Egyértelműsítés és „mozaikfordítás" a MetaMorpho rendszerben

Gröbler Tamás

MorphoLogic kft.
1126 Budapest, Orbánhegyi út 5.
grobler@morphologic.hu

Kivonat: A gépi fordító rendszerekben több szinten is szükség van egyértelműsítésre. A MetaMorpho rendszerben különös jelentősége van az egyértelműsítésnek akkor, amikor a teljes mondat szintaktikai elemzése sikertelen, de számtalan – jellemzően egymást átlapoló – részelemzés keletkezik, amelyekből össze kell állítani a teljes mondat lehető legjobb fordítását. Ezt az eljárást nevezzük *mozaikfordításnak*. A cikk összefoglalja azokat a szintaktikai elemzést kiegészítő heurisztikákat, amelyeket a legjobb mozaikfordítás előállításának érdekében alkalmazunk.

1. Bevezetés

A nyelvi elemzés szempontjából általában a többértelműség két típusát szokás megkülönböztetni: a lexikális és a strukturális többértelműséget. Ezek kezelésének hagyományos, „tankönyvi" receptje szerint a lexikális többértelműség feloldása a szintaktikai elemzés előtt, a strukturális egyértelműsítés pedig az után következik.

A MetaMorpho angol-magyar gépi fordítórendszerben [3] ettől eltérő megközelítést alkalmazunk. Az egyes mondatok fordításakor abból indulunk ki, hogy az általunk kézben tartott és bármikor fejleszthető nyelvtanra épülő szintaktikai elemzés sikere, vagyis a teljes mondat szerkezetének feltárása előnyt élvez minden más egyértelműsítéssel szemben. Ritkán előfordul, hogy egy mondatot többféleképpen is sikeresen elemez meg a nyelvtan, de ilyenkor az összes megoldást jónak tartjuk. Ezért a szintaktikai elemzés előtt nem végzünk egyértelműsítést, és ha a teljes mondat elemzése sikeres, akkor utána sem.

Vizsgálatunk tárgya tehát az az eset, amikor a teljes elemzés nem áll elő, hanem a mondat egyes részeire épülő elemzések sokaságából kell összeválogatni azokat, amelyek

a.) a legjobb fordítást eredményezik,
b.) nem átfedők,
c.) lehetőleg lefedik a teljes mondatot.

Ezt az eljárást nevezzük *mozaikfordításnak*, amely során nagymértékben felhasználjuk az egyes egyértelműsítési eljárások eredményeit.

2. Egyértelműsítés a MetaMorpho rendszerben

A MetaMorpho fordítórendszer jelenleg a következő egyértelműsítési eljárásokat tartalmazza:

- szófaji egyértelműsítés
- jelentés szerinti egyértelműsítés
- a lexikális egységek határainak egyértelműsítése
- szintaktikai egyértelműsítés

Míg az utóbbi két eljárás alapvetően a nyelvtan részeként működik, a szófaji és a jelentés szerinti egyértelműsítést külön modulok (POS tagger, ill. WSD modul [2]) végzik, amelyek a hozzájuk kapcsolódó *szűrőkön* keresztül épülnek be az elemzés folyamatába [4].

A nyelvtan számos olyan mechanizmust tartalmaz, amelyek a szintaktikai szerkezetet egyértelműsítik, és ezáltal biztosítják, hogy a tipikusan több ezer létrejött elemzési tényből általában csak néhány tucat gyökérelem marad, amelyek közül válogatni kell. Ilyen mechanizmus például a tények egymás közötti „ölési mechanizmusa" és a gyökérszimbólumok kitüntetése a tények között [3].

A nyelvtan többszintűsége lehetővé teszi a szintenkénti egyértelműsítést is, például a többszavas lexikális egységek (szóösszetételek, számok, dátumok, földrajzi, intézmény- és személynevek) felismerését, és ezáltal a lexikális elemek határainak egyértelműsítését. (A korábbi változatban ezt a feladatot is egy szűrő végezte.)

A jelentés-egyértelműsítő szűrőt [2] mutatja be. Ez annyiban különbözik a többitől, hogy az elemzések számát nem csökkenti, hanem az egyes részelemzésekből generálható többes fordítás lehetőségét szünteti meg.

A mozaikfordítás előállításához szükséges további egyértelműsítéseket az utolsó fejezetben bemutatott szűrők végzik.

3. A mozaikfordítás előállítása

A mozaikfordítást a gyakorlatban úgy valósítjuk meg, hogy a forrásnyelvi szintaktikai elemzés eredményét *szűrőkön* engedjük keresztül [4]. Mivel bizonyos esetekben szükségünk lehet az összes részelemzésre, a legjobb fordítás érdekében először egy *rendező szűrőt* alkalmazunk, amely az egyértelműsítő modulok segítségével az alábbiak szerint állítja sorba a részelemzéseket.

A rendező szűrő bármely két részelemzésről képes eldönteni, hogy melyiket részesíti előnyben. Az összehasonlítandó két elemzést az előre megadott rendezési szempontok szerint, azok szigorú sorrendjében hasonlítja össze. A szempontok sorrendjében később következő szempontokat csak akkor vesszük figyelembe, ha az előttük álló egyik szempont szerint sem eldönthető, hogy melyik részelemzés a jobb. Az egyes rendező szűrők működését a következő fejezet mutatja be.

A mozaikfordításokat egy külön szűrő állítja össze a már rendezett részelemzésekből a b.) és c.) követelményeknek megfelelően. Az egyes lefedéseket pontozzuk aszerint, hogy az előzetes rendezés szerint mennyire jó mozaikokból állnak. Egy lefedés pontszámát úgy számítjuk ki, hogy összeadjuk a benne szereplő részelemzéseknek (mozaikoknak) a rendezés során kialakult sorszámát. Ezáltal a kevés és „jó" mozaik-

ból álló fordítás pontszáma lesz a legkisebb. Mivel a lehetséges lefedések száma exponenciálisan függ a részelemzések számától, ezt a válogatást legalább részben „mohó" módon végezzük, vagyis a legjobb elemzésektől indulva csak előre rögzített számú lefedés pontszámát számítjuk ki. Ha csak a legjobb mozaikfordításra vagyunk kíváncsiak, akkor a legalacsonyabb pontszámú lefedést választjuk.

4. Egyértelműsítő szűrők

Az alábbi példákban a MetaMorpho rendszerben jelenleg megvalósított szűrőket mutatjuk be. A szűrők működését egy-egy példával is illusztráljuk. A példákban olyan mondatokat kell vizsgálnunk, amelyek teljes elemzése valami miatt meghiúsul. Mivel ezek általában hosszúak, az áttekinthetőség kedvéért csak azokat a kifejezéseket tüntetjük fel, amelyeken a szűrő hatása a legjobban megfigyelhető. Az egyértelműsített elemzés bemutatására a magyar fordítás látszott a legalkalmasabbnak. A fordításban szögletes zárójel jelöli az egyes mozaikok határát.

4.1. Szófaj-egyértelműsítő szűrő

Ehhez a szűrőhöz külön modulként implementáltuk Brill [1] szófaji egyértelműsítőjét (POS tagger). A szűrő az egyértelműsítés által kijelölt szófajokból épülő részelemzéseket részesíti előnyben. Ha két részelemzés azonos tartományt fed le, akkor az „győz", amelyik kevésbé tér el az egyértelműsített szófaj-sorozattól. A nem azonos tartományt lefedő részelemzéseknél azonban óvatosan csak az olyan eltérést vesszük figyelembe, ahol az egyik szófaj ige (de nem gerund vagy perfect alak), a másik pedig névszó (főnév, melléknév vagy határozószó).

	szűrő nélkül	szűrővel
angol	*the dog barks*	
magyar	*[a kutyakérgek]*	*[a kutya ugat]*

4.2. Tagmondat-kiválasztó szűrő

Sokszor azt szeretnénk, hogy ha a teljes mondat nem is áll össze, a felismert tagmondatok mindenképpen szerepeljenek a mozaikfordításban. Ezek helyett azonban sokszor más részelemzések is előjönnek, és ilyenkor a szófaji egyértelműsítés is tévedhet. Ezért az önállóan tagmondatot alkotó részelemzéseket előnyben részesítjük.

	szűrő nélkül	szűrővel
angol	*the plane lands*	
magyar	*[a repülőgépföldek]*	*[a repülőgép leszáll]*

4.3. Töredék-szűrő

Mivel semmiképpen sem szeretnénk, hogy bármely szó is kimaradjon az elemzésből, minden szóhoz tartozik egy részelemzés, amely végső esetben „beugrik" az esetleg kimaradó pozícióba. Az ilyen részelemzéseket töredékeknek hívjuk. A töredék-szűrő gondoskodik róla, hogy a töredékek a lista aljára kerüljenek, és csak akkor kerüljenek be a mozaikfordításba, ha nincs más elemzés. A példában a római számokból képzett töredéket bírálja felül a személyes névmásból létrejött főnévi csoport.

	szűrő nélkül	szűrővel
angol	*I*	
magyar	*[I]*	*[én]*

4.4. Pozicionális szűrő

Ha a részelemzések sorrendjét a fenti szűrők egyike sem tudja meghatározni, akkor azok egymáshoz viszonyított helyzete alapján állapítjuk meg a sorrendet. Ebben az esetben a leghosszabb elemzéseket részesítjük előnyben, ezek közül pedig a balrább kezdődőket választjuk.

	szűrő nélkül	szűrővel
angol	*the train stations in poor countries*	
magyar	*[a vonat] [állomásoztat] [] [szegény országok]*	*[a vonatállomások szegény országokban]*

5. Összefoglalás

A bemutatott egyértelműsítési eljárások segítségével elértük, hogy a mozaikfordítás minősége jelentősen javult. A továbbiakban a legtöbb lehetőséget a szófaji egyértelműsítésben látjuk. Ennek finomításától és más POS tagger algoritmusok kipróbálásától további javulást remélünk.

Irodalom

1. Brill, E. (1992): 'A simple rule-based part of speech tagger'. *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, Trento, Italy.
2. Miháltz M. (2004): 'Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő rendszerrel'. *ebben a kötetben*
3. Prószéky G., Tihanyi L. (2002): 'MetaMorpho: A Pattern-Based Machine Translation Project'. *Translating and the Computer 24*, ASLIB, London.
4. Prószéky G., Tihanyi L., Ugray G. (2004): 'Moose: A Robust High-Performance Parser and Generator'. *9th EAMT Workshop*, Malta.