Morfológiai egyértelműsítés nyelvfüggetlen annotáló módszerek kombinálásával

Laki László János, Orosz György

MTA-PPKE Magyar Nyelvtechnológiai kutatócsoport,
Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
1083, Budapest, Práter utca 50/a
e-mail:{laki.laszlo, oroszgy}@itk.ppke.hu

Kivonat Írásunkban megvizsgálunk két szófaji egyértelműsítő modult, s arra következtetésre jutunk, hogy bizonyos esetekben a két rendszer hibái nagyon távoliak. Bemutatjuk, hogy egy esetleges kombináció milyen eredményekkel kecsegtethet, illetve ismertetünk két egyszerű összetételi technikát, melyek segítségével készített nyelvfüggetlen rendszer a morfológiai tudást használó társával pontosság tekintetében versenyképes.

1. Bevezetés

A szófaji egyértelműsítés a számítógépes nyelvfeldolgozás egyik alapfeladata. A feladat megoldására számos szabadon elérhető nyelvfüggetlen rendszer használható, melyek többsége valamilyen statisztikai tanuló algoritmust használ. Egy-egy eszköz nagyon alacsony hibarátája több, mint kívánatos, hiszen egy szövegfeldolgozási láncban a többi elemző algoritmus ennek kimenetére épít, ezt használja.

Jelen írásunkban először ismertetünk két szófaji egyértelműsítő rendszert: a PurePos [1] eszközt és egy statisztikai gépi fordításon alapuló PoS-taggert [2]. Közelebbről megvizsgálva az általuk hibásan osztályzott szavakat, azt találtuk, hogy a rendszerek által vétett hibák közötti átfedés nagyon alacsony. Ebből az észrevételből kiindulva, megvizsgáltuk, hogy milyen lehetőségek nyílnak a két rendszer tudásának kombinálására. Megmutatjuk, hogy csupán a két nyelvfüggetlen rendszer kombinációját használva, jobb eredményt érhetünk el, mint egy harmadikkal való egyszerű szavazásos kombinációt használva. Eredményeinkből az is kiolvasható, hogy a prezentált nyelvfüggetlen metódus címkézési pontosságban versenyképes lehet a PurePos morfológiai elemzővel segített változatával.

2. A használt eszközök

Magyar nyelvre egy szabadon elérhető statisztikai alapon működő, de mégis hibrid rendszer a PurePos, mely integrált morfológiai elemzőt tartalmazó rejtett Markov-modellezésen alapuló teljes egyértelműsítő rendszer. A rendszer a Brants

[3] és Halácsy et al. [4] által ismertetett algoritmusokra épít, különös tekintetettel a morfológiai elemző teljes integrációjára. Az egyszerű simított trigram modellnek köszönhetően magas precizitással és alacsony tanítási idővel rendelkezik. Az eszköz Java nyelven íródott, így szükség esetén könnyen módosítható. Megmutattuk [1], hogy azon esetekben, amikor lehetőség van morfológia használatára, kis méretű tanítóanyag esetén is jelentős növekedést ér el mind a szófaji címkézés, mind pedig a lemma egyértelmű meghatározása esetén is.

Egy korábbi írásunkban [2] megvizsgáltuk a statisztikai gépi fordítórendszer (SMT) szófaji egyértelműsítő és szótövesítő eljárásként való alkalmazhatóságát (HuLaPos). Itt sikerült megmutatnunk, hogy minimális előfeldolgozással viszonylag kis méretű tanítóhalmaz esetén is jó minőségű PoS-tagger állítható elő. Ez többnyire annak volt köszönhető, hogy a szófaji egyértelműsítés feladata nagyságrendekkel kisebb komplexitású a szóösszekötő rendszer számára, mint egy természetes nyelvi fordítási feladat, valamint a kifejezés alapú gépi fordítórendszer döntése során képes figyelembe venni a szavak mindkét oldali környezetét is. Az SMT-módszer leggyengébb pontja a szótárban nem szereplő szavak elemzése. Egy szógyakoriságon alapuló módszerrel sikerült az OOV[1] szavak környezetének súlyát megnövelni és ezzel a rendszerre gyakorolt negatív hatását csökkenteni.

Cikkünkben felhasználjuk még az OpenNLP [5] eszközkészletben elérhető maximum entrópiás és perceptron tanulásos algoritmusokat is. Az említett eljárások nagy népszerűségnek örvendenek, mivel a tanuló algoritmusában használt jellemzők könnyen adaptálhatóak egy-egy új feladatra. Ezen módszerekre igaz még, hogy a nagy számosságú jellemzőhalmaz miatt a tanítási idejük nagyságrendekkel nagyobb rejtett Markov-modellezésen alapuló társaiknál.

A PurePos esetén láttuk a morfológiai tudás nagyon értékes tud lenni – különös tekintettel agglutinativ nyelvek esetén – de sajnos csak korlátozott számú nyelvre érhető el szabadon morfológiai elemző. Továbbá egy új elemző létrehozása nagyon időigényes, és nyelvész szakértők bevonását igényli, így felmerülhet az igény olyan általános célú módszerekre, melyek csupán a tanító halmazt használva magas pontossággal képesek szófaji egyértelműsítésre.

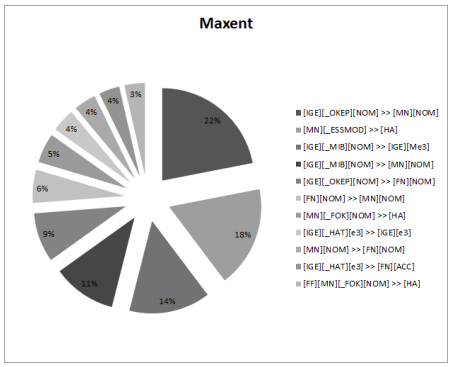
3. Az összetett rendszer

3.1. Motiváció

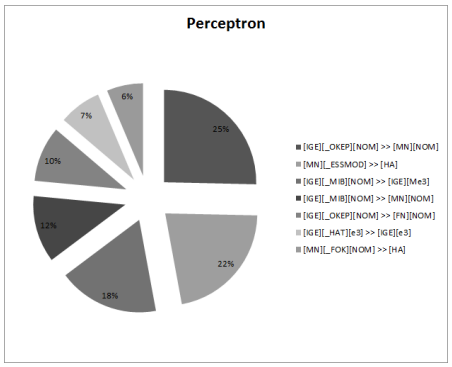
Megvizsgáltuk a négy rendszer hibáit[2] (1. és 2. ábra), s azt találtuk, hogy bár a PurePos pontossága általában magasabb társainál, de az általa vétett hibák átfedése az SMT-alapú HuLaPos rendszerrel alacsony átfedésben van. Ezen kívül nagy számban előfordulnak olyan hibák is, melyeket az OpenNLP valamely eljárása javított helyesen. Viszont az is megfigyelhető, hogy a maxent és perceptron tanulásos algoritmusok hibái jelentős részben egybeesnek. Továbbá, az Orosz

[1] Az egyértelműsítő által korábban nem látott események.

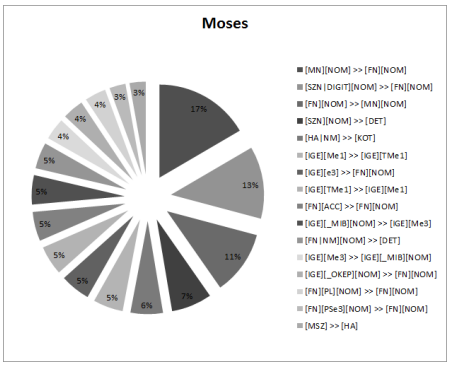
[2] A prezentált hibák az egyes taggerek által vétett hibatípusok legjellemzőbb 40%-át tartalmazzák `helyes címke >> tippelt címke` formátumban.



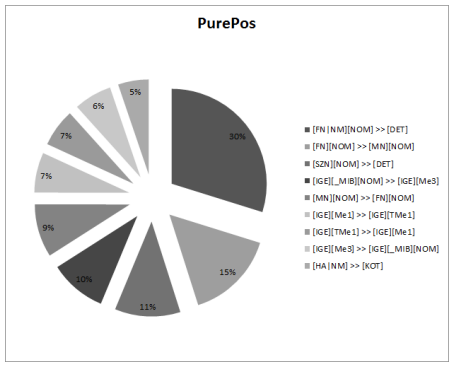
(a) A maxent tanulás gyakori hibái



(b) A perceptron tanulás gyakori hibái

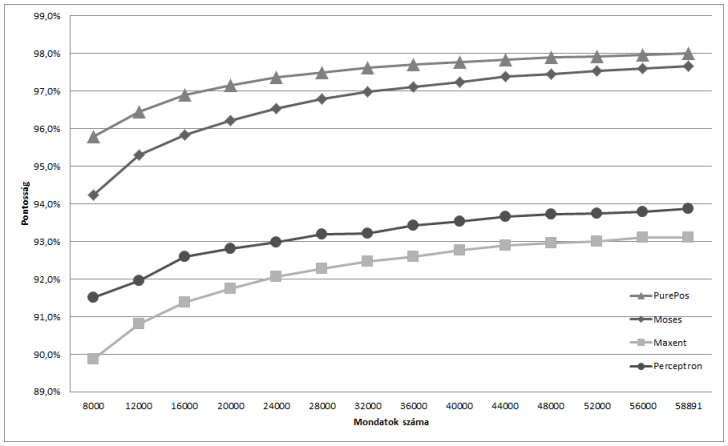


(c) A HuLaPos rendszer gyakori hibái



(d) A PurePos rendszer gyakori hibái

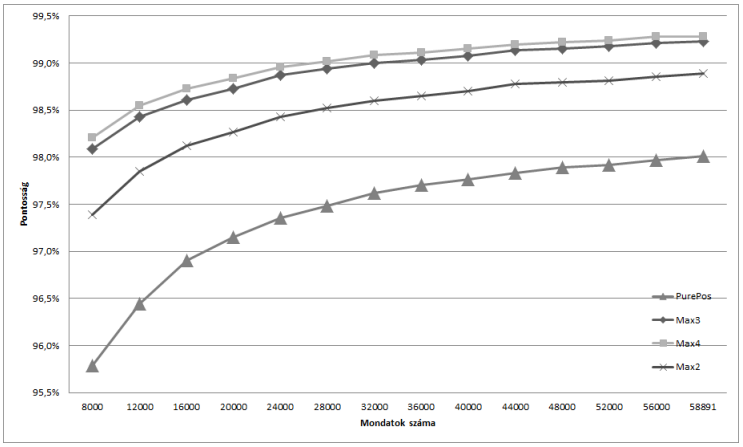
1. ábra: A szófaji egyértelműsítő rendszerek leggyakoribb hibáinak összetétele.



2. ábra: A szófaji egyértelműsítő rendszerek eredményessége a tanítóanyag méretének függvényében.

által ismertetett eszköz hibáinak legnagyobb része a határozott névelő – mutató névmás; számnév – határozatlan névelő ambiguitási osztályok rossz címkézése, míg a gépifordító-rendszer sokszor a számára ismeretlen szavakat nem tudja a megfelelő morfoszintaktikai osztályba sorolni.

A fenti hibaanalízisből merítve megvizsgáltuk, milyen maximális együttes tudással rendelkezhet egy olyan rendszer, mely az egyes rendszerek összetételéből állhat. Vizsgálatunkat a Szeged Korpuszon [6] HuMor [7,8] tagekre konvertált változatán végeztük, annak 10%-át elkülönítve tesztelési célra, míg a többin inkrementálisan tanulva vizsgáltuk, hogy hogyan változik az egyes címkézők pontossága a tanítóanyag méretének változásával. A 3. ábrán megfigyelhető, hogy a két, a három, illetve a négy rendszer szignifikánsan jobban teljesít a többinél és hogy legalább egyike a fennálló hibák legalább 44,24; 61,26; 63,90 százalékáról rendelkezik helyes információval.



3. ábra: A PoS taggerek aggregált szófaji egyértelműsítő képessége.

A továbbiakban a jelen előzetes felmérésben legjobban teljesítő kettő, illetve három rendszer összetételével foglalkozunk.

4. Kombináció

Két – hagyományos értelemben vett – osztályozó algoritmus kombinációja esetén a kutatónak „csupán” az egyes esetekhez tartozó megfelelő jellemzőhalmazt és az összetételi algoritmust kell megválasztania. Esetünkben – bár a PoS taggelés is tekinthető osztályozási problémának – a helyzet összetettebb, mert az egyes események – ami a szó és a hozzá tartozó morfoszintaktikai címke – nem függetlenek egymástól. Továbbá ezen elven alapul a legtöbb szófaji egyértelműsítő

módszer is, nevezetesen egy mondathoz tartozó legvalószínűbb címkesorozat keresése az egyes szavakhoz tartozóak helyett. Így két út áll előttünk: az összetétel alapjaként tekinthetjük az egyes mondatokat, így a két rendszer valódi kimenete között választva, vagy a tokenszintű címkézési hibákat javítjuk. A hibaanalízisben részletezetteknek megfelelően, jelen írásunkban a második eshetőséget vizsgáljuk.

Az elsőként elkészített kombinációs technika egy egyszerű többségi szavazáson alapuló algoritmus volt. Páratlan számú résztvevőt használva a három előzetesen legjobban teljesítőt választottuk: a PurePos, az SMT-alapú és a perceptron tanulásos algoritmust használtuk. A szavazás azon fázisában, amikor a három rendszer nem tud dönteni, a legjobbnak vélt PurePos rendszer szavazatát tekintjük helyesnek. Ezzel az egyszerű módszerrel relatív 12,05%-os javulást értünk el szószintű pontosságot tekintve.

Következő lépésként az előzetesen két legjobban teljesítő rendszert kombináltuk. Két osztályozó között a többségi szavazás nem tud működni, így az alábbi algoritmust alkalmaztuk: a két címkéző mondatonként végzi az annotálást, majd a szavakat egyesével megvizsgálva, ha egy szónál egyetértés van a taggerek között, akkor azt elfogadjuk, ellenben egy gépi tanulásos algoritmus a korábban látott hibák alapján eldönti, hogy mely egyértelműsítő szavazatát részesítse előnyben. A hagyományos szófaji címkék tanulásához szükséges tanítóhalmaz mellett, elkülönítettünk egy ezzel diszjunkt, a címkézők hibáinak tanulásához szükségeset is. Így kutatásunkat a Szeged Korpusz egy részén képzett 50000 mondat méretű tanítóanyagon végeztük, melyen felül még 5000 mondatot használtunk a másodszintű tanításra, melyhez az alábbi szószintű jellemzőket találtunk: a szó, a megelőző szó, a kettővel megelőző szó, következő szó, kettővel rákövetkező szó, PurePos-címketipp, HuLaPos-címketipp, PurePos címketippje a következő szóra és a megelőző szóra, tartalmaz-e kötőjelet, tartalmaz-e pontot, nagybetűvel kezdődik-e, maximum 10 hosszú suffixek.

1. táblázat: A egyes kombinációs algoritmusokkal elért pontosság.

	NaiveBayes	PRISM	IB1
Pontosság	98,48%	98,23%	**98,51%**

2. táblázat: A kombinációs módszerek eredményessége a kiindulási rendszerek tükrében.

	HuLaPos	PurePos	PurePos(M)	Max2	Comb3	IB1
Pontosság	97,40%	**97,82%**	98,53%	98,77%	98,08%	**98,51%**

A 8000 mondatból álló optimalizálásra szánt halmazon – ami 133752 tokenből és 3566 másodszintű eseményből[3] áll – megvizsgáltunk, miként teljesít néhány,

[3] Azon esetek, amikor a két tagger tippje nem egyezik.

a WEKA [9] keretrendszeren keresztül elérhető algoritmus, melyek eredményességéről a 1. táblázatban számolunk be. A legjobbnak vélt IB1 [10] algoritmussal való kombináció pontosságát egy új teszthalmazon összevetettük a már meglévő egyértelműsítőink eredményével (2. táblázat). (A táblázatban a PurePos(M) a morfológiai tudást alkalmazó rendszert, a Max2 a HuLaPos és a nyelvfüggetlen PurePos maximális tudását, a Comb3 a három rendszerből álló egyszerű szavazást, míg az IB1 a két elemből álló összetételt jelöli.) Azt találtuk, hogy az így készített – előzetes nyelvi tudást nélkülöző – rendszer, szószintű pontosságot tekintve megelőzi a három rendszerből álló egyszerű többségi szavazást, sőt hibák számát tekintve versenyképes a PurePos morfológiát tartalmazó változatával is. A teszthalmazon mérve csupán relatív 1,22%-os a két eljárás közti hibák relatív különbsége.

5. Összefoglalás

Cikkünkben bemutattunk két csak statisztikai módszeren alapuló szófaji egyértelműsítő rendszert és azok jellemző hibáit, melyekből kiindulva megvizsgáltuk azok kombinációjának lehetőségét. Megmutattuk, hogy a két eszköz együttes tudása jelentősen meghaladhatja az önálló rendszerekét. Az elérhető tudás kihasználása érdekében tett erőfeszítésünk eredményeként ismertettünk két összetételi technikát. Az utóbb prezentált rendszer nemcsak hogy meghaladja a három rendszerből álló szavazásos összetétel eredményeit, de bemutattuk, hogy egyes esetekben olyan más eljárásokkal is versenyképes, melyek integrált nyelvi tudással dolgoznak.

Eredményeink bizakodásra adnak okot, így jövőbeni tervünk, hogy megvizsgáljuk, miként lehetséges a két algoritmusra alkalmazott összetételi technikát kiterjeszteni három vagy több rendszerre.

Hivatkozások

1. Orosz, Gy., Novák, A.: PurePos – an open source morphological disambiguator. In Sharp, B., Zock, M., eds.: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science, Wroclaw (2012) 53–63
2. Laki, L.J.: Investigating the Possibilities of Using SMT for Text Annotation. In: SLATE 2012 - Symposium on Languages, Applications and Technologies, Braga, Portugal, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2012) 267–283
3. Brants, T.: TnT - A Statistical Part-of-Speech Tagger. In: Proceedings of the sixth conference on Applied natural language processing. Number i, Universität des Saarlandes, Computational Linguistics, Association for Computational Linguistics (2000) 224–231
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, Prague, Czech Republic, Association for Computational Linguistics (2007) 209–212
5. Baldridge, J., Morton, T., Bierner, G.: The OpenNLP maximum entropy package (2002)

6. Csendes, D., Csirik, J., Gyimóthy, T.: The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. In: Proceedings of the 5th International Workshop on Linguistically Interpreted Corpora LINC 2004 at The 20th International Conference on Computational Linguistics COLING 2004. (2004) 19–23
7. Novák, A.: Milyen a jó humor? In: Magyar Számítógépes Nyelvészeti Konferencia 2003, Szeged (2003) 138–145.
8. Prószéky, G., Novák, A.: Computational Morphologies for Small Uralic Languages. In: Inquiries into Words, Constraints and Contexts., Stanford, California (2005) 150–157
9. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software. ACM SIGKDD Explorations Newsletter **11**(1) (2009) 10
10. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Machine Learning **6**(1) (1991) 37–66
11. Kuba, A., Felföldi, L., Kocsor, A.: POS tagger combinations on Hungarian text. In: 2nd International Joint Conference on Natural Language Processing, Jeju Island, Republic of Korea, Association for Computational Linguistics (2005)
12. 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)