HuLaPos2 – Fordítsunk morfológiát

Laki László[1,2], Orosz György[1,2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
1083 Budapest, Práter utca 50/a
e-mail:{laki.laszlo,orosz.gyorgy}@itk.ppke.hu

Kivonat Jelen munkánkban bemutatunk egy gépi fordításon alapuló nyelvfüggetlen teljes morfológiai egyértelműsítő rendszert, ami egyidejűleg végzi a szótövesítést és a morfológiai egyértelműsítést. Annak érdekében, hogy demonstráljuk a módszer hatékonyságát, több különböző nyelv legjobb rendszerével hasonlítottuk össze. A legtöbb nyelv esetén rendszerünk jobban teljesít szófaji egyértelműsítés tekintetében, valamint a szótövesítés pontossága hasonló az általunk összehasonlított rendszerekével.

1. Bevezetés

A nyelvtechnológiai feldolgozási lánc fontos elemei a morfológiai elemzés és egyértelműsítés. Az utóbbi komponens feladata, hogy egyértelműen meghatározza a szavak szótövét, és megállapítsa azok morfoszintaktikai (PoS) címkéit. Az első, erre a célra létrehozott eszközök angol nyelvű szövegek elemzésére szolgáltak, melyek azonban egymást követően végezték a PoS címkézést és a szótövesítést. Így az ezek alapján létrehozott újabb rendszerek is ezt a sémát követték. Következésképp kevés olyan eszköz létezik, amely teljes morfológiai egyértelműsítést végez, ami elengedhetetlen morfológiailag gazdag nyelvek elemzése esetén. Továbbá csak néhány olyan eljárás létezik, amely grammatikailag nagyon különböző nyelvek esetében is ugyanolyan magas pontossággal képes működni. Bár az egyes nyelvspecifikus eszközök sokszor magas pontosságot produkálnak, de a legtöbbször csak egy-egy nyelv nagy teljesítményű elemzésére korlátozódik a tudásuk.

A tanulmány célja egy Moses SMT[3] rendszeren alapuló nyelvfüggetlen morfológiai elemző rendszer bemutatása, amely különböző típusú nyelvek esetén végez teljes morfológiai egyértelműsítést úgy, hogy pontossága felveszi a versenyt a nyelvfüggő társai eredményeivel.

Dolgozatunk első részében ismertetjük a létrehozott rendszer (HuLaPos2) felépítését, majd bemutatjuk az általa elért eredményeket összehasonlítva azokat hat különböző nyelv state-of-the-art egyértelműsítő eredményeivel.

[3] Statisztikai gépi fordító

2. Kapcsolódó munkák

Az első általánosan elterjedt statisztikai taggerek rejtett Markov-modellen alapultak, úgymint a TnT [1] vagy a HunPos [2]. Ezzel párhuzamosan Ratnaparkhi [3] bemutatott egy maximum entrópián alapuló megközelítést, amit számos nyelv esetében sikerrel alkalmaztak (pl. a Stanford tagger [4] különböző adaptációi, vagy a `magyarlanc` [5]). Ezeken kívül számos más felügyelt tanulásos módszer is jól teljesít különböző nyelvek esetében: úgymint Brill transzformációalapú módszere [6], az SVMTool [7] Support Vector Machine alapú modellje, vagy a TreeTagger [8] döntési fákat használó algoritmusa.

Mora és Sánchez [9] voltak az elsők, akik SMT módszert használtak szófaji egyértelműsítésre, de ők a rendszert csak az angol nyelv PoS taggelésére tervezték, lemmatizálásra nem. Munkájukban a tanítóanyagban nem előforduló szavak (OOV) kezelésére egy szógyakoriságon alapuló modellt és egy 11 elemből álló szuffixum listát alkalmaztak.

Hasonló megközelítést használtunk egy korábbi munkánkban [10], ahol a fenti metódust magyar nyelvre alkalmaztuk. A Mora és Sánchez által angol nyelvre optimalizált algoritmus jelentős mértékben alulmaradt a legjobb magyar elemzőkhöz képest (pl. a morfológiai elemzővel kiegészített PurePos [11]). Ez többek között azzal is magyarázható, hogy a magyar nyelv agglutináló tulajdonságaiból adódóan fejlettebb módszerek szükségesek a jelentős számú OOV tokenek kezelésére. Ebben a tanulmányban a Laki-rendszer továbbfejlesztett változatát mutatjuk be.

3. Elméleti háttér

3.1. Kifejezésalapú statisztikai gépi fordítás

A gépi fordítórendszer leképezést biztosít két nyelv között függetlenül attól, hogy ezek természetes vagy mesterséges nyelvek. A statisztikai gépi fordító algoritmusok párhuzamos kétnyelvű korpuszokból gépi tanulásos módszerek segítségével tanulják meg a transzformációhoz szükséges modelleket.

Ha W egy mondat a forrásnyelvi szövegből, melynek a helyes fordítása $\hat{T}$, akkor a fordítási feladat a következőképpen formalizálható:

$$\hat{T} = \underset{T}{argmax}\ P(T|W) = \underset{T}{argmax}\ P(W|T)P(T) \tag{1}$$

ahol $P(T)$ a nyelvi modell és $P(W|T)$ fordítási modell. Míg az első modell a lefordított szöveg olvashatóságára (folyékonyságára) ad becslést, addig a második modell a fordítás minőségét becsüli. A statisztikai gépi fordítás egyik gyakran használt változata a kifejezésalapú fordítás, melynek alapja, hogy a fordítandó W mondatot kifejezésekre bontjuk $W = w_1 w_2 \dots w_N = w_0^N$, amiket különkülön lefordítunk. A lefordított részek legjobb kombinációját véve kapjuk a célnyelvi mondatot ($T = t_0^N$). A kifejezések fordítását a párhuzamos tanító anyagból számolt $\phi(w_i^{i+k_1}|t_i^{i+k_2})$ valószínűségi eloszlás alapján végzi a rendszer. Ezek használatával a (1) a következőképpen fejthető ki:

$$\underset{T}{argmax}\ P(W|T)P(T) = \underset{T}{argmax}\left[\prod_{i=0}^{N}\ \phi(w_i^{i+k_1}|t_i^{i+k_2})P(t_i|t_{i-1}^{i-j})\right] \qquad (2)$$

3.2. Morfológiai egyértelműsítés mint gépi fordítási feladat

A szófaji címkézés feladatára számos módszer létezik, melyek közül a legelterjedtebbek a rejtett Markov-modellezésen (HMM) alapulók. Ennek működése a következőképpen (vö. (3)) írható le formálisan: ha W az elemzendő szöveg egy mondata, mely helyes elemzésének címkesorozata $\hat{T}$, akkor ennek valószínűsége maximális a címkeátmenet-modell $P(T)$ és a lexikai-modell $P(W|T)$ szorzatát tekintve. A legtöbb rendszer (így pl. a TnT és a HunPos is) az első valószínűségi értéket egy másodrendű modellel becsli, ami lényegében egy címkékre épülő trigram modell: $P(t_i|t_{i-1}^{i-2})$. A lexikai-modell becslésére pedig legtöbbször maximum likelihood becslést alkalmaznak, ami a szavakhoz rendelt morfoszintaktikai címkék relatív gyakoriságából tevődik össze: $P(w_i|t_i)$.

$$\hat{T} = \underset{T}{argmax}\ P(W|T)P(T) = \underset{T}{argmax}\left[\prod_{i=0}^{N}\ P(w_i|t_i)P(t_i|t_{i-1}^{i-2})\right] \qquad (3)$$

Összevetve a (1) és (3) egyenleteket láthatjuk, hogy a statisztikai gépi fordítás feladata könnyen megfeleltethető a morfológiai címkézés HMM módszerének. A megfeleltetés lépései: az SMT nyelvi modellje a címkeátmenet-valószínűség modell, míg a fordítási modell a lexikai modellnek felelnek meg. A leképezésen túl az is megfigyelhető még, hogy az SMT-n alapú megközelítés egy általánosabb keretrendszert biztosít a feladat megoldására

Motivációnk a nyílt forráskódú Moses SMT toolkit [12] keretrendszert használatára a következők voltak:

1. A Moses tanítási lánca gyors a valószínűségi modellek létrehozását illetően.
2. A leggyakrabban alkalmazott HMM alapú elemzőkkel szemben a Moses rendszer által létrehozott fordítási modell nemcsak egy-egy szó lehetséges elemzését tartalmazza, hanem a hosszabb kifejezésekéit is, ami lehetővé teszi az elemző számára, hogy a szöveg hosszabb részeit is egy egységként kezelje.
3. A címkeátmenet-valószínűség modell (a nyelvmodell) építése során nemcsak az azt megelőző két szó elemzését veszi figyelembe, hanem akár a mondatban szereplő összes megelőzőét, valamint a létező egyik legjobb simító algoritmust, a módosított Kneser-Ney simítást [13] használja.
4. A dekóder a beam-search algoritmus egy hatékony és gyors változatát az úgynevezett verem dekódolást alkalmazza. A módszer legnagyobb előnye, hogy az elemzést a dekódoló működésének köszönhetően a szavak tetszőleges sorrendjében végezheti, szemben a HMM-alapú elemzők szigorúan balról jobbra történő működésével.
5. A dekódolás folyamatába egyszerűen integrálható morfológiai guesser vagy elemző.

4. A rendszer bemutatása

Ebben a fejezetben áttekintjük azokat a legfontosabb módosításokat, amelyek megkülönböztetik az eredeti SMT rendszert a morfológiai egyértelműsítőtől (egy részletesebb leírás a [14] cikkünkben olvasható).

A szuffixumokat használó ragozó nyelvek esetén (mint például a magyar vagy a török) a szótövek egyszerűen leírhatók olyan rekordokkal, melyek megadják azt a szükséges transzformációt, amit el kell végezni egy adott szón, hogy megkapjuk annak szótövét. Egy ilyen rekord: $\langle cut, paste \rangle$, ahol a *cut* a sztringről eltávolítandó karakterek számát adja meg, a *paste* pedig az a karaktersorozat, amit illeszteni kell a „csonka szó" végére, hogy megkapjuk a szótövet. Ezt az ötletet használva az elemzőnk a morfoszintaktikai címkék mellett képes még reprezentálni a szótöveket is.

Másrészt természetes nyelvek esetében az SMT rendszer szóösszekötője gépi tanulásos algoritmusokat használ a fordítási frázispárok meghatározásához. Ez a mi esetünkben a feladat felesleges bonyolítása, mivel a morfológiai egyértelműsítéshez egy egyértelmű monoton megfeleltetésre van szükség, mely a tokeneket az elemzéseikhez rendeli. Ezért a HuLaPos2 rendszerben a Giza++ algoritmust monoton leképezéssel helyettesítettük.

Harmadrészt, a Moses dekóder legnagyobb előnye, hogy hosszabb kifejezéseket is képes egy egységként fordítani, de itt a frázisok maximális hossza és a nyelvi modell mérete nagyban befolyásolja a rendszer minőségét. Ezért szükséges ezen paramétereinek finomhangolása, amihez az optimális beállításokat – minden nyelvre külön-külön – empirikusan határoztuk meg.

Végül az adathiány által okozott problémák elkerülése érdekében a számjegyek generikus szimbólumokkal lettek helyettesítve a tanítóhalmazban és a bemeneti szövegben egyaránt. Az SMT rendszer legnagyobb hiányossága, hogy a tanítóhalmazban nem szereplő szavakat figyelmen kívül hagyja, és semmilyen elemzést sem ad hozzájuk. Ennek kiküszöbölésére rendszerünkbe – a PurePos és HunPos rendszerekhez hasonlóan – egy trie-alapú suffix-guessert építettünk, amely elemzési javaslatokat ad az OOV szavakra. Ez az algoritmus a tanítóhalmazban ritkán előforduló szavak végződései alapján képes megbecsülni, hogy egy szó az egyes *(szótő-transzformáció; címke)* elemzésekkel milyen valószínűséggel címkézhető. Ennek a módszernek további előnye, hogy az elemzések valószínűségének számítása – a TnT-hez hasonlóan – különböző hosszúságú toldalékok simított interpolált modellje alapján történik. Ráadásul ez az algoritmus megoldást nyújt az SMT rendszer azon gyengeségére, miszerint az OOV szavakat tartalmazó szegmensek elemzése során a dekódoló csak az unigram modelleket használhatja. Mivel ez a modul arra hivatott, hogy a ritkán előforduló szavakat kezelje, ezért ilyen tulajdonságú szavakon kell betanítani. A ritka szavak esetén a használt küszöbértéket empirikusan határoztuk meg: a legmagasabb pontosságot általában akkor értük el, amikor ez az érték 2 volt, azaz a guesser csak hapaxokon volt tanítva. A javasló komponens a következő módon lett a dekódolóba integrálva: A Moses képes a kifejezések fordítása közben előre definiált fordítási javaslatokat is figyelembe venni. Ezzel az egyszerű módszerrel a tanítóhalmazban nem szereplő szavakhoz hozzárendeljük a guesser javaslatait, mint előfordítás.

5. Eredmények

A HuLaPos2 rendszert több különböző nyelvhez (magyar, szerb, horvát, bolgár, portugál és angol) elérhető legjobb pontossággal teljesítő egyértelműsítő rendszerekkel hasonlítottuk össze. A tanító- és a teszthalmazt a kapcsolódó publikációkban leírt módon (részletesen lentebb) definiáltuk. A rendszerek pontosságának részletes összehasonlítását a 1-es és 2-es táblázatokban foglaltuk össze, ahol az első táblázatba gyűjtöttük össze azokat a rendszereket, amelyek teljes morfológiai egyértelműsítést csinálnak, míg a második táblázatban szereplők csak morfológiai egyértelműsítést végeznek.

1. táblázat. A HuLaPos2 rendszer minőségének összehasonlítása más rendszerekével a szófaji egyértelműsítés, szótövesítés, valamint a teljes morfológiai egyértelműsítés tekintetében

Nyelv	Rendszer	Szószíntű pontosság		
		címkézés	szótövesítés	teljes
magyar (MSD)	HuLaPos2	**99,57%**	**97,24%**	96,84%
	PurePos	96,74%	96,35%	94,76%
magyar (HUMor)	HuLaPos2	**99,18%**	98,23%	97,62%
	PurePos	96,50%	96,27%	94,53%
	PurePos+MA	98,96%	**99,53%**	98,77%
horvát	HuLaPos2	**93,25%**	96,21%	90,77%
	HunPos+CST	87,11%	**97,78%**	–
szerb	HuLaPos2	**92,28%**	92,72%	86,51%
	HunPos+CST	85,00%	**95,95%**	–

Magyar nyelv esetében a legjobb egyértelműsítő rendszer a PurePos [11], ami egy HMM-alapú teljes morfológiai egyértelműsítő, melybe morfológiai elemző van integrálva. Az eredmények összehasonlításához a Szeged Korpuszt [15] választottuk, melynek két változatán teszteltük rendszerünket: az eredeti MSD-kódolással készültet, és egy HuMor [16] címkékre automatikusan átírtat. A HuLaPos2 rendszert a PurePos rendszer morfológiai elemzőt használó, valamint anélkül működő (tehát nyelvfüggetlen) változataival hasonlítottuk össze. Az eredmények megmutatták, hogy a HuLaPos2 az összes mért esetben jobb eredményt ért el a PurePos morfológiai elemző nélküli változatával szemben, és szófaji címkézés esetén pontossága meghaladja a PurePos morfológiai elemzős változatát.

Szerb és horvát nyelvre Agić és munkatársai [17] készítettek szófaji címkéző és szótövesítő alkalmazást 2013-ban. A rendszert a HunPos és a CST szótövesítő [18] kombinációjából építették fel, és a SETimes.HR [17] korpuszon tanították. Az 1. táblázat eredményeiből látható, hogy PoS taggelés esetén a HuLaPos2 teljesítménye szignifikánsan meghaladja Agićék rendszerét, míg a szótövesítésben elért eredmény is közelít annak eredményességéhez. A különbség a javasló algo-

ritmus működéséből ered: a CST rendszerben a szótő-transzformációk nemcsak szuffixumok lehetnek, hanem a tetszőleges helyű változások is. Ezzel szemben a HuLaPos2 által használt guesser csak a szóvégi változást képes kezelni.

Georgi Georgiev és munkatársai [19] létrehoztak egy morfológiai lexikonnal és nyelvtani szabályokkal kiegészített irányított tanuláson alapuló szófaji egyértelműsítő rendszert bolgár nyelvre. Eszközüket a BulTreeBank korpuszon [20] tanították és tesztelték. A 2. táblázat eredményeiből látható, hogy a HuLaPos2 teljesítménye nagymértékben meghaladja a nyelvtani tudással nem rendelkező tisztán statisztikai módszereket használó rendszerek minőségét. Annak ellenére, hogy rendszerünk semmilyen nyelvspecifikus eszközzel nincs támogatva, jobban teljesít, mint a morfológiai lexikont használó eszköz, valamint pontossága megközelíti Georgiev által készített legjobb rendszerét (irányított tanulás + lexikon + szabályok).

2. táblázat. A HuLaPos2 rendszer minőségének összehasonlítása olyan rendszerekkel, amelyek csak szófaji egyértelműsítést csinálnak

Nyelv	Rendszer	Címkézés pontossága
bolgár	TnT	92,53%
	gépi tanulás	95,72%
	gépi tanulás + morf. lexikon	97,83%
	HuLaPos2	**97,86%**
	gépi tanulás + morf. lexikon + szabályok	97,98%
portugál	**HuLaPos2**	**93,20%**
	HMM-alapú PoS tagger	92,00%
angol	TnT	96,46%
	PBT (Mora and Sánchez [9])	96,97%
	HuLaPos2	**97,08%**
	Stanford tagger 2.0	97,32%
	SCCN [21]	97,50%

A HuLaPos2 rendszert teszteltük még morfológailag egyszerűbb nyelvek esetében is, mint a portugál és az angol. Mindkét esetben csak a PoS tagger eredményességét tudtuk összehasonlítani (2. táblázat), mivel az elérhető korpuszok nem tartalmazták a szavak lemmáit.

Portugál nyelvre a Maia és Xexéo [22] által 2011-ben készített HMM-alapú rendszert vettük összehasonlítási alapul. Ez az eszköz a Floresta Sintá(c)tica Treebank-en [23] lett tanítva, melyből az első 10% volt a teszthalmaz, a fennmaradó 90% pedig a tanító halmaz. Ugyanezekkel a beállításokkal a HuLaPos2 pontossága több mint 1%-kal felülmúlta a portugál címkéző eredményeit.

Ami az angol nyelvet illeti, a Penn Treebank [24] WSJ korpuszát használtuk az általánosan bevált elosztásban.[4]

[4] http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)

A 2. táblázat a HuLaPos2 és a másik négy rendszer által elért eredményeket mutatja. Megfigyelhető, hogy a HuLaPos2 meghaladja a TnT és a Mora és Sánchez-féle [9] rendszerek által elért értékeket. Az eredmények vizsgálatánál fontos még figyelembe venni, hogy algoritmusunk a tanítóanyagon kívül semmilyen más lexikai adatbázist, vagy előzetes tudást nem használ, így elmondható, hogy annak teljesítménye a maga nemében kiemelkedő.

6. Konklúzió

Írásunkban bemutattunk egy, a Moses keretrendszeren alapuló, nyelvfüggetlen teljes morfológiai egyértelműsítő rendszert. Ez az eszköz egyidejűleg végzi a szófaji egyértelműsítést és a szótövesítés feladatát egy trie-alapú suffix-guesser segítségével, amely hatékonyan kezeli a morfológiailag gazdag nyelvekre jellemző OOV szavak problémáját. A HuLaPos2 hat különböző nyelv legjobb rendszerével lett összehasonlítva. Szófaji egyértelműsítés tekintetében rendszerünk (az angol nyelv kivételével) jobb eredményt ér el a vizsgált taggerekhez képest. Mindemellett szótövesítés esetén is versenyképesnek bizonyult a nyelvfüggő vetélytársakkal szemben. Az angol nyelv esetén a HuLaPos2 meghaladja a közismert TnT rendszer eredményeit, valamint megközelíti az elérhető legjobb rendszer minőségét.

Köszönetnyilvánítás

Ez a projekt a TÁMOP–4.2.1./B–11/2-KMR-2011-0002 és a TÁMOP–4.2.2./B–10/1-2010-0014. támogatásával készült.

Hivatkozások

1. Brants, T.: Tnt - a Statistical Part-of-Speech Tagger. In: Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA (2000)
2. Halácsy, P., Kornai, A., Oravecz, C.: HunPos: An open source trigram tagger. In: Proceedings of the 45th Annual Meeting of the ACL, Stroudsburg, Association for Computational Linguistics (2007) 209–212
3. Reynar, J.C., Ratnaparkhi, A.: A maximum entropy approach to identifying sentence boundaries. In: Proceedings of the fifth conference on Applied natural language processing. ANLC '97, Stroudsburg, PA, USA, Association for Computational Linguistics (1997) 16–19
4. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13. EMNLP '00, Stroudsburg, PA, USA, Association for Computational Linguistics (2000) 63–70
5. Zsibrita, J., Vincze, V., Farkas, R.: Ismeretlen kifejezések és a szófaji egyértelműsítés. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Tudományegyetem (2010) 275–283

6. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational Linguistics **21** (1995) 543–565
7. Giménez, J., Màrquez, L.: SVMTool: A general POS tagger generator based on Support Vector Machines. In: In Proceedings of the 4th International Conference on Language Resources and Evaluation. (2004) 43–46
8. Schmid, H.: Improvements In Part-of-Speech Tagging With an Application To German. In: In Proceedings of the ACL SIGDAT-Workshop. (1995) 47–50
9. Gascó I Mora, G., Sánchez Peiró, J.A.: Part-of-Speech tagging based on machine translation techniques. In: Proceedings of the 3rd Iberian conference on Pattern Recognition and Image Analysis, Part I. IbPRIA '07, Berlin, Heidelberg, Springer-Verlag (2007) 257–264
10. Laki, L.: Investigating the Possibilities of Using SMT for Text Annotation. In Simões, A., Queirós, R., da Cruz, D., eds.: 1st Symposium on Languages, Applications and Technologies. Volume 21 of OpenAccess Series in Informatics (OASIcs)., Dagstuhl, Germany, Schloss Dagstuhl–Leibniz-Zentrum für Informatik (2012) 267–283
11. Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria (2013) 539–545
12. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, Association for Computational Linguistics (2007) 177–180
13. James, F.: Modified Kneser-Ney smoothing of n-gram models. Technical report (2000)
14. Laki, L.J., Orosz, Gy., Novák, A.: HuLaPos 2.0 – Decoding morphology. In: 12th Mexican International Conference on Artificial Intelligence, Mexico City, Mexico (2013)
15. Csendes, D., Csirik, J., Gyimóthy, T. In: The Szeged Corpus: A POS Tagged and Syntactically Annotated Hungarian Natural Language Corpus. Volume 3206 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2004) 41–47
16. Novák, A.: What is good Humor like? In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2003) 138–144
17. Agić, Ž., Ljubešić, N., Merkler, D.: Lemmatization and Morphosyntactic Tagging of Croatian and Serbian. In: Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, Sofia, Bulgaria, Association for Computational Linguistics (2013) 48–57
18. Jongejan, B., Dalianis, H.: Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Suntec, Singapore, Association for Computational Linguistics (2009) 145–153
19. Georgiev, G., Zhikov, V., Simov, K.I., Osenova, P., Nakov, P.: Feature-rich part-of-speech tagging for morphologically complex languages: Application to bulgarian. In Daelemans, W., Lapata, M., Màrquez, L., eds.: EACL, The Association for Computer Linguistics (2012) 492–502
20. Chanev, A., Simov, K., Osenova, P., Marinov, S. In: The BulTreeBank: Parsing and conversion. Volume 309 of Current Issues in Linguistic Theory. John Benjamins, Amsterdam & Philadelphia (2007) 321–330

21. Søgaard, A.: Semisupervised condensed nearest neighbor for part-of-speech tagging. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2. HLT '11, Stroudsburg, PA, USA, Association for Computational Linguistics (2011) 48–52
22. Maia, M.R.d.H., Xexéo, G.B.: Part-of-speech tagging of Portuguese using hidden Markov models with character language model emissions. Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology (2011) 159–163
23. Freitas, C., Rocha, P., Bick, E.: Floresta Sintá(c)tica: Bigger, thicker and easier. In: Proceedings of the 8th international conference on Computational Processing of the Portuguese Language. PROPOR '08, Berlin, Heidelberg, Springer-Verlag (2008) 216–219
24. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A.: Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics **19**(2) (1993) 313–330