Korpusz-egyértelműsítés – a morfoszintaxison túl

Nagy Viktor

MTA Nyelvtudományi Intézet, Budapest
nagyv@nytud.hu

Kivonat A korpuszok morfoszintaktikai egyértelműsítése nem elegendő ahhoz, hogy a korpusz minden szövegszaváról megállapítsuk, mely lexémához tartozik, a probléma megoldásához szemantikai egyértelműsítés is szükséges. A dolgozat bemutatja, hogy a más nyelvekre már sikeresen alkalmazott statisztikai algoritmusok biztató eredményeket produkáltak már kis tanulómintán is.

Kulcsszavak: statisztikai alapú szójelentés-egyértelműsítés, homonímia-egyértelműsítés, felügyelt tanulás, naiv Bayes-osztályozás, döntési lista algoritmus

1. Bevezető

A napjainkban használatos korpuszoknak egyszerre kell teljesíteniük azt a két követelményt, hogy nyelvileg interpretáltak, illetve megfelelő méretűek legyenek. Bizonyos méret felett már a minimális annotáció sem végezhető el emberi munkával. A 150 millió szavas Magyar Nemzeti Szövegtár (MNSZ) [1] morfoszintaktikai elemzése és egyértelműsítése (tagging) – tanítás után – teljesen automatikusan zajlott. A morfoszintaktikai egyértelműsítés a korpusz szövegszavaihoz hozzárendeli a szóalak alapalakját (lemmáját), szófaját és morfoszintaktikai kategóriáját. Ez az eljárás a szövegszó morfoszintaktikai környezetét veszi figyelembe, vagyis a környezetet a konkrét szövegszóktól elvonatkoztatva morfoszintaktikai kódsorozatként látja, ezért nem tud olyan jelenségeket megkülönböztetni, amelyeknek szemantikája különbözik, morfoszintaktikai disztribúciója viszont azonos [2].

Az ilyen mélységű elemzés általában nem elegendő annak meghatározására, hogy a szövegszó mely lexémához tartozik, mivel a (lemma, szófaj) párhoz nem mindig rendelhetünk egyértelműen egyetlen lexémát, vagyis a magyar lexikográfiai gyakorlat által önállónak tekintett szótári egységet. Ez a szófajon belüli homonímia esetén figyelhető meg: míg a morfoszintaktikai egyértelműsítés el tudja dönteni, hogy a *tűz* alak a *tűz*^{1} IGE vagy a *tűz*^{2} FN előfordulása-e egy adott kontextusban, az *ül* (1 nyugszik', 2 'ünnepel') vagy a barát (1 'társ', 2 'szerzetes') homonimáit nem tudja szétválasztani, mivel az utóbbiak morfoszintaktikai környezete megegyezik.[1]

További problémát jelent az az eset, amikor egy szóalak több különböző, de azonos szintaktikai viselkedésű lemmához tartozhat. A *sejtette* lehet a *sejtet* IGE

[1] A homonimitás eldöntéséhez az ÉKsz-t vettük alapul, és a homonimaszámozás is az ÉKsz számozását követi.

vagy a *sejt* IGE tárgyas múlt idejű egyes szám 3. sz. alakja, a *lappal* pedig a *lap* FN vagy a *lapp* MN/FN eszközhatározós alakja. Ilyenkor az egyértelműsítő nem tud a szóalakhoz egyetlen lemmát rendelni. A többértelmű lemmák egy része mégis mutat valamelyest kontextusbeli különbséget. A *bán* IGE és a *bánik* IGE különböző vonzatkerettel rendelkezik, tehát elviekben meg lehet őket különböztetni csupán a morfoszintaktikai környezet alapján. Az MNSZ-ben használt taggelési eljárás azonban Markov-modellre épül, ezért szükségszerűen kicsi a figyelembe vett környezet mérete, ami lehetetlenné teszi a mondatban szétszórt vonzatok figyelembe vételét.

Ahhoz, hogy a morfoszintaktikai egyértelműsítés fent említett korlátain túllépjünk, mindenképpen az elvont morfoszintaktikai kategóriák mögé kell nézni a környezetben, és mivel az eldöntendő problémák egy része szemantikai, a szemantikai egyértelműsítés problémáját kell megoldani.

2. Statisztikai alapú szójelentés-egyértelműsítés felügyelt tanulás alapján

A szójelentés-egyértelműsítés (word sense disambiguation, WSD) célja annak megállapítása, hogy egy szó melyik lehetséges jelentésével fordul elő egy adott kontextusban. Az egyértelműsítésnek azt a fajtáját vesszük most figyelembe, amelynél a feladat szóelőfordulások előre felállított jelentésosztályokba sorolását jelenti. Fel kell tételeznünk, hogy a szavak jelentésosztályainak száma véges, és egymástól jól elkülöníthetők, de hogy pontosan mennyi is van, azaz milyen aprólékos a jelentésfelbontás, az alkalmazásról alkalmazásra változik. Az eredeti feladatunk szerint a jelentésosztályok lexémák, vagyis – a probléma típusától függően – a szó lehetséges lemmái vagy lehetséges homonimái vagy ezek kombinációi. Ez utóbbira példa az *érnek* igei alak, amely a *ér*^{1} 'érint', *ér*^{2} 'értéket képvisel, érvényes' és *érik* lexémát is képviselheti. Mivel a morfoszintaktikai egyértelműsítő által adott kimenetből indulunk ki, az *ér*^{3} főnévi lexéma már nem jöhet szóba.

Az eljárás, amelyet a fent vázolt egyértelműsítési problémára alkalmaztunk, statisztikai alapú, és felügyelt tanulást igényel. Minden egyértelműsíteni kívánt jelenséghez egy tanulókorpuszt kell konstruálni, amelyben kézzel megjelöljük, hogy a jelenség (többértelmű szóalak) egyes előfordulásai mely jelentésosztályba (mely lexémához) tartoznak. A szó környezetéből jegyeket vonunk el. A jegyek és a jelentésosztályok együttes előfordulásaiból egy tanuló algoritmus olyan modellt alkot, ami alapján a szó eddig nem látott előfordulásait jelentésosztályokba tudja sorolni. Tanulóalgoritmusnak a naiv Bayes-osztályozást és a döntési listás eljárást választottuk, amelyek a felügyelt szójelentés-egyértelműsítés népszerű megközelítései ([3]; [4], [5]).

3. A környezeti jegyek

A következő környezeti jellemzőket vettük figyelembe:

– tartalmas lemma jelenléte egy K méretű ablakban,
– szóalak jelenléte egy k méretű ablakban,
– inflexiós kategória jelenléte egy k méretű ablakban,
– szomszédos szóalakok,
– szomszédos szópárok (bigrammok),
– a többértelmű szó alakja.

Kétféle ablakméretet szoktak megkülönböztetni. A környezet által tartalmazott szemantikai információt egy szélesebb (mindkét irányban $K \approx 50$ méretű) ablakban keresik. Az itt található tartalmas szavak utalnak a kontextus témájára. Természetesen a szavak konkrét alakjától el kell vonatkoztatni, ezért kell a lemmákat figyelembe venni. Egy szűkebb (mindkét irányban $k \approx 3$ méretű) ablakban keresik a szintaktikai jellemzőket és az állandósult szókapcsolatokat. Mint a bevezetőben említettük, nem minden szintaktikai jellemzőt vett figyelembe a morfoszintaktikai egyértelműsítő, ezért érdemes a közeli inflexiós kategóriákat is felvenni a jegyek közé.

Az egyértelműsíteni kívánt szó alakját is érdemes a környezet jellemzőjének tekinteni. Ezt olyan homonímiák motiválják, amelyeknek tagjai nem pontosan ugyanúgy ragozódnak, vagyis egyes alakjaik egyáltalán nem homonimak. A *karja* alakból a *kar*^{1} 'végtag' lexémára, a *kara* alakból a *kar*^{2} testület' lexémára lehet következtetni, pedig mindkettő a *kar* FN lemma alakja. A morfológiai elemzőben hiába van meg ez a tudás, a kimenetből (lemma + morfológiai kategóriák) ez a megkülönböztetés eltűnik. Ezért ezeket a szabályokat ugyanúgy meg kell tanulnia az egyértelműsítő eljárásnak, mint a többi típusú jegyekből történő következtetést.

4. A felhasznált eljárások

4.1. A naiv Bayes-osztályozás

A naiv Bayes-osztályozás[3] azt az s jelentést rendeli egy C kontextusban megjelenő szóhoz, amelyre $P(s|C)$ maximális. Ennek kiszámításához felhasználja a Bayes-szabályt, miszerint $P(s|C) = \frac{(C|s)}{P(C)}P(s)$, és az úgynevezett Bayes-féle naiv feltevést, hogy a C környezet dekomponálható szavakra vagy egyéb jegyekre (f_j), amelyek egymástól független valószínűségi eseményeket alkotnak. Így a döntés a következőképpen írható le:

$$s = \arg\max_{s_i} P(s_i) \prod_{f_j} P(f_j|s_i)$$

A $P(f|s)$ valószínűségek a tanulókorpuszban található relatív gyakoriságok alapján becsülhetők. A döntésben minden jegy részt vesz, különböző mértékben.

4.2. Döntési listák

A döntési listákat [4] használta először jelentésegyértelműsítési célokra. Az eljárás kivonja a tanulóadatokból a jegyeket, és az alábbi formula alapján súlyokkal látja el őket:

$$w(s_i, f) = log\left(\frac{P(s_i|f)}{\sum_{j\neq i} P(s_j|f)}\right)$$

$s_1 \dots s_n$ a lehetséges jelentések, f a jegy, $P(s|f)$ az s jelentés tanulóadatokban megfigyelt valószínűsége feltéve f jegy jelenlétét. A döntési eljárás a következő szerint hajtódik végre:

if f_i **then** s_k
else if f_j **then** s_l
...
else s_m

A feltételek vizsgálata a jelentés-jegy párok súlya szerint csökkenő sorrendben történik. Amennyiben nem születik döntés, a leggyakoribb jelentés kerül kiválasztásra. Látható, hogy az algoritmus minden esetben egyetlen, a legerősebbnek vélt jegy alapján hozza meg a döntést, mégis eredményesnek bizonyult.

5. A tanulókorpusz

A Értelmező Kéziszótár gyakorisági adatainak összeállításakor készült egy kézzel annotált korpusz. Ez olyan szóalakok konkordanciáit tartalmazza, amelyeket lefed a kéziszótár címszóállománya, és a jelen dolgozatban vizsgált többértelműségeket mutatják. Minden többértelműségi problémára (azaz homonimapárra – kar^1/kar^2 vagy lexémapárra – *sejt*/*sejtet*) készítettünk egy kétszáz elemű véletlen mintát az MNSZ-ből, és kézzel megjelöltük, hogy a többértelmű szóalakok mely lexémához tartoznak. Ezek közül 40 olyan homonímiát választottunk ki, amely kétszeres többértelműséget mutat, és amelynek mintájában mindkét jelentés képviselve van.

6. Kiértékelés

A tanulókorpuszon három algoritmust értékeltünk ki a 3. szakaszban meghatározott jegyekkel. Az első mindig a többértelműségen belüli relatív leggyakoribb jelentést rendelte az esetekhez, ez szolgált viszonyítási alapul. A másik kettő a fent említett Bayes-osztályozó és a döntési listák eljárás. A kiértékelést tízszeres keresztellenőrzéssel (cross-validation) végeztük: A tesztmintát tíz egyenlő részre osztottuk. A tanulás során mindig más részminta maradt ki a tanulómintából, a kiértékelés pedig a kimaradt részmintán zajlott. Az 1. táblázat bemutatja a kapott eredményeket. A három algoritmus által a különféle többértelműségeken elért pontosság rendre az Alap, Bayes és a DL oszlopban van felsorolva.

1. táblázat. A három eljárás pontossága a tesztadatokon – részletek

Lemma	Szófaj	Alap %	Bayes %	DL %
érem/érme	fn	86,43	87,86	86,43
gém	fn	77,78	95,56	94,44
kar	fn	51,67	87,22	75,00
karton	fn	91,00	91,00	91,00
méh	fn	54,44	83,33	83,33
passz	fn	61,18	91,76	92,94
prímás	fn	52,63	93,16	89,47
rák	fn	64,12	81,76	85,29
trópus	fn	85,26	87,37	86,84
hetes	mn	66,32	80,53	74,21
napos	mn	72,63	86,84	92,11
bán/bánik	ige	56,88	76,25	91,25
hajt	ige	78,46	80,00	80,00
megbíz/megbízik	ige	84,00	84,00	84,00
megtör/megtörik	ige	61,00	65,50	64,50
nyúl/nyúlik	ige	77,65	78,24	81,76
olt	ige	73,53	79,41	81,18
túl	hsz	71,11	83,34	83,56
Átlag a teljes mintán		*71,00*	*83,54*	*83,98*

A két vizsgált algoritmus átlagos pontossága közel egyforma, 83,54, illetve 83,98 százalékos volt, de vannak esetek, amikor az egyik sokkal gyengébben teljesít a másiknál (és viszont). Mindkettő minden esetben legalább olyan eredményt produkált, mint az alapalgoritmus, sőt általában jobbat. Minél nagyobb a többségi jelentés aránya, annál kisebb a Bayes- és a DL-algoritmus „előnye". Az alapalgoritmus pontossága a többségi jelentés relatív gyakoriságát közelíti. Amikor ez meghaladja a 80%-ot, a többi algoritmus nem tud szignifikánsan jobb eredményt elérni, azok is a többségi jelentést választják, mert a ritkábbik jelentésből kevés példa van a mintában.

Érdemes megvizsgálni, hogy a különböző típusú jegyek mekkora sikerrel vettek részt a döntési listákkal hozott döntésben. A *prímás* érte el a legjobb eredményt a viszonyítási alaphoz képest (70%-kal jobbat). A 2. táblázat mutatja a legerősebb jegyeket. Eszerint itt a bővebb környezet szavai adtak jobb támpontot az algoritmusnak. A 3. táblázat szerint viszont a *bánik*/*bán* eldöntésében leginkább a szóalak és a szűk környezete nyújtott segítséget. Az összes eset jegyeit megvizsgálva azt kapjuk, hogy várakozásunkkal ellentétben a szűk környezetben található esetkategóriák egyáltalán nem bizonyultak jelentésmegkülönböztető jegyeknek.

Az MNSZ-ben sok olyan homonim pár van, amelynek egyik tagja olyan ritka, hogy néhány százas véletlen mintában való előfordulásának nagyon kicsi az esélye (pl. *barát*[2] szerzetes', *fok*[1] 'eszköz tompa fele'). Kétséges, hogy volna értelme a mintát addig növelni, amíg elegendő előfordulást nem kapunk. Ezeket a jelentéseket nem lehet statisztikai megközelítéssel felismerni.

Összefoglalva elmondhatjuk, hogy a naiv Bayes-osztályozás és a döntési listák eljárás hasonló pontossággal teljesített, és már relatív kis méretű tanulókorpuszon is jó eredményt kaptunk. A minta növelésével várhatóan még lehet javítani a pontosságon.

2. táblázat. A *prímás* példa tanult jegyei

Jegy	Súly	Döntés
bő k.: egyház	5.70	1
bő k.: érsek	5.67	1
bő k.: bíboros	5.67	1
bő k.: zenekar	4.94	2
...		

3. táblázat. A *bán*/*bánik* példa tanult jegyei

Jegy	Súly	Döntés
alak: bánom	5.25	bán
alak: bánnak	5.14	bánik
alak: bánja	4.78	bán
szűk k.: kell	4.70	bánik
szűk k.: vele	4.61	bánik
...		

Hivatkozások

1. Váradi, T.: The Hungarian National Corpus. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas (2002) 385–389
2. Oravecz, C., Dienes, P.: Efficient stochastic Part-of-Speech tagging for Hungarian. In: Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002, Las Palmas (2002) 710–717
3. Gale, W., Church, K., Yarowsky, D.: A method for disambiguating word senses in a large corpus. Computers and the Humanities **26** (1992) 415–439
4. Yarowsky, D.: Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In: Meeting of the Association for Computational Linguistics. (1994) 88–95
5. Agirre, E., Martinez, D.: Exploring automatic word sense disambiguation with decision lists and the web (2000)