Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő modullal

Miháltz Márton

MorphoLogic Kft.
1126 Budapest, Orbánhegyi út 5.
mihaltz@morphologic.hu

Kivonat. A számítógépes jelentés-egyértelműsítés során egy adott nyelven többértelmű lexikai elemekről kell eldönteni, hogy adott előfordulásuk kontextusában az ismert jelentéseik közül melyekkel szerepelnek. Ennek a feladatnak speciális esete, amikor a megkülönböztetendő jelentéseket nem egy egynyelvű szótár meghatározásai, hanem egy másik nyelven lehetséges különböző fordításaik alapján határozzuk meg. A cikkben bemutatott, Naiv Bayes osztályozó algoritmussal működő supervised egyértelműsítő rendszer egy angol-magyar fordítóprogram támogatásához készült. A jelenleg prototípus implementációban működő, 38 különböző többértelmű angol főnevet kezelő rendszer 84%-os átlagos pontossággal egyértelműsít.

1 Bevezetés

A számítógépes jelentés-egyértelműsítésnek (Word Sense Disambiguation, WSD) nevezett feladatban a számítógép feladata annak eldöntése, hogy egy adott nyelv lexikálisan többértelmű tételei egy adott szövegben, adott kontextusban az előzetesen meghatározott lehetséges jelentéseik közül melyekkel szerepelnek. Ebben a jelenleg igen aktívan kutatott problémában az egyik fő nehézséget a megkülönböztetendő jelentések megfelelő leírása jelenti. A többértelműség különböző formái különböző fokú nehézség elé állítják egy adott nyelv beszélőit: míg homonim alakok (pl. *körte*) megkülönböztetésekor az adott nyelvet anyanyelvként beszélő két ember közel minden esetben egyetért, addig a poliszémia termékeny, finom jelentésárnyalatbeli különbségeket produkáló eseteiben az egyetértés még anyanyelvi beszélők között is akár 85%-os vagy alacsonyabb is lehet ([4]). Számítógépes rendszereknél sem magasabb ez az érték: a 2001-es Senseval-2 nyilvános WSD versenyen főneveknél az angol lexikális minta feladatban legjobban teljesítő rendszer pontossága a jelentések durva felbontású, kevésbé megkülönböztető leírásával 76,6%, míg az árnyaltabb, több jelentést megkülönböztető, finomabb felbontású jelentés-leírással 69,5% volt ([5]).

Természetesen a jelentések leírásának módját egy adott WSD rendszer megvalósításában főképpen a rendszer konkrét célja határozza meg. Az alábbiakban bemutatott angol nyelvű jelentés-egyértelműsítő rendszerben a különböző, általában egynyelvű szótárak értelmezései alapján meghatározott lehetséges jelentéseket leképeztük azok különböző magyar fordításaira. Mivel az esetek nagy részében több különböző angol

jelentés is ugyanazzal a magyar fordítással fejezhető ki, az azonos fordítású jelentések összefogásával sikerült elérni egy durvább felbontású jelentés-leírást, ahol kevesebb, egymástól jobban elkülönülő alternatíva megkülönböztetésére van szükség.

A bemutatandó rendszer a MorphoLogic Kft. által fejlesztett *MetaMorpho* angol-magyar megértés-támogató fordítóprogram ([8]) támogatását szolgálja. A fordítórendszerben a többértelmű szavak különböző magyar fordítású jelentéseinek megkülönböztetésére van szükség, amikor a program nem képes csupán szintaktikai információk alapján egyértelműsíteni. Az alábbi két példában az előfordulási kontextus szemantikai tartalmának feldolgozására van szükség ahhoz, hogy a többértelmű angol *party* főnévhez a két esetben a megfelelő magyar fordításokat rendelhessük:

1. The **party** that won the elections four years ago did not make it into Parliament this time.
2. The **party** yesterday celebrated her birthday at one of the finest restaurants in town.

2 A jelentés-egyértelműsítő rendszer

2.1 Az osztályozó algoritmus

A bemutatandó jelentés- (vagy fordítás-) egyértelműsítő rendszer felügyelt (supervised), statisztikai osztályozó algoritmussal működik, így a betanításához szükség van kézi munkával, emberi erővel annotált tanítópéldákra. A statisztikai elven működő algoritmus mellett főképpen hatékonysági megfontolások miatt döntöttünk, hiszen a memória-alapú tanuló-rendszerekben, pl. a „lusta" KNN algoritmusban minden tanítópéldát állandóan „észben kell tartania" a rendszernek. Emellett pedig az ilyen rendszerben a különböző jegyek megfelelő súlyozása elengedhetetlen a helyes működéshez ([5]).

A rendszer az elterjedt, egyszerű *Naiv Bayes* osztályozó algoritmussal működik, mely a Bayes-tétel alapján, a különböző jegyek értékeinek egyes jelentésekre, mint osztályokra vett együttes feltételes valószínűségei alapján választja ki azt a jelentést, ami az adott kontextusban a legvalószínűbb. Az elnevezésben a „naiv" szó arra utal, hogy az algoritmus feltételezi, hogy az egyes kontextuális jegyek független valószínűségi változók. Ez természetesen természetes nyelvi szövegekben általában nem teljesül, azonban így is meglepően jó eredmények érhetők el ezzel a módszerrel ([3], [4]).

Az egyértelműsítő rendszer számára a tanuló algoritmuson felül alapvető fontosságú a többértelmű szavak előfordulási kontextusait reprezentáló jegyek helyes megválasztása is. Ezeknek a jegyeknek a tanítópéldákon megfigyelt értékeivel megy végbe a tanulás. Az ismeretlen jelentésű szó kontextusát szintén ilyen jegyekkel reprezentáljuk, illetve az online működés során ilyen jegyekből alkotott vektort osztályoz a Naiv Bayes algoritmus.

A bemutatott rendszer jelenleg a [3] és [5] munkákban bemutatott kontextuális jegyek egy részét használja, melyek két csoportba oszthatók. A csak az egyértelműsítendő szó mondatából kikerülő *lokális jegyek* főként a kontextus szintak-

tikai tulajdonságait, gyakori kollokációkat, módosítókat stb. reprezentálnak. A felhasznált lokális jegyek a következők:

- a többértelmű szó felszíni alakja (ahogy a szövegben előfordul)
- funkcionális szavak a többértelmű szó körül 2+2-es méretű ablakban
- tartalmas szavak tövei a többértelmű szó körül 3+3-as méretű ablakban

A jegyek másik csoportjába a teljes rendelkezésre álló kontextus topikjára, az aktuális szójelentés szemantikai tartományára jellemző *globális jegyek* tartoznak. Ezt az információt a tágabb kontextusban található, az adott tartalmas szóval leggyakrabban előforduló szavak előfordulási arányából alkotott vektorral reprezentáljuk (a sz0vegbeli lineáris sorrend és az egyértelműsítendő szótól való relatív távolság itt nem számít). A rendszerben paraméterezni lehet, hogy (a tanítópéldák alapján) az első hány leggyakoribb tartalmas szót, illetve hány, a többértelmű szó mondatát megelőző mondatot vegyen a kontextusból figyelembe.

Noha a rendszer jelenleg az összes ismert főnév esetében ugyanazokat a jegyeket használja, a rendszer implementációja lehetőséget biztosít arra is, hogy a későbbiekben a különböző szavakhoz automatikus jegy-kiválasztással ([5]) külön-külön megállapíthassuk a legoptimálisabb jegyhalmazokat.

2.2 A rendszer működése a fordítóprogramban

A Naiv Bayes osztályozóval működő egyértelműsítő rendszert a prototípus szinten a WEKA Java-s programrendszerrel ([10]) teszteltük, jelenleg folyik a saját fejlesztésű (C++ nyelven írt) motor tesztelése. Az osztályozó számára a lokális és globális jegyeket diszkrét, illetve numerikus értékkészletű attribútumokra képeztük le. A diszkrét attribútumok értékkészlet-halmaza a lokális jegyek tanítópéldákon megfigyelt értékekből áll, a numerikus attribútumok pedig egyenként jelölik, hogy a globális kontextus hányat tartalmaz a fontosnak ítélt tartalmas szavak közül. A numerikus attribútumok értékeit kernel sűrűség közelítéssel (kernel density estimator) modelleztük, amely nem feltételez normál eloszlást, és szignifikánsan javít a Bayes osztályozó pontosságán ([2]).

A rendszer számára kézzel annotált tanítókorpuszokhoz a Senseval ([1], www.senseval.org), és az Open Mind Word Expert ([6]) projektek jóvoltából jutottunk. A korpuszokat elsőként közös XML formátumba konvertáltuk, majd a nyelvi előfeldolgozási lépések következtek: bekezdésekre, mondatokra és szavakra bontás, morfológiai elemzés (HuMor angol morfológiai elemző), szófaji egyértelműsítés (MorphoLogic saját transzformációs szabályalapú POS-tagger), tövesítés. A tanítópéldák közül eltávolítottuk azokat, amelyek lexikálisan felismerhető, több szóból álló összetételeket tartalmaztak, mivel ezeket a fordítóprogram külön minták alapján, a WSD modul támogatása nélkül képes lefordítani. A feldolgozott korpuszokból kinyert kontextuális jegy-vektorokkal tanítottunk be Naiv Bayes osztályozókat, jelenleg 38 különböző többértelmű angol főnév egyértelműsítéséhez.

A MetaMorpho rendszerben az egyértelműsítő modul feladata az előfeldolgozott mondatokban (illetve bekezdésekben) az ismert többértelmű szavakon egy nyelvtani jegy értékének specifikálása, mely azok aktuális jelentését határozza meg. Ezen a ponton a jelentés-azonosító jegyek még az eredeti, többnyire a Princeton WordNet-ből ([7]) származó finom felbontású azonosítókat kapják értékül. A fordítóprogramban ezután történik a szintaktikai elemzés, ahol a szintaktikai szabályok már támaszkod-

hatnak a jelentés-azonosító jegyek értékeire is. Az elemzési fák felépülése után, az elemzéssel párhuzamosan kiválasztott generáló szabályok alkalmazásával létrejön a mondatok magyar fordítása. A magyar többértelmű főnevek generáló szabályaiban vannak az angol jelentések és magyar fordítások közötti leképezések elágazásszerűen rögzítve. Generáláskor a rendszer a jelentés-azonosító jegy WSD modul által beállított értéke alapján választja ki a megfelelő fordításokat. Ennek a megoldásnak két előnye van. Egyrészt, a magyar fordítások nincsenek „beledrótozva" az egyértelműsítő motorba, a működő rendszer nyelvtanában könnyen módosíthatjuk-javíthatjuk az angol jelentésekről való leképezéseket. Másrészt a megfelelő generáló szabályok megírásával nyitva áll az út az angolról más nyelvre fordító, jelentés-egyértelműsítő rendszerrel támogatott fordítóprogram létrehozása előtt is.

Az angol jelentésekről magyar fordításokra való leképezésben természetesen felmerülnek problémák is. Bizonyos többértelmű angol szavak egyes angol jelentései nem adhatók vissza egyetlen magyar fordítással, ilyenkor az eredeti jelentések tanítópéldáinak szétválasztására, magyar fordításokkal való újracímkézésére van szükség. A megvizsgált 38 többértelmű angol főnév között 4 olyat találtunk, ahol egyes jelentések esetén a fentiek szükségesnek bizonyultak volna (eddig azonban idő hiányában ezt nem tudtuk elvégezni, ezeket a jelentéseket kihagytuk a jelenlegi rendszerből).

Előfordult olyan eset is, amikor azért kellett kihagyni rendelkezésre álló szavakat, mert az összes angol jelentésük visszaadható volt egyetlen magyar fordítással (2 ilyen eset).

3 A rendszer értékelése

A jelentés-egyértelműsítő modul pontosságának megállapításához megvizsgáltuk a WSD-ben szokásos pontosság (precision) értékeket a jelenleg rendelkezésre álló 38 főnév tanítókorpuszán 10-szeres keresztellenőrzéssel (10-fold stratified cross validation). A pontosság a helyesen klasszifikált példák százalékos arányát jelenti az összes klasszifikált példához képest. A 10-szeres keresztellenőrzés során az egyes korpuszokat a jelentések eloszlási arányának megtartásával 10 egyenlő részre osztottuk, majd az egyes kis részeken végighaladva, a maradék részeken pedig tanítva teszteltük a rendszer pontosságát, és átlagoltuk az eredményt. Az alapszint minden esetben az adott szó leggyakoribb jelentésének relatív gyakorisága.

A teszteket elvégeztük mind az eredeti angol jelentésazonosítókkal, mind a magyar fordításokra leképzett és összevont azonosítókkal is. Az angol jelentés-azonosítók fölött történt egyértelműsítés eredményeit hely hiányában tételesen nem tudjuk közölni. Ennek átlaga a 38 főnév felett 76,39%-os pontosság 10-szeres keresztellenőrzéssel (64,15% átlagos alapszint). A magyar fordításokra történő egyértelműsítés eredményei az 1. Táblázatban láthatók.

Az egyértelműsítő algoritmus hibája a teljes tanítókorpuszon ellenőrizve angol jelentéseknél 1,90%, magyar fordításoknál 1,48%.

Az 1. Táblázatból leolvasható, hogy a megvizsgált 38 főnév esetében az egyértelműsítő pontossága általában meghaladta az alapszint értékét. Ez alól kivétel 5 eset, amikor egyenlő volt vele, illetve további 5 eset, amikor alacsonyabb volt annál (ekkor azonban csak átlagosan 1,77%-kal maradtak el az értékek az alapszinttől).

Az angol jelentések számának csökkentése a magyar fordításokra való leképezéssel átlagosan 7,86%-ot javított az egyértelműsítés pontosságán. 14 esetben nem változott így a pontosság, és csak egyetlen esetben csökkent (ebben az esetben mind az eredeti, mind az összevont jelentésazonosítók használatával a pontosság elmaradt az alapszinttől).

A tanítópéldák teljes mennyisége és a legritkább jelentéshez tartozó példák száma úgy tűnik együttesen befolyásolják a rendszer pontosságát. Az alapszint értékétől kevesebb, azzal egyenlő, vagy azt csak minimálisan meghaladó pontosságú esetekben az összes példák száma kb. 200-nál kevesebb, a legkevésbé reprezentált jelentéshez pedig kb. 20-nál kevesebb példa tartozott. Ennél magasabb együttes értékek jó eredményeket adtak.

4 További munka

A 38 főnév átlagos egyértelműsítési pontossága, ha a finom felbontású angol jelentésazonosítókat tekintjük, mindössze 0,21%-kal tér el a 2001-es SensEval-győztes rendszer pontosságától (ld. 1. rész). Ez a pontosság már a jelenlegi, hosszú jövőbeni fejlesztés előtt álló WSD rendszer esetében is alkalmas arra, hogy használható legyen egy olyan interaktív, emberi intelligenciára is építő rendszerben, mint a MetaMorpho megértés-támogató fordítóprogram.

A közeli jövőben szeretnénk a rendszerbe beledolgozni a WSD közösség számára rendelkezésre álló, szabadon hozzáférhető, illetve megvásárolható angol tanítókorpuszok anyagát is. Az egyértelműsítő szótárának további bővítéséhez szükség lehet egyéb szemantikailag annotált anyagok (pl. SemCor, Extended WordNet, MEANING projekt stb.) feldolgozására is. Ontológiák, szótárak felhasználásával, bootstrapping módszerek segítségével állíthatunk elő további tanítópéldákat ([3]), valamint hasznosak lehetnek ebből a szempontból a szinkronizált kétnyelvű (párhuzamos) korpuszok is.

Szintén a közeli jövőben szeretnénk a rendelkezésre álló korpuszok többértelmű igei és melléknévi anyagát is a rendszernek megtanítani. Az eltérő szófajok esetében valószínűleg különböző kontextuális jegyek bizonyulhatnak megfelelőnek, szükség lesz a 2.1 részben már említettekhez hasonló automatikus jegy- és paraméteroptimalizáció alkalmazására külön-külön az egyes többértelmű tételeknél.

A MetaMorpho rendszer szintaktikai elemzőjének fejlődésével szeretnénk továbbá a jelentés-egyértelműsítő modul számára a tanuláshoz és az egyértelműsítéshez magasabb nyelvi szintű információkat is rendelkezésre bocsátani. A tulajdonnév- és főnévi csoport-felismerés eredményeiből, illetve más morfoszintaktikai és szintaktikai információkból származó jegyek használatától a jelentés-egyértelműsítés pontosságának további javulását várjuk.

Bibliográfia

1. Edmonds, P. Kilgariff, A.: Introduction To The Special Issue On Evaluating Word Sense Disambiguation Systems. Journal of Natural Language Engineering 8 (4) (2002), 279-291

2. John, H. G., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers, in Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, San Mateo (1995)
3. Leacock, C., Miller, G. A., Chodorow, M.: Using Corpus Statistics and WordNet Relations for Sense Identification. Computational Linguistics, Special Issue on Word Sense Disambiguation. (1998)
4. Manning, C. D., Schütze, H: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA (1999)
5. Mihalcea, Rada Word sense disambiguation with pattern learning and automatic feature selection. Journal of Natural Language Engineering (special issue on evaluating word sense disambiguation systems, 8 (4) 279-291 (2002)
6. Mihalcea, R., Chklovski, T.: Building a Sense Tagged Corpus with Open Mind Word Expert. Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions (2002)
7.. Miller, G. A., Beckwith R., Fellbaum, C., Gross D., Miller, K. J.: Introduction to WordNet: an on-line lexical database. International Journal of Lexicography 3(4) (1990) 235 – 244
8. Prószéky, G., Tihanyi, L.: MetaMorpho: A Pattern-based Machine Translation Project. Proceedings of the 24th 'Translating and the Computer' Conference. London, UK, 19–24 (2002)
9. Vancsa, L.: A „BLEU" automatikus kiértékelési eljárás alkalmazása angol-magyar fordítóprogram gyakori, folyamatos minősítésére. Magyar Számítógépes Nyelvészet Konfeencia, Szeged (2003)
10. Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann, San Francisco (2000)

1. Táblázat. A jelentés-egyértelműsítő validációjának eredményei a különböző többértelmű tételeken (átlagos pontosság 10-szeres keresztellenőrzés után)

Főnév	Jelentések száma		Tanítópéldák sz.		Alapszint	Pontosság
	Angol	Magyar	Összes	Legrit-kább magy. jelentés-hez		
arm	5	4	787	16	56,67%	93,27%
art	4	2	108	3	97,22%	97,22%
authority	3	3	257	18	54,09%	68,09%
bank	4	2	398	7	98,24%	98,74%
bar	7	4	337	7	54,01%	60,53%
bum	5	2	118	20	83,05%	80,51%
chair	8	3	191	11	87,96%	87,43%
chance	6	4	615	21	65,37%	77,40%
chapter	3	2	137	45	67,15%	85,40%
child	7	2	180	66	63,33%	68,89%
church	3	2	183	76	58,47%	75,96%
circuit	6	4	184	25	43,48%	76,63%
day	2	2	192	67	65,10%	76,04%
degree	4	2	485	124	74,43%	96,29%
dyke	4	2	86	13	84,88%	87,21%
facility	3	2	37	2	94,59%	94,59%
fatigue	4	2	104	11	89,42%	93,27%
feeling	3	2	149	11	92,62%	90,60%
grip	5	2	218	17	92,20%	93,12%
hearth	3	2	96	17	82,29%	82,29%
holiday	4	2	83	3	96,39%	96,39%
image	7	2	512	219	57,23%	86,52%
lady	4	2	134	11	91,79%	92,54%
letter	3	2	927	140	84,90%	92,23%
line	5	4	4157	374	53,43%	84,94%
mouth	2	2	169	9	94,67%	93,49%
operator	2	2	119	31	73,95%	78,15%
party	2	3	623	108	42,05%	88,28%
performance	2	2	353	131	62,89%	88,95%
plane	4	3	474	2	96,41%	97,05%

post	3	3	141	18	63,12%	80,14%
process	2	2	302	70	76,82%	76,82%
report	3	3	335	42	67,76%	81,79%
restraint	6	4	89	2	44,94%	74,16%
sense	4	3	136	16	50,74%	55,88%
spade	5	3	89	4	71,91%	85,39%
stress	3	2	115	14	87,83%	85,22%
term	5	3	125	15	70,40%	80%
Átlag:					73,47%	84,25%