Főnévicsoport-azonosító módszerek főnévicsoport-szinkronizációs célokra

Pohl Gábor

Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

Kivonat: A MorphoTM frázisalapú kísérleti fordítómemóriában alapvető fontosságú a főnévi csoportok (NP) automatikus azonosítása és szinkronizációja (*alignment*). Az előző években bemutattunk egy gyors, szótári és szófaji információkra építő NP-szinkronizáló módszert, most csak a főnévi csoportok azonosításával foglalkozunk. A MorphoTM rendszerben a forrásoldali NP-ket minden esetben mély szintaktikai elemzéssel választjuk ki, ebben a cikkben a célnyelvi NP-azonosítás lehetséges módszereit hasonlítjuk össze. A mérési eredmények azt mutatják, hogy az NP-ket fordításaik alapján meghatározó sekély elemzést használó módszerünk megfelel az elvárásoknak, illetve hogy a pontosság növelhető, ha a kiválasztott NP-jelölteket mély szintaktikai elemzővel (de a teljes mondat elemzése nélkül) ellenőrizzük.

1 Bevezetés

Cikkünkben főnévi csoportok (NP) azonosításának módszereit főnévicsoport-szinkronizációs alkalmazásban való felhasználhatóságukat vizsgálva hasonlítjuk össze. Az NP-azonosítás módszereinek szinkronizációs feladatban való vizsgálatát a MorphoLogic-nál fejlesztett MorphoTM [1] angol-magyar kísérleti fordítómemória fejlesztése motiválja.

A MorphoTM rendszer két fő ponton tér el a piacon termékként elérhető nagyrészt nyelvfüggetlen fordítómemóriáktól. Egyrészt a keresett forrásmondathoz az adatbázisban leginkább hasonlót nem csak karakteralapú, felszíni karaktersorozatokat összehasonlító hasonlósági mértéket használva keresi, a hasonlóság számítása során morfológiai és szintaktikai információra is épít. Másrészt nem csak teljes szegmenseket (mondatokat) keres az adatbázisában. Hasonló NP-ket illetve lehetséges mondatvázakat (a mondatban az NP-ket szimbolikus NP helyekre cserélve kapott struktúrát nevezzük így) is keres, majd a leghasonlóbbak tárolt fordításaiból a megfelelő morfológiai alakokat generálva épít javasolt fordítást.

A fordítómemória fedését növeli, hogy eredetileg különböző mondatok NP-it illetve mondatvázait kombinálva is képes fordításokat javasolni. (Azaz a MorphoTM valójában egy egyszerű minta-alapú gépi fordítórendszer.) A különböző emberi fordításokat automatikusan kombináló módszer hibás fordításokat eredményez, ha az NP-k fordítása a mondat vázától vagy más mondatbeli NP-től függ, ugyanakkor a javasolt

fordítás részei még ilyenkor is segíthetik a fordítómemóriát használó fordítót. Úgy gondoljuk, hogy a teljes szegmenseket tároló fordítómemóriáknál magasabb fedés még akkor is hasznos, ha az elérhető pontosság biztosan alacsonyabb.

Az adatbázisban tárolni kívánt NP-k azonosítását és fordításaikkal való összerendelését (szinkronizációját, párhuzamosítását) nem bízhatjuk a fordítómemóriát használó fordítóra, mert az NP-k megjelölésére és összerendelésére fordított munkaidő a későbbiekben valószínűleg nem térülne meg a fordítómemória fedésének növekedése révén. Az NP azonosítás és szinkronizáció feladatát tehát gépi algoritmusokra bízzuk.

Az előző években kidolgoztunk egy gyors, lexikai jegyeket használó NP szinkronizációs módszert [2], amely párhuzamosított korpuszon tanított osztályozót alkalmaz [3], így az NP-k kiválasztására alapvetően különbözően viselkedő (hibázó) módszereket is választhatunk, csak egy kézzel NP-szinkronizált tanítókorpuszt kell készíteni a választott NP-azonosító módszerhez. Cikkünkben a korábban kidolgozott NP-szikronizációs módszerünkkel csak érintőlegesen foglalkozunk, célunk a MorphoTM rendszerben az NP-k azonosítására leginkább alkalmas módszer mérésekkel történő kiválasztása, pontosabban az NP-k fordításbeli azonosítására legalkalmasabb módszert keressük, mivel a forrásoldali mondatokban mindenképp teljes mondatelemzéssel választjuk ki a főnévi csoportokat.

2 Az NP-azonosítás hibái

Automatikus módszerekkel, a vizsgált mondat megértése nélkül (szemantikai információk nélkül) az NP-k helyes azonosítása elvileg sem lehetséges, ahogyan erre az alábbi 1. példa is rámutat.

[I] saw [the man] in [the street].
[I] know [the girl in the garden]. (1. példa)

1. példa: Az automatikus NP-azonosító módszerek általában nem tudják helyesen kiválasztani a fenti angol példamondatok maximális méretű főnévi csoportjait. A maximális főnévi csoportokat zárójelekkel jelöltük.

Az automatikus NP-azonosító módszerek minimalizálni próbálják a hibákat, természetesen a teljes siker reménye nélkül. A fő kérdés, hogy egy adott alkalmazásra mely módszer a legalkalmasabb, illetve hogy az egyes módszerek hibái milyen következményekkel járnak az adott alkalmazásban. A hibákat tehát nem az NP azonosítás közvetlen kimenetében, hanem az NP-k azonosítására építő alkalmazás kimenetében kell majd keresnünk.

3 Az NP-azonosító és szinkronizáló módszerekkel szembeni elvárások a MorphoTM rendszerben

A MorphoTM rendszerben az NP-azonosítás és szinkronizáció minősége és sebessége egyaránt fontos. A minőség mérésére a szokásos fedés és pontosság értékeket használjuk.

A pontosság azért fontos, mert az adatbázisban tárolt hibásan azonosított és/vagy szinkronizált NP-párok, újra és újra megjelennek a javasolt fordításokban, amíg kézzel el nem távolítjuk őket az adatbázisból (jelenleg a MorphoTM rendszerben semmilyen automatikus megoldás nincs az adatbázis tisztítására).

Az NP-azonosító és szinkronizáló módszerek fedése több okból is fontos. Egyrészt minél több NP-párt tárolunk a memóriában, annál nagyobb lesz a fordítómemória fedése, másrészt minél több NP-párt azonosítunk egy mondatpárban, annál általánosabb lesz a párosított NP-ket szimbolikus NP-helyekre cserélve kapott mondatváz.

Fordítómemória alkalmazás esetében a sebesség különösen fontos. Az új mondatpárokat kevesebb, mint 1 másodperc alatt úgy kell tárolni a memóriában, hogy a frissen tárolt fordítás már a következő mondat javasolt fordításában is megjelenhessen. A hagyományos fordítómemóriák gyorsan tárolják a fordításokat, így a fordítók egy fejlettebb fordítómemóriától is jogosan elvárják ugyanezt. Az NP-k szinkronizációjára kifejlesztett módszerünk [2, 3] elég gyors (hosszú mondatpárokon is kevesebb, mint 10 ms alatt lefut egy 4 éves PC-n), azonban ez az NP-azonosításra használt módszerekről már nem mondható el, utóbbiak futtatása csak nehezen vagy egyáltalán nem fér be az egy másodperces időkeretbe.

4 Célnyelvi NP-azonosító módszerek

A MorphoTM rendszerben a forrásnyelvi mondatok NP-it a teljes mondat mély elemezésével választjuk ki. A tárolt fordítás NP-inek azonosítására több lehetőségünk is van, ezeket hasonlítjuk most össze. Megvizsgáljuk a teljes mondat mély szintaktikai elemzésének lehetőségét, az NP-ket fordításaik alapján sekély elemzővel meghatározó módszerünket valamint a két módszer kombinálásának lehetőségeit.

4.1 A célnyelvi mondatok mély elemzése

A mély elemzés során a teljes mondatot lefedő elemzési fát keres az elemző, vagy ha ilyet nem talál, a mondat kiesebb részeit külön fákkal fedi le. A mély elemzés memória- és számításigényes. A módszer előnye, hogy amennyiben a teljes mondatot lefedő elemzési fát talál, nagy pontosságú NP-azonosításra képes. Sajnos azonban számos hátránnyal is számolnunk kell:

- A MorphoTM rendszerben a magyar mondatok teljes elemzése túl sokáig (>1s) tartana. (A forrásnyelvi angol mondatok elemzése is 1 másodperc körüli időt vesz el, így nagyon gyors módszer kellene a magyar elemzéshez.)

- A teljes mondatot fedő elemzési fát sajnos a mondatok többségénél még nem talál az elemző, így a fedés még nem elég magas.
- A gyakori elemzési hibák csökkentik az NP-azonosítás pontosságát. A teljes mondatot lefedni nem képes elemzési fák gyakran tartalmaznak hibásan azonosított NP-ket.
- A forrás és célnyelvi mondatok elemzésére használt elemzők különböző hibákat ejtenek, ami megnehezíti a pontos szinkronizációt.

Az előbbiek miatt sajnos a teljes mondatelemzés továbbra sem valódi lehetőség a MorphoTM rendszerben.

4.2 Fordítás által támogatott sekély elemzés

A korábbi években kifejlesztettünk egy speciális NP-azonosító módszert, amely párhuzamos szövegekben a mondatpárok forrásoldalán megbízható módszerrel (a gyakorlatban a mondat teljes elemzésével) meghatározott NP-k párjelöltjeit jelöli meg a fordításban.

Módszerünk a megbízhatóan kiválasztott forrásnyelvi NP-k szavait leképezi a célnyelvi mondat szavaira. Tövesített szótári megfejeletetést, szófaji megfeleltetést alkalmazva, illetve hasonló alakú szavakat keresve [4] minden egyes nem grammatikai funkciót betöltő szó összes lehetséges párját megkeressük a célnyelvi mondatban. A pusztán grammatikai funkciót betöltő szavakhoz (pl. névmás, névelő) nem keresünk párt. Mivel egy forrásnyelvi szó több megfelelője és akár többször is előfordulhat a célnyelvi mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a célnyelvi mondatra. A legrövidebb illeszkedés szavait NP-váznak nevezzük. Természetesen a találatok között más szavakat is tartalmazhat az NP-váz. Az NP-vázat ezek után egyszerű szintaktikai szabályok (sekély nyelvtan) szerint, a forrásnyelvi főnévi csoport le nem fedett szavainak szófaját is figyelembe véve teljes célnyelvi főnévi csoporttá bővítjük. A bővítés során először a célnyelvi mondatban az illeszkedő szavak közötti meg nem feleltetett szavakat próbáljuk szófajuk alapján a forrásnyelvi NP meg nem feleltetett szavaival összerendelni, majd balra, illetve szükség esetén jobbra bővítjük a célnyelvi főnévi csoportot. (A bővítés preferált iránya nyelvfüggő, illetve függ attól, hogy a le nem fedett szavak között hány főnév van.)

Az algoritmus több előnyös tulajdonsággal bír:

- Az NP-azonosítás nagyon gyors (<10 ezredmásodperc egy 4 éves PC-n futtatva).
- Az algoritmus szinte teljesen nyelvfüggetlen, a nyelvpárfüggő szófajmegfeleltetési tábla, illetve a sekély nyelvtani szabályok meghatározása egy új nyelv esetében kevesebb, mint egy nap alatt megvalósítható.
- Nagy kétnyelvű vagy automatikusan gyűjtött szótárral magas fedés érhető el.

A módszer gyengéje megkérdőjelezhető pontossága. A hibásan kiválasztott (hibásan bővített) NP-ket a szinkronizáló algoritmusnak kell elvetnie.

4.3 NP azonosító módszerek kombinációi

Az NP-ket a fordításuk alapján sekély elemzéssel meghatározó módszerünk pontosságának növelése érdekében megpróbáltuk a módszert úgy kombinálni mély elemzéssel, hogy a mély elemzés előnyét (a nagy pontosságot) átvegyük, de a számításigény (azaz a futási idő) alacsony maradjon. Alapötletként megvizsgáltuk, hogyan lehetne csökkenteni az elemzési időt a mély nyelvtan módosítása nélkül. A rendelkezésünkre álló MetaMorpho elemezőt [5] vizsgálva azt tapasztaltuk, hogy legfeljebb 10 szavas bemenet esetén az elemzés megfelelően gyors. A kérdés ezek után csak az, mennyire várhatók pontos eredmények, ha nem a teljes mondatot elemezzük, hanem csak NP jelölteket ellenőrzünk a mély elemzővel (azaz azt vizsgáljuk, hogy az egész elemzett mondatrész lehet-e főnévi csoport).

Két lehetséges kombinációját próbáltuk ki az NP-k fordításalapú meghatározásának és mély nyelvtannal való ellenőrzésének:

- az NP-váz bővítésekor a mély elemzőt használva,
- a mély elemzőt csupán a fordításalapú sekély elemzés eredményének ellenőrzésére használva.

Az első esetben megvizsgáltuk, hogy balra, illetve jobbra maximum két szóval bővítve az NP-vázat a mély elemző által elfogadható NP-t kapunk-e. A bal oldali bővítést részesítettük előnyben, ha a baloldali bővítés eredményes volt, jobboldali bővítést nem kíséreltünk meg.

A második esetben csak azokat a sekély nyelvtannal kiválasztott NP jelölteket választottuk ki, amelyeket a mély elemző elfogadott.

5 Kiértékelési módszer

Tavaly az NP-azonosító módszerek elméleti összehasonlítását már elkezdtük [3], azonban még nem állt módunkban méréseket végezni. A puszta elméleti fejtegetésnél mérnökként sokkal fontosabbnak tartjuk a legegyszerűbb mérést is, persze csak akkor, ha a mérés során azt mérjük, amit kell és ahogyan kell.

Három NP-azonosító módszert hasonlítunk most össze (a fordításuk alapján sekély nyelvtannal NP-ket meghatározó módszerünket, az előbbi eredményét mély elemzővel ellenőrző módszert, illetve az NP-váz bővítését mély elemzőre bízó módszert). Az összehasonlítás célja annak megállapítása, hogy melyik módszer alkalmasabb NP-szinkronizációs alkalmazásban való felhasználásra, így elsősorban nem a nyers kimenetüket vizsgáljuk, hanem azt hogy milyen szinkronizációs eredmények érhetők el velük.

NP-szinkronizáló módszerünk többféle skaláris lexikai jegyet (*feature*) határoz meg minden egyes összehasonlított NP-párhoz, majd a jegyértékek egyszerű normalizálása után egy korpuszon tanított osztályozó dönti el, hogy a vizsgált NP-párjelöltet fordításként tároljuk-e a fordítómemória adatbázisában. Az osztályozási elfogultság (*bias*) elkerülése érdekében minden vizsgált NP-azonosító módszerhez külön be kell tanítanunk az osztályozót, így minden vizsgált módszerrel az elérhető legjobb eredményt fogjuk kapni.

6 Mérés

A mérésekhez egy kisméretű, 100 mondatpárból álló párhuzamos korpuszt választottunk. A korpusz mondatpárjai különböző szövegekből származnak, kiválasztásuknál nem vettünk figyelembe különleges szempontokat. A korpusz angol oldalán a mondatok átlagos hossza 14 szó.

Az angol mondatok NP-it a MetaMorpho angol elemzővel [5] azonosítottuk, majd a mondatpárok magyar oldalán a három vizsgált NP-azonosító módszert alkalmazva három angol-magyar párhuzamos korpuszt készítettünk.

A magyar főnévi csoportok vázának meghatározásakor egy 116000 szó- és kifejezéspárt tartalmazó angol-magyar szótárat használtunk. Mély elemzőként a MetaMorpho magyar elemzőt használtuk.

A főnévi csoportok automatikus azonosítása után a három párhuzamos korpuszban kézzel szinkronizáltuk a kiválasztott főnévi csoportokat angol fordításaikkal, így három tanító- és tesztkorpuszt készítetünk az NP-szinkronizáló módszerünk osztályozójának tanítására és tesztelésére.

A kézi szinkronizálás során a következőképp jártunk el:

- Csak teljesen megfeleltethető NP-ket rendeltünk egymáshoz, ha az egyik NP fordítása csak része volt a másik NP-nek, nem rendeltük őket egymáshoz.
- Amennyiben az NP-k az adott mondatpárban egymás fordításai voltak, akkor is rögzítettük őket, ha a mondatváztól függően más mondatpárban nem lehettek volna egymás fordításai.
- Egymáshoz rendeltük azokat az NP-ket, amelyek kis mértékben, de csak grammatikai funkciót betöltő szavakban különböztek, és a mondatokat egészben vizsgálva egymás fordításának tekinthettük őket (pl. *this family – a család*).

Az automatikus NP szinkronizáció minőségét a három tesztkorpuszon 10-szeres keresztkiértékeléssel mértük. A szinkronizáló módszerünkben a tavalyi méréseik szerint legmegfelelőbb logisztikus regressziós osztályozót alkalmazva mértük a szinkronizáció pontosságát.

7 Mérési eredmények

Mindhárom vizsgált NP-azonosító módszer esetében megvizsgáltuk a kiválasztott NP-jelöltek számát, a helyesen kiválasztott NP-jelöltek számát, az adott NP-azonosító módszerrel készített korpuszon tesztelve az NP-szinkronizáló algoritmus döntési pontosságát (a helyes döntések számát, a hamis pozitív, illetve hamis negatív NP-párokat), valamint a fordítómemória adatbázisába kerülő helyes illetve hibás NP-párok számát. Természetesen az utóbbi két érték függ az előzőektől, de segítenek a MorphoTM rendszerben legmegfelelőbb NP-azonosító módszer kiválasztásában.

Az I. táblázat a nyers NP-azonosítási eredményeket mutatja, a II. táblázatban az NP-szinkronizáció eredményeit rögzítettük, a III. táblázat pedig az egyes módszereket alkalmazva a fordítómemória adatbázisába helyezett helyes illetve hibás NP-párok számát mutatja.

Az eredményekből tisztán kiolvasható, hogy az NP-vázat mély elemzővel bővítő módszer rosszabb az NP-ket fordításaik alapján sekély elemzővel meghatározó módszernél, mind nyers NP-azonosítási eredményeiben, mind a vele elérhető NP-szinkronizációs eredményeket tekintve.

Az NP-ket fordításaik alapján sekély elemzővel meghatározó módszer NP-jelöltjeit mély elemzővel ellenőrizve magasabb pontosság érhető el, azonban lényegesen alacsonyabb fedés árán.

Az I. és II. táblázatban az NP-azonosítás pontosságát valamint a helyes szinkronizációs döntések számát összevetve a szinkronizáló algoritmus futtatásának haszna is megfigyelhető. A szinkronizációs algoritmus eredményesen veti el a rosszul kiválasztott NP-ket, de futtatásának haszna csökken, ha a bemenetén a helyesen kiválasztott NP-k aránya magas. (Más kérdés, hogy ha nem olyan NP-azonosító módszereket használnánk, amelyek eleve egy fordításbeli NP-hez keresnek párt, akkor az NP-szinkronizáló algoritmusra mindenképp szükség lenne. Jelen esetben csak szűrőnek használjuk az egyébként általánosan használható módszert.)

I. táblázat: Nyers NP-azonosítási eredmények

Módszer	Kiválasztott NP-pár	Helyes párok	Pontosság
TGSNPP	**228**	186	82%
NPS+DP	196	139	71%
TGSNPP+DP	130	115	**88%**

TGSNPP = NP-ket fordításaik alapján sekély elemzővel azonosító módszer, NPS+DP = az NP-vázat elemzővel bővítő módszer, TGSNPP+DP = TGSNPP eredményének ellenőrzése mély elemezővel. A TGSNPP fedése nagy (=sok kiválasztott NP-párjelölt), A TGSNPP+DP módszer pontossága nagy. Az NPS+DP módszer pontossága meglepően alacsony.

II. táblázat: Szinkronizációs eredmények

Módszer	Helyes döntés	Hamis pozitív	Hamis negatív
TGSNPP	86%	11%	3,1%
NPS+DP	83,7%	9,7%	6,6%
TGSNPP+DP	**90%**	**7,7%**	**2,3%**

A TGSNPP+DP módszer magas szinkronizációs pontosságot eredményez. Az NPS+DP módszer mindkét másiknál rosszabbul szerepel. (A módszerek azonosítóit lásd az I. táblázatnál.)

III. táblázat: adatbázisba került NP-párok

Módszer	Helyes párok	Helytelen párok	Pontosság
TGSNPP	**179**	25	87,7%
NPS+DP	126	19	86,9%
TGSNPP+DP	112	10	**91,8%**

A TGSNPP+DP kombinált módszer érte el a legnagyobb pontosságot, de fedése viszonylag alacsony. Az NPS+DP módszer a mély elemzőt nem használó TGSNPP-nél is rosszabb eredményt ért el. (A módszerek azonosítóit lásd az I. táblázat ismertetőjében.)

8 Összegzés

Korpuszalapú méréseket végeztünk annak érdekében, hogy megtaláljuk az NP-szinkronizációs feladatra legalkalmasabb NP-azonosító módszert.

A mérések ismét igazolták, hogy a fordítás alapján sekély nyelvtannal NP-ket azonosító módszerünk akár önmagában is megfelelő módszer NP-szinkronizációs feladatokra. A mérések alapján kijelenthetjük, hogy a forrásnyelvi NP-kből a célnyelvi mondatra leképezett NP-vázakat mély elemzővel bővítve, a fordítás által támogatott sekély elemzőt használó módszernél rosszabb eredményeket érünk el, ugyanakkor a mély elemzőt a pontosság növelése érdekében használhatjuk a sekély elemzés által kiválasztott NP-jelöltek ellenőrzésére.

Az NP-jelöltek ellenőrzésére a MetaMorpho magyar elemzőt használva a fordítómemória adatbázisába hibásan felvett NP-párok arányát 12,3%-ról 8,2%-ra sikerült csökkenteni, ami már lényeges különbség, főképp mivel eddig nem dolgoztunk ki megoldást az adatbázis automatikus tisztítására.

Sajnos a pontosság növelését csak a fedés jelentős csökkenése árán tudtuk elérni. Az adatbázisba felvett helyes párok száma 37,4%-kal csökkent. Az alacsonyabb fedés okozta problémára megoldás lenne, ha a mély elemző által el nem fogadott NP-jelölteket is felvennénk az adatbázisba, viszont a fordítások ajánlásakor az elemző által elfogadott párokat részesítenénk előnyben.

A mérések azt is megmutatták, hogy a jelenleg csak a párjelöltek szűrésére használt NP-szinkronizáló módszerünk [2, 3] minden esetben jobb eredményt ért el a minden párjelöltet elfogadó baseline módszernél.

Bibliográfia

1. Hodász G., Pohl G.: MetaMorpho TM: a linguistically enriched translation memory. In *International Workshop, Modern Approaches in Translation Technologies* (szerk.: Hahn, W.; Hutchins, J.; Vertan, C.), Borovets, pp. 26-30, 2005.
2. Pohl G.: English-Hungarian NP-alignment in MetaMorpho TM. In *Proceedings of EAMT 2006 (11th Annual Conference of the European Association for Machine Translation)*, pp. 69-74, 2006.
3. Pohl G.: A MorphoTM főnévicsoport-szinkronizáló módszereinek továbbfejlesztése és vizsgálata. In *IV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2006)*, pp. 190-201, 2006.
4. Simard, M., Foster, G. & Isabelle, P. (1992): Using Cognates to Align Sentences in Bilingual Corpora. In: *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine translation*, (TMI92), Montreal, pp. 67-81, 1992.
5. Prószéky, G.: Translating While parsing. In *A Man of Measure* (szerk.: M. Suominen et al.), The Linguistic Association of Finland, Turku, pp. 449-459, 2006.