Javaslat szemantikailag annotált többnyelvű tanítókorpuszok automatikus előállítására jelentés-egyértelműsítéshez párhuzamos korpuszokból

Miháltz Márton[1], Pohl Gábor[2]

[1]MorphoLogic, Orbánhegyi út 5, 1126 Budapest
mihaltz@morphologic.hu
[2]Pázmány Péter Katolikus Egyetem Információs Technológiai Kar
1083 Budapest, Práter utca 50/A
pohl@itk.ppke.hu

A cikkben bemutatunk egy kísérletet, melynek célja, hogy automatikus módszerekkel annotált tanítókorpuszokat állítsunk elő angol-magyar, illetve magyar-angol fordítórendszerben működő jelentés-egyértelműsítő modul számára. A tanítópéldákat, melyekben a forrásnyelven többértelmű (tehát több lehetséges fordítással rendelkező) szavakat célnyelvű fordításaikkal látunk el, nagyméretű, mondatszinten szinkronizált párhuzamos korpuszból nyerjük ki. Az annotáló algoritmus kétnyelvű szótárak és statisztikus heurisztikák alkalmazásával működik.

Egy olyan szabályalapú gépi fordítórendszerben, mint a MetaMorpho megértéstámogató fordítóprogram [6], jelentős kihívást jelent a többértelmű lexikális elemek kezelése. A forrásnyelvi nyelvtani elemzés során csak korlátozott mértékben van lehetőség a forrásnyelven többértelmű, következésképpen a célnyelven is általában több különböző fordítással rendelkező szavak egyértelműsítésére. Többértelmű főnévi, melléknévi és gyakran igei elemeknél a rendszernek szüksége van külső segítségre, melyhez egy statisztikai gépi tanuláson alapuló jelentés-egyértelműsítő alrendszert fejlesztettünk [3]. Ez a modul a forrásnyelven többértelmű szó eredeti kontextusában (a fordítási egység bekezdésében) megfigyelt szemantikai és szintaktikai információk alapján hoz döntést a legvalószínűbb célnyelvi fordításról.

Minden, a forrásnyelven többértelmű szóhoz külön osztályozót használunk, melyek annotált tanítópéldákból betanított modelleken alapulnak. Megfelelő tanítópéldák előállítására korábban angol nyelvű lexikális erőforrásokkal (Princeton WordNet) annotált korpuszokat használtunk, melyekben az angol jelentés-címkéket magyar fordításokra képeztük le. Mivel ilyen korpuszok véges mennyiségben állnak csak rendelkezésre, a rendszer felskálázásához további megoldásokra van szükség. Az egyik lehetőség angol korpuszokból kigyűjtött példák kézi annotálása a többértelmű szavak magyar fordításaival. A kézi annotálás azonban rendkívül időigényes, és ezért költséges folyamat.

Egy másik, kedvezőbb alternatíva párhuzamos korpuszok felhasználása. Mivel a jelentés-egyértelműsítő modulnak a mi esetünkben eleve célnyelvi fordításokkal annotált tanítópéldákra van szüksége, a kétnyelvű, mondatszinten szinkronizált szövegben a többértelmű szavakat a másik oldalon megtalálható fordításaikat azonosítva juthatunk megfelelő tanítóanyaghoz ([1], [5]).

A Hunglish projektben elkészített Hunglish kétnyelvű angol-magyar párhuzamos korpusz 44,6 millió angol, illetve 34,6 millió magyar szövegszót tartalmaz [7]. A szabadon felhasználható, nagypontosságú mondatszintű illesztéssel ellátott korpuszt szeretnénk felhasználni mind többértelmű angol, mind többértelmű magyar szavak előfordulásainak automatikus annotálásához.

A korpusz angol oldalán automatikus szófaj-egyértelműsítőt (POS-tagger) [2] alkalmazunk, mivel a MetaMorpho rendszerben egy adott szóalak különböző szófajú előfordulásaihoz külön jelentés-egyértelműsítő modell betanítása szükséges. A többjelentésű szó lehetséges fordításait tövesítve keressük a fordításban, ha több változatot is találunk, a mondatpárt nem egyértelműként jelöljük meg. Az ilyen mondatpárok esetleg elhagyhatók, kézzel egyértelműsíthetők, vagy ha túl gyakoriak (gyakori többértelmű igék esetében), automatikus módszerrel is megpróbálhatjuk egyértelműsíteni őket: a szó (szavak) környezetét szótárral, illetve szószintű szinkronizációs algoritmusokkal leképezve a mondatpárban. Ha nem találunk egy keresett szóhoz fordítást, megvizsgáljuk, hogy ismert kifejezés részét képezi-e a mondatban (ezekkel nem foglalkozunk). A szavak ismeretlen fordítású előfordulásait tartalmazó mondatpárokban megpróbálunk statisztikai módszerekkel [4] valószínű fordításokat keresni, a jó találatokkal bővítjük a lehetséges fordítások halmazát, majd megismételjük az eljárást.

Az algoritmus eredményeinek kiértékelésére a következő metodológiát tervezzük. Az angol-magyar irány ellenőrzéséhez kiválasztunk 10, a British National Corpusban gyakori és a WordNet szerint többértelmű szót, és ezekhez a Hunglish korpuszban 50-50 véletlenszerűen kiválasztott példában meghatározzuk a magyar szövegben a fordításaikat (ha vannak). A magyar-angol irányban hasonló módon hozunk létre kiértékelő halmazt, csak a WordNet helyett egy magyar-angol középszótár segítségével kiválasztva a többértelmű, a Magyar Nemzeti Szövegtárban gyakori szavakat. Ezeken a halmazokon futtatva az algoritmust meghatározzuk a humán annotációhoz képesti pontosságot és a lefedettséget.

Bibliográfia

1. Diab, M.: Relieving the data acquisition bottleneck for Word Sense Disambiguation. In Proceedings of ACL (2004)
2. Giménez, J., L. Márquez: SVMTool: A general POS tagger generator based on Support Vector Machines. In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). Lisbon, Portugal. 2004 .
3. Miháltz, M.: Angol-magyar gépi fordítórendszer támogatása jelentés-egyértelműsítő modullal. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004)
4. Och, F. J., Ney, H.: Improved Statistical Alignment Models. In Proceedings. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China, October 2000.
5. Specia, L., M. G. Volpe Nunes, M. Stevenson: Exploiting Parallel Texts to Produce a Multilingual Sense Tagged Corpus for Word Sense Disambiguation. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria (2005)
6. Tihanyi, L.: A MetaMorpho projekt 2004-ben. II. Magyar Számítógépes Nyelvészeti Konferencia, Szeged (2004)
7. Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón: Parallel corpora for medium density languages. In Proceedings of Recent Advances in Natural Language Processing (RANLP-05), Borovets, Bulgaria (2005)