Statisztikai alapú tulajdonnév-felismerő magyar nyelvre

Farkas Richárd[1], Szarvas György[1]

[1] MTA-SZTE, Mesterséges Intelligencia Tanszéki Kutatócsoport,
6720 Szeged, Aradi vértanúk tere 1., Hungary,
{rfarkas, szarvas}@inf.u-szeged.hu

Kivonat: Ebben a cikkben bemutatunk egy döntési fa alapú statisztikai tulajdonnév-felismerő rendszert magyar nyelvre. A modellt a Szeged Korpusznak az MTI honlapjáról származó, gazdasági rövidhíreket tartalmazó szegmensén tanítottuk és teszteltük, s vizsgáltuk annak pontosságát különböző méretű és összetételű tanuló halmazok felhasználása esetén. A feladathoz csak numerikusan kódolható információkat használtunk fel (nem használtuk fel a szóalakot), melyek között előfordultak speciálisan a magyar nyelv tulajdonneveinek helyesírására vonatkozó előírásai is, de a feladat során célunk volt a gazdasági hírekben előforduló, nagy számú idegen eredetű tulajdonnév azonosítása is. A kísérletek során legjobb pontosságot mutató modell 89,6%-os F mértéket ért el.

1 Bevezetés

A tulajdonnevek azonosítása (és kategorizálása) a folyó szövegben meghatározó fontosságú számos számítógépes nyelvfeldolgozó alkalmazás során. Példaként tekinthetjük a különböző információkinyerő rendszereket, ahol a tulajdonnevek általában jelentős információt hordozó szerepet töltenek be a szövegben, vagy a gépi fordítási alkalmazásokat, ahol értelemszerűen más módon kell kezelni emberek, szervezetek neveit, mint a szöveg többi részét.

Számos más nyelven eredményesen alkalmaztak különböző gépi tanulási eljárásokat tulajdonnév-felismerésre, sőt sok esetben ezek az eljárások egyszerre több célnyelven is hatékonynak bizonyultak [1]. Magyar nyelvre is készült már alkalmazás, a MorphoLogic Kft. HumorESK [6] nyelvi elemzője nyelvészek által összeállított szakértői szabályokon alapul.

E cikk célja egy statisztikai tulajdonnév-felismerő rendszer bemutatása, mely meglehetősen jó pontossági mutatóival igazolja, hogy más nyelvekhez hasonlóan magyar szövegekben is eredményesen alkalmazhatók tanuló modellek tulajdonnevek felismerésére. A feladatra a C4.5 [7] döntésifa-tanuló algoritmust használtunk, mert a fa struktúrájában kódolt döntések könnyen értelmezhetők a felhasználó által, valamint az egyes leveleken keletkező osztályozások jól mérhető pontossági adatokkal szolgálnak arra nézve, melyek azok a szegmensei a tulajdonnevek sokaságának, ahol további vizsgálatokkal a modell finomításra szorulhat.

A következő fejezetben (2) röviden ismertetjük a tulajdonnév-felismerési problémát, majd az általunk felállított tanuló modell bemutatása következik (3). Ezt követően ismertetjük az elvégzett kísérletek eredményeit (4), valamint röviden értékeljük azokat, felsorolva a modellel kapcsolatos további teendőket (5).

2 Tulajdonnév felismerés

A tulajdonnevek felismerése tekinthető egy taggelési feladatként, ahol minden szóra (tokenre) egy folyó szövegben a cél: megjelölni, hogy az adott nyelvi elem része-e egy tulajdonnévnek, és ha igen, akkor milyen kategóriába sorolható.

Kísérleteink során a CoNLL konferenciákon [1] is használt <helységnév, személynév, szervezet, egyéb> kategóriákat használtuk, egyrészt a jobb összemérhetőség végett, másrészt ez a 4-es osztályozás jól illeszkedik az információkinyerési alkalmazások céljaihoz, melybe a modellt beépíteni akarjuk. Az egyetlen lényegi eltérés az ott definiált osztályozáshoz képest, hogy az <egyéb> kategóriába angol és más nyelveken általában beleveszik a különböző, mennyiségeket jelölő kifejezéseket is, míg mi ezeket kihagytuk a modellből (egyrészt mert ezek a magyar nyelvben általában nem minősülnek tulajdonnévnek, másrészt ezek azonosításával a Szegedi NLP csoportban egy másik modul foglalkozik [5]).

Legtöbbször megkülönböztetik az osztályozás során a tulajdonnevek kezdő tokenjeit, és a tulajdonnév részét képező belső szóelemeket. Ennek elsősorban akkor van jelentősége, amikor a szövegben egymást követően több, azonos kategóriába tartozó tulajdonnév található, mert ilyenkor ezek segítségével állapítható meg azok kezdőpozíciója. A mi esetünkben ettől a megkülönböztetéstől eltekintettünk, azaz célunk csak annak eldöntése volt, hogy az adott token része-e tulajdonnévnek, vagy sem. A későbbiekben természetesen akár egy szakértői szabályrendszerrel, akár a tanuló modellbe való beépítéssel szükséges lesz a szókezdő tokenek azonosítása is.

A tanuláshoz a Szeged Korpusz gazdasági hírekből álló részét használtuk fel. A tanuló halmaz tehát a korpusz 200 ezer szóból álló szegmense volt, amely a teljes mondatszintaxisra nézve tartalmaz bejelöléseket, és amelyet a megfelelő tulajdonnévi osztályok kódjaival is elláttunk. Modellünkben a szintaktikai jegyekből csak a szófaji kódokat, esetragokat használtuk fel, a Humor elemzőre, egy végződéstippelő program eredményeire (ismeretlen szavakon), valamint POS taggerre támaszkodva. A 200156 szövegszóból 25382 képezi tulajdonnév részét, ezek a következő megoszlást mutatják a különböző kategóriák között:

1541 db helységnév, 19982 db szervezet, 2124 db személynév, 1735 db egyéb tulajdonnév

3 A tanuló modell

A modellhez minden szóhoz magára a szóra és a környezetére vonatkozó, numerikusan kódolható információkat gyűjtöttünk le (ezek egy részénél a [3] cikkben ismertetett modellt vettük alapul), ezek szerepeltek az osztályozásnál az adott elem attribútumaiként. Magát a szóalakot nem használtuk fel, mint attribútumot.

Az általunk használt jellemzők rendre a következők:

- Szófaji kód (magára a szóra, és +/-4 szavas környezetére)
- Esetrag
- Kezdőbetű típusa (magára a szóra és +/-4 szavas környezetére)
- Tartalmaz-e számjegyet (a szó belsejében)
- Tartalmaz-e nagybetűt (a szó belsejében)
- Tartalmaz-e írásjelet (a szó belsejében)
- Modateleji szó-e
- Idézőjelek közt szerepel-e a mondatban
- Szóhossz
- Arab vagy római szám-e
- A Szószablya [4] gyakorisági szótárban a kisbetűs és összes előfordulás hányadosa
- A Szószablya gyakorisági szótárban a mondatközi nagybetűs és összes nagybetűs előfordulás hányadosa
- Szerepel-e a szó valamely szótárunkban (településnevek, országnevek, keresztnevek, cégnevek utótagjait, földrajzi nevek utótagjait, valamint tulajdon-névben gyakori kisbetűs szavakat tartalmazó szótárakat használtunk)

A fenti jellemzők felhasználásával minden szóalakhoz 37 különböző attribútum tartozott, valamint a megfelelő osztály kódja. Az osztályozásra C4.5 döntésifa-tanuló algoritmust alkalmaztunk, melyet a WEKA programcsomagból [8] használtunk fel.

A módszer értékeléséhez egy meglehetősen jól működő naiv algoritmust használtunk, ami a következőképpen működött: *Minden nagybetűs szóra, amely tulajdonnév, a <szervezet> osztályt rendelte.* Ez az egyszerű eljárás a validációs halmazon 71.9%-os precision, 69.8% recall értékeket ért el (70.8% F mérték). A jó eredmény annak köszönhető, hogy hozzáadtuk azt az információt, hogy mi tulajdonnév (tehát a mondat eleji nagybetűs szavak nem okoztak tévesztést), ami sok információt adott a modellhez, másrészt az alapul vett szövegekben a <szervezet> osztály dominálja a másik 3 kategóriát, ami kedvez a valószínűségi taggernek.

4 Kísérletek

A tanuló modell kiértékelésére 3 különböző kísérletet végeztünk, melyek eredményei láthatóak az 1. táblázatban. A kísérletek során vizsgáltuk a tanult modell pontosságát, és a döntési fa méretét a tanuló halmaz nagyságának függvényében. Az adathalmaz 96 db XML-fájlból állt, melyből validációs célokra a korpuszból véletlenszerűen válogatott 9 db fájlt használtunk (A teszthalmaz 5, a tanuló halmaz 82 fájlból állt).

Az első kísérletben a rendelkezésre álló, megközelítőleg kétszázezer példából csak annak kis részét (mintegy 9-10%-ot) használtuk fel a tanuláshoz, melyen 10-szeres keresztvalidációt végeztünk, a halmaz kis mérete miatt. A kapott döntési fa 155 csúcsot és 78 levelet tartalmazott, a validációs halmazon 82.8% precision, 86.7% recall, 84.7% F mérték[1] eredményeket ért el. Ezek az eredmények meghaladják az összehasonlításhoz használt naiv algoritmus eredményeit, ami elsősorban a <szervezet> osztály sikeres felismerésének köszönhető, a másik 3 kategórián a pontosság gyengébb.

A második kísérletben a korpuszból előállt példák felét használtuk fel a tanuláshoz, ezúttal keresztvalidáció nélkül (elegendően nagy példahalmaz), egy rögzített teszt-halmazt használva a modell értékelésére. Ezúttal a döntési fa 433 csúcsból, 217 levélből állt, és 85.4% precision, 86.9% recall, 86.1% F mérték pontosságot ért el. A javuló eredményeket döntően 3 kategórián mérhető pontosságnövekedés okozta, míg meglepő módona leggyengébb pontosságot adó osztály felismerése romlott (egészen 1%-ra, azaz gyakorlatilag nem ismert fel pontosan elemet a fa ebbe a kategóriába, az <egyéb> tulajdonnevek döntően <szervezet> cimkét kaptak).

Harmadik esetben a teljes rendelkezésre álló adathalmazt használtuk. Az eredményül kapott döntési fa 839 csúcsból, 420 levélpontból állt, és 91% precision, 89.7% recall, 90.3% F mérték pontosságot mutatott. Mivel ezek voltak az eddigi legjobb eredmények, 10 különböző futtatást végeztünk (más-más tanuló, teszt, validációs halmazokon). Az eredmények átlagban nem sokkal maradtak el az előzőleg használt futtatás eredményeitől (89.6%-os átlagos pontosság, 1.86% szórással).

1. táblázat: A különböző vizsgálatok eredményei osztályonként, és átlagban:

	1. Kísérlet	**2. Kísérlet**	**3. Kísérlet**	**3. Kísérlet (10 futás átl.)**
	F measure			
Helységnév	57.5%	64.2%	68.5%	74.2%
Szervezet	90.2%	92.3%	94.5%	93.8%
Személynév	65.3%	67.6%	74.4%	76.5%
Egyéb	11.0%	1.0%	54.0%	59.5%
Átlagos pontosság	**84,7%**	**86,1%**	**90.3%**	**89.6%**
Javulás a baseline-hoz képest	**19,6%**	**21,6%**	**27.5%**	**26.6%**

[1] Precision: Helyesen osztályozott tulajdonnevek és az összes tulajdonnévnek osztályozott példa hányadosa
Recall: Helyesen osztályozott tulajdonnevek és az összes tulajdonnév token hányadosa
F mérték: A Precision és Recall értékek harmonikus közepe

5 Konklúzió, további lehetőségek

Az eredmények alapján elmondható, hogy a tulajdonnevek felismerésének statisztikai módszerekkel való megközelítése magyar szövegekben is eredményes lehet. Természetesen a tanuló modellt számos további attribútummal lehetne finomítani (pl. az adott szó kapott-e már tulajdonnév címkét az aktuális szövegben, és ha igen, milyet), valamint a döntési fák által előállított osztályozások alacsonyabb pontossággal rendelkező osztályait további vizsgálatoknak is alá lehetne vetni. További feladat a rendszer felkészítése az egymás melletti, azonos típusú tulajdonnevek szeparálására (szakértői szabályokkal, vagy a modellbe építve), valamint tervezzük az eredmények összevetését a HumorESK program tulajdonnév felismerő funkciójával, a két rendszer kombinálhatóságának vizsgálatát.

Bibliográfia

1. Conference on Computational Natural Language Learning (CoNLL-2003, 2002): Language-Independent Named Entity Recognition. http://cnts.uia.ac.be/signll/conll.html (2003)
2. Csendes Dóra, Csirik János, Gyimóthy Tibor: The Szeged Corpus: A POS Tagged and syntactically Annotated Hungarian Natural Language Corpus. In Sojka et al., 41–47 (2004)
3. Curran, James R., Clark, Stephen: Language Independent NER Using a Maximum Entropy Tagger. Proceedings of CoNLL-2003, 164–167, Edmonton, Canada (2003)
4. Halácsy P., Kornai A., Németh L., Rung A., Szakadát I., Trón V.: A szószablya projekt – www.szoszablya.hu. MSZNY 2003, 298–299, Szeged, Magyarország (2003)
5. Miháczi András, Németh László, Rácz Miklós: Magyar szövegek természetes nyelvi előfeldolgozása. MSZNY 2003, 38–44, Szeged, Magyarország (2003)
6. Prószéky Gábor: Syntax As Meta-Morphology. Proceedings of COLING-96, Vol.2, 1123–1126. Copenhagen, Denmark (1996)
7. Quinlan, J. R.: C4.5: Programs for machine learning, Morgan Kaufmann. (1993)
8. Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, USA (2000)