Hunglish mondattan – átrendezésalapú angol–magyar statisztikai gépifordító-rendszer

Laki László János[1,2], Novák Attila[1,2], Siklósi Borbála[2]

[1] MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
[2] Pázmány Péter Katolikus Egyetem,
Információs Technológiai Kar,
Budapest, Práter u. 50/a,
e-mail:{laki.laszlo,siklosi.borbala,novak.attila}@itk.ppke.hu

Kivonat A napjainkban népszerű frázisalapú statisztikai gépifordító-rendszerek az egymáshoz hasonló szerkezetű és a nem nagyon gazdag ragozó morfológiával bíró nyelvpárok esetében látványos eredményeket értek el az utóbbi évek során. Azon nyelvpárok esetében azonban, ahol jelentős szórendi és strukturális különbségek vannak a két nyelv között, az eredmények messze elmaradnak a várakozásoktól. Az utóbbi kategóriába tartozik az angol-magyar nyelvpár is. Cikkünkben egy olyan angol-magyar statisztikai gépifordító-rendszer létrehozására tett kísérletünket írjuk le, amelyben a két nyelv közötti strukturális különbségeket úgy próbáltuk áthidalni, hogy az angol forrásnyelvi mondatok szintaktikai elemzését felhasználva, azokat automatikusan a nekik megfelelő magyar mondatok szerkezetének jobban megfelelő szórendűvé alakítottuk. A korlátozott mértékű tanítóanyag és a magyar ragozó jellege miatt fennálló adathiány-probléma megoldása érdekében szó- helyett morfémaalapú fordítórendszert hoztunk létre.

Kulcsszavak: SMT, morfológiai elemzés, átrendezés

1. Bevezetés

Az informatika fejlődése új lehetőségeket nyitott meg többek közt a nyelvészetben. A humán nyelvtechnológia egyik legfontosabb feladata, hogy leküzdje a soknyelvűség okozta akadályokat és nehézségeket, illetve támogassa globalizálódó világunk különböző nyelveinek megértését. Ennek megvalósításában nyújt nagy segítséget a gépi fordítás.

Az első ilyen rendszerek előre definiált szabályok, illetve transzformációk alapján működtek. A szabályalapú gépi fordítás hátránya, hogy a különböző nyelvi sajátosságok nem írhatók le mindent lefedő szabályrendszerrel. A statisztikai módszeren alapuló gépi fordítás (SMT) a számítógépre bízza a szabályrendszer felépítését, ami egy párhuzamos kétnyelvű korpusz felhasználásával történik.

Azokra a nyelvekre, melyek szintaktikailag hasonlóak és morfológiailag nem túl komplexek, a frázisalapú SMT módszerei viszonylag jó eredménnyel működnek. Ezzel ellentétben az ilyen szempontból egymástól távolabb eső nyelvpárok

(pl. angol-magyar) esetén jelentős lemaradás van. Több tanulmány bemutatta azt is, hogy az ilyen esetekben csupán a tanító korpusz növelése nem elegendő a minőség számottevő javításához. A magyar nyelv szabad szórendje és szóalaki sokfélesége miatt nem is lehetséges olyan korpusz létrehozása, amely minden nyelvi jelenséget elég jól lefedne. Ezért célunk egy olyan hibrid fordítórendszer létrehozása volt, amely amellett, hogy kihasználja a statisztikai gépi fordítás előnyeit, igyekszik csökkenteni a szórendi különbségekből és a magyar nyelv morfológiai sokszínűségéből adódó problémákat.

2. Gépi fordítás angol-magyar nyelvpárra

Bár a gépi fordítás területén vannak jelentős eredmények, de a feladat korántsem tekinthető megoldottnak. Különösen igaz ez az egymástól távol álló nyelvpárok esetén, mint például a magyar és az angol. A legfőbb probléma a két nyelv jelentősen eltérő struktúrájában rejlik, emberként is leginkább csak az egyik nyelven megfogalmazott mondat jelentésével azonos jelentésű mondatot tudunk megfogalmazni a másik nyelven, ez azonban nem nyelvtani megfeleltetést jelent. Ez a probléma az alkalmazott fordítási módszer kiválasztásánál éppen úgy jelen van, mint a fordítás folyamán, vagy a kiértékelésnél. A tapasztalatok alapján a gépi fordítás minőségét három tulajdonság befolyásolja a forrás- és célnyelvek megválasztásának függvényében: a szórendbeli eltérés mértéke, a célnyelv nyelvtani összetettsége és a két nyelv közötti történeti kapcsolat [1].

2.1. A magyar nyelv sajátosságai

A magyar és az angol nyelvek mind történetileg, mind nyelvtanilag egymástól távol álló nyelvek, illetve célnyelvként a magyart tekintve, ennek összetett volta sem kérdéses. A legfontosabb nehézséget a magyar nyelvre jellemző gazdag morfológia, és a két nyelv jelentősen eltérő szórendje jelenti, amelyet a magyarban egész más tényezők határoznak meg, mint az angolban.

A magyar nyelvre jellemző agglutináló (ragozó) jelleg toldalékok halmozását is lehetővé teszi. Szintén jellemző a többféle alakváltozat mind a szótövek, mind a toldalékok terén, a gazdag esetrendszer és az irányhármasság (honnan? hol? hová?) a helyhatározók használatában. Kevés az igeidő, és hiányzik az indoeurópai nyelvekre jellemző birtoklást kifejező ige („én birtoklok valamit" helyett „nekem van valamim"). A magyar nyelv megkülönbözteti a határozatlan („alanyi") és a határozott („tárgyas") ragozást: olvasok, olvasom; a főnévi igenév pedig ragozható (látnom, látnod, látnia stb.) . A magyar az indoeurópai nyelveknél sokkal ritkábban használja a határozatlan névelőt (egy); a páros szerveket (pl. kéz, láb, szem, fül) és a több birtokos egy-egy birtokát is egyes számban mondja (pl. élik az életüket, nem pedig életeiket) a számneves névszói csoportok pedig alakilag egyes számúak, és így is egyeztetjük őket az igével.

Egy egyszerű magyar mondatban általában az alany az első, az ige a második és végül a tárgy az utolsó elem. A magyar pro-drop nyelv, az alanyi és tárgy névmások is rendszerint hiányoznak a mondatból . Mindemellett, a szórend nem

feltétlenül kötött, ugyanis azt elsősorban nem szintaktikai szabályok, hanem pragmatikai tényezők határozzák meg (ugyanakkor semleges mondatok esetében rendkívül összetett szintaktikai megszorítások is). Így például gyakran előfordul a tárgy-alany-állítmány vagy az állítmány-alany-tárgy sorrend is, hiszen a ragozás egyértelműen utal az elemek mondatbeli szerepére. A kihangsúlyozni kívánt információ rögtön a ragozott ige elé helyezendő.

A magyarral ellentétben az angol főleg izoláló nyelv, vagyis a mondatokban a fő nyelvtani funkciókat a szavak sorrendje határozza meg. Szórendje sokkal kötöttebb, mint a magyar nyelv esetén, jellemzően alany-állítmány-tárgy alakú. Ugyanakkor számos példáját hordozza a flexiónak (a tő megváltoztatásával járó ragozás), főleg a rendhagyó esetekben. Így bár a nyelv flektálónak tekinthető, lassan tart az izoláló felé, hiszen például a mai angolban már nincsenek esetek, az eredeti esetragok lekoptak [2].

2.2. Statisztikai gépi fordítás

Napjaink nemcsak legelterjedtebb módszere a statisztikai gépi fordítás (SMT), hanem egyben a legtöbb lehetőséget magában rejtő és az egyik legjobban kutatott irányzat is. Bár a statisztikai módszereket nyelvfüggetlen megoldásoknak tekinthetjük, azonban mégis szükségesnek látjuk a nyelvspecifikus problémák kezelését, melyet elő- és utófeldolgozási lépésekként építettünk be a rendszerbe.

A statisztikai gépi fordítás alapötlete a kétnyelvű párhuzamos korpuszból tanult statisztika alapján való fordítás, illetve az így nyert fordítási lehetőségek célnyelvi korpuszból épített, a célnyelvet jellemző modell alapján történő kiértékelése.

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

Ezt a folyamatot a Bayes-tétel segítségével lehet leírni két valószínűségi változó szorzataként. Ezeket fordítási- és nyelvmodellnek nevezzük. Az a mondat lesz a rendszerünk kimenete, amelyik a legjobban hasonlít a fordítandó (forrásnyelvi) mondathoz.

3. Átrendezési szabályok alkalmazása előfeldolgozási lépésként

A fent részletezett nyelvi különbségek áthidalása végett a cikkben bemutatott rendszerben olyan előfeldolgozási lépéseket alkalmaztunk, melyeknek célja a forrásnyelvi (angol) szöveg mondatainak a célnyelvi (magyar) mondatokhoz hasonló alakra hozása. Ehhez első lépésként az angol mondatokra szófaji egyértelműsítés és szintaktikai elemzés után a mondatban megjelenő függőségi relációkat is meghatároztuk. Így olyan gazdag információkkal kiegészített mondatokat kaptunk,

melyek birtokában megfogalmazhatók olyan szabályok, amelyek a mondatok magyar megfelelőjében szereplő szerkezetekkel párhuzamos formára hozzák azokat. Így a fordítórendszer tanítása során a nyelvpárt az alaprendszerénél jobban reprezentáló statisztikák jönnek létre. Mivel a statisztikai módszer alapját a kétnyelvű párhuzamos mondatokban szereplő szavak megfeleltetésére épített valószínűségek képezik, ezért a szóösszerendelés minősége alapjaiban meghatározza a végső fordítás minőségét is.

A két nyelv közelítése a morfémák szavakba szerveződése szempontjából hatékonyan csökkentheti a szóösszerendelési hibák számát. Más nyelvekkel (pl. az angol-német nyelvpárral) kapcsolatban publikált eredmények pedig azt mutatják, hogy a szórend szabályalapú megváltoztatása csökkenti a dekódolás során a fordításból kimaradt szavak számát.

Az alkalmazott szabályaink csak azokat a szórendi eltéréseket szüntetik meg, amelyek a két nyelv között szabályszerűen fellépnek (pl. elöljárók vs. esetragok / névutók), nem volt célunk ugyanakkor a magyar „szabad szórend"-ből adódó eltérések eltüntetése.

A szabályok az angol mondatok szófajilag egyértelműsített elemzését, közvetlen összetevős és függőségi elemzését használják. A függőségi relációkból az elemzés után kiválasztjuk a releváns kapcsolatokat, amelyek mentén alkalmazzuk a megfelelő szabályt. Nagyon egyszerű példa az angol „in my house" kifejezés, mely az átrendezés és összevonások után „house_my_in" formára alakult, amely megfelel a magyar „házamban" alaknak. Az ilyen rövid szókapcsolatok során a szabályok alkalmazása nem jelent nagy problémát, azonban hosszabb mondatok esetén az egymáshoz kapcsolódó részek egészen távol is eshetnek, több függőségi kapcsolatban is érintettek lehetnek. Hasonló módon kerültek beszúrásra olyan morfológiai elemek, melyek az eredeti angol mondatban nincsenek explicit módon jelölve (pl. tárgyrag), a magyar megfeleltetés miatt azonban szükségesek. Természetesen figyelembe vettük az összetartozó szerkezeti egységeket, ezeket az átrendezés során is egységként kezelve, egyben helyeztük át.

Három fő csoportba sorolható átrendezési szabályokat alkalmaztunk:

3.1. Szórendet és morféma összevonást/felbontást tartalmazó szabályok

Ezek a szabályok a függőségi relációk meghatározása után, a közvetlen összetevős szerkezetet is figyelembe véve alakítják át a szavak sorrendjét, ezzel egyidőben vonják is össze azokat, amikor szükséges. Olyan szabályok kerülnek végrehajtásra, mint a passzív, a segédigés, a prepozíciós és birtokos szerkezetek átalakítása, az angolban hátravetett módosítók előremozgatása, és még néhány, ritkábban előforduló szabály. Fontos az átrendezési szabályok végrehajtásának sorrendje is, mivel nem csak szavakat, hanem nagyobb egységeket helyezünk át. Az alábbi mondatban két szabályt hajtunk végre:

A „living in the city” prepozíciós szerkezet a PARTMOD[1] (merchant, living), PREP[1](living, in) és a POBJ[1](in, city) relációk mentén kerül átalakításra. Először a prepozíció kerül rá annak gyerekére, majd az így kapott összevont szót helyezzük át az ezt megelőző főnévi szerkezet elé. Így kialakul az „a város_ban élő” magyar fordításnak már egyértelműen megfeleltethető szórend. Hasonlóan járunk el a „the sons of the merchants” esetén a megfelelő relációk használatával, melynek eredményeként a „kereskedők fiai” magyar szintaktika szerinti alakra jutunk. Ezt látható az 1. táblázatban.

1. táblázat. Példamondat I.

Eredeti mondat:	The/DT sons/NNS of/IN the/DT many/JJ merchants/NNS living/VBG in/IN the/DT city/NN ./.
Átrendezett mondat:	the/DT city/NN_in/IN living/VBG many/JJ merchants/NNS sons/NNS_of/IN ./.

Bár általában az angol oldalon szükséges a szavak számának a csökkentése azok összevonásával, így a magyarnak megfelelő toldalékok létrehozásával, mégis vannak esetek, amikor új szavakat kell beszúrni az átrendezések során az angol mondatba. Mivel az ilyen eseteknél nem tudjuk előre meghatározni az oda illő magyar szót, mivel az az aktuális szövegkörnyezettől függ, ezért csupán egy karaktersorozat kerül beillesztésre, melynek konkrét realizációját a fordítás kell hogy meghatározza. A 2. táblázatban az xxx/xxx jelöli a „lévő” magyar szó pozícióját a mondatban, valamint néhány további átrendezési példát is tartalmaz.

2. táblázat. Példamondat II.

Eredeti mondat:	That/DT is/VBZ the/DT account/NN at/IN the/DT largest/JJS bank/NN in/IN Bern/NNP ./. ”/”
Átrendezett mondat:	That/DT is/VBZ the/DT Bern/NNP_in/IN xxx/xxx largest/JJS bank/NN_at/IN xxx/xxx account/NN ./. ”/”
Eredeti mondat:	Only/RB I/PRP ’m/VBP allowed/VBN to/TO ./.
Átrendezett mondat:	Only/RB allowed/VBN_P_they/P3 I/PRP_acc/ACC to/TO ./.

3.2. Átrendezést nem tartalmazó, csupán a morfológiai összetételt változtató szabályok

Az angol mondatokban sok olyan információ nincs jelen, amely a magyar oldalon toldalékokként szerepelnek. Ezekre azonban a függőségi relációk alapján tudunk

[1] A függőségek teljes listája itt olvasható: http://nlp.stanford.edu/software/dependencies_manual.pdf

következtetni. Így például az angolban jelöletlen tárgyrag a megfelelő relációk mentén meghatározható. Az ilyen esetekben beszúrtuk ezeket a morfémákat az angol mondatba.

Így lett a *„while/IN giving/VBG a/DT present/NN ./."* mondatból *„while/IN giving/VBG a/DT* ***present/NN_acc/ACC*** *./."*

Előfordulnak továbbá olyan esetek is, amikor az angol különálló szóként jelöli a magyar toldaléknak megfelelő morfémákat, amelyeket így rácsatoltunk a megfelelő szóra. Ezek az összevonások nem nagyobb szerkezetek átrendezését jelentik. Például a birtokos névmás esetén, ha a birtok tárgya is szerepel a mondatban, akkor azt csak ennek megfelelően hozzákapcsoljuk ahhoz. Így lett a *„my/PRP$ own/JJ country/NN"* mondatból *„own/JJ country/NN_my/PRP$"*.

3.3. Redundanciák feloldása, utófeldolgozás

Ezek a szabályok elsősorban az első két csoportba tartozó átrendezések mellékhatásai miatt szükségesek. Például előfordulhat, hogy az átrendezés után két névelő kerül egymás mellé, ilyen esetekben az egyiket törölni kell. Ide tartozik még a birtokos *'s* rácsatolása a megfelelő szóra. Ezen kívül még néhány apró módosítást láttunk szükségszerűnek (például pénznemek áthelyezése a számérték utánra).

3. táblázat. Példamondat III.

Eredeti mondat:	John's cat
Függőségi relációk:	poss(cat, John)
	possessive(John, 's)
Átrendezett mondat:	John/NNP cat/NN_'s/POS

4. Felhasznált eszközök

4.1. Korpusz

Az elérhető angol-magyar párhuzamos korpuszok többsége nem alkalmas egy általános SMT-rendszer betanítására, mivel csupán egy-két terület terminológiáját tartalmazzák. Munkánk során ezért az elérhető legnagyobb és témáját tekintve legáltalánosabb párhuzamos korpuszt, a BME MOKK és az MTA Nyelvtudományi Intézete készített Hunglish korpuszt [3] használjuk. Ez a korpusz egyidejűleg több területről tartalmaz szövegeket: szépirodalom, magazin, jog, filmfeliratok. A rendszer betanítása során nehézséget jelent azonban, hogy az egyes részek minősége meglehetősen változó. Az így létrejött korpusz mérete 1202205 párhuzamos mondatpár.

4.2. Szintaktikai és függőségi elemző

Az előfeldolgozás első lépéséhez szükség volt egy robusztus szófaji egyértelműsítőre, szintaktikai és függőségi elemzőre az angol mondatok átrendezéséhez, illetve a morfémaalapú fordítás miatt a magyar nyelv elemzésére.

Magyar nyelvre a PurePos[4] automatikus morfológiai annotáló eszközt használtuk. A teljes tanító anyag magyar oldalát ezzel elemeztük, az összetett morfológiával rendelkező szavak esetén ezeket felbontottuk elemi egységekre azért, hogy az angol oldalon külön szóként, a magyar oldalon azonban csak toldalékként megjelenő morfémák is megfeleltethetőek legyenek egymásnak. Mivel a morfológia önmagában tartalmaz minden információt a szóalakok eredeti voltáról, ezért a szavak jelentésének megfeleltetésére elegendő azok szótövét figyelembe venni, így mivel az egyes szavakhoz tartozó szótő alakok előfordulási gyakorisága jóval nagyobb a korpuszban, ezért biztosabb statisztikát kaptunk, mint a teljes szóalakokra való statisztika építése során. Természetesen az így betanított morfémaalapú fordítórendszer fordítási eredménye is szótő+címkék alakú eredményt hoz létre, ezért ezeket a fordítás után vissza kell alakítani, amihez a Humor [5] szóalakgeneráló modulját használtuk.

Angol nyelvre a Stanford elemzőt [6] használtuk, mely az egyik gyakran használt szabadon hozzáférhető angol szintaktikai elemző. Az elemző hozzáférhető változatát a Penn Wall Street Journal Treebank egy töredékén tanították. Az elemzés minősége sokkal fontosabb számunkra, mint a gyorsasága, hiszen az elemzés és a szóösszekötés offline, csak a tanítás során egyszer történik (illetve a fordítandó szöveget kell még elemeznünk), ezért úgy döntöttünk, hogy az elemző lexikalizált változatát alkalmazunk. Ez valamivel jobb elemzést eredményezett, mint az alapváltozat, de még így is nagyon sok olyan eset fordult elő, melyeket az elemző nem tudott megfelelően kezelni.

A Stanford parser sorba kapcsolt elemekből álló rendszer, amelynek első szófaji egyértelműsítő komponense önmagában meglehetősen sok hibát generál, amelyet azután minden további komponens csak továbbiakkal tetéz. A korábbi komponensek hibáit a láncban később következők soha nem javítják, inkább a legnyakatekertebb megoldásokkal próbálnak a kapott inputhoz alkalmazkodni. Ezek a rosszul elemzett és rossz helyre csatolt szavak és kifejezések az egész rendszerben kritikus problémát jelentenek, hiszen az átrendezések ez alapján az elemzés alapján történnek. Ez azt jelenti, hogy ha egy eleve rosszul elemzett szöveget rendezünk át, akkor az így kapott hibás átrendezés inkább ront, mint javít a fordítás minőségén.

Az első ilyen hibaforrás a helytelen POS-címkék használata az elemző által ismeretlen szavak, vagy az ismeretlen kontextusban megjelenő ismert szavak esetében. A legtipikusabb hiba a főnevek, melléknevek és igék összetévesztése, amely szinte minden esetben végzetes következményekkel jár az elemzés egészére. Erre látható példa a 4. táblázatban.

Mivel mind a szintaktikai, mind a függőségi elemzés ilyen félrevezető információkon alapul, a hiba továbbterjed a rendszerben és az 5. táblázatban látható hibákat eredményez.

4. táblázat. Példamondat IV.

100/CD million/CD **sound/NN** good/JJ to/TO me/PRP ./.
For/IN airline/NN personnel/NNS ,/, we/PRP **cash/NN** personal/JJ **checks/VBZ** up/RP to/TO / 100/CD ./.

5. táblázat. Példamondat V.

-/: 100/CD million/CD sound/NN good/JJ to/TO me/PRP ./.
For/IN airline/NN personnel/NNS ,/, we/PRP cash/NN personal/JJ checks/VBZ up/RP to/TO $/$ 100/CD ./.
-/: me/PRP_to/TO xxx/xxx 100/CD million/CD sound/NN good/JJ ./.
airline/NN personnel/NNS_For/IN ,/, cash/NN personal/JJ up/RP_checks/VBZ_we/PRP 100/CD_$/$_to/TO ./.

5. A MOSES keretrendszer

A statisztikai gépi fordítás területén legelterjedtebben a frázisalapú fordítást végző nyílt forráskódú MOSES nevű keretrendszert [7] használják, amely mind a tanítás, mind a dekódolás feladatára megoldást jelent. Mindemellett tartalmaz olyan segédprogramokat is, amelyek a nyelvmodellépítést és az automatikus kiértékelést is elvégzik. Ezt használtuk az itt leírt rendszerünk létrehozásához.

A MOSES alkalmas arra, hogy úgynevezett faktoros fordítórendszert hozzunk létre benne. A faktoros fordításba lehet a szövegben szereplő szavak puszta alakjánál mélyebb legalábbis morfoszintaktikai szintű információt belevinni. A fordítási faktorok olyanok lehetnek, mint a szóalakok felszíni alakja, töve alapvető szófaja, morfoszintaktikai jegyei. Faktoros fordítás esetén a keretrendszerben több fordítási, generálási és kontextuális nyelvmodellt hozhatunk létre, amelyeknek valamilyen kombinációját használja a rendszer, és így elvben képes lehet a korlátozottan rendelkezésre álló hiányos nyelvi adatok alapján is a sima szóalak alapú alaprendszernél jobb fordítások létrehozására olyan esetekben, ahol némi absztrakcióra van szükség az adott fordítás létrehozásához, mert pontosan azokat a szavakat nem látta a rendszer a tanítóanyagban, amelyekre az adott fordításhoz szükség lenne.

Sajnos azt találtuk, hogy a *MOSES-ben jelenleg létező konkrét faktorosfordítás-implementáció igazán nem alkalmas arra, hogy a magyarhoz hasonlóan gazdag morfológiájú nyelvekhez a rendelkezésünkre álló véges tanítóhalmaz alapján a sima szóalak alapú alaprendszernél jobb fordításokat hozzon létre. Ezért az itt leírt kísérleteink során egy alternatív megoldást próbáltunk létrehozni a morfológiai gazdagság és a két nyelv szerkezeteinek eltérő jellegéből adódó problémák (kötött morfémák a magyarban vs. izoláló szószerkezetek az angolban) kezelésére: morfémaalapú fordítót hoztunk létre.

6. Eredmények

Írott szövegek fordítása emberi olvasatra készül, ezért minden fordításnak a célja az, hogy emberek számára olvasható, érthető, az eredeti szöveggel azonos tartalmú fordítást hozzon létre. Mivel azonban az emberi kiértékelés lassú és drága, ezért elterjedt módszer a gépi fordítás minőségének vizsgálatakor annak automatikus kiértékelése, melyre több metrika is létezik. Ezek mindegyikének alapja az, hogy a géppel létrehozott fordítási eredményt ember által létrehozott referenciafordításhoz hasonlítják. Bár mindegyik metrikának vannak erősségei, és különböző szempontokat részesítenek előnyben a fordítás eredményének vizsgálatakor, önmagában egyik sincs mindig összhangban a fordítások emberi értékelésével. Munkánk során annak több fázisában végeztünk automatikus kiértékelést is a BLEU metrika szerint, de néhány esetet emberi kiértékeléssel is megvizsgáltunk, ami igazolta azt, hogy az automatikusan mért alacsonyabb értékek nem feltétlenül jelentenek rosszabb minőségű fordítást.

A rendelkezésünkre álló eredeti korpuszból a tanítás előtt félretettünk háromszor 1000 mondatból álló halmazokat a kiértékeléshez. Ezen kívül további vizsgálatokat végeztünk olyan teszthalmazon, amely a tanítóanyagban nem szereplő stílusú és témájú szövegeket (híreket) tartalmazott. Több rendszer eredményét mértük a különböző előfeldolgozási lépések hatásának értékelése céljából. Az alaprendszer a párhuzamos korpuszból minden előfeldolgozás nélkül tanított modell alapján fordított. A második fázis a morfémaalapú fordítás, ahol a forrásoldalon alkalmaztuk az elemzést és az átrendezést is, de a fordítás után a kapott morfémaalapú magyar mondatban nem generáltuk vissza a teljes magyar szavakat. Természetesen ebben az esetben sokkal magasabb BLEU-értéket kaptunk, de ez nem összehasonlítható a többi esettel, amelyekben szóalapú BLEU-értékeket kaptunk, így csak annak vizsgálatára alkalmas, hogy a morfémák milyen sikerrel kerültek bele a fordításba. A harmadik rendszer pedig a visszagenerált szövegen mért eredmény. Az egyes fázisok százalékban mért minőségét a 6. táblázat foglalja össze.

6. táblázat. A rendszerek eredményeit összefoglaló táblázat

	BLEU-érték		
Név	Baseline	Morf. elem. ford.	Generált ford.
test1	15,82%	64,14%	12,61%
test2	14,60%	57,39%	13,95%
test3	15,04%	57,84%	12,98%

Az eredményeken az látszik, hogy az alaprendszer BLEU-értéke a legmagasabb mindegyik teszthalmaz esetén. Ezek a különbségek azonban nem feltétlenül fejezik ki a valós minőségbeli, különbséget az egyes rendszerek által előállított fordítások között. Ennek oka, hogy a BLEU algoritmus minden eset-

ben egyszerűen a rendelkezésünkre álló egyetlen referenciafordítás szóalakjaihoz hasonlítja a rendszer által létrehozott fordítást. Mivel a hasonlításnál csak a szavak felszíni alakját veszi figyelembe, ezért teljesen mindegy, hogy egy egészen más, a fordításba egyáltalán nem illő szó került az eredménybe, vagy csak valamilyen ragozási hiba, esetleg szinonima szerepel.

A méréseknél kitűnik továbbá, hogy az egészen más témájú és stílusú hírkorpuszon is működik a rendszer, sőt ennek eredményére az átrendezési szabályok nagyobb hatással voltak, mint a tanítókorpuszhoz hasonló tesztek esetén.

6.1. Emberi kiértékelés

A különböző rendszerek által létrejött fordítások emberi vizsgálata során könnyen belátható, hogy a fent szereplő alacsony BLEU-értékek nincsenek összhangban a valódi minőséggel. A lefordított mondatok nagy része közelebb állt az eredeti mondat jelentéséhez az átrendezést és generálást alkalmazó rendszer esetén.

7. táblázat. Példamondat VI.

Eredeti mondat:	Nayla arrived then and the argument about the climb had begun .
Átrendezett mondat:	Naylum/[NNP] arrive/[VB] [Past] then/[RB] and/[CC] the/[DT] <zone> climb/[NN] about/[IN] </zone> xxx/[xxx] argument/[NN] <zone> begin/[VB] [PPart] have/[VB] [Past] </zone> ./[.]
Morféma alapú magyar fordítás:	Nayla/[FN] odaér/[IGE] [Past] [e3] ,/[PUNCT] a/[DET] haladás/[FN] [DEL] vitatkozik/[IGE] [INF] kezd/[IGE] [Past] [e3] ./[PUNCT]
Generált magyar:	Nayla odaért , a haladásról vitatkozni kezdett .
Baseline fordítás:	nayla odaért , és az a mászni kezdett .

6.2. Hibajelenségek

Az automatikus kiértékelő módszer hiányosságai mellett számos egyéb, a későbbiekben javítható probléma is megfigyelhető a fordítás minőségének ellenőrzésekor.

- Az angol szófajiegyértelműsítő-rendszer hibái: ha egy szó rossz szófaji címkét kap a fordítandó mondatban, akkor mivel a tanítás során a fordítási modellben a helyes címke a gyakoribb (ami kellően nagy korpusz esetén elvárható), ezt az alakot nem fogja tudni lefordítani még akkor sem, ha egyébként a szó önmagában gyakran előfordul. Ugyanakkor előfordulhat az is, hogy egy szónak többféle szófajú fordítása is szerepel a fordítási modellben, melyek közül a szövegkörnyezettől függően több is lehet helyes. Ezért ha az aktuálisan fordítandó mondatban rossz címke szerepel, akkor az annak megfelelő hibás fordítás kerül az eredménybe.

- Szintén az angol forrásszöveg hibás elemzése okozhat olyan hibát, ami a függőségi relációknál jelenik meg, így az átrendezési szabályok is helytelenül

8. táblázat. Példamondat VII.

Eredeti mondat:	For 50 years , barely a whisper .
Átrendezett mondat:	50/[CD] <zone> year/[NN] [PL] For/[IN] </zone> ,/[,] barely/[RB] a/[DT] whisper/[VB] ./[.]
Morféma alapú magyar fordítás:	50/[SZN_DIGIT] év/[FN] [PL] [TER] ,/[PUNCT] alig/[HA] egy/[DET] suttog/[IGE] [e3] ./[PUNCT]
Generált magyar:	50 évekig , alig egy suttog .
Baseline fordítás:	50 éve , alig egy suttogás .

hajtódnak végre. Ekkor olyan kifejezések kerülhetnek rossz helyre, melyek eredeti állapotukban jobbak voltak, s helyes elemzés esetén ott is maradtak volna. A Sinbad nem tulajdonnévként, ezzel szemben a valójában számnév Thousand tulajdonnévként címkézése alapvetően hibás szintaktikai elemzéshez vezetett, amelynek következtében az Ezeregy éjszaka fordítása (és Szindbádé is) zátonyra futott.

- Mivel a dekódolás során az egy szóhoz tartozó, de külön egységként megjelenő morfémákat bár külön tokenként, de egy egységként kezeltük (zónák), a fordítás során az ezeken átívelő frázisok nem érvényesültek. A zónahatárok lazább kezelése megoldhatná ezt a problémát.

- A tanító és a tesztkorpuszok minősége jelentős mértékben befolyásolja a fordítás minőségét is. Ez nemcsak amiatt jelent problémát, hogy bizonyos kifejezéseket hibásan tanul meg, hanem az automatikus kiértékelés során is sokszor hibás referenciafordításhoz végzi a hasonlítást. Ezért bár az eredeti mondat fordításának megfelel a létrejött fordítás is, ezekben az esetekben semmiképpen nem hasonlítható a referenciához.

- Mivel a fordítás során a fordítandó kifejezések alapegységei a morfémák, ezért ezek előfordulhatnak rossz szó mellé kerülve is, hiszen ugyanaz a toldalékmorféma egy mondaton belül többször is előfordul, a fordítási modell pedig több, az adott mondatban akár nem megfelelő szóhoz is hozzákapcsolhatja ezeket. Így a generálás során a toldalékok nem feltétlenül kerülnek a megfelelő szóra, illetve a kívánt helyen nem jelennek meg. Úgy látjuk, hogy a morfémaalapú fordítás alapvető problémát jelent már a tanító anyagban szereplő szóösszerendelések (illetve a mi esetünkben morféma-összerendelések) számára is, amelyek alapján a fordítóban használt frázistábla készül, ugyanis a hosszabb mondatokban ugyanaz a funkcionális morféma számos példányban előfordulhat, és a rendszerben használt Giza++ szópárosító algoritmus ezeket nem jól párosítja össze.

7. Összegzés

Cikkünkben bemutattunk egy olyan frázisalapú angol-magyar nyelvpárra készült hibrid fordító rendszert, melyet az automatikus statisztikai modellek használata mellett elő- és utófeldolgozási lépésekkel egészítettünk ki. Ezeknek célja az angol nyelvű mondatok átalakítása a magyarhoz jobban hasonlító szerkezetekké.

Ezekkel a transzformációkkal sikerült olyan fordításokat létrehozni, melyeknek bár az automatikus kiértékelés során mért minősége nem javult az alaprendszerhez képest, emberi olvasatra mégis sokszor sokkal jobbak annál. Számos olyan jelenség helyesen fordítható ezzel a módszerrel, melyet a hagyományos statisztikai gépifordító-rendszer nem tud kezelni. Bemutattuk azokat a hibajelenségeket is, melyeknek megoldása a további terveink része, s ezen kritikus pontok feloldása után további javulást várhatunk, ami jelentős áttörést jelentene az angol-magyar gépi fordítás területén.

Köszönetnyilvánítás

Ez a projekt a TÁMOP: 4.2.1.B – 11/2/KMR-2011–0002, valamint a MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport támogatásával készült. Továbbá köszönetet szeretnénk mondani Orosz György kollégánknak segítségért.

Hivatkozások

1. Birch, A., Osborne, M., Koehn, P.: Predicting Success in Machine Translation. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, Hawaii, Association for Computational Linguistics (2008) 745–754
2. Siklósi, B., Prószéky, G.: Statisztikai gépi fordítás eredményének javítása morfológiai elemzés alkalmazásával (2009) Msc diplomaterv.
3. Halácsy, P., Kornai, A., Németh, L., Sass, B., Varga, D., Váradi, T., Vonyó, A.: A hunglish korpusz és szótár. In: III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, SZTE (2005) 134–142
4. Orosz, G., Novák, A.: Purepos – an open source morphological disambiguator. In: Proceedings of the 9th International Workshop on Natural Language Processing and Cognitive Science., Wroclaw, Poland (2012)
5. 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
6. de Marneffe, M.C., MacCartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: LREC-06. (2006)
7. 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