mezőkön alapuló módszerrel;
- doménadaptációs technikák segítségével csökkenthető a domének közti távolság az angol és magyar nyelvű félig kompozicionális szerkezetek esetében;
- szintaxisalapú megközelítés segítségével a félig kompozicionális szerkezetek teljes halmaza azonosítható;

- abban az esetben, ha az adott doménre elérhető jól működő szintaktikai elemző, akkor a félig kompozicionális szerkezetek automatikus azonosítására a szintaxisalapú megközelítés ajánlott, egyébként a szekvenciajelölésen alapuló módszer.

A fentieken kívül az értekezés eredményeit a számítógépes nyelvészet más területein, illetve más tudományterületeken is hasznosítani lehet. Összetett főnevek kontextusukban való automatikus azonosítása számos számítógépes nyelvészeti alkalmazás számára hasznos lehet, mint például információkinyerés és -visszakeresés, terminológiakinyerés, gépi fordítás vagy dokumentumosztályozás. A gépi fordítás esetében tudnunk kell, hogy egy adott összetett kifejezés egy szemantikai egységet jelöl, ezért részeit nem fordíthatjuk külön-külön. Ezért szükséges az összetett kifejezések automatikus azonosítása az automatikus fordítás előtt. Másrészről a félig kompozicionális szerkezetek automatikus azonosítása eseménykinyerő rendszerek építése során elengedhetetlen lehet, mivel azok gyakran egy eseményt jelölnek, és ezért szükséges egy egységként kezelni azokat.

A jövőben szeretnénk továbbfejleszteni rendszereinket az egyes jellemzők hatásainak részletesebb elemzésével. Szintén tervezzük meglévő módszereink adaptálását más összetett kifejezések automatikus azonosítására, mint például angol vonzatos igék (phrasal verbs), valamint azok angol és magyar nyelveken túli kiterjesztését.

Továbbá javítani kívánjuk meglévő módszereinket új, nyelvspecifikus jellemzők megvalósításával. Annak érdekében, hogy egy nyelvfüggetlen gépitanuló-megközelítést is létrehozhassunk, a jövőben szeretnénk a meglévő jellemzőket általánosítani.

Véleményünk szerint az értekezésben ismertetett összetett kifejezések automatikus azonosítására szolgáló módszerek jól hasznosíthatók számos számítógépes nyelvészeti feladat megoldása során, valamint újfajta megközelítések kidolgozásában.

Hivatkozások

Artiles, Javier; Borthwick, Andrew; Gonzalo, Julio; Sekine, Satoshi; Amigó, Enrique. 2010. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Task. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*.

Calzolari, Nicoletta; Fillmore, Charles; Grishman, Ralph; Ide, Nancy; Lenci, Alessandro; MacLeod, Catherine; Zampolli, Antonio. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pp. 1934–1940, Las Palmas.

Kim, Su Nam. 2008. *Statistical Modeling of Multiword Expressions*. Doktori értekezés, University of Melbourne, Melbourne.

Nagy, István; Farkas, Richárd; Jelasity, Márk. 2009. Researcher affiliation extraction from homepages. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09, pp. 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nagy, István; Vincze, Veronika. 2013. English Nominal Compound Detection with Wikipedia-Based Methods. In Matousek, Václav; Mautner, Pavel; Pavelka, Tomás (szerk.), *Proceedings of the 16th International Conference on Text, Speech and Dialogue, TSD 2013*, Lecture Notes in Computer Science, pp. 225–232. Springer, Berlin / Heidelberg, September.

Nagy T., István; Berend, Gábor; Vincze, Veronika. 2011. Noun compound and named entity recognition and their usability in keyphrase extraction. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Nagy T., István; Vincze, Veronika; Farkas, Richárd. 2013. Full-coverage Identification of English Light Verb Constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 329–337, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Nagy T., István. 2009. Összetett rendszer vállalkozások címeinek webről történő automatikus összegyűjtésére [Complex system for automatic detection of addresses of companies from Web]. In *XXIX. Országos Tudományos Diákköri Konferencia OTDK Informatikai szekció*. Debrecen.

Nagy T., István. 2012. Person attribute extraction from the textual parts of web pages. *Acta Cybernetica*, 20(3):419–440.

Sag, Ivan A.; Baldwin, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pp. 1–15, Mexico City, Mexico.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011a. Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 116–121, Portland, Oregon, USA, June. ACL.

Vincze, Veronika; Nagy T., István; Berend, Gábor. 2011b. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Vincze, Veronika; Nagy T., István; Farkas, Richárd. 2013a. Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 255–261, Sofia, Bulgaria, August. Association for Computational Linguistics.

Vincze, Veronika; Nagy T., István; Zsibrita, János. 2013b. Learning to detect English and Hungarian light verb constructions. *ACM Trans. Speech Lang. Process.*, 10(2):6:1–6:25, June.