A MorphoTM főnévicsoport-szinkronizáló módszereinek továbbfejlesztése és vizsgálata

Pohl Gábor

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

Kivonat: A MorphoTM (korábbi nevén MetaMorpho TM) olyan fordítómemória, amely nemcsak teljes mondatpárokat, hanem főnévi csoportokat (NP), illetve a mondatban ezeket szimbolikus NP-helyekre cserélve kapott mondatvázakat tárol fordításaikkal együtt adatbázisában. A főnévi csoportokat automatikusan határozzuk meg és szinkronizáljuk fordításaikkal. Cikkünk első felében összehasonlítjuk a magyar főnévi csoportok meghatározására kínálkozó módszereket, a MetaMorpho magyar szintaktikai elemzőt, illetve a főnévi csoportokat fordításaik alapján szótári megfeleltetésekkel és sekély nyelvtannal meghatározó, korábban kidolgozott módszerünket. Cikkünk második felében bemutatjuk, hogy hogyan váltottuk le a főnévi csoportok hasonlóságának meghatározására eddig használt heurisztikus képletet gépi tanulással meghatározott osztályozóra.

1 Bevezetés

A MorphoLogicnál fejlesztett MorphoTM (korábbi nevén MetaMorpho TM[72]) egy olyan EBMT-alapú[73], nyelvi tudásra is építő fordítómemória (TM), amely nem csak egész mondatpárokat, hanem a mondatnál kisebb részek párjait is tárolja. Jelenleg a teljes mondatpárokon kívül főnévi csoportokat (NP), illetve a mondatban ezeket szimbolikus NP helyekre cserélve kapott mondatvázakat tárolunk fordításaikkal együtt a fordítómemória adatbázisában. A keresés során, amennyiben a keresett mondathoz kellőképp hasonló forrásoldallal rendelkező teljes mondatpár nem található a

[72] A névváltoztatást a MetaMorpho fordítórendszer és a MorphoTM fordítómemória jobb megkülönböztethetősége indokolta. A MorphoTM fordítómemóriában továbbra is felhasználjuk a MetaMorpho fordítórendszerhez kifejlesztett szintaktikai elemezőt és nyelvtanokat, azonban a rendszer jelentős része független a MetaMorpho gépi fordítórendszertől, amely egyébként tartalmaz fordítómemória jellegű bővíthetőséget, ezért is szükség volt a névváltoztatásra.

[73] EBMT (Example Based Machine Translation) – minta alapú gépi fordítás. A fordítórendszer emberi fordításokból származó mintákból készít fordításokat, a tisztán statisztikai fordítórendszerektől abban különbözik, hogy tárolt nyelvi tudásra is épít.

memóriában, viszont megtalálhatók a keresett mondat vázához és főnévi csoportjaihoz kellőképp hasonló forrásoldallal rendelkező mondatrész párok, akkor az utóbbiak fordításoldalából – a megfelelő morfológiai alakok generálásával – építünk javasolt fordítást [1][2][3].

A főnévi csoportok fordításaikkal való összerendelését (szinkronizációját, párhuzamosítását) nem bízhatjuk a fordítómemóriát használó fordítóra, mert a főnévi csoportok megjelölésére és összerendelésére fordított munkaidő nem feltétlenül térülne meg a későbbiekben a fordítómemória fedésének növekedése révén. Ezen kívül a fordítómemória motor más fordítómemória rendszerekbe (pl. MemoQ) való beépítését is nehezítené, ha a felhasználói felületen a szokásos funkcióktól eltérőeket is megkövetelne. A főnévi csoportokat tehát automatikus módszerekkel határozzuk meg és szinkronizáljuk fordításaikkal.

Az automatikus főnévicsoport-szinkronizáció – teljes pontosságot biztosító módszer hiányában – alapvető hibaforrásként jelenik meg a rendszerben, a hibásan tárolt párokat később plusz munkával kell eltávolítani a memóriából, ezért az alkalmazott módszerekkel szembeni legfontosabb elvárás, hogy magas pontossággal határozzák meg a főnévicsoportpárokat. Emellett a magas fedés annyiban fontos, hogy ha csak kevés főnévi csoporthoz tudunk párt rendelni, akkor a fordítómemória fedése nem lesz sokkal nagyobb egy csak teljes mondatok kezelésére képes fordítómemóriáénál. A páratlan főnévi csoportokat a mondatváz részeként tároljuk (1. példa), hiszen ezeket a mondatvázból kiemelve nem tudnánk automatikusan meghatározni a maradék fordítását. Ez azt eredményezi, hogy alacsony fedésű módszerek alkalmazása esetén a mondatvázak a bennük tárolt párnélküli főnévi csoportoktól speciálisabbak lesznek, így kevésbé remélhetjük, hogy egy későbbi fordítási feladat során felhasználhatók lesznek.

> [I] have read [his new book on bread baking] and [I] am going to try [one of his recipes].
>
> Elolvastam [a kenyérsütésről szóló új könyvét] és ki fogom próbálni [egy receptjét]. (1. példa)

1. példa: Az angol *I* személyes névmáshoz nem található a magyar fordításban neki megfeleltethető főnévi csoport, illetve tegyük fel, hogy az automatikus módszerrel nem sikerült egymáshoz rendelni a mondatok utolsó főnévi csoportjait. (A példákban a maximális méretű főnévi csoportokat szögletes zárójelek határolják.) Ekkor a memóriába a teljes mondatpáron kívül az *NP1 := his new book on bread baking = a kenyérsütésről szóló új könyv* főnévicsoportpár, illetve az *I have read [NP1] and now I'm going to try one of his recipes = Elolvastam [NP1] és ki fogom próbálni [egy receptjét].* mondatvázpár kerül. A szimbolikus NP helyek megőrzik az eredeti főnévi csoportok morfológiai tulajdonságait, jelen esetben a *book* egyes számú voltát, illetve hogy a *könyvét* egyes számú, tárgy esetű.

A pontosság és a fedés mellett nagyon fontos az alkalmazott főnévicsoport-meghatározó és szinkronizáló módszerek sebessége, hiszen a fordító jogosan várhatja el, hogy a tárolt mondatpárok főnévi csoportjainak fordításai akár már a következő mondat fordításakor is megjelenjenek a javaslatok között, hiszen a hagyományos fordítómemóriák is gyorsan tárolják, és azonnal elérhetővé is teszik a fordításokat.

A MorphoTM rendszerben a korábbiakban két módszert javasoltunk a főnévi csoportok meghatározására [4][5]. Első ötletként a tárolt mondatpár angol és magyar oldalán is mondatelemzővel határoztuk volna meg a főnévi csoportokat, ezt a mód-

szert azonban a MetaMorpho magyar nyelvtanának [6][7] akkori alkalmatlansága miatt el kellett vetnünk, és kidolgoztunk egy módszert a magyar főnévi csoportok angol párjaik alapján, szótári és szófaji megfeleltetésekkel valamint sekély nyelvtani szabályokkal történő meghatározására. A MetaMorpho magyar elemző fejlődését figyelembe véve ebben a cikkünkben a korábbi ismertetésnél bővebben is összehasonlítjuk a két módszert, részletesebben elemezve az egyes módszerek előnyeit és hátrányait, illetve bemutatjuk a két módszer ötvözésének lehetőségeit, rámutatva arra, hogy a főnévi csoportok fordítás alapján történő meghatározása akkor is hasznos lehet, ha mindkét nyelvhez jó, de sajnos nem tökéletes, illetve lassú mondatelemzővel rendelkezünk.

Cikkünk második felében bemutatjuk, hogy a főnévicsoportpárok hasonlóságának meghatározására kidolgozott szótáralapú módszerünkben [4][5] a heurisztikus súlyozást hogyan váltottuk ki gépi tanulás alkalmazásával.

Végül a mérési eredmények ismertetése után a MorphoTM továbbfejlesztésével kapcsolatos terveinkről is szót ejtünk majd.

2 Főnévi csoportok automatikus meghatározása

A főnévi csoportok pontos és gyors meghatározása alapvető fontosságú feladat a MorphoTM rendszerben. A főnévi csoportok jó minőségű szinkronizálásához pontos főnévicsoport-meghatározó módszer szükséges azért, hogy a szinkronizáló algoritmusnak már lehetőleg csak párokat kelljen egymáshoz rendelni, és ne kelljen a hibásan meghatározott főnévi csoportokból adódó hibák szűrésével foglalkoznia. Tökéletes főnévicsoport-meghatározó módszert feltételezve elég lenne néhány egymásnak megfeleltethető szó, hogy egymáshoz rendeljük az egyes főnévi csoportokat, enélkül azonban a szinkronizáló módszernek kell elkerülnie az adatbázis hibás párokkal való feltöltését. Kicsit szabatosabban úgy fogalmazhatnánk, hogy minél pontosabban tudjuk meghatározni a főnévi csoportokat, annál nagyobb fedésűre hangolhatjuk a szinkronizáló algoritmusunkat.

A pontosság mellett a fordítómemória egésze (és nem a főnévicsoport-szinkronizáló algoritmus) szempontjából a fedés is fontos, hiszen önmagában nem sokat érünk egy tökéletesen pontos, módszerrel, ha az a főnévi csoportoknak csak kis részét képes felismerni.

A pontosság és fedés mellett azonban sajnos nem feledkezhetünk meg arról sem, hogy gyors módszerre van szükség. Tökéletes módszer híján egyszerre mindhárom cél sajnos nehezen közelíthető (nincs új a nap alatt).

A következőkben a MorphoTM rendszerben alkalmazott főnévicsoport-meghatározó módszereket fogjuk bemutatni és összehasonlítani, majd a módszerek ötvözésére vonatkozó javaslatot fogunk bemutatni.

2.2 Főnévi csoportok meghatározása szintaktikai elemzővel

A főnévi csoportok meghatározásának legalapvetőbb módszere szintaktikai elemző alkalmazása. A MorphoTM rendszerben az angol mondatokat a MetaMorpho angol nyelvtanát használva elemezzük. Amennyiben több elemzés is születik, jelenleg csak az elsőt vizsgáljuk. Amennyiben nem áll össze a teljes mondat elemzése akkor a

lehetséges részfákból a MetaMorpho heurisztikus gyökérválogató[74] módszereivel választunk egy elemzést, így a szinkronizáció során már csak egyetlen elemzés főnévi csoportjaihoz keresünk párt.

A MetaMorpho magyar nyelvtanának fejlődése lehetővé teszi, hogy a magyar mondatok főnévi csoportjait is szintaktikai elemzővel válasszuk ki, azonban egyelőre a következő pontban bemutatott módszerünket használjuk.

2.2 Főnévi csoportok meghatározása fordításaik ismeretében

Tavaly módszert mutattunk [4] főnévi csoportok fordításaik alapján szótárral, illetve sekély nyelvtannal történő meghatározására, jelenleg a MorphoTM rendszerben ezt a módszert használjuk a magyar főnévi csoportok meghatározására. A módszert most részletesen nem ismertetjük újra, csak a leglényegesebb elemeit foglaljuk össze.

Az angol elemzővel automatikusan meghatározott főnévi csoportokhoz rendelhető magyar főnévi csoportokat az angol főnévi csoportok szavait és kifejezéseit a magyar szövegre leképezve próbáljuk meghatározni. Az angol főnévi csoport nem csak grammatikai funkciót betöltő szavainak lehetséges fordításait tövesített szótári kereséssel (többszavas kifejezéseket is keresve), illetve hasonló alakú szavakat (*cognate*) keresve [8] próbáljuk a magyar mondatban megtalálni. Mivel egy angol szó több megfelelője és akár többször is előfordulhat a magyar mondatban, a lehetséges találatok közül azt választjuk ki, amelynek szavai a lehető legrövidebben illeszkednek a magyar mondatra. Természetesen a találatok között más szavakat is tartalmazhat a kijelölt illeszkedés. Az illeszkedést ezek után egyszerű szabályok szerint, az angol főnévi csoport le nem fedett szavainak szófaját is figyelembe véve teljes magyar főnévi csoporttá bővítjük.

A módszer előnye, hogy más nyelvekhez is könnyen elkészíthető, mindössze egy morfológiai elemző, egy szótár és egy főnévi csoportok meghatározására használható sekély nyelvtani szabályrendszer szükséges hozzá (természetesen a nyelvpár másik oldalán továbbra is szükség lenne szintaktikai elemzőre).

2.3 Magyar főnévicsoport-meghatározó módszerek összehasonlítása

Az összehasonlítás főbb szempontjairól (pontosság, fedés, sebesség) már írtunk a 2. szakasz elején, most lássuk, hogyan felelnek meg az egyes feltételeknek a fenti módszerek.

Legkönnyebben a módszerek sebessége illetve erőforrásigénye mérhető. A szintaktikai elemzőt nem használó módszer néhány ezredmásodperc alatt lefut hosszabb mondatpárokon is, ezzel szemben a MetaMorpho magyar szintaktikai elemző sajnos egyelőre nem mondható gyorsnak. Hosszú, összetett mondatoknál az elemzés egy átlagos személyi számítógépen akár percekig is eltarthat (közben a viszonylag kötetlen szórendből és a nyelv egyéb sajátosságaiból adódóan sok esetben 10 milliónál is több lehetséges csomópontot azonosít az elemző). Érthető, hogy a nyelvtan fejleszté-

[74] A MetaMorpho gépi fordítórendszerben a gyökérválogató módszereket mozaikfordítások készítésére használják, így akkor is képes a rendszer valamiféle fordítást adni, ha a teljes forrásmondatot nem tudta egyetlen fával lefedni.

sénél jelenleg elsődleges szempont a pontosság és fedés növelése, de a későbbiekben a fejlesztőknek az erőforrásigénnyel is szembe kell majd nézniük.

A fedés terén a szinkronizációs alkalmazásunk esetében egyelőre szintén a szintaktikai elemzőt nem igénylő megoldás tűnik jobbnak. 100 tesztmondatunkból 1-re valamiért nem futott le a magyar elemző, 17 mondat esetében pedig semmilyen elemzés nem született. A 82 elemzéssel rendelkező tesztmondatból (csak az első elemzéseket nézve) 16 esetében a mondat ismeretlen szavai következtében egyáltalán nem talált főnévi csoportot az elemző (a valóságban minden mondatban volt legalább egy főnévi csoport). A maradék 66 mondat esetében sem talált meg minden főnévi csoportot az elemző, illetve sokszor nem találta meg a maximális méretű főnévi csoportokat, csupán részeiket. Az MetaMorpho angol elemzőhöz viszonyítva ezek az eredmények még javulhatnak (feltehetően fognak is, hiszen a magyar elemző fejlesztése folyamatosan tart, az egy évvel ezelőtti állapothoz képest hatalmas javulást tapasztaltunk).

Az elemzőt nem igénylő megoldás a szinkronizáció szempontjából előnyös módon keresi a magyar főnévi csoportokat, azokat próbálja meg kijelölni, amelyeket a szintén szótári és szófaji megfeleltetéseket figyelő szinkronizációs módszer feltehetően egymáshoz rendel majd. A fedést azonban a módszer esetleges pontatlansága ronthatja, hiszen nem az a kérdés, hogy hány főnévicsoport-jelöltet talál a módszer, hanem, hogy hány valódi főnévi csoportot.

A pontosság terén egyértelműen jobbnak bizonyult a MetaMorpho magyar nyelvtana. A 2. példán látható, hogy az elemzőt nem használó módszer sok esetben rosszul határozza meg a főnévi csoport határait.

EN: *The Ombudsman has wide powers of investigation.*
HU: *Az Ombudsman széleskörű vizsgálati jogkörrel rendelkezik.*

```
NP_EN1>  The Ombudsman
NP_HU1>  Az Ombudsman

NP_EN2>  wide powers of investigation
NP_HU2>  Az Ombudsman széleskörű vizsgálati jog-
körrel
```

(2. példa)

2. példa: Az angol mondatban a MetaMorpho angol elemzővel talált főnévi csoportok és a magyar oldalon szintaktikai elemzőt nem használó módszerrel hozzájuk rendelt főnévi csoportok. Látható, hogy a második magyar főnévi csoport határát nem tudta pontosan meghatározni a módszer. (Ez a probléma orvosolható lenne, ha – balról jobbra haladva – a szükséges hasonlósági értéket elérő főnévi csoportok által lefedett szavakat foglaltnak jelölnénk, és nem használnánk őket más főnévi csoportban, ez a módszer azonban más problémákat is felvet, például, ha egy főnévi csoport egy másik részeként és önállóan is szerepel a mondatban, akkor a hosszabbat esetleg nem tudjuk így azonosítani.) A MetaMorpho magyar elemző helyesen azonosítja mindkét főnévi csoportot.

Az elemzőt nem használó módszert a 100 tesztmondatunkban 1 alkalommal megzavarta, hogy a tövesített szótári megfeleltetésnél eddig nem vizsgáltuk a szavak szófaját, így egy mondatpárban a *the tag* (= *a dögcédula*) főnévi csoporthoz a *jelölték* magyar szót rendelte az algoritmus (a *tag=jelöl* pár is szerepelt a szótárban).

Olyan esetek is előfordultak, ahol segített az elemzőt nem használó módszer nagyobb flexibilitása. Néhány esetben az angol elemző a mondat szabad bővítményeit helytelenül a főnévi csoporthoz csapta, ekkor egy szabályos főnévi csoportokat kere-

ső magyar elemzővel nem tudtunk volna párt találni ezekhez a „főnévi csoportokhoz”, azonban egy kis csalással a magyar főnévi csoporthoz hozzácsapva a szabad bővítményt már jó mondatvázpárt tudtunk az adatbázisba helyezni. Ezt a „kis csalást” az elemzőt nem használó módszerünk automatikusan elvégezte.

2.4 Javaslat a magyar főnévicsoport-meghatározó módszerek ötvözésére

A főnévicsoport-meghatározó módszereinket kétféleképp is ötvözhetjük. Egyrészt felhasználhatjuk az angol mondat főnévi csoportjait arra, hogy a lehetséges magyar elemzések közül olyat válasszunk, amelynek főnévi csoportjai tartalmazzák az angol főnévi csoportokhoz tövesített szótári megfeleltetéssel, illetve hasonló szavakat keresve hozzárendelt magyar szavakat. Ezzel az elemzési időt nem tudnánk csökkenteni, az elemzések pontossága azonban tovább nőhetne.

Másrészt leválthatjuk a magyar mondatban kijelölt főnévicsoport-vázat teljes főnévi csoporttá bővítő egyszerű szabályrendszert a MetaMorpho elemzőre. Ennek a megoldásnak előnye, hogy pontosabb eredményt tudunk majd elérni vele az eddigieknél, ugyanakkor remélhetjük, hogy az elemzési idővel se lesznek gondjaink, mivel az elemző csak hosszú mondatokra lassú, néhány szavas főnévi csoportokat gyorsan elemez. Így a néhány lehetséges főnévi csoport ellenőrzése se tartana sokáig. Kérdés viszont, hogy mennyire fogja pontosan meghatározni a főnévi csoportok határait a teljes mondat ismerete nélkül a MetaMorpho magyar nyelvtana.

3 A főnévi csoportok hasonlóságának meghatározása

A főnévi csoportok meghatározása után el kell dönteni, hogy a lehetséges párjelöltek közül melyeket rögzítsük párként az adatbázisban. Ha a mondatpár mindkét oldalán szintaktikai elemzővel jelöltük ki a főnévi csoportokat, akkor az összes lehetséges párosítást meg kell vizsgálnunk; ha az egyik nyelv esetében – szintaktikai elemzőt nem használva – a mondat fordításának főnévi csoportjai alapján határoztuk meg a főnévicsoport-jelölteket, akkor csak azt kell megvizsgálnunk, hogy ezek tényleg eléggé hasonlítanak-e párjaikra.

A feladat mindkét esetben az, hogy egy párjelölthöz egyetlen, a hasonlóságot jól jellemző skalár értéket rendeljünk. (A mindkét oldalon elemzőt használó módszer esetében lehetséges, hogy egy főnévi csoport több másikhoz is hasonlít, az ilyen helyzetek feloldását tavalyi cikkünkben [4] ismertettük.)

A MorphoTM rendszerben a hasonlóság mérésére a tavaly kifejlesztett, szótári és szófaji megfeleléseket kereső módszerünket [4] és ennek most bemutatott újabb változatát használjuk. A következőkben ezeket a módszereket fogjuk bemutatni és értékelni.

3.1 Szótáralapú és szófaji megfeleltetés

A hasonlósági vizsgálat során az összehasonlított két főnévi csoport szavait egymás után többféle módon is megpróbáljuk egymásnak megfeleltetni, majd az egyes mód-

szerek által lefedett tokenek számából számítjuk ki a hasonlóságot jellemző skalár értéket.

Először tövesített szótári keresést alkalmazunk: a forrásnyelvi főnévi csoport szavainak lehetséges töveit keressük egy speciális, tövesített indexet és találatlistát tartalmazó szótárban, majd a találatok közül csak azokat hagyjuk meg, amelyek a forrásoldalra illeszthetők és fordításuk minden szavának legalább egy lehetséges töve megtalálható a fordításbeli főnévi csoportban. A szótár segítségével többszavas kifejezéseket is keresünk. A főnévicsoportpárban így lefedett tokenek számát ettől kezdve **D**-vel jelöljük.

A szótári megfeleltetés után, a főnévicsoportpár le nem fedett, nagybetűt vagy számot tartalmazó szavai között hasonló alakúakat (*cognate,* [8]) keresünk. A főnévicsoportpárban így lefedett tokenek számát ettől kezdve **C**-vel jelöljük.

A korábban le nem fedett szavakat ezután szófajaik alapján próbáljuk egymáshoz rendelni. A szófajuk alapján megfeleltetett tokenek számát ettől kezdve **P**-vel jelöljük.

Végül a lefedetlen szavak közül kiválogatjuk a pusztán grammatikai funkciót betöltőket, ezeket az összehasonlítás során kisebb súllyal vehetjük majd figyelembe, hiszen nem jelentenek lényegi különbséget a két főnévi csoport között. A pusztán grammatikai funkciót betöltő szavak számát **F**-fel jelöljük.

A hasonlóság mértékét kiszámító módszerekben szükségünk lesz még a két főnévi csoport szavainak (most tokenjeinek) számára, ezt **W**-vel jelöljük.

(Az egyes megfeleltetési lépéseket korábbi cikkünkben [4] bővebben ismertettük, itt most ezért nem tértünk ki a részletekre.)

3.2 Heurisztikus hasonlósági érték

Eddig a MorphoTM rendszerben az előzőekben ismertetett tulajdonságjegyekből az 1. képletben meghatározott heurisztikus hasonlósági értéket alkalmaztuk, azokat a főnévicsoportpárokat tekintve tárolandó párnak, amelyeknél a hasonlósági érték meghaladott egy küszöbértéket.

$$h = \frac{1 \cdot D + 0{,}9 \cdot C + 0{,}3 \cdot P - 0{,}1 \cdot F}{W - F} \qquad \text{(1. képlet)}$$

1. képlet: Az eddigiekben alkalmazott *h* hasonlósági érték számításának módja. Azokat a frázispárokat tekintettük tárolandó párnak, ahol a *h* hasonlósági érték meghaladta a 0,67 küszöbértéket. A képletbeli együtthatókat néhány mondatpáron, kísérletezéssel állapítottuk meg.

3.3 Gépi tanulással meghatározott osztályozó

Bár a korábbi heurisztikus képlet (1. képlet) a gyakorlatban jónak tűnt, úgy döntöttünk, hogy kis korpuszt építve megvizsgáljuk, hogyan válthatnánk ki egy gépi tanulással meghatározott (azaz empirikus alapokon nyugvó) osztályozóval.

Az eredmények mérhetőségén kívül a gépi tanulás mellett szólt az is, hogy így a későbbiekben lehetőségünk lesz az eddigiek mellett újabb hasonlóságvizsgálati módszerek kipróbálására is, az egyes módszerek összevetése egyszerűen megoldható lesz.

Az alábbiakban bemutatjuk, hogyan építettünk egy kis tanító illetve tesztkorpuszt, hogyan normalizáltuk a 3.2 pontban felsorolt tulajdonságjegyeket (*feature*), majd bemutatjuk a WEKA gépitanuló-rendszerben [9] különböző osztályozókkal elért eredményeket.

3.3.1 Korpuszkészítés

A tanító illetve tesztkorpusz építésekor elsődleges célunk az volt, hogy a tényleges osztályozási feladatot szimuláljuk. A tesztkorpuszhoz mindenféle szempontok nélkül 100 angol-magyar mondatpárt választottunk ki, a mondatok átlagos hossza 14 szó volt. A mondatpárokból úgy építettünk tesztkorpuszt, hogy az angol oldalon a MetaMorpho angol elemzővel kijelöltük a főnévi csoportokat, majd lehetséges párjaikat a 2.2 pontban ismertetett elemzőt nem igénylő algoritmussal határoztuk meg, mivel jelenleg ezt a módszert használjuk a magyar főnévi csoportok kijelölésére. Ezek után minden egyes párt megvizsgáltunk, és kézzel megjelöltük, hogy helyesnek találjuk-e, majd automatikusan meghatároztuk a 3.1 pontban ismertetett tulajdonságjegyeket. Így egy olyan korpuszt kaptunk, amely a kinyert tulajdonságjegyek mellett tartalmazta, hogy tárolandó párnak tekintjük-e a tulajdonságjegyekkel jellemzett frázispárt.

Elvetettük azokat a párokat, ahol a főnévi csoportok közül az egyik csak a másik egy részének fordítását tartalmazta, azaz csak teljesen megfeleltethető párok egymáshoz rendelését fogadtuk el.

Azokban az esetekben, amikor a mondat fordítása más főnévi csoportokkal esetleg nem lett volna jó, de a főnévi csoportok összerendelése helyes volt, a főnévicsoportpárokat elfogadtuk, egy ilyen mondatpárt találunk a 3. példában is.

EN: *There were pictures of castles and lakes and pretty girls on the walls.*
HU: *A falakra kastélyok és tavak és szép lányok képeit ragasztották.*

```
NP_EN1>  pictures of castles and lakes and pretty girls
NP_HU1>  kastélyok és tavak és szép lányok képeit

NP_EN2>  the walls
NP_HU2>  a falakra
```

(3. példa)

2. példa: Ha a főnévi csoportok összerendelése helyes, akkor is elfogadjuk őket, ha mondatváz fordítása más főnévi csoportokkal esetleg rossz lenne. A mondatvázak jelen esetben az angol oldalon *There were [NP1] on [NP2]*, illetve a magyar oldalon *[NP2] [NP1] ragasztották*, amelyek csak ritkán feleltethetők meg egymásnak. A későbbiekben a mondatvázfordítások megfelelőségét is érdemes lehet vizsgálni.

Abban az esetben is elfogadtuk a főnévicsoportpárt, amikor a főnévi csoportok a mondatban egymás fordításai voltak, de a fordító apró, a lényeges szavakat nem érintő változtatást ejtett, például az angol *this family* főnévi csoportot magyarul *a családnak* fordította. Ezt azért engedtük meg, hogy jobb mondatvázpárt tárolhassunk az adatbázisban. Ehelyett a későbbiekben a mondatvázpárok megfelelőségét is érdemes lehet vizsgálni.

A korpuszt a WEKA rendszer által használt ARFF formátumban készítettük el, az olvashatóság és módosíthatóság érdekében kommentben rögzítve az egyes mondatpárokat és a belőlük kinyert főnévi csoportokat.

Az elkészült korpusz 186 helyes és 42 helytelen összerendelést tartalmazott.

3.3.2 Az adatok normalizálása

Az előző pontban ismertetett mintahalmazunkat vizsgálva azt találtuk, hogy a szótári megfeleltetéssel (D), a hasonló alakú szavakat keresve (C) és a hasonló szófajú szavakat keresve(P) megfeleltetett szavak száma – a várakozásnak megfelelően – szinte lineáris korrelációban áll a szavak számával (W). Ez után megvizsgáltuk, hogy az 1. (heurisztikus) képletben alkalmazott normalizálást is (vagyis a pusztán grammatikai funkciót betöltő, meg nem feleltetett szavak kihagyását), amely kis mértékben jobbnak bizonyult a kis korpuszunkon, a később bemutatott osztályozók megbízhatóságát 1-3%-kal növelte. Az így kapott normalizált tulajdonságjegyeket a 2. képletben rögzítettük.

$$\frac{D}{W-F}, \frac{C}{W-F}, \frac{P}{W-F}, \frac{F}{W-F} \qquad \text{(2. képlet)}$$

2. képlet: A főnévi csoportok hasonlóságának meghatározására használt tulajdonságjegyek normalizálása. Az empirikus képlet egybevág az elméleti meggondolásainkkal, a pusztán grammatikai funkciót betöltő szavak elhagyhatók az összehasonlítás során, illetve hosszabb főnévicsoport-párokban a szavak számával egyenesen arányosan több szót tudunk lexikai módszerekkel egymásnak megfeleltetni.

3.3.4 Tanulóalgoritmusok

A WEKA rendszerben elérhető osztályozó algoritmusok közül többet is kipróbáltunk. Próbálkoztunk többrétegű perceptronokból épített neurális hálóval (MLP), amelyet a szokásos *back-propagation* eljárással tanítottunk; próbálkoztunk RBF hálóval, ahol a radiális bázisfüggvényeket *K-means* klaszterezéssel állapítottuk meg, illetve kipróbáltuk a C4.5 döntési fa WEKA rendszerbeli J48 nevű implementációját. A komolynak tekinthető tanulóalgoritmusok mellett két egyszerűbb osztályozót is kipróbáltunk, egy lineáris regressziós modellre épülőt illetve egy logisztikus regressziós osztályozót.

A lineáris regressziós modell, az előzőekben ismertetett normalizálással gyakorlatilag az 1. képlet együtthatóinak korpuszalapú meghatározását jelenti.

A logisztikus regressziós modell hasonlít a lineáris regressziósra, ugyanakkor a paraméterek beállításakor a helyes döntések illetve a helytelen döntések arányát maximalizálja (3. képlet), illetve előnye, hogy kimenete valószínűség, azaz közvetlenül használható egy főnévicsoportpár hasonlóságának értékelésekor (4. képlet).

$$\ln\left(\frac{P(\text{fordítás})}{1-P(\text{fordítás})}\right) = \alpha\frac{D}{W-F} + \beta\frac{C}{W-F} + \gamma\frac{P}{W-F} + \delta\frac{F}{W-F} + \varepsilon \qquad \text{(3. képlet)}$$

3. képlet: A logisztikus regressziós osztályozó a tanulás során a helyes döntések meghozásának valószínűségét maximalizálja a görög betűkkel jelölt együtthatók beállításával.

$$P(\text{fordítás}) = \frac{1}{1 + \exp(-\alpha \frac{D}{W-F} - \beta \frac{C}{W-F} - \gamma \frac{P}{W-F} - \delta \frac{F}{W-F} - \varepsilon)} \qquad \text{(4. képlet)}$$

4. képlet: A logisztikus regressziós osztályozó kimenete annak a valószínűsége, hogy a vizsgált frázispár egymás fordítása.

3.3.5 Eredmények

Az egyes osztályozókat tízszeres keresztkiértékeléssel (*10-fold cross-validation*) tanítottuk és teszteltük. Azt vizsgáltuk, hogy a 228 mintából hány esetben döntenek helyesen. Viszonyítási alapként (*baseline*) tekinthetjük azt, hogy minden párjelöltet párnak tekintünk, így az osztályozónk a korpusz adottságainak megfelelően 186 esetben hozna helyes döntést.

Ha csak az osztályozók döntési pontosságát vizsgáljuk, akkor egy helytelen pár rögzítését egy helyes pár fel nem vételével azonos hibának tekintjük. A C4.5 döntési fa kivételével ez nem okoz gondot, a többi osztályozó kimenete egy 0 és 1 közötti valós érték, amelyet nem feltétlenül a 0,5 értéknél kell elvágnunk, például a logisztikus regressziós osztályozó valószínűségi kimenetét tekintve dönthetünk úgy is, hogy csak a 80% valószínűséggel fordításnak tekinthető párokat rögzítjük az adatbázisban.

A viszonylag bonyolult MLP, RBF és C4.5 osztályozóknál jobban szerepeltek a regressziós osztályozók, amelyek nagyjából azonos eredményeket értek el. A tanított osztályozók mellett a teljes tesztanyagon (azaz nem keresztkiértékeléssel) megvizsgáltuk a tavaly meghatározott képlettel elérhető osztályozási pontosságot is. Az eredményeket az 1. táblázatban foglaltuk össze.

osztályozó	**# helyes döntés**	**# helyes pár el nem fogadása**	**# helytelen pár elfogadása**
baseline	186	0	42
MLP + back-propagation	191	14	23
RBF + K-means	188	12	28
C4.5 (J48)	193	4	31
lineáris regresszió	**196**	8	24
logisztikus regresszió	**196**	7	25
régi heurisztikus képlet	184	14	30

1. táblázat: Az egyes osztályozókkal elért eredmények.

Látható, hogy a tavaly meghatározott heurisztikus képlet sajnos a baseline-nál is rosszabbul teljesített; érdekes azonban, hogy a tőle csak együtthatóiban különböző lineáris regressziós osztályozó lett a legjobb, nagyjából azonos eredményt érve el a logisztikus regressziós osztályozóval. A tanulással meghatározott együtthatók esetében a legfőbb különbség az volt, hogy a szófaji megfeleltetés jóval nagyobb pontszámot kapott.

Érdekes volt, hogy a korpusznak csak az első 50 mondatpárján tanítva és tesztelve az osztályozókat sokkal jobb eredményt értek el a regressziós modellek, a bonyolultabb osztályozók viszont a teljes korpuszon mérténél is jobban lemaradtak, feltehetően a tanítóminták kis számából adódóan. Ez azt is jelentheti, hogy a jelenleginél nagyobb korpuszon ezek az osztályozók behozhatják mostani lemaradásukat. A nagyobb korpuszból véletlenszerűen választva 50 mondatpárt a regressziós osztályozók eredményei már jobban hasonlítottak a teljes korpusznál mértekre, a bonyolultabb osztályozók viszont hasonlóan rosszul teljesítettek.

A korpusz növelésén túl azt is érdekes lesz megvizsgálni, hogy ha mindkét nyelv esetében elemzővel kiválasztott főnévi csoportokat vetünk össze, milyen eredményeket tudunk majd elérni az egyes módszerekkel.

4 Összefoglalás

Cikkünkben bemutattuk az elmúlt egy év a MorphoTM rendszer főnévicsoport-szinkronizáló moduljával kapcsolatos legfőbb eredményeit. A MetaMorpho magyar nyelvtanával pontosan tudnánk főnévi csoportokat keresni, de az elemzési sebesség sajnos ezt még nem teszi lehetővé, és a nyelvtan fedése is sokat javulhat még. A főnévi csoportok meghatározásának kérdése után bemutattuk, hogyan váltottuk le gépi tanuló módszerek alkalmazásával az eddig használt heurisztikus főnévicsoport-hasonlóságot mérő képletünket (pontosabban csak együtthatóit, hiszen egy hasonló struktúrájú képlet bizonyult a legjobbnak).

5 További tervek

A bemutatott módszereink további vizsgálata érdekében tervezzük egy nagyobb főnévicsoport-szinten párhuzamosított korpusz építését.

Főnévi csoportokon kívül lehetőségünk lenne más mondatrészeket is tárolni a fordítómemória adatbázisában (pl. melléknévi csoportok, határozói szerkezetek), ezekkel eddig nem foglalkoztunk, de úgy gondoljuk, hogy kezelésüket a főnévi csoportokéhoz hasonlóan tudnánk megoldani.

A 3. példában rámutattunk arra, hogy esetleg érdemes lehetne a mondatvázak hasonlóságát is vizsgálni, hogy csak feltehetően helyes fordításokat tároljunk az adatbázisban. Felvetődik azonban a kérdés, hogy a helyességvizsgálattal hány valójában megfelelő párt vetnénk el.

A maximális méretű főnévi csoportokon belüli kisebb főnévi csoportok szinkronizációját is hasznosnak tartjuk. Az 1. példa főnévi csoportjaira visszatekintve láthatjuk, hogy viszonylag kis többletmunkával megsokszorozhatnánk az adatbázisba felvett főnévi csoportok számát.

A MetaMorpho magyar nyelvtanának fejlődésével lehetőségünk lesz a jelenlegi szómegfeleltetéseket használó módszereken túl elemzésifa-szinkronizáló módszerek vizsgálatára is.

Bibliográfia

[1] Hodász G., Grőbler T., Kis B.: Translation Memory as a Robust Example-based Translation System. In Proceedings of the Ninth EAMT workshop, University of Malta, Valletta, pp. 82-89, 2004.

[2] Hodász G.: Nyelvi hasonlóságon alapuló intelligens keresés fordítómemóriában. In *II. Magyar Számítógépes Nyelvészeti Konferencia* (szerk.: Alexin Z, Csendes D.), Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 108-116, 2004.

[3] 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.

[4] Pohl G.: Angol–magyar szótáralapú főnévicsoport-szinkronizáció és fordításalapú főnévicsoport-meghatározás. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 125-133, 2005.

[5] Pohl G.: English-Hungarian NP Alignment in MetaMorpho TM. In *EAMT 11th Annual Conference, 2006,* (CD-ROM)

[6] Tihanyi L.: A MetaMorpho fordítóprogram projekt 2005-ben. In *III. Magyar Számítógépes Nyelvészeti Konferencia* (szerk.: Alexin Z, Csendes D.), Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 99-107, 2005.

[7] Merényi Cs.: A MetaMorpho magyar-angol gépi fordító rendszer ige vonzatkereteit működtető nyelvtan. In *III. Magyar Számítógépes Nyelvészeti Konferencia*, Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged, pp. 108-115, 2005.

[8] 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

[9] Witten, I. H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco, 2005.