Automatikusan előállított proto-szótárak közzététele

Héja Enikő[1], Takács Dávid[1]

[1] MTA Nyelvtudományi Intézet
{eheja,takdavid}@nytud.hu

A három éve folyó EFNILEX projekt célja (ld. Héja, 2010) annak vizsgálata, hogy a modern nyelvtechnológiai eszközök mennyiben alkalmasak a szótárkészítés támogatására. Jelen demonstráció célja, hogy bemutassa az automatikusan előállított prototípus szótárak (a továbbiakban proto-szótárak) lekérdezhető változatát.

A proto-szótárak újdonságát az adja, hogy párhuzamos korpuszokon automatikusan, szóillesztéssel állítjuk elő őket. Bár már majdnem két évtizede használnak különféle statisztikai algoritmusokat forrásnyelvi és célnyelvi szópárok kinyerésére, hogy így bővítsék a gépi fordítás bemenetéül szolgáló szótárakat (pl. Wu, 1994), érdekes módon a lexikográfusok között a mai napig sem eldöntött kérdés, hogy használhatóak-e a párhuzamos korpuszok emberi felhasználásra készülő szótárak előállítására.

Az így létrejövő szótárak természetesen több ponton is lényegesen különböznek a hagyományos, lexikográfusok által létrehozott szótáraktól. A legfontosabb különbség, hogy a proto-szótárak alapstruktúrájában más típusú adatokkal találkozunk: a proto-szótárak mikrostruktúrája kevésbé kidolgozott, de a fordítási jelölteken kívül korpuszgyakorisági adatokat valamint az illesztő algoritmus által kalkulált fordítási valószínűséget ($P(szó_{cél}|szó_{forrás})$) is tartalmazza. Nagymennyiségű természetes nyelvi kontextus áll rendelkezésre, valamint könnyen kiszámíthatóak a fordított irányú proto-szótár fordítási valószínűségei is ($P(szó_{forrás}|szó_{cél})$) is. A proto-szótár hátránya, hogy utószerkesztési munkálatok hiányában szükségszerűen tartalmaz hibás jelentésmegfeleltetéseket is. Általánosan elmondható, hogy a proto-szótár fedése és pontossága fordítottan arányosak: a fent említett paramétereken alapuló szűréssel növelhető a jó fordítási jelöltek aránya, ennek az ára viszont a szótár fedésének a csökkenése.

Célunk egy olyan online felület fejlesztése, amely kiaknázza a módszer előnyeit és minimálisra csökkenti a hátrányait. Fedés és pontosság vonatkozásában ez azt jelenti, hogy a lekérdező felülettel a proto-szótárak személyre szabhatóak lesznek: a fedés-pontosság görbe különböző pontjai eltérő felhasználói igényeknek feleltethetőek meg. Pl. egy kezdő nyelvtanuló esetében az alapszókincsre van szükség, és az is elvárás, hogy a célnyelvi megfelelő a legjobb (legtöbbet használt) fordítás legyen. Ebben az esetben tehát a proto-szótárat úgy vágjuk, hogy a gyakoribb szavakat vesszük csak figyelembe mind a forrásnyelvi, mind a célnyelvi oldalon, és a fordítási párok közül is csak azokat, amelyeknek magas a fordítási valószínűsége. Ezzel szemben egy fordító képes a rossz fordítások közül a jót kiszűrni, különösen, ha rendelkezésére állnak a javasolt fordításokat támogató párhuzamos szövegrészletek, így az ő esetében egy nagyobb lefedettségű, ám alacsonyabb pontosságú proto-szótár megfelelő. Ezért

követelmény, hogy az online felületen a felhasználó határozhassa meg, hogy a proto-szótár melyik szeletével kíván dolgozni.

A proto-szótár paramétereinek beállításával határozható meg a szótár mérete. Eddigi kiértékelési eredményeink szolgálhatnak ugyan némi fogodzóul arra nézve, hogy hogyan érdemes ezeket a paramétereket beállítani, ám ezzel pont a valódi testreszabás lehetőségét veszítjük el: sokkal célszerűbb lehetővé tenni, hogy a felhasználó egyénileg kisérletezhesse ki, melyek a számára optimális paraméterbeállítások. Célunk annak biztosítása, hogy ezt ne közvetlenül a paraméterek beállításával kelljen megtennie, hanem a felhasználó számára jelentéstelibb módon, fordítási példákon keresztül lehessen elvégezni.

Nevezzük fordítási sornak egy forrásnyelvi kifejezéshez tartozó fordítási valószínűség mentén rendezett célnyelvi kifejezések összességét. A fordítási sorokból általában az első néhány tétel jó megfelelő, míg a fordítási valószínűség csökkenésével egyre több rossz fordítási jelölt is megjelenik, ám jó fordítások még mindig felbukkanhatnak. Az online felület lehetővé teszi, hogy a felhasználó válassza ki, hogy melyik az az utolsó fordítás, ami számára még elfogadható. A fordítási sorban az összes korábban megjelenő fordítást jónak, a későbbieket pedig rossznak tekintjük. Több fordítási sor vágásával létrejön egy tanító adathalmaz, amelyre egy vágófüggvényt illesztve automatikusan határozzuk meg a felhasználó számára legmegfelelőbb paraméterbeállításokat.

A ritkán használt fordítások értelmezésénél nyújt segítséget a nagymennyiségű természetes példamondat, amely a kérdéses fordításra kattintva kilistázható.

A felület kialakításánál célunk, hogy a rendelkezésünkre álló információkat vizuálisan reprezentáljuk. A fordítási jelölteket szófelhőben illetve szógráfban is megjelenítjük. Az ábrázoláshoz az alábbi változók közül választhatunk: oda- és visszirányú fordítási valószínűség, forrásnyelvi és célnyelvi szó relatív gyakorisága.
Hipotézisünk szerint ezek mentén a paraméterek mentén a fordítási jelöltek különböző osztályokba sorolhatók, aszerint, hogy milyen szemantikai viszony áll fenn a fordítási pár két tagja között, illetve a fordítási jelöltek jelentése szerint. Például, ha mindkét irányú fordítási valószínűség magas és a gyakoriságok megközelítőleg megegyeznek, a fordítási jelöltek nagy valószínűséggel jól meghatározott, konkrét dolgokra referáló kifejezések lesznek (pl. terminusok, tulajdonnevek.) Ezzel szemben, ha az odairányú fordítási valószínűség magas, de a célnyelvi kifejezés sokkal gyakoribb, valószínű, hogy a célnyelvi kifejezés jelentése sokkal általánosabb, illetve a forrásnyelvi kifejezés használata jelölt. Pl. egy magyar-litván párhuzamos tesztkorpuszban a magyar *tüzetes* szó 5-ször fordul elő, míg a litván *įdėmiai* 100-szor úgy, hogy a fordítási valószínűségük magas: 0.76. Valóban, egy angol-litván szótár alapján a litván szó jelentése sokkal általánosabb: *attentively*, *carefully* – ’figyelmesen’, ’óvatosan’, ’gondosan’ jelentései egyaránt lehetnek.

Bibliográfia

1. Héja, E.: The Role of Parallel Corpora in Bilingual Lexicography. In: Proceedings of the LREC2010 Conference, La Valletta, Malta, May 2010, pp. 2798-2805.
2. Wu, D. (1994), Learning an English-Chinese Lexicon from a Parallel Corpus. In: Proceedings of AMTA'94. 206–213.