Hunglish: nyílt statisztikai magyar-angol gépi nyersfordító

Halácsy Péter*, Németh László*, Kornai András**, Rung András*, Szakadát István*, Trón Viktor***

Kivonat A Budapesti Műszaki Egyetem Média Oktató és Kutató Központjának vezetésével 2004 júliusában indult Hunglish projekt[1] egy szabadon felhasználható, statisztikai gépi nyersfordítót, illetve fordítástámogató rendszert hoz létre, magyar nyelvű szövegek angolra való átültetéséhez. A gépi fordító tanításához egy kétnyelvű illesztett párhuzamos korpuszt hozunk létre. A projekt lezárása után nemcsak a kifejlesztett szoftvereket, hanem a korpuszt és az ez alapján épített/javított kétnyelvű magyar–angol szótárat is szabadon hozzáférhetővé tesszük bárki számára.

1. Bevezetés

A globális szolgáltatók szemszögéből a helyi nyelv használata elengedhetetlen termékeik és szolgáltatásaik új piacokra történő bevezetéséhez és elterjesztéséhez – különösen a termékleírások és az információ-szolgáltatások követelnek állandó fordítási munkát. A lokális piacok, a nemzeti kultúrák szemszögéből tekintve azonban más összefüggések válnak fontossá! Az információáramlás és az ebből fakadó gazdasági előnyök biztosítása érdekében elsősorban arra van szükség, hogy a helyben rendelkezésre álló információ globálisan elérhető legyen. A magyar viszonyokra vetítve tehát kulcsfontosságúnak tartjuk azt, hogy a magyar termékek, szolgáltatások és általában magyar nyelven elérhető információk minél hatékonyabban és minél szélesebb körben válhassanak ismertté. Ahhoz, hogy magyar nyelvű információ más nyelven is elérhető legyen, tömérdek fordítási munkára van szükség. Miután az angol nyelv mind a gazdasági életben, mind az információáramlásban központi szerepet kap, úgy gondoljuk, hogy a magyar nyelvből való gépi fordítás szempontjából az angol a kulcsfontosságú célnyelv. A projekt elsődleges célja így egy magyar-angol nyersfordító rendszer építése.

Nem tekintjük célunknak a magas szintű, netán irodalmi igényű gépi fordítást. Célunk olyan rendszer elkészítése, melynek kimenete egynyelvű információ-visszakereső (IV, angolul information retrieval) rendszerek bemeneteként szolgál-

* Budapesti Műszaki Egyetem Média Oktató és Kutató Központ, `{hp, nemeth, runga, szakadat}@mokk.bme.hu`

** MetaCarta Inc., `andras@kornai.com`

*** International Graduate College, Saarland University and University of Edinburgh, `v.tron@ed.ac.uk`

1 A projekt indulását az Informatikai és Hírközlési Minisztérium ITEM 2003 pályázatán elnyert összeg biztosítja.

hat. A többnyelvű IV rendszerek kutatásai, különösen az Amerikai Szabványügyi Hivatal (NIST) által évente megrendezett TREC konferencia „keresztnyelvi IV" (cross-language information retrieval) vizsgálatai világossá tették, hogy az IV rendszerek maguk sem képesek a finom árnyalatok megkülönböztetésére, és lényegében ugyanazt a teljesítményt nyújtják gyengébb minőségű (pl. beszédfelismerésből származó, 25-30%-ban hibás) szövegeken, mint a hibátlan nyelvtanú, választékosan megírt anyagokon. Ez annyit jelent, hogy nyersfordítás bizonyos használati helyzetekben ugyanolyan hasznos, mint egy igényes emberi fordítás.

A projekt végeredményeként egy működőképes nyersfordító szolgáltatás prototípusa fog elkészülni. A szoftvereket, vagyis a fordítóprogram kódját és a munka során kifejlesztett eszközkészletet, valamint a felépített adatbázisokat, a kétnyelvű illesztett korpuszt és a kétnyelvű szótárat szabadon hozzáférhetővé tesszük. A munka során kidolgozott módszereket és technológiát publikációk, illetve használati kézikönyvek formájában kiadjuk. A projekt eredményeit ezáltal bárki elérheti, felhasználhatja, illetve továbbfejlesztheti, vagy a technológiára építve önálló szolgáltatást indíthat.

Az eredményekhez való szabad hozzáférés a projekt egyik kulcsfontosságú eleme, amellyel számos célunk van. Egyrészt így látjuk biztosítva, hogy a támogatás megszűnésével a fejlesztések tovább folytatódhatnak, akár a jelen projekt résztvevőitől teljesen függetlenül is. Másrészt, minden olyan kutató- és fejlesztőcsoport munkáját támogatni kívánjuk, amely valamilyen módon a magyar nyelvtechnológiával foglalkozik. A projekt olyan alapvető fontosságú technológiai megoldásokat és adatforrásokat tesz hozzáférhetővé, melyek mind további alapkutatásokhoz, mind gyakorlati alkalmazások fejlesztéséhez elengedhetetlenek.

2. A projekt céljai

A gépi fordítás lényegében a számítógép megjelenésével egyidős vállalkozás; az első ilyen célú programot 1947-ben fejlesztették ki Weaver és munkatársai. A gépi fordítás nehézségeit összegző ALPAC jelentés[2] megállapításai sok tekintetben máig érvényesek, és emiatt nem meglepő, hogy a gépi fordítás alkalmazási köre meglehetősen korlátozott. Köztudomású, hogy a gépi fordító rendszerek kimenete kézi utószerkesztés nélkül emberi kommunikációra nem alkalmas, az automatikus fordítások gyakran kifejezetten komikus hatást keltenek. Éppen ezért jelen projekt célja sem az elsődlegesen emberi fogyasztásra szánt végleges fordítás, hanem csak a gépi vagy utószerkesztői felhasználásra szánt nyersfordítás.

Ehhez a főcélhoz vezető munkálataink során a projekt több olyan részeredményt is felmutat majd, amelyek önmagukban is jelentős nyelvtechnológiai hozzájárulásként tekinthetőek:

- magyar-angol szótár: szabad felhasználású, gyakorisági információkat is tartalmazó elektronikus magyar-angol szótár
- a statisztikai alapú szótárak előállításához, karbantartásához és javításához szükséges infrastruktúra
- párhuzamos korpusz: szabad felhasználású, mondatonként illesztett magyar-angol párhuzamos szövegkorpusz

– nyersfordító: szabad forrású rejtett Markov modell alapú nyersfordító technológia

A nyersfordítás legfontosabb eszköze a kétnyelvű szótár. Immár harminc éve vannak forgalomban olyan fordítástámogató rendszerek, melyek elsősorban a szavak szótári kikeresésének munkáját automatizálják. Projektünk *első célja egy jogtiszta, szabadon felhasználható magyar-angol szótár publikálása*, amelyet az egyéni felhasználók és a szoftverfejlesztő közösség szabadon bővíthet tovább. Ehhez komoly hozzájárulás Vonyó Attila közismert kétnyelvű gépi szótára. Amennyiben a magyarországi K+F-támogatási rendszer keretében további angol-magyar rendszerek is épülnek, és amennyiben az alkotók hajlandók ezek szóanyagát is nyílt forráskódúvá tenni (ideértjük nemcsak a kutatási, hanem a kereskedelmi célra való továbbfelhasználás korlátozás nélküli engedélyezését is), annyiban rendszerünk szótára ezekkel tovább bővíthető.

A szótári ekvivalencián alapuló (nyers)fordításnak ragozott szavak és szótári tételek problémáján kívül két alapproblémával kell megküzdenie. Az első probléma a célnyelv és tárgynyelv nyelvtani eltérései. Esetünkben ez különösen nagy problémaként jelentkezik az angol és a magyar nyelvi rendszer jelentős különbségei miatt. Amit az angol tipikusan szórendiséggel fejez ki (pl. az alany/állítmány/tárgy megkülönböztetést) azt a magyar ragokkal érzékelteti. Miután célunk elsősorban a gépi IV-t támogató nyersfordítás, a probléma nagyobb részét – elsősorban az angol szórend finomságainak algoritmizálását – mi figyelmen kívül hagyhatjuk, hiszen az információ-visszakereső rendszerek eleve a szöveg sorrendiségét elhanyagoló „szózsák” (angolul bag of words) modelleken alapulnak.

Egy másik probléma a szótári többértelműség. Például a magyar *nap* szó egyszerre jelenti az égitestet és az időegységet, amelyet az angol nyelv két külön szóval fejez ki (*sun*, illetve *day*). Miután egy magyar szónál átlagban három angol ekvivalenssel is lehet számolni, egy hétszavas magyar mondat lefordítása 3^7 (tehát több mint kétezer) variánst kínál. Erre a problémára megoldást nyújt a szövegkörnyezetben található információ, például abban a kifejezésben, hogy 'a nap és bolygói' a *nap* szó egyértelműen a *sun*, míg abban, hogy 'egy esős nap' egyértelműen a *day* fordítást kaphatja. Világos, hogy az ilyen környezettől függő valószínű fordítások megtalálásához szükséges, hogy a szövegkörnyezet által nyújtott információt pontosan meg tudjuk ragadni és azt elvszerűen integráljuk a potenciális ekvivalensek kiválasztásának folyamatában.

A nyelvi elemek egymás környezetében való megjelenésének statisztikai elméletét még a múlt század elején alkotta meg A. A. Markov. Ma ennek az elméletnek különféle változatai léteznek: a Markov-láncok (angolul Markov chains) és az ún. rejtett Markov modellek (HMM, angolul Hidden Markov Model) a nyelvtechnológia számos ágának alapvető eszközei, ezek közül külön kiemeljük a beszédfelismerést és a HMM alapú gépi fordítást[1]. A Markov modellezés nyelvtechnológiai használhatóságát a franciától a kínaiig már számos nyelvhez készült alkalmazás bizonyítja. *A projekt második célja tehát a rejtett Markov modell technológiának alkalmazása a szótári többértelműség problémájának megoldására.*

A statisztikai módszer – bár kétségkívül eredményesebb, mint a hagyományos szabályrendszereken alapuló GF – azért nem csodaszer. Legfontosabb gyengesége abban áll, hogy a rendszer megépítése kifejezetten sok adatot igényel. A statisztikai alapú gépi fordítás alapvető adatforrása a párhuzamos korpusz. A párhuzamos korpusz olyan szövegminta, amely egy adott tartalmat két nyelven jelenít meg és a nyelvi egységek (például mondatok) sorrendileg illesztve vannak egymáshoz. *A projekt harmadik célja magyar-angol párhuzamos korpusz létrehozása.*

Párhuzamos kétnyelvű szövegkorpusz készítésének bevett módja szépirodalmi szövegek és igényes műfordításaik gyűjtése és illesztése. Ez a statisztikai alapú GF módszerhez szükséges adatmennyiségnek csupán töredékét (néhány száz megabyte-ra tehető anyagot) képes nyújtani. Ennél nagyobb baj, hogy az elérhető irodalmi jellegű források (pl. a Biblia vagy Orwell 1984 című regénye) a gyakorlati (nyers)fordításhoz nem megfelelőek. Mivel a gyakorlati gépi fordítás legfontosabb célszövegei üzleti, technológiai és jogi tartalmak, elengedhetetlen, hogy a szövegkorpusz ezeknek a területeknek a jellemző szakszókincsét minél nagyobb mennyiségben tartalmazza. A cél nem lehet Mikszáth angolra fordítása, hiszen ilyesmire vállalkozni automatizált módszerrel egyszerűen sarlatánság lenne. Praktikus lehet viszont, hogy a magyarul kiírt tenderek angol nyelven is elérhetőek legyenek, ami lehetővé tenné a beszállítók körének növekedését, és a magyar vevő potenciálisan több és jobb ajánlat közül választhatna. A korpusz előállításánál így elsősorban nem a szépirodalmi szövegekre, hanem a világhálón található többnyelvű szerverekre koncentrálnánk (l. [3]). Előzetes becsléseink szerint ettől egy nagyságrenddel nagyobb, és persze gyakorlati szempontból sokkal hasznosabb, párhuzamos korpusz várható.

Hivatkozások

1. Brown, Peter F. , Della Pietra, Stephen, Della Pietra, Vincent J., Mercer, Robert L.: The Mathematic of Statistical Machine Translation: Parameter Estimation. In Computational Linguistics 19 (1994) 263–311.
2. ALPAC 1966: Languages and machines: computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council. (Publication 1416.).
3. Resnik, Philip: Mining the Web for Bilingual Text. Proceedings of the International Conference of the Association of Computational Linguistics. Maryland. (1999)