Javában taggelünk

Novák Attila[1], Orosz György[2], Indig Balázs[2]

[1]MorphoLogic Kft., 1116 Budapest, Kardhegy utca 5.
novak@morphologic.hu
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai Kar,
oroszgy@itk.ppke.hu
dlazesz@gmail.com

Kivonat: A szófaji egyértelműsítés (POS tagging) a számítógépes nyelvfeldolgozás egyik alapfeladata. A feladat megoldására számtalan algoritmus sok különböző programozási nyelven megírt implementációja létezik. Az egyes szövegszavakhoz rendelendő morfológiai címkék megállapítása azonban csak az egyik részfeladat, amelyet a szöveg morfológiai annotációjakor el kell végezni: a címkén kívül a szótövet is azonosítani kell. A nem túl gazdag morfológiájú analitikus angol nyelv esetében egy szófaji egyértelműsítő és egy egyszerű tövesítő egymás után kapcsolása elfogadható eredményt ad. A magyarhoz hasonló ragozó nyelvek esetében azonban sokkal jobb eredményt kapunk, ha a szófaji egyértelműsítést és a szótő megállapítását egyaránt elvégző morfológiai elemzőt tartalmazó integrált eszközt használunk.

1 Bevezetés

Cikkünkben egy olyan új nyílt forráskódú eszközt mutatunk be, amely egyszerre végzi el a szófaji egyértelműsítést és a szótő megállapítását, tehát teljes egyértelműsített morfológiai annotációt ad. Az eszköz szófaji egyértelműsítő algoritmusa a TnT és HunPoS taggerekben implementált rejtett Markov-modell (HMM) algoritmuson alapul. Emellett tartalmaz egy olyan felületet, amelynek használatával morfológiai elemző illeszthető hozzá, amely nemcsak a tanítóanyagban nem látott szavak morfológiai címkéjének egyértelműsítését teszi sokkal pontosabbá, hanem a szavak szótövét is megadja. Az eszközt Java nyelven implementáltuk.

2 A korpusz reprezentativitása

Ha a magyarhoz hasonló agglutináló nyelveket az angollal hasonlítjuk össze abból a szempontból, hogy egy adott méretű korpusz milyen arányban tartalmazza az adott nyelv lehetséges szóalakjait, akkor azt tapasztaljuk, hogy míg egy azonos méretű korpuszban sokkal több különböző szóalak szerepel az agglutináló nyelvek esetében, mint az angolban, ezek ugyanakkor mégis sokkal kisebb részét fedik a korpuszban szereplő szótövek lehetséges alakjainak. A korpusz tehát sokkal kevésbé reprezenta-

tív a szókincs szempontjából, mint az angol esetében. 10 millió szavas korpuszméret esetében például az angolban általában 100 000-nél kevesebb különböző szóalakot találunk, ugyanakkor a magyarban jóval 800 000 feletti a különböző szóalakok száma. Ugyanakkor míg az angolban egy nyílt szóosztályba tartozó szónak legfeljebb 4–6 alakja van, a magyarban több száz vagy több ezer különböző alakot kapunk attól függően, hogy a produktív szóképzés eseteivel is számolunk-e. Természetesen a sokkal több lehetséges szóalak azt jelenti, hogy a lehetséges szófaji címkék száma is jóval magasabb a magyar esetében (több ezer szemben az angol néhány tucat címkéjével). Ezért egy magyar korpusz a szóalakok szintjén több szempontból is sokkal hiányosabban reprezentálja a nyelvet, mint az angol esetében: a szövegekben szereplő lemmák lehetséges ragozott alakjainak túlnyomó többsége teljesen hiányzik; az előforduló szóalakok is sokkal kevesebbszer szerepelnek; sokkal kevesebb példa van az adott konkrét morfológiaicímke-sorozatokra, sőt a lehetséges címkék nagy része egyáltalán nem szerepel a korpuszban.

A tanítóanyagban nem látott szavak kezelésére (illetve pl. a maximum entrópia modellt használó taggerek esetében a tanítóanyagban látott szavak esetében is) a szófaji egyértelműsítő eszközök általában tartalmaznak valamilyen mechanizmust, amely a szavak végződéseit vizsgálja a címke megjósolásához. A magyar esetében az előforduló hosszú toldaléksorozatok miatt jóval hosszabb szóvégek figyelembevételére van szükség, mint a nem agglutináló nyelvek esetében (ez különösen így van, ha a ragok mellett bizonyos produktív képzőket is azonosítani szeretnénk).

3 A morfológiai elemző hatása

A magyarhoz hasonló nyelvek esetében a rendszer tanítóanyagában nem szereplő szóalakok nagy része olyan szó, amelynek más ragozott alakjai előfordulnak a tanítóanyagban. Oravecz és Dienes [5], valamint Halácsy és mtsai. [4] bemutatták, hogy morfológiai elemző felhasználásával az általa ismert szóalakok esetében sokkal pontosabban meg lehet állapítani a tanítóanyagban nem szereplő szavak címkéjét, mint pusztán a tanítóanyagon betanított nyelvfüggetlen szóvégződés-felismerővel. Az utóbbi téves javaslatait a morfológiai elemző kimenetével megszűrve a tanítóanyagban nem látott szavakra a szófaji egyértelműsítés pontossága hatékonyan javítható. A morfológiai elemző pontosságot javító hatása annál jelentősebb, minél kisebb a rendelkezésre álló kézzel egyértelműsített tanítóanyag.

Az imént idézett eredmények nem olyan rendszerrel készültek, amely valóban integrált morfológiai elemzőt tartalmazott volna, hanem az annotálandó szövegen offline lefuttatott morfológiai elemző által visszaadott címkéket táblázat formájában betöltve szimulálták a morfológiai elemző hatását. Ez a fajta megoldás azonban nem használható bizonyos alkalmazásokban, például ha a taggert webszolgáltatásként szeretnénk üzemeltetni.

Többek között ezért döntöttünk úgy, hogy olyan eszközt implementálunk, amely integrált morfológiai elemzőt tartalmaz. A morfológiai elemzőt nemcsak arra használjuk, hogy a tanítóanyagban nem látott szavak címkézésének pontosságát javítsuk, hanem szükségünk van rá a szótövek megállapításához is. A morfológiai elemző számára sem ismert szavak kezelése (legfőképpen a szótövük megállapítása) morfo-

lógiai guesser (toldalékelemző) beépítésével oldható meg. Ezért az eszköz két csatolófelületet tartalmaz: egyet a morfológiai elemző, egyet pedig a guesser illesztésére.

4 Az optimális tő kiválasztása

A morfológia és főleg a sokkal lazább megszorításokkal dolgozó guesser gyakran több olyan lehetséges tőjelöltet is visszaad, amely a tagger által választott címkével kompatibilis. Sokszor tehát nem triviális a helyes szótő kiválasztása. A magyarban az egyik ilyen többértelműségi osztály az az azonos tövű ikes–iktelen igepároké. A lexikális *tör/törik, (fel)dolgoz/dolgozik* típusú párok mellett a produktív *-z/-zik* képzőpár szinte korlátlan mennyiségben hozza létre az ilyen típusú többértelműségeket. Emellett a két ragozási paradigma lényegében csak abban az egyetlen E/3 jelen idejű kijelentő módú alakban tér el, amely a lemmát adja, az összes többi igealak többértelmű a tő szempontjából, ezért egyben ez a leggyakoribb olyan tőtöbbértelműség-típus, amely a morfológiai elemző által felismert szóalakok körében fellép.

A tő egyértelműsítésére legegyszerűbb alapmodellként egy egyszerű unigram modellt használtunk. Ebben a modellben a szóalakként leggyakrabban előforduló alakot választjuk a lehetséges tövek közül. Ennek az egyszerű modellnek előnye, hogy nincs szükség a statisztika alapját képező korpusz semmiféle annotációjára. Ezért nem kell a rendelkezésünkre álló annotált korpuszra szorítkoznunk, hanem tetszőleges méretű anyagot használhatunk, még maga az annotálandó szöveg is hozzáadható a statisztika alapját képező anyaghoz. Ez a modell magyarra elég jó teljesítményt ad az ismeretlen szavak túlnyomó részét adó névszók esetében, mert ezeknek a leggyakoribb alakja a toldalékolatlan alanyeset.

Az egyik leggyakoribb többértelműségi osztály, ahol az egyszerű tőválasztási algoritmus hibázik, a magas hangrendű ikes–iktelen igepárok esete (ahol az *–ik* nélküli ige tárgyas). Ezeknek az *–ik* végű alakja is többértelmű: T/3 alanyú határozott tárgyas alak is lehet, és az ennél az igeosztálynál sokszor gyakoribb az *–ik* nélküli lemmánál (pl. a *nevezik* alak 4-szer olyan gyakori, mint a *nevez*). Ezt a problémát részben lehet kezelni egyrészt úgy, hogy a morfológiai elemzőben letiltjuk a *nevez*-hez hasonló gyakori igék produktív képzéssel előállított felbontását (ezzel a *név+ezik = nevezik* képzett alakot). Emellett az egyszerű unigram szóalak-gyakorisági modell annotált korpuszból vett adatokkal nyelvspecifikus módon kombinálva, illetve a tövek megelemzése után a tagger által választott elemzéssel inkompatibilis tövek kiszűrésével a tőmeghatározás pontossága növelhető.

5 Morfológiailag annotált korpusz építése nulláról

Azon nyelveknek jelentős része, amelyekre nem léteznek kézzel annotált tanítóanyagok, a magyarhoz hasonlóan bonyolult morfológiával rendelkezik. Ezen nyelvekre morfológiailag annotált egyértelműsített korpusz létrehozására egy olyan iteratív eljárás tűnik a leghatékonyabb módszernek, amelynek során morfológiai elemző létrehozását követően a rendelkezésre álló korpusz egy kis részhalmazát elemeztetjük,

és ezt kézzel egyértelműsítve a taggert betanítjuk. A korpusz következő részletét az így betanított taggerrel előegyértelműsítjük, majd az annotációt kézzel javítjuk, ezt a folyamatot addig ismételve, amíg elegendő annotált korpuszhoz nem jutunk. Nulláról épített annotált korpuszok esetében a minimális méretű tanítóanyag miatt a korábban vázolt adathiány-probléma még súlyosabb. Minél kevesebb tanítóanyag áll rendelkezésre, annál jelentősebb az integrált morfológiai elemző jótékony hatása az automatikus morfológiai címkézés pontosságára. Az annotáció kézi javítása is sokkal hatékonyabban végezhető, ha a morfológiai elemző egyéb elemzései is rendelkezésre állnak a tagger által választott elemzés mellett, és egyszerűen választani lehet az elemzések közül, mint ha ténylegesen mindig kézi javítgatásra van szükség.

Az iteratív korpuszannotációs eljárás használhatóságának fontos feltétele, hogy a tagger újratanítása ne vegyen igénybe túlzottan hosszú időt. A betanítás sebességének szempontjából a rejtett Markov-modell alapú szófaji címkéző eszközök nagyságrendekkel felülmúlják a bonyolultabb maximum entrópia vagy CRF-alapú algoritmusokat, amelyeknek betanítási ideje jóval hosszabb. (Konkrétan a HMM-alapú HunPoS [4] betanítása a Szeged korpuszon [6] kevesebb, mint egy percet vesz igénybe, szemben a maximum entrópia alapú OpenNLP hat órás betanítási idejével ugyanazon a gépen.) Mindemellett a HMM-alapú eszközök számos nyelvre – többek között magyarra is – az egyértelműsítés pontosságában is élen járnak.

Bár a magyar nyelvre már létezik egy olyan nyelvspecifikus eszköz, amely tartalmaz morfológiai elemzőt, és platformfüggetlen implementációval rendelkezik: a magyarlanc [7], ennek azonban nyelvspecifikus mivolta mellett komoly hátránya az alapjául szolgáló Stanford POS tagger nagy erőforrásigénye és a betanítás lassúsága.

6 Az új eszköz

Az elérhető HMM-alapú megoldások nem tartalmaznak beépített morfológiai elemzőt. A népszerű és megengedő licenszű HunPos tagger kiegészíthető lenne a kívánt funkcionalitással, de az implementációjához használt programozási nyelv csekély ismertsége ennek (és a tagger integrálásának) korlátját jelenti. Egy, az iparban elterjedtebb nyelv használata könnyebb szerves integrációt tesz lehetővé olyan nyelvfüggetlen keretrendszerekhez, mint az UIMA vagy a GATE. Ezért döntöttünk egy új, a tanítási sebességét tekintve jól használható, nyelvfüggetlen, morfológiai elemzővel könnyen integrálható szófaji egyértelműsítő implementációja mellett. Az új, nyílt forráskódú, Java nyelven implementált, rejtett Markov modellen alapuló POS-tagger, melynek alapjául a TnT [1] és a HunPos rendszerek szolgálnak, a korábban említett problémák kiküszöbölése érdekében a szófaji egyértelműsítés és a szótövezés problémáját egy feladatként kezeli. A rendszer képes morfológiai elemző és guesser aktív használatára a szófaji egyértelműsítés közben, továbbá az elemzés kimenetét a szótő meghatározására is felhasználja. Az eszközt olyan alkalmazásprogramozási felülettel láttuk el, amelyen keresztül egyszerűen illeszthető hozzá tetszőleges morfológiai elemző. Mivel gyakran az egyértelműsített taghez tartozó tő sem egyértelmű (különösen nem az azoknak a szóalakoknak az esetében, amiket a morfológiai elemző nem ismer, hanem a lehetséges töveiket a guesser állítja elő), olyan

mechanizmussal is kiegészítettük a rendszert, amely a lehetséges többértelmű tövek közül is hatékonyan választ.

Bibliográfia

1. Brants, T.: TnT – A Statistical Part-of-Speech Tagger. In: Proceedings of the sixth conference on Applied natural language processing (2000)
2. Farkas, R., Szeredi, D., Varga, D., Vincze, V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (2010) 349–353
3. Halácsy, P., Kornai, A., Oravecz, Cs., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC (2006) 2245–2248
4. Halácsy, P., Kornai, A., Oravecz, Cs.: HunPos: an open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (2007) 209–212
5. Oravecz, Cs., Dienes, P.: Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: Third International Conference on Language Resources and Evaluation (2002) 710–717
6. Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., Csirik, J.: Hungarian Dependency Treebank. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (2010)
7. Zsibrita, J., Nagy, I., Farkas, R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 394–395