Nyelvi tudásra épülő fordítómemória

Hodász Gábor[1], Grőbler Tamás[2]

[1]Pázmány Péter Katolikus Egyetem
Információs Technológiai Kar
Budapest
hodasz@morphologic.hu
[2]MorphoLogic Kft.
Budapest
grobler@morphologic.hu

Kivonat. A cikkben bemutatásra kerülő MetaMorpho TM rendszer olyan fordítástámogató eszköz, amely a hagyományos fordítómemória-funkciókat nyelvi intelligenciával kiegészítve a jelenlegi rendszereknél többször ajánl fordítást, és azok jobban közelítik a kívánt minőségű fordítást. A fordítás egységei a mondatnál kisebb szegmensek (főnévi szerkezetek és az ezeket tartalmazó mondatvázak), amelyeket a forrás- és célnyelvi elemzők állítanak elő. Az adott bemeneti mondathoz hasonló szegmenseket „nyelvi intelligencián" alapuló távolság segítségével keressük, és a megszülető új fordításokat mint szabályokat tároljuk, amelyek a gépi fordítás minőségét folyamatosan javítják.

Bevezetés

Napjainkban a leginkább elterjedt fordítástámogató eszközök a fordítómemóriák (translation memory, TM), amelyek a fordítási munkák során keletkező párhuzamos szegmensek eltárolásával nyújtanak segítséget a fordítónak. A hagyományos fordítómemóriában a szegmensek a teljes mondatok, amelyek nyelvi tudás nélkül kerülnek tárolásra, és közülük karakter-alapú távolság alapján választ a rendszer az aktuális mondathoz hasonlót, majd ez alapján ajánl fordítást a felhasználónak. Bár a legtöbb rendszer felismeri és kezeli a nem fordítandó és a szabadon behelyettesíthető elemeket (számokat, dátumokat stb.), valamint rendelkezik terminológia-kezeléssel, azonban nem képesek kezelni a pusztán morfológiai különbségeket, vagy a szintaktikailag hasonló, de karakteresen különböző mondatokat.

A szabály alapú gépi fordító rendszerek (rule based machine translation, RBMT) a fordítómemóriákkal szemben a fordítási folyamat emberi beavatkozás nélküli automatizálását célozzák meg. Nyelvi tudással rendelkeznek, elemző algoritmusok és fordítási szabályok segítségével végzik a fordítást, általában kötött nyelvpárokra. A nyelvi kétértelműségek és az elemző algoritmusok tökéletlensége mellett a szabálybázis korlátozott bővíthetősége is gátat szab ezen rendszerek pontosságának.

Nagao a '80-as években új megközelítést javasolt a szabály alapú fordítás hibáinak megoldására: a példa alapú fordítást (Example Based Machine Translation, EBMT) [1]. Az ötlet alapja az a pszicho-lingvisztikai megfigyelés volt, hogy a fordítási

folyamat során az emberi fordító is használja a szónal nagyobb, de mondatnál kisebb szerkezeteket. A megfigyelések szerint minél tapasztaltabb a fordító, annál nagyobb egységeket használ [2]. Az EBMT rendszerekben a fordítás alapjai a szónál nagyobb, de mondatnál kisebb nyelvtani egységek, amelyeket párhuzamos korpuszból állít elő a rendszer [3]. A fordítási folyamat során az egyes visszakeresett mintákból szabályok állítják össze a célnyelvi mondatot. A kutatások célja, hogy megtalálják az optimális utat a két szélsőnek tekinthető megközelítés között. Több szerző kimutatta, hogy a fordítás minősége függ a rendszer nyelvi tudásának mennyiségétől [4]. Így az alkalmazott nyelvtechnológiai algoritmusok pontosságának és hatékonyságának növelése szintén tárgya a jelen kutatásoknak.

A cikkben bemutatásra kerülő MetaMorpho TM rendszer szintén a két alapvető iránynak az ötvözését tűzi ki célul, így az EBMT rendszerek közé sorolható [5].

A MetaMorpho TM működésének vázlata

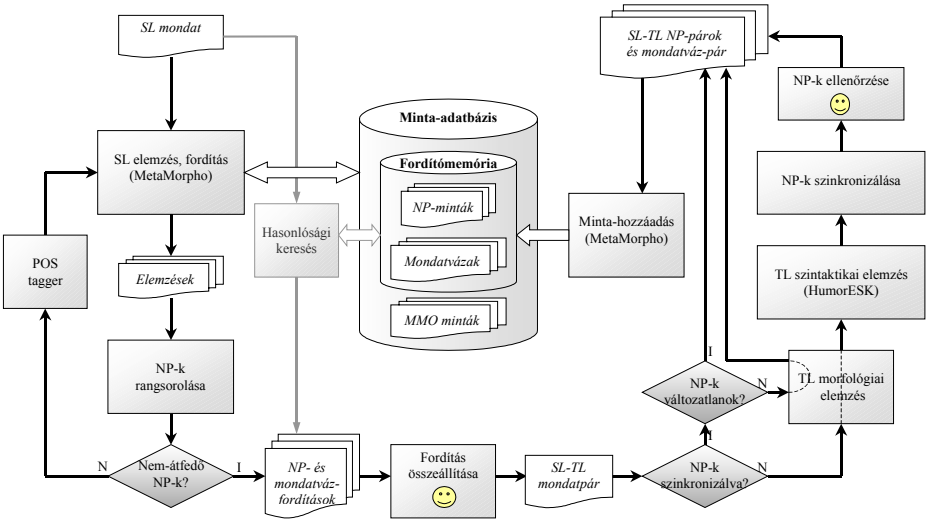
A MetaMorpho TM alapvetően fordítómemóriaként működik, azaz fejlett eszközökkel támogatja az emberi fordítót, valamint lehetőséget ad az adatbázis bővítésére. A fordítói munka során a fordított mondatok feldolgozásával bővül a szabálybázis, és emellett lehetőség van minták párhuzamos korpuszból való automatikus felvételére is.

A fordítás előállítása

A fordítási munka során az emberi fordító felügyeli a fordítás folyamatát: módosítja vagy elfogadja a rendszer által felkínált fordításokat.

A folyamat lépései a következők (1. ábra):

- A beérkező forrásnyelvi mondatot a morfoszintaktikai elemző lemmákra bontja.
- A főnévi csoportok (NP-k) és a többi lemmából álló mondatváz alapján az „intelligens” kereső hasonló mondatot, illetve hasonló NP-ket keres az adatbázisban.
- A találatokat megfelelően szűrve és rangsorolva előáll a célnyelvi szegmensfordítás-jelöltek listája.
- A jelöltekből, illetve az opcionálisan gépi fordítással előállított célnyelvi szegmensekből összeáll az eredeti mondat felajánlott fordítása.
- A felajánlott fordítás(oka)t a felhasználó elfogadhatja vagy módosíthatja.
- A módosított célnyelvi szegmensek elemzése és forrásnyelvi párjukkal való eltárolásuk révén új szabályokkal bővül a fordítómemória.



1. ábra:A MetaMorpho TM működése

A fordítómemória bővítése

A fordítómemória bővítése során új szabályokat adhatunk hozzá a szabálybázishoz.

A fordítás folyamán az emberi fordító által javított célnyelvi mondatok elemzése és a szegmens forrásnyelvi párjával való összerendelése révén olyan párhuzamos korpuszelem jön létre, amelyből előállítható a megfelelő formában a fordítási szabály. A rendszer jelenlegi verziójában a mondatpárokból kiemeljük a főnévi csoportokat, amelyekből önálló szabály-minta keletkezik. Ugyancsak szabály-minta keletkezik a mondat fennmaradó részéből, az ún. mondatvázból, amelyben az NP-k helyét üres hely jelzi. A későbbi fordítások során ezek a helyek más NP-kkel is kitöltődhetnek, amennyiben azok kielégítik a megfelelő feltételeket. Azért, hogy az eredeti mondat egyértelműen visszaállítható legyen, a mondatvázból születő szabályban eltárolásra kerül az eredeti NP-k azonosítója is.

Párhuzamos korpuszból való automatikus szabályfelvétel esetén a fentebb vázolt folyamat emberi beavatkozás nélkül megy végbe. A külön modulként kifejlesztendő mondatszinkronizáló (aligner) által előállított párhuzamos mondatokból a nyelvi elemző előállítja az NP-ket és a mondatvázat. Az így létrejövő mintákat mint szabályokat hozzáadjuk a szabálybázishoz.

A szabály alapú fordító néhány jellemzője

A MetaMorpho TM intelligens fordítómemória a MorphoLogic MetaMorpho nevű szabály-alapú fordítórendszerére [6] támaszkodik. A MetaMorpho rendszerben a forrásnyelvi mondat elemzésének egyes lépéseivel egy időben párhuzamosan létrejön a megfelelő célnyelvi struktúra is. Így egy MetaMorpho szabály minden esteben egy forrás (angol) és egy célnyelvi (magyar) részből áll. Példa egy szabályra:

```
*NX=approach+to:12
EN.NX[ct=COUNT] = N(lex="approach") + PPOBJ(lex="to")
HU.NX = PPOBJ[case=GEN] + N[lex="megközelítés"]

;example: This is a really nice approach to religion.
```

A MetaMorpho szabályrendszerének másik jellemzője, hogy a szabálybázis homogén: nem különböztetjük meg a szótár-szerű és a szintaxis-szerű szabályokat.

A fenti két tulajdonság lehetővé teszi, hogy a fordítómemóriába kerülő szabályok bármilyen szintű nyelvtani struktúrát írjanak le, legyen az egyetlen főnév és fordítása, vagy egy mondatváz, amelyben üres helyek jelzik a főnevek helyét és a vonatkozó megkötéseket.

A szabályok az elemzés-fordítás folyamán egyszerű unifikációs nyelvtan szerint működnek. Az egyes szabályokban a különböző jegyek (megkötések) határozzák meg a szabály specifikusságát. A szabályban a nem kitöltött megkötések a konkrét mondat elemzése során kerülnek kitöltésre. Így a fordítómemória működése folyamán egy korábban eltárolt minta akkor is releváns lehet az aktuális mondat fordításában, ha előzőleg más morfológiai jegyekkel szerepelt. Ehhez az szükséges, hogy a szabályok eltárolásakor meghatározzuk a kellő megszorításokat, a többi jegyet azonban kitöltetlenül hagyjuk. Így a fordítómemória által megtalált korábbi fordítás csak abban az esetben lesz jelölt, ha kielégíti a szükséges megszorításokat. A nem szükséges megszorítások (pl. szám, személy, idő stb.) pedig a célnyelvi mondat generálása során az aktuális forrásnyelvi megfelelőik szerint kerülnek kitöltésre. Ez a megközelítés lehetővé teszi, hogy a hagyományos fordítómemóriával szemben például az angol ’go’ igének különböző idejű alakjait (pl. ’went’, ’has gone’ stb.) annak ellenére megtalálja a rendszer a minták között, hogy közöttük a karakter-alapú távolság igen nagy.

A felhasználónak felajánlott fordítások a memória által visszaadott hasonló minták (főnévi szerkezetek és mondatvázak) összeillesztésével és utófeldolgozásával állnak elő. Amennyiben a fordítandó mondat nem minden eleme található meg a memóriában, úgy a felhasználó opcionálisan kérheti a fenmaradó részek gépi fordítását.

Minták és szabályok

A MetaMorpho TM rendszer átmenetet képez a minta (memória) alapú és a szabály alapú fordítástámogató illetve fordító eszközök között. A fentebb leírt tulajdonságai miatt egyesíti a minta és a szabály fogalmát. A fordítómemória egy eleme egyaránt

párhuzamos korpuszelem és fordítási szabály, és a felépített memória egyaránt tekinthető párhuzamos annotált korpusznak, valamint szabálybázisnak. Egy feltöltött memória alkalmas lehet korpusznyelvészeti célokra, például terminológia-keresésre, glosszáriumépítésre stb. Ugyanakkor a szabályok a szabályalapú fordítás minőségét javíthatják.

Hasonlósági keresés

A korábban eltárolt minták közötti keresés fontos lépése a fordítómemória működésének. A hagyományos, ma kereskedelmi forgalomban kapható fordító-memóriák az egyes minták közötti karakter-alapú távolság alapján keresnek az adatbázisban. A MetaMorpho TM jövőben kifejlesztendő, nyelvi intelligenciával ellátott kereső algoritmusa egyaránt képes felismerni az azonos lexikai elem más morfológiai alakjait, valamint a kissé különböző szerkezetű szintaktikai egységeket is.

A dinamikus programozáson alapuló algoritmus három szinten vizsgálja a szegmensek hasonlóságát:

1. felszíni alak
2. morfológiai jegyek
3. szófaj

Az egyes szinteken a különböző típusú különbségek különböző büntető pontokat eredményeznek. Az összehasonlítást „alulról felfelé" végezzük, azaz amennyiben a *n*-dik szinten egyezést találunk, tovább vizsgáljuk az *(n-1)*-dik szinten. Amennyiben a két szegmens nem azonos számú lemmából áll, úgy a büntetés függ attól, hogy a hiányzó vagy más értékű lemma az adott szerkezet feje vagy más eleme.

Várakozásaink szerint ez a hasonlósági keresés és az ennek megfelelően indexelt adatbázis nagyobb számban fog megfelelő fordítás-javaslatot adni, és a javaslatok jobban közelítik a kívánt fordítást.

Implementáció

A fentiekben felvázolt MetaMorpho TM intelligens fordítómemóriát első lépésben angol-magyar nyelvpárra valósítjuk meg. A rendszer moduljait C++ nyelven implementáljuk. Az egyes modulok megvalósítják a különböző elemző illetve generáló funkciókat: mondatszegmentálás, szószegmentálás, morfológiai elemzés stb. A modulok közös adatstruktúrát dolgoznak fel, és bármely feldolgozási lépés eredménye XML formátumban is megjeleníthető.

A szintaktikai szabályokat a [7]-ben leírt modell szerint valósítjuk meg.

A fordítómemória funkciót relációs adatbázis valósítja meg. Az egyes táblákban tároljuk a szabályokat valamint a közöttük lévő leszármazási kapcsolatokat. A hasonlósági keresés és index jelenleg még implementálásra vár.

Összefoglalás és további munkák

Jelen munkánkban bemutattuk a MetaMorpho TM intelligens fordítómemóriát, mely az irodalomban leírt EBMT elv szerint a szónál nagyobb, a mondatnál kisebb nyelvi egységek feldolgozásával, tárolásával és összeillesztésével nyújt hatékony segítséget az emberi fordító számára. A morfológiailag elemzett főnévi szerkezetek, valamint az ezeket üres helyekként tartalmazó mondatvázak eltárolásával és nyelvi távolságon alapuló hasonlósági keresésével a hagyományos fordítómemóriáknál várhatóan több korábbi fordítást tud felajánlani a rendszer, valamint az egyes fordítási egységek összeillesztésével a kapott fordítás jobban közelíti a kívántat.

Az elvégzendő további munkák között szerepel a mintákat tároló relációs adatbázis továbbfejlesztése, a hasonlósági keresés implementálása, valamint a célnyelvi oldalon az egyes fordítási egységek összeállítása helyes mondatokká. Szükséges továbbá a felhasználó által megadott célnyelvi mondat nyelvi elemzésének megvalósítása, amely lehetővé teszi a memória bővítését, előállítva a párhuzamos elemzett mintákat (szabályokat). A nyelvi motort a fordítómemória-rendszerekhez hasonló integrált fordítástámogató szoftverré kívánjuk fejleszteni.

Irodalom

1. Nagao, M, ’A framework of a mechanical translation between Japanese and English by analogy principle’, In A. Elithorn and R. Banerji (eds.) (1984), Artificial and human intelligence, 173-180. Amsterdam: North-Holland.
2. Gerloff, P. ’Identifying the Unit of Analysis in Translation’, in Færch & Kasper (eds.) (1987) *Introspection in Second Language Research*, Clevedon: Multilingual Matters, pp. 135-158.
3. Turcato, D. & F. Popowich, ‘What is Example-Based MT?’ *Proceedings of the Workshop on Example-Based Machine Translation.* (2001) http://www.eamt.org/summitVIII/workshop-papers.html
4. McTait, K., ‘Linguistic Knowledge and Complexity in an EBMT System Based on Translation Patterns’. (2001) *Proceedings of the Workshop on Example-Based Machine Translation.* http://www.eamt.org/summit VIII/workshop-papers.html
5. Prószéky, G. and L. Tihanyi, ‘MetaMorpho: A Pattern-Based Machine Translation Project’. (2002) *Translating and the Computer 24,* ASLIB, London.
6. Hegedűs, B. ’Természetes nyelvű információk számítógépes feldolgozása’, (2003) Diplomaterv, Budapesti Műszaki és Gazdaságtudományi Egyetem, Számításelméleti és Információtechnológiai Tanszék, Budapest.
7. Prószéky ‘Syntax As Meta-morphology’, (1996) *Proceedings of COLING-96,* Vol.2, 1123–1126. Copenhagen, Denmark.