Statisztikai gépi fordítási módszereken alapuló egynyelvű szövegelemző rendszer és szótövesítő

Laki László János[1]

Pázmány Péter Katolikus Egyetem, ITK,
1083, Budapest, Práter u. 50/a,
e-mail: laki.laszlo@itk.ppke.hu

Kivonat Jelen munkában az SMT módszer alkalmazhatóságát vizsgáltam szófaji egyértelműsítő és szótövesítő feladat megoldására. Létrehoztam egy alaprendszert, illetve további lehetőségeket próbáltam ki a rendszer eredményeinek javítására. Megvizsgáltam, milyen hatást gyakorol a célnyelvi szótár méretének változtatása a rendszer minőségére, továbbá megoldást kerestem a tanító halmazban nem szereplő szavak elemzésének megoldására.

Kulcsszavak: Statisztikai Gépi Fordítás (SMT), szófaji egyértelműsítés (POS tagging), szótövesítés, Szeged Korpusz, OOV

1. Bevezetés

Az informatika fejlődése szinte az összes tudományág számára új lehetőségek halmazát nyitotta meg, és ez nincs másképp a nyelvészetben sem. Napjaink számítógépei segítségével képesek lettünk óriási méretű szöveges anyagok gyors és hatékony kezelésére, feldolgozására. A szövegek szintaktikai és/vagy szemantikai információval történő jelölése, valamint a szavak szófaji elemzése rendkívül fontos feladat a számítógépes nyelvészet számára. A szófaji egyértelműsítés problémája korántsem megoldott, annak ellenére, hogy sokféle rendszer létezik ennek implementálására. A legelterjedtebbek a gépi tanuláson alapulnak, melyek maguk ismerik fel a szabályokat a különböző nyelvi jellemzők segítségével. További nehézséget jelent azonban ezen jellemzők meghatározása, hiszen a különböző sajátosságok nehezen fogalmazhatók meg.

Ezzel szemben a statisztikai gépi fordító (SMT) rendszerek előzetes nyelvi ismeret nélkül képesek a fordításhoz szükséges szabályok felismerésére. Kézenfekvő megoldásnak tűnik SMT rendszerek alkalmazása szövegelemzésre. Munkám során az ebben rejlő lehetőségeket vizsgáltam a szófaji egyértelműsítés és szótövesítés feladatának megoldására.

2. A szófaji egyértelműsítés

Szófaji egyértelműsítés az a folyamat, amely a szövegben található szavakat általános lexikai jelentésük és kontextusuk alapján megjelöli a megfelelő POS cím-

kével. Egy helyesen címkézett mondatban minden szóhoz pontosan egy címke van rendelve. Ennek ellenére a szófaji egyértelműsítés sokkal komplexebb feladat egy szó és címkéjének listájából való kikereséshez képest, mivel egy szónak több szófaji alakja is lehet.

Erre a feladatra létrehozott első megoldások előre megírt szabályrendszerek segítségével elemezik a szöveget. A probléma ezekkel a rendszerekkel a szabályok létrehozásának magas költsége volt. Napjaink elterjedt rendszerei gépi tanuláson alapuló módszereket használnak, amelyek különböző nyelvi jellemzők segítségével maguk ismeri fel a szabályokat, ám a megfelelő jellemzők meghatározása szintén nehéz feladat. A különböző nyelvi sajátosságok nehezen fogalmazhatók meg és állíthatók össze olyan teljes, mindent magába foglaló szabályrendszerré, mely a számítógép számára feldolgozható. Ilyen nyelvi sajátosságok lehetnek például a nyelvek közötti fordítás szabályai, valamint a morfológiai elemzés.

A szófaji egyértelműsítők teljesítményének egyik nagyon fontos tényezője a tanítóhalmazban nem szereplő szavak (OOV: out-of-vocabulary) elemzése. Az OOV szavak elemzése nagyban függ az elemzendő nyelvtől. Például az angol nyelv esetében nagy valószínűséggel az OOV szavak tulajdonnevek lesznek. Ezzel szemben más nyelvek esetében – mint a magyar vagy a mandarin kínai – az OOV szavak főnevek és igék is lehetnek.[1]

2.1. A szótövesítés

Lemmatizálás számítógépes nyelvészeti szempontból az az algoritmikus folyamat, amelyik meghatározza egy szó szótári alakját. Napjainkban több megvalósítás is létezik ezen feladat megoldására (például: HUMOR [2]), de ezek általában bonyolult módszereket alkalmaznak. Ezzel szemben az SMT rendszeren alapuló szótövesítés előzetes nyelvtani ismeret nélkül végzi el ezt a feladatot.

2.2. Létező megvalósítások

Oravecz és Dienes 2002-ben készítették el az első magyar nyelvű sztochasztikus POS-taggert. A rendszer MSD-kódokat használ és 98.11%-os pontosságot ért el [3].

Halácsy et al. létrehoztak egy maxent modellen alapuló szófaji egyértelműsítőt. Csoportjával 2007-ben létrehozták a HunPOS nevű rendszert, ami napjaink legjobb magyar nyelvű POS-taggerjének számít. A rendszer MSD-kódokat használ és 98.24%-os pontosságot ért el [4].

3. Statisztikai gépi fordítás

A statisztikai nyelvfeldolgozás elterjedt alkalmazása a gépi fordítás. A statisztikai gépi fordító (SMT) módszer nagy előnye a szabályalapú fordítással szemben, hogy az architektúra létrehozásához nem szükséges a nyelvek grammatikájának ismerete. A rendszer tanításához csupán egy kétnyelvű korpuszra van szükség, amelyből statisztikai megfigyelésekkel nyerjük ki a szabályokat. A fordítás során

az egyetlen, amit biztosan tudunk, az a mondat, amit le szeretnénk fordítani (forrásnyelvi mondat). Ezért a fordítást úgy végezzük, mintha a célnyelvi mondatok halmazát egy zajos csatornán átengednénk, és a csatorna kimenetén összehasonlítanánk a forrásnyelvi mondattal.

$$\hat{E} = \underset{E}{argmax}\ p(E|F) = \underset{E}{argmax}\ p(F|E) * p(E) \tag{1}$$

Az a mondat lesz a rendszerünk kimenete ($\hat{E}$), amelyik a legjobban hasonlít a fordítandó (forrásnyelvi) mondatra. Ez a hasonlóság lényegében egy valószínűségi érték, amely a nyelvi modellből $p(E)$ és a fordítási modellből $p(F|E)$ számolható. Lásd az 1. egyenletben.

4. A POS-Tagging probléma mint SMT-probléma

Amint a bevezetőben már említettem, a szövegelemzés is megfogalmazható fordítási feladatként. Egy tetszőleges mondat (F) szófaji elemzése ($\hat{E}$) megfogalmazható a következő egyenlettel:

$$\hat{E} = \underset{E}{argmax}\ p(E|F) = \underset{E}{argmax}\ p(F|E) * p(E) \tag{2}$$

ahol $p(E)$ a címkék nyelvi modellje és $p(E|F)$ a fordítási/elemzési modell. A fordítási feladathoz hasonlóan a forrásnyelvi mondatot kifejezések halmazának tekintjük, ahol minden frázist a címkék egy halmazára „fordítunk”. Egy természetes nyelvek közti fordításhoz képest a szófaji egyértelműsítés egyszerűbb az SMT-rendszerek számára, hiszen nincs szükség a mondatban elhelyezkedő szavak sorrendjének megváltoztatására. A fordítás során a forrásnyelvi és célnyelvi oldal szavainak száma is megegyezik, azaz a rendszer nem végez elembeszúrást és törlést.[1,5] Ezen tulajdonságok miatt az SMT-rendszer jól alkalmazható megvalósításnak tűnik szófaji egyértelműsítésre.

5. Munkám során alkalmazott rendszerek

A következő fejezetben bemutatom a munkám során alkalmazott keretrendszereket.

5.1. MOSES

Több módszert is megvizsgáltam, melyek képesek párhuzamos korpuszból információt kinyerni. Végül az IBM modellek mellett döntöttem, mivel hatékony, viszonylag pontos, és a feladatnak nagyon jól megfelelő algoritmusnak bizonyultak. Ezért kezdtem használni a Moses keretrendszert [6,7,8], amely implementálja ezeket a modelleket. Ebben a rendszerben megtalálható a párhuzamos korpusz előfeldolgozása, a fordítási és nyelvi modellek létrehozása, a dekódolás, valamint a BLEU-metrikára való optimalizálás.

5.2. Joshua

Másfelől a Joshua keretrendszert [9] használtam, mely nem pusztán szó- vagy frázisszintű statisztikai valószínűségi modelleket használ, hanem bizonyos nyelvtani jellemzők előfordulását is figyelembe veszi. A Joshua rendszer további nagy előnye, hogy képes ezen generatív szabályok közti fordításra oly módon, hogy megadhatóak a szabályok mind a forrásnyelvre, mind a célnyelvre, valamint az is definiálható, hogy mekkora valószínűséggel transzformálhatók át a szabályok egymásba.

5.3. Korpusz

Az SMT-rendszer tanításához szükséges kétnyelvű párhuzamos korpuszt, a Szeged Korpusz 2.0-t használtam. A korpusz előnyei, hogy a szavak MSD-kódolású POS-címkéi mellett azok szótövei is szerepelnek benne, általános témájú, valamint készítői kézzel ellenőrizték annak helyességét. Hátránya, hogy viszonylag kis méretű. Mivel a szófaji címkék elemszáma korlátozott, ezért elvben kisebb méretű korpuszban is elég nagy gyakorisággal szerepelhetnek. [10,11]

5.4. Kiértékelő módszerek

A rendszer minőségének kiértékeléséhez a BiLingual Evaluation Understudy (BLEU) módszert használtam, amely egy gyakran alkalmazott módszer az SMT-rendszerek minőségének vizsgálata. Lényege, hogy a fordításokat referenciafordításokhoz hasonlítja, majd hozzájuk egy 0 és 1 közötti valós értéket rendel. Ezt BLEU-értéknek nevezzük. Tanulmányomban ennek százalékosított formáját használom. [12]

Másfelől egy Levenshtein távolságon alapuló automatikus módszer segítségével kiszámítottam az elemző rendszer pontosságát a mondatok és a tokenek szintjén egyaránt.

6. Eredmények

6.1. Az alaprendszer létrehozása

Az első betanítás. Mint már korábban említettem, az SMT-rendszer betanításához egy párhuzamos korpusz szükséges. A Szeged Korpusz 2.0-ből állítottam elő az általam használt forrásnyelvi és célnyelvi korpuszokat. Az előbbibe az eredeti, elemzetlen és tokenizált mondatokat tettem, míg az utóbbiba a mondatban szereplő szavak szótövei, valamint azok POS-címkéi kerültek. Az így kapott rendszer eredményei az 1. táblázatban szerepelnek.

A kiértékelésénél szembetűnt a rendszer egy súlyos hibája, miszerint az elemzett korpuszban egymás után szerepelnek a szavak szótövei, amikhez hozzákapcsolódnak az elemzést tartalmazó címkék, de a több tagból álló kifejezések esetében (pl.: többtagú tulajdonnevek, igei szerkezetek) a címke csak a kifejezés utolsó szaván, vagy utána helyezkedik el. Az egy szófaji egységbe tartozó kifejezések

1. táblázat. A 6.1. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.97%	90.29%	9.71%
JOSHUA	90.96%	91.02%	8.08%

jelölésének hiánya a statisztikai módszerben félrevezető fordítási modellt eredményez. Ennek köszönhetően a rendszer az elemzett szöveghez véletlenszerűen hozzáad címkéket, ezért gyengébb eredményt ért el.

Az önálló POS-címkék eltávolítása. Az eredmény javítása érdekében minden önálló címkét hozzácsatoltunk az előtte álló szóhoz, így kaptuk a 2. táblázatban látható eredményeket.

2. táblázat. A 6.1. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.97%	90.80%	9.20%
JOSHUA	90.96%	90.72%	9.28%

A 2. táblázatból látszik, hogy változatlan BLEU-értékek mellett a rendszer pontossága 0,5–0,6 százalékkal javult. Ezt annak köszönhetjük, hogy nem kerültek a fordításba felesleges elemek. Ennek ellenére a többtagú kifejezések fordítása továbbra sem megoldott.

A többtagú kifejezések kezelése. Többtagú kifejezések esetében a nehézség abból adódik, hogy mivel a rendszer szavakat elemez, így az összetett kifejezések részeit is külön-külön címkézi. Célom, hogy az elemző egy egységként kezelje a többtagú kifejezéseket. A probléma megoldásához elengedhetetlen ezeknek a kifejezéseknek az összekapcsolása például a tulajdonnevek felismerésével. Nem volt célom ilyen rendszer kifejlesztése, viszont az elmélet igazolása érdekében összekötöttem a korpuszban ezeket a kifejezéseket. A tanítás után a 3. táblázatban látható eredményt kaptam.

Az 1500 mondatos teszthalmazból számszerűsítve 506 mondat elemzése volt teljesen helyes és 994-ben volt valamilyen hiba. Első ránézésre ez rossznak tűnhet, de ha az eredményt címkék szintjén is megvizsgáljuk, sokkal jobb arányt kapunk, hiszen 24557 helyes és csak 2343 helytelen elemzést kaptam. Láthatjuk, hogy a 6.1 rendszerhez képest a többtagú kifejezések összekötése és egyként kezelése javított a rendszer pontosságán, annak ellenére, hogy rosszabb BLEU-eredményt kaptam.

3. táblázat. Az alaprendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.76%	91.29%	8.71%
JOSHUA	90.77%	91.07%	8.93%

Az eredmények mélyebb vizsgálatából kiderül, hogy a helytelen annotációnak két oka lehet. Az első, amikor a szó nem szerepel a tanító halmazban (out-of-vocabulary, OOV), ekkor a rendszer elemzetlenül adja vissza a forrásnyelvi kifejezést. Ez 1697 esetben fordult elő. A helytelen annotációk másik típusa, amikor az SMT rendszer helytelen címkét rendel az adott szóhoz (646 eset). Ennek további két csoportja lehet: egyrészt, amikor a megfelelő szófaji címkét megtalálja, viszont a mélyebb szintű elemzés során hibázik; másrészt amikor teljesen rosszul elemzi a szót.

A 4. táblázatban egy példamondat olvasható a 6.1. rendszer kimenetéből.

4. táblázat. Példamondat az alaprendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	ez_[pd3-sa] a_[tf] lobbyerőt_[x] és_[ccsw] képesség_[nc-sa] a_[tf] diplomáciai_[afp-sn] erőfeszítés_[nc-pp] kívül_[st] mindenekelőtt_[rx] a_[tf] magyarországi_[afp-sn] multinacionális_[afp-pn] adhat_[vmcp3p—y] ._[punct]
SMT elemző:	ez_[pd3-sa] a_[tf] lobbyerőt és_[ccsw] képesség_[nc-sa] a_[tf] diplomáciai_[afp-sn] erőfeszítéseken kívül_[st] mindenekelőtt_[rx] a_[tf] magyarországi_[afp-sn] multinacionális_[afp-pn] adhat_[vmcp3p—y] ._[punct]

Továbbiakban ezt a rendszert fogom alaprendszernek tekinteni. A továbbiakban vizsgált rendszereknél kikötés lesz, hogy a fent említett hibákat elhagyjam, vagyis ne álljanak önmagukban címkék, illetve a többtagú kifejezések össze legyenek kötve.

6.2. A célnyelvi szótár méretének csökkentése

Csak szófaji egyértelműsítés. Az SMT-rendszer tulajdonságaiból következik, hogy egy megfelelő korpuszból bármilyen szabály betanítható. Mivel az általam használt korpusz mérete korlátos, a rendszer minőségének javulása többek között elérhető az annotációs feladat komplexitásának csökkentésével. Ebben az esetben ezt úgy érhetem el, ha az elemzendő szöveget a POS-címkék „nyelvére" fordítom.

Ezt munkám során úgy valósítottam meg, hogy az elemző rendszeremből elhagytam a szótövesítést, és csak a szófaji egyértelműsítést alkalmaztam. Mivel ezáltal csak a szavak POS-tag-jeire fordítok, a célnyelvi oldal szótári elemeinek száma nagy mértékben csökken. Az alaprendszer esetében 152694 elemből állt a célnyelvi szótáram, ezt csökkentettem le 1128 elemre. Így a fordítási feladat bonyolultságát csökkentve egy relatíve pontos rendszer hozható létre kis korpuszból is. Másrészről a szótövek elhagyásával csak címkék halmazára fordítok, ezáltal az egyes címkék nagyobb súllyal szerepelnek, mind a fordítási, mind pedig a nyelvi modellben. A tanítás után az 5. táblázatban látható eredményt kaptam.

5. táblázat. A 6.2. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	89.01%	91.46%	8.54%
JOSHUA	88.57%	91.09%	8.91%

A rendszer eredményeit vizsgálva kiderült, hogy a BLEU-érték további csökkenésének ellenére a rendszer pontossága jobb lett. Itt már az 518 teljesen helyes mondat mellett 982 mondat volt helytelen (0.8%-os javulás az alaprendszerhez képest). Tokenek szintjén 24603 volt helyes és 2297 volt helytelen (0.17%-os javulás). Ebből a rendszer által nem elemzett szavak száma 1699, amely változatlan az alaprendszerhez képest. Ezekből az eredményekből világosan látszik, hogy a rendszer minőségének javulása abból adódik, hogy az alaprendszer által elrontott 646 elemzés az új rendszerben 598-ra csökkent. Az eredmények mélyebb vizsgálata során szembetűnt, hogy e mögött a 48 darabos javulás mellett több eddig helyes elemzés romlott el. Ilyen hiba például a határozószók és a kötőszók keverése, valamint a kötőszók és a mutató névmások tévesztése. A 6. táblázatban egy példamondat olvasható a 6.2. rendszer kimenetéből.

6. táblázat. Példamondat a 6.2 . rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] lobbyerőt [ccsw] [nc-sa] [tf] [afp-sn] erőfeszítéseken [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]

A POS címkék egyszerűsítése. Az előző (6.2) fejezet eredményeiből kiindulva megvizsgáltam, hogy a célnyelvi szótár további csökkentése milyen hatást gyakorol a rendszer minőségére. Annak érdekében, hogy megvizsgáljam a rendszer működését a lehető legegyszerűbb körülmények között, hogy az elemzési mélységet nagy mértékben csökkentettem.

Ezt a következő rendszer segítségével tanulmányoztam oly módon, hogy csak a fő szófaji címkéket (az MSD-kód első karaktereit) hagytam meg a célnyelvi szótárban. Ebben az esetben a célnyelvi szótár 14 elemből áll. A tanítás után a 7. táblázatban látható eredményt kaptam.

7. táblázat. A 6.2. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	79.57%	92.20%	7.80%

A rendszer kiértékeléséből kiderült, hogy az eddig megfigyelt tendencia folytatódik. Tehát amíg a BLEU-érték csökkent, a rendszer pontossága növekedett. Ebben az esetben a rendszer 553 mondatot elemzett helyesen, miközben 947-et rontott el. Ez a 6.2. rendszerhez képest 2.3%-os, míg az alaprendszer (6.1) esetében 3.1%-os növekedést jelent mondatok szintjén. Tokenek tekintetében 24803 volt helyes és 2097 volt helytelen elemzés, ami 0.74%-os javulás a 6.2. rendszerhez képest, illetve 0.88% az alaprendszerhez képest. A 8. táblázatban egy példamondat olvasható a 6.2. rendszer kimenetéből.

8. táblázat. Példamondat a 6.2. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a lobbyerőt és képességet a diplomáciai erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	p t x c n t a n s r t a a v p
SMT elemző:	p t lobbyerőt c n t a erőfeszítéseken s r t a a v p

Konklúzió. A fent elért eredmények rendkívül bíztatóak, mivel egy viszonylag kisméretű korpusz esetén is az elemző rendszerek pontossága 90% feletti. Érdemes megfigyelni, hogy a 6.2. rendszer szótára két nagyságrenddel kevesebb elemet tartalmaz (1128 darab címke) az alaprendszeréhez képest (152 694 darab címke), ennek ellenére pontossága csupán 0.17%-al javult. Továbbá megfigyelhető, hogy a 6.2. rendszer csupán 14 címkéből álló szótára esetén (ami négy nagyságrend-

del való csökkentést jelent az alaprendszerhez képest) is csak 0.88%-os javulás mutatkozott.

Értékelésem szerint ez a 0.88%-os minőségjavulás nem áll arányban azzal a hatalmas információveszteséggel, amely a rendszerek célnyelvi szótárméretének csökkentésével jött létre. További tanulság, hogy a célnyelvi szótár méretének változtatásától függetlenül az OOV szavakat (1698 darab) egyik rendszernek sem sikerült elemeznie. Ebből arra a következtetésre jutottam, hogy a rendszer eredményének további javulása érdekében megoldást kell találnom a tanítóhalmazban nem szereplő szavak kezelésére.

6.3. Az OOV szavak kezelése

Az első, legkézenfekvőbb megoldás a korpusz növelése. A tanító halmazban minél több token fordul elő, annál pontosabb lesz a rendszer. A magyar nyelv agglutináló tulajdonságából adódóan, azért, hogy minden token megfelelő számban forduljon elő a korpuszban, nagyon nagy méretű korpuszra lenne szükség. A következő fejezetben egy olyan módszert vizsgálok, amely alkalmas lehet az OOV szavak kezelésére.

Sima szöveg esetén. Mivel az OOV szavak elemzéséhez a tanító halmazból semmilyen információt nem nyertünk ki, szükségünk van ezen szavak további vizsgálatára. Ebben segítségünkre lehet az ismeretlen szavak kontextusa. A nyelvi sajátosságok, valamint a zárt és nyílt szóosztályok miatt az OOV szavak nagy valószínűséggel csak egy-két szófaji osztályból kerülnek ki. Az előző rendszerek megfigyelése alapján elmondható, hogy a szótárban nem szereplő szavak túlnyomórészt főnevek.

Guillem és Joan Andreu módszere alapján [1] ezt a problémát úgy próbálom meg kiküszöbölni, hogy azokból a szavakból, melyek a tanító halmazban egy bizonyos küszöbértéknél gyakrabban fordulnak elő, egy szótárat hozok létre. Azokat a szavakat, amelyek nem kerülnek be ebbe a szótárba, egy tetszőleges (az esetemben „UNK") kifejezésre cserélem ki. Így ez a szimbólum nagy gyakorisággal kerül be az elemzendő szövegbe. Feltételezésem szerint, mivel az OOV szavak csak egy-két szófaji osztályból kerülnek ki, a környezetükben lévő szófaji szerkezetek nagyon hasonlóak lesznek. Mivel az SMT rendszer kifejezés alapú fordítást végez, figyelembe veszi mind az elemzendő szavak, mind a címkék környezetét is. Ennek segítségével tudja meghatározni az „UNK" szimbólum elemzését.

Kulcsfontosságú kérdés a megfelelő gyakorisági szint kiválasztása, hiszen ettől függ, hogy mennyi „UNK" szimbólum kerül a korpuszba. Egyrészt, ha túl nagy ez a szám, akkor túl sok token cserélődik ki az „UNK" szimbólumra, emiatt a környezet vizsgálatából sem kapunk megbízható elemzést, hiszen abban is előfordulhat nagy valószínűséggel „UNK". Másrészt viszont ha túl kicsi, akkor túl sok ritka szó marad a szótárban, ezzel nem tudjuk megfelelő mértékben kihasználni a módszer előnyét. Rendszeremben ezt a gyakorisági küszöböt 10-re választottam.

A fentiek alapján felépített rendszer betanítása után a 9. táblázatban látható eredményt kaptam.

9. táblázat. A 6.3. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	88.71%	85.74%	14.26%

Szembetűnő változás, hogy a rendszer eredménye nagymértékben romlott. Csupán 294 mondatot sikerült teljesen hibátlanul elemeznie a rendszernek, míg 1206-ban fordult elő valamilyen hiba. Tokenek szintjén 23064 volt helyes és 3836 volt helytelen. A 10. táblázatban egy példamondat olvasható a 6.3. rendszer kimenetéből.

10. táblázat. Példa mondat a 6.3. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a unk és unk a diplomáciai unk kívül mindenekelőtt a magyarországi unk unk .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct] [pd3-sa] [tf] [nc-sa] [ccsp] [vmis3p—y] [tf] [afp-sn] [nc-pn] [st] [rx] [tf] [afp-sn] [nc-pn] [nc-sa—s3] [punct]

A magyar nyelvű szövegben a főnevek és az igék különböző ragozott formái találhatók meg, melyek kis korpusz miatt nagy valószínűséggel az általam alkalmazott küszöb alá esnek. Ez magyarázza, hogy a korpuszban szereplő mondatok többségében a főnevek és az igék helyére is az „UNK” szimbólum kerül, ami a szóösszekötő munkáját nehezíti meg. Ez okozta, hogy a rendszer elrontotta az eddig helyes mondatelemzéseket is, ráadásul előfordult, hogy összekeverte a szavak sorrendjét az elemzés során.

Szótövek esetén. Az előző rendszer hibáinak kiküszöbölésére megvizsgáltam, hogyan befolyásolja a rendszer eredményét, ha a gyakoriságot nem a szövegben megtalálható szavakra, hanem azok szótöveire vizsgálom. Ettől azt vártam, hogy így csak azokat a szavakat/szótöveket cserélem „UNK”-ra, amelyek előfordulása tényleg nagyon alacsony. A két rendszer összehasonlításának érdekében ebben az esetben is 10-re választottam a küszöbértéket. A 11. táblázatban látható eredményt kaptam.

Az eredmények elemzése során az előző rendszer (6.3) eredményéhez képest viszonylag nagy javulás figyelhető meg, bár ez az alaprendszer (6.1) eredményét még mindig nem éri el. A rendszer 450 helyes mondat mellett 1050-et ront el. Tokenek szintjén 24190 volt helyes és 2710 volt helytelen.

11. táblázat. A 6.3. rendszer eredménye

Rendszer	BLEU-érték	Helyes	Helytelen
MOSES	90.87%	89.93%	10.07%

A fent említett változtatások hatására valóban csak az igazán ritka szavak lettek lecserélve „UNK”-ra. Ezek többsége nagyrészt főnév, és már alig van köztük ige. Ezzel párhuzamosan viszont az igék esetében egyre gyakoribb jelenség, hogy az elemző OOV szóként elemezte őket. Ez abból adódik, hogy ragozott formájuk nem szerepel a tanító halmazban megfelelő súllyal. A 12. táblázatban egy példamondat olvasható a 6.3. rendszer kimenetéből.

12. táblázat. Példamondat a 6.3. rendszer eredményéből

Rendszer	Fordítások
Sima szöveg:	ezt a unk és képességet a unk erőfeszítéseken kívül mindenekelőtt a magyarországi multinacionálisok adhatnák .
Referencia elemzés:	[pd3-sa] [tf] [x] [ccsw] [nc-sa] [tf] [afp-sn] [nc-pp] [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]
SMT elemző:	[pd3-sa] [tf] [nc-sa] [ccsw] [nc-sa] [tf] [afp-sn] erőfeszítéseken [st] [rx] [tf] [afp-sn] [afp-pn] [vmcp3p—y] [punct]

7. Összefoglalás

Kutatásom során az SMT-rendszer lehetőségeit vizsgáltam a szófaji egyértelműsítés és a lemmatizálás feladatainak megvalósítására. Megfigyelésem szerint ezek a problémák megfogalmazhatók a sima szövegről elemzett szövegre való fordításként is. Az erre a célra használt rendszerek pontossága elérheti akár a 92%-ot is. Annak ellenére, hogy ez az eredmény nem éri el a napjaink legjobb POS-tagger rendszerének szintjét, az általam felépített rendszer teljesen automatikusan ismeri fel a szabályokat, és nincs szükség előzetes szövegfeldolgozásra. Másrészt ez a rendszer párhuzamosan végzi az annotálás és a lemmatizálás feladatát. Az itt elvégzett kísérletekkel bebizonyítottam, hogy a célnyelvi szótár méretének csökkentése csak minimális javulást okoz a rendszer pontosságában, viszont óriási információveszteséget eredményez.

Az eredmények azt is megmutatják, hogy tisztán statisztikai alapú módszerek nem elegek ezen feladatok megvalósítására, hanem szükség lenne valamiféle hibridizációra is. Az eredmények a jövőre nézve biztatóak, célom a további lehetőségek vizsgálata.

Hivatkozások

1. 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
2. Prószéky, G., Kis, B.: A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. ACL '99, Stroudsburg, PA, USA, Association for Computational Linguistics (1999) 261–268
3. Oravecz, C., Dienes, P.: Efficient Stochastic Part-of-Speech Tagging for Hungarian. In: Proc. of the Third LREC, pages 710–717, Las Palmas, Espanha. (2002) ELRA.
4. Halácsy, P., Kornai, A., Oravecz, C., Trón, V., Varga, D.: Using a morphological analyzer in high precision POS tagging of Hungarian. In: Proceedings of LREC 2006. (2006) 2245–2248
5. Laki, L.J., Prószéky, G.: Statisztikai és hibrid módszerek párhuzamos korpuszok feldolgozására. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Egyetem (2010) 69–79
6. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
7. Koehn, P.: Moses - A Beam-Search Decoder for Factored Phrase-Based Statistical Machine Translation Models. (2009)
8. 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
9. Li, Z., Callison-Burch, C., Dyer, C., Ganitkevitch, J., Khudanpur, S., Schwartz, L., Thornton, W.N.G., Weese, J., Zaidan, O.F.: Joshua: an open source toolkit for parsing-based machine translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation. StatMT '09, Stroudsburg, PA, USA, Association for Computational Linguistics (2009) 135–139
10. Csendes, D., Hatvani, C., Alexin, Z., Csirik, J., Gyimóthy, T., Prószéky, G., Váradi, T.: Kézzel annotált magyar nyelvi korpusz: a Szeged Korpusz. In: I. Magyar Számítógépes Nyelvészeti Konferencia, Szegedi Egyetem (2003) 238–247
11. Farkas, R., Szeredi, D., Varga, D., Vincze, V.: MSD-KR harmonizáció a Szeged Treebank 2.5-ben. In: VII. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Szegedi Egyetem (2010) 349–353
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. ACL '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 311–318