Új módszerek az emberi fordítás számítógépes támogatásában

Kis Balázs, Lengyel István

MorphoLogic Kft.
{kis,lengyel}@morphologic.hu

A piacon jelenleg élesen elkülönül a fordítástámogató eszközök és a gépi fordítóprogramok kategóriája. A szerzők megvizsgálják a fordítástámogató eszközök lehetőségeit, az elkülönülés okát, és javaslatot tesznek a két csoport közelítésére, a szinergia kiaknázására. A cikk ismerteti a szerzők elképzelését az ideális fordítástámogató csomagról: az intelligens fordítómemóriáról, amely a statisztikai hasonlóságkeresésen kívül számítógépes nyelvészeti eszközöket is felhasznál, a csoportmunkát támogató, a terminológiát rugalmasan kezelő terminológiakezelőről, a szöveg terminológiai előkészítését részben automatizáló terminuskeresőről és az egész rendszert egybefogó fordítási munkafolyamat-automatizálási rendszerről. Megvizsgálja annak előnyeit és hátrányait, hogy a fordítómemória üresen kerül a fordítóhoz, és foglalkozik a fordítómemória-adatbázisok fejlesztésének lehetőségével.

1. A számítógépes fordítástámogatás szükségessége

A fordítók munkájuk során számtalanszor kerülnek olyan helyzetbe, hogy rutinfeladatokat kell végrehajtaniuk. A gépi fordítással szemben a számítógépes fordítástámogatás nem az emberi intelligencia kiváltását, hanem annak kiegészítését, hatékonnyá tételét célozza meg. A fordítástámogatási szoftverek célja a rutinfeladatok automatizálása, ezáltal az egy fordítási egység lefordításával/célnyelven történő véglegesítésével[1] töltött átlagos idő csökkentése és a fordítás minőségének javítása. A fordítás minőségének értékelése a fordítástudományi szakirodalomban vitatott kérdés (Klaudy 2003), de abban minden szerző egyetért, hogy a terminológiai, stiláris stb. konzisztencia alapvető ismérve a jó fordításnak.

A jelenleg elérhető fordítástámogató (CAT – *Computer Assisted Translation*, számítógéppel támogatott fordítás) eszközök három kategóriába sorolhatók:

1. Fordítómemória: a fordító, illetve a fordítóközösség korábbi fordításainak újrahasznosítására;

[1] A fordítási egység célnyelven történő véglegesítése alatt a szerzők a szöveg előkészítésével kezdődő, a fordítást, lektorálást, esetleg korrektúrázást és olvasószerkesztést, nyomdai előkészítést magába foglaló folyamatot értik, amelybe beletartozik a szöveggel kapcsolatos projektmenedzsment is.

2. Terminológiakezelő rendszer: a fordítás témakörének megfelelő terminológia hatékony megkeresésére és szótárazására;
3. Munkaszervező eszköz: a csoportmunkában végzett fordítás szétosztására, összegyűjtésére, továbbítására, mérésére és egyéb szervezésére.

2. A fordítómemória

A fordítómemória működése azon a feltételezésen alapul, hogy a forrásnyelven íródott egyforma mintákat egyforma módon kell lefordítani a célnyelvre. Ez a feltételezés legtöbb esetben jogos, kivétel, amikor egy adott regiszter a forrás- vagy célnyelven nem létezik. Előnye, hogy a fordítócsoportok outputját is egységesíti, kollektív tudást hoz létre a meglévő fordítások hasznosítása révén. A terminológiakezelést támogató szoftverek azon a feltételezésen alapulnak, hogy vannak olyan kifejezések, amelyek egy adott nyelvről egy másik nyelvre egyértelműen fordíthatók az adott szövegkörnyezetben. Éppen ezért az ilyen szoftverek nem csupán a szócikket tartalmazzák, hanem annotáció révén meghatározható bennük az adott szócikk érvényességi tartománya – azon szövegek típusa, amelyekben az adott kifejezés terminusnak tekinthető. A munkaszervező eszköz a fordítók munkáját közvetlenül nem könnyíti: a fordításszervezők tapasztalatai alapján alakult ki, és az ő munkájuk minél szélesebb körű automatizálását tűzi ki célul.

A piaci forgalomban jelenleg kapható fordítástámogató programok fejlesztése piacvezérelt módon történik, amelynek lényege, hogy olyan terméket készítsenek, amely minél szélesebb réteg által használható. Az ilyen szoftverek éppen ezért nyelvfüggetlenek: így a termék potenciális vásárlói bázisa nem csak egy nyelv vagy nyelvpár fordítóira terjed ki. E megközelítés hátránya, hogy nyelvi elemzés nélkül a fordítómemória funkcióját (az aktuális forrásszöveg szegmenseivel megegyezők vagy hozzájuk hasonlók keresése a korábbi fordítások adatbázisában) csak részben tudja betölteni. A hasonlóságok keresése csak statisztikai módon történhet, amelybe bizonyos fokú intelligenciát a fuzzy logika visz, hiszen lehetővé teszi az alulspecifikált összehasonlításokat. A jelenleg kapható fordítómemóriák egyike sem lép túl a szöveg stringként történő kezelésén, a hasonlóságok keresése is string alapon történik, a morfológiai és grammatikai információ absztrakt kezelése nem jelenik meg. A legelterjedtebb nyelvek (angol, francia) esetében ez a megközelítés a nyelvi információ explicit megjelenése miatt jó hatásfokkal működik, de a ragozást használó nyelveknél nem: a *sajt* és a *hajt* között ugyanakkora a hasonlóság, mint az *írom* és az *írod* között – 1 karakter. Az előbbi nyelvek esetében a viszonylag kötött szórend miatt a szavak távolsága elég sok információt hordoz, míg a kevésbé kötött szórendet alkalmazó nyelvek esetében a szavak távolságát nem elég figyelni: például a „*Vettem egy zöld kerékpárt.*" alapján a nyelvi elemzést nem támogató fordítómemória nem képes javaslatot adni a „*Pisti vett tegnap a régi biciklije helyett egy nagy, rikítóan piros, váltós férfi kerékpárt.*" mondatra. Megfelelően nagy szótárak nélkül azonban a jól támogatott nyelvek esetében sem lehet felismerni például az idiomatikus helyzeteket, ezért szükség van az idiómák olyan szabályokként történő értelmezésére, amely felülbírálja a többi nyelvtani szabályt. Ha azonban egy idióma ragozott formában szerepel a mondatban, a hagyományos fordítómemóriák ismét csődöt mondanak.

A fenti példákból látható, hogy a szórendet szemantikai szerepben felhasználó és ragozási sorokat alkalmazó nyelvek esetében a hatékony hasonlóságkeresés csak morfológiai és bizonyos szintű grammatikai elemzés révén valósítható meg.

Felismertük azt a *tényt*, hogy az eredeti szegmenshez hasonló szegmenst már fordítottak a program segítségével. Most vagy megelégszünk annyival, hogy megjelenítjük a fordító számára a hasonló szövegre eltárolt fordítást, vagy hozzáigazítjuk azt a jelenlegi forrásszegmenshez: felruházzuk a célszegmenst azokkal a nyelvtani tulajdonságokkal, amelyek a forrásszegmensre jellemzőek voltak. Ha például a forrásszegmens felszólító módú és E/2-re vonatkozik, szükség esetén átalakítjuk a tárolt fordítást felszólító módra, E/2-re. Ha a fordítónak nem kell vesződnie az apró nyelvtani módosításokkal, időt takarítunk meg a számára.

A jelenleg kapható fordítómemóriák előnye és hátránya egyszerre az, hogy üres adatbázissal érkeznek a felhasználóhoz. Így minden fordítómemória tartalma szubjektív, a világnak azt a szegmensét tükrözi, amellyel a fordító a gyakorlata során eddig találkozott. Ennek egyaránt vannak előnyei és hátrányai.

Előny, mert:

- A fordító/megbízó fordításaiból „tanul" csupán, ezért a fordító számára a lehető legmegfelelőbb találatokat adja, a fordító stílusától nem tér el.
- Biztosítja a fordítások konzisztenciáját fordító szintjén.
- Lehetőséget ad az egyéniség kibontakozására.

Hátrány, mert:

- Sok időt vesz igénybe az adatbázis feltöltése, azaz a fordítómemória hasznossá válásának elérése.
- A fordító stílusát konzerválja – hiába tanul meg a fordító később szebben fordítani, a memóriából a régi fordításai jönnek elő.
- Rögzülnek a fordító félrefordításai, konzisztens félrefordítás lehetséges.
- Sok időbe kerül a régi fordítások forrás- és célszegmenseinek összepárosítása, az *alignment* (elrendezés) művelete.
- Nehezen hozható összhangba több, addig külön dolgozó fordító munkája és stílusa, ha mindannyian használtak korábban is saját fordítómemória-adatbázisokat.
- Nem garantálható az egy szakterületen kialakult fordítási normákhoz való alkalmazkodás.

A fenti összefoglalóból látható, hogy az előnyökhöz képest többségben vannak a hátrányok, ezért érdemes lenne a fordítómemóriákat eleve adatbázissal együtt adni.

Egyes nagy megbízók már ma is ellátják a fordítókat a fordítási megbízás kezdetén fordítómemória-adatbázissal, azonban ez még nem tekinthető gyakorlatnak, hiszen a megbízók általában nem kapják meg a befejezett fordításuk fordítómemória-adatbázisát, maguk pedig nem építenek ilyen adatbázist.

A fordítómemória-adatbázis (ami végső soron egy szinkronizált korpusz) kiadása és értékesítése általában szerzői jogi problémákba ütközik, de gondos előkészítéssel mégis lehetséges úgy összeállítani jó minőségű szövegeket, hogy azok ne legyenek ellentétesek senki érdekével, ugyanakkor reprezentálják az adott szakterületen kialakult, normaként elfogadott tudást.

3. Terminológiakezelés

A jelenleg szokásos terminológiakezelő rendszerek nagy hátránya, hogy szabványszerűen kezelik a terminológiát, vagyis a terminus technicusokat egyértelműnek tekintik. A szerzők fordítói és terminológusi tapasztalatai szerint azonban a terminológia legfőbb attribútuma nem az egyértelműség, hanem adott nyelvi tartalom témaspecifikus megformálása, illetve az általános nyelvhasználatban is előforduló szavaknak, kifejezéseknek az általános használattól eltérő jelentéssel (eltérő kontextusban, esetleg eltérő szintaxissal) való használata. A terminológia így sem nem feltétlenül nominális, és nem is egyértelmű (még egy tárgykörön belül sem): szociolingvisztikai tény, hogy adott tárgykör terminológiája minden nyelven önállóan fejlődik, sokszor a szabványosítási folyamatoktól függetlenül vagy éppen azok ellenére.

A terminológiakezelés esetében a számítógépes fordítástámogatás szempontjából a terminológia a szerzők által javasolt definíciója: *Terminológia mindaz, amelynek inkonzisztens fordítása a fordítás érthetőségét rontja.* Ez a definíció megengedi, hogy egy adott nyelven terminusnak minősülő kifejezés fordításait ne tekintsük minden esetben, minden nyelven terminológiának, azaz ne rontsuk a fordítás egészét olyan, az adott nyelven idegenül hangzó fordításokkal, amelyeket csak azért fordítunk következetesen, mert a forrásnyelvi szöveg e szempontból következetes. Megengedi két kultúra szaknyelvében vagy nyelvében a szemantikai háló eltéréseit. Az ilyen szempont figyelembe vételével megalkotott szöveg a nyelvi elemek egyértelmű leképezése helyett a kontextus leképezését, a forrásnyelvi, az adott kultúrát figyelembe vevő kontextus újbóli, célnyelvi létrehozását jelenti. Például az angolszász jogrend, az ún. common law kifejezéseinek terminológiaként történő magyarítása teljességgel értelmetlen, mivel az angolszász jogrend alapjaiban különbözik a magyartól, és az egyes kifejezések használata – főleg, ha azok a jelenleg a magyar jogban használt kifejezések új jelentéssel való felruházása, angol terminusokkal történő megfeleltetése – azt a téveszmét keltené a magyar olvasóban, hogy az angol jogrendnek sok közös pontja van a magyarral. Felhozhatnánk még azt a példát is, hogy a tengerhajózásnak a tengeri nagyhatalmak nyelveiben sokkal kiterjedtebb terminológiája van, mint a magyarnak, egész egyszerűen az ország földrajzi körülményei miatt, vagy azt, hogy a számviteli beszámolók jó fordítása (azaz olyan fordítás, amely más számviteli környezetben – országban – élő emberek számára is egyértelmű), elképzelhetetlen a számvitel ismerete nélkül.

A terminológiát nyelvpárokra bontva kezelni hatékonyabb, mint többnyelvű terminológia esetében feltételezni, hogy egy kifejezést minden nyelven terminológiaként kell kezelni. A terminológia mind szűk, mind tág értelemben kontextusfüggő: szűk értelemben a szöveghez illeszkedik, tág értelemben pedig a célnyelvi kultúrához és a szöveg fogadójához, annak ismereteihez, tudásához. Mindezt figyelembe kell venni a terminológiakezelés során, ha a profi fordítók igényeit is kielégítő fordítástámogató eszközt kívánunk fejleszteni.

A terminológia megalkotása jó esetben csoportmunka révén alakul ki, ezért fontos, hogy a számítógépes terminológiakezelő eszköz képes legyen terminológiai fórumként is működni. A jelenlegi terminológiakezelők nem képesek státusokat megkülönböztetni egy adott terminusra. Megfelelően kifinomult jogosultságkezeléssel a fordítási folyamat minden résztvevője beleszólhat, javaslatokat tehet a terminusok kialakítására – például jelöljük 1-gyel azokat a fordításokat, amelyeket a fordító javasol, 2-vel

azokat, amelyeket egy másik fordító is elfogad, 3-mal azokat, amelyeket egy nyelvi lektor, 4-gyel azokat, amelyeket egy szaklektor, 5-tel azokat, amelyeket egy szakma több képviselője is elfogad. Az Európai Unió fordítási intézményeiben ugyan megoldották a terminológiaalkotás folyamatának szabályozását, de intézményközi megállapodás nincs, ezért mind a mai napig előfordul, hogy pl. az Európai Parlament és az Európai Bizottság két külön kifejezést használ olaszul egy francia kifejezésre. Fontosnak tartjuk egy olyan terminológiakezelő kifejlesztését, amelyben nem csak a végleges terminológia tárolása oldható meg, hanem a terminológiai javaslattétel és a viták is a rendszeren belül bonyolíthatók le.

A terminológiakezelő és a fordítómemória egyesítése szintén fontos kérdés. A piacon kapható terminológiakezelők ugyan együttműködnek a fordítómemóriákkal (általában a rendszerek mindkét alkalmazást tartalmazzák), de ezek sem alkalmaznak morfológiai elemzést, így nem képesek például a ragozott szavak felismerésére, csak akkor, ha azok külön szótári bejegyzésként vannak eltárolva.

4. Munkaszervezés

A munkaszervező (projektmenedzsment) eszköz ugyan szűk értelemben nem tekinthető nyelvtechnológiai eszköznek, de mivel a fordításnak vagy a fordítás véglegesítésének teljes folyamatán keresztülnyúlik, a fordítástámogatás alapvető eleme, amely a gerincét biztosítja a teljes folyamatnak. A jó munkaszervező eszköz megfelelően skálázható és bővíthető, támogatja az egyéni munkát is, de a csoportmunka előkészítési és ellenőrzési funkciói is bele vannak építve.

Csoportos fordításra általában a rövid határidők miatt van szükség. Ilyen esetben alapvető követelmény, hogy a fordításon ne lehessen észrevenni, hogy az nem egy fordító munkája. Még a jó fordítók között sem általános, hogy jól dolgoznak csoportban is, mivel a stílusuk, szóhasználatuk, a világ szegmenseiről alkotott képük különbözik. A csoportos fordítás támogatása nem merül ki a terminológiakezelésben, mint ahogyan azt a piacon kapható CAT-eszközök feltételezik. A fordítás előkészítése során rendkívül fontos a terminológia *felismerése:* annak meghatározása, hogy milyen szavakat, kifejezéseket kell terminológiának tekinteni. Ez jelenleg úgy történik, hogy egy vezető fordító vagy terminológus a fordítás előkészítése során végigolvassa az eredeti szöveget, kijelöli annak terminusait, és meghatározza a célnyelvi megfelelőit. Ez a művelet azonban időigényes, rövid szövegek esetében jól működik, de a legjobb terminológusok kapacitása sem haladja meg napi 100 oldal előkészítését. Szükség van egy olyan eszközre, amely a szöveget „átolvassa", és felismeri a szövegben található terminusokat.

A terminusok felismerése azonban nem egyszerű feladat, a közhiedelemmel ellentétben nem elegendő csak az adott szöveg szavainak gyakorisága. A terminuskeresés két módszere a statisztikai és a determinisztikus-heurisztikus módszer. A determinisztikus-heurisztikus módszerrel azokat a kifejezéseket keressük, amelyek környezetében nagy valószínűséggel terminológia szerepel, például „.... alatt azt értjük, hogy ....", „definíció: ....", „.... nevet adták neki" stb. A statisztikai módszer lényege a gyakorisági alapon történő keresés, de a kritikus gyakoriság meghatározása azért nehéz feladat, mert ez az érték szakterületenként és célközönségenként változó. Jelenleg olyan

eszközt fejlesztünk, amely minden szöveg esetében – lehetőség szerint – négy korpusszal dolgozik: egy forrásnyelvi általános, egy forrásnyelvi szaknyelvi, egy célnyelvi általános és egy célnyelvi szaknyelvi korpusszal, és ha létezik ilyen, egy kétnyelvű általános és szaknyelvi szótárral. Alapfeltevésünk, hogy a fordító számára az a terminológiai szójegyzék a legnagyobb segítség, amely olyan kifejezésekre ad egyértelmű fordítást, amilyen nem szerepel a szótárakban vagy amilyen több értelemben szerepel a szótárakban, de az adott szövegben csak egy értelemben alkalmazható. Az algoritmus alapja, hogy kiszámoljuk, hogy a potenciális terminus milyen gyakorisággal szerepel a forrásnyelvi általános korpuszban és a szakkorpuszban, kiszámítjuk ugyanezt az értéket a szótári bejegyzések lehetséges fordításai alapján a célnyelvre is, és ha az egyik fordítás esetében ez az érték kiugró, azt a kifejezést terminusnak tekintjük. A rendszer azonban csak jó korpusszal és szótárakkal működőképes, amelyek építése erőforrás-igényes munka, ezért a szakterületekre jellemző „terminusküszöbértékek" kiszámítása csak hálózati szolgáltatásként képzelhető el. A küszöbérték utána a felhasználó által finomítható. Az ideálisnál alacsonyabb küszöbérték esetén olyan kifejezéseket is terminusnak minősít az eszköz, amelyek következetes fordítására esetleg nincs feltétlen szükség, magasabb küszöbérték esetén pedig előfordulhat, hogy nem talál meg a rendszer olyan kifejezéseket, amelyek a terminológia részét kellene, hogy képezzék. A terminológiagyűjtés végső fázisában a statisztikai és a determinisztikus-heurisztikus módszerrel egymás találatai verifikálhatók. Az ilyen eszköz megkönnyíti a terminológus dolgát, hiszen viszonylag jó terminológiai konzisztencia garantálható rövid időn belül. A fordítási minőség-javító funkciója legszembetűnőbb a rendkívül hosszú szövegek nagyon rövid idő alatt, sok fordítóval történő fordítása esetén.

A munkaszervező keretrendszerbe egyéb eszközök is beépülhetnek, amilyenek például a kollokációellenőrzés, a terminológiai konzisztencia ellenőrzése, a hivatkozások eredethű fordításának ellenőrzése stb.

Irodalomjegyzék

AUSTERMÜHL, Frank (2001): Electronic Tools for Translators. Manchester: St. Jerome.

CASTELLVÍ, M. Teresa Cabré – BAGOT, Rosa Estopà – PALATRESI, Jordi Vivaldi: Automatic Term Detection: A Review of Current Systems. In: *Bourigault, Didier – Jacquemin, Christian – L'Homme, Marie-Claude (eds.): Recent Advances in Computational Terminology.* John Benjamins, Amsterdam-Philadelphia, 2001. pp. 53–88.

ESSELINK, Bert (2001): A Practical Guide to Localization, Amsterdam & Philadelphia: John Benjamins. 488 pp.

JACQUEMIN, Christian (2001): Spotting and Discovering Terms through Natural Language Processing. The MIT Press, Cambridge, MA, USA–London.

KIS, Ádám–KIS, Balázs (2003): A Prescriptive Corpus-based Technical Dictionary. Development of a multi-purpose technical dictionary. In: *Proceedings of COMPLEX 2003*, Budapest.

KLAUDY Kinga (2003): Fordítástechnikai minimum (kézirat). Budapest-Miskolc.

PRÓSZÉKY Gábor (2002): Nyelvi technológiák és gépi fordítás. *In: Emberi és gép nyelv, beszéd és hallás* (megjelenés alatt)

PRÓSZÉKY Gábor–KIS Balázs (1999): Számítógéppel – emberi nyelven. SZAK Kiadó, Bicske. 344 pp.