

A MetaMorpho projekt 2004-ben

Tihanyi László

MorphoLogic
1126 Budapest Orbánhegyi út 5.
tihanyi@morphologic.hu

Kivonat: A MetaMorpho fejlesztések idén, a projekt ötödik évében elérték a piacképes szintet. A programnak, mely a tervek szerint a kezdő, a haladó és hivatásos fordítói igényeket különböző termékekkel kívánja kiszolgálni, először a nyelvet nem vagy csak kezdőszinten beszélőknek ajánlott megértéstámogató változata, a MoBiCAT került a piacra. Ez a cikk folytatása az előző MSZNY konferencián elhangzott beszámolónak [1], és összefoglalja a 2004-es év legfontosabb fejlesztéseit.

1 Bevezetés

A idén június elején piacra került *MoBiCAT*, egy olyan internetes angol-magyar megértéstámogató fordítóprogram, amely a fordítást egy központi szerverrel végezteti. A MoBiMouse-hoz hasonló felülettel rendelkező programot Microsoft Word, Internet Explorer, Outlook és Adobe Acrobat alkalmazásokban használhatjuk. A fordítást az egérrel való rámutatással kezdeményezhetjük, és az eredeti mondat a fordításával együtt egy buborékban jelenik meg. A program ingyenesen letölthető és kipróbálható. Továbbiak a programról a www.mobicat.hu oldalról tudhatók meg.

Ugyancsak nyár elejére készült el a *morphoWAP*, a MetaMorpho mobiltelefonokról elérhető WAP-os kliense. A fordítás mobilszolgáltatatóktól független, és jelenleg mindenki díjmentesen használhatja. Az eszköz egy-egy szó esetén szótárprogramként, mondatokra fordítóprogramként működik. A *morphoWAP* beépített angol helyesírás-ellenőrző programot is tartalmaz. A szolgáltatás a <http://www.metamorpho.hu/wap> címen érhető el wapos telefonokról.

Az idei fejlesztések harmadik tervezett alkalmazása az egész weboldalakat fordító *morphoWEB* fordítóprogram. Ez a program az angol oldalak magyar változatát az eredetivel azonos megjelenésben állítja elő az. Az oldalon lévő hivatkozásokat feloldja, tehát azokon tovább haladva ezeket is magyarul olvashatjuk.

A fokozatos minőségi javulás mögött számos nyelvészeti és szoftverfejlesztési újítás húzódik meg, ezek közül a fontosabbakat a következő fejezetekben mutatjuk be.

2 Nyelvészeti fejlesztések

Idén nyelvi fejlesztésünk eljutott arra a szintre, amikor a fő problémát már nem a szerkezetek leírása és felismerése, hanem a megfelelő minták érvényre juttatása kezdte okozni. A projektben kezdetektől fogva alapvető elv volt, hogy az elemzések számát a lehetséges legkisebb értéken tartjuk. Teljes lefedettség esetén minden eredményt, részleges elemzéskor pedig a részelemzések sorozatát jelenítjük meg. [2]

A többértelműség mindkét esetben problémát jelent, hiszen teljes elemzéskor a sok mondat lesz nehezen áttekinthető, részleges elemzéskor pedig minthogy önkényesen választjuk ki a többértelmű részelemzések egyikét, és fűzzük össze a teljes fordítást, a helyes elemzés gyakran meg sem jelenik.

A megfelelő minták érvényre juttatásának érdekében az eddig homogén lexikont és grammatikát szintekre tagoltuk. A szabályok leírásmódja és működési elve nem változott, de a megfelelő szintek olyan szűrőkként működtethetők, amelyek már csak az általuk megengedett eredményeket engedik tovább. Az új eljárással nemcsak a nyelvi minőség javulását, hanem a felesleges elemzések elkerülésével jelentős sebességbeli növekedést is el tudtunk érni. A fejlesztéséhez természetesen a szintaktikai elemző programot is megfelelően át kellett alakítani.

Egy másik jelentős fejlesztés az entitások, különösen a nagybetűs tulajdonnevek (néventitások) kezelésének megoldása volt.

További nyelvészeti fejlesztések voltak a koordináció kezelése, a globális tulajdonságok kezelése (cím, stb.), a címnyelvtan, az alárendelő mellékmondatok valamint a kötőmód helyes kezelése, és természetesen a folyamatos szótárbővítés.

3 Programfejlesztések

A programfejlesztés idejének legfőbb eredménye, hogy létrejött egy interneten át egy egyszerre több felhasználót kiszolgálni képes, stabil szerver program. A szerver jogosultságkezelő védelmi modullal van ellátva, amely a fizetős használat mellett ingyenes kipróbálásra is lehetőséget ad.

Készült egy mintagyártó felhasználói program (RuleBuilder), amely képes arra, hogy forrásszövegekből olyan mintákat állítsunk elő, amelyet a rendszer a sajátjaival egyenértékű módon tud kezelni. Az eltárolt minták az SQL adatbázisban a nyelvi adatok mellett egy domainkezelő segítségével szakterületi és egyéb jelentésmegkülönböztető tulajdonságokkal is elláthatók. Az adatbázis közösen tárolja a felhasználói mintákat, de alapértelmezésben mindenki csak a saját mintáját látja. A rendszer felügyelőjének azonban lehetősége van, hogy bizonyos mintákat az egész közösséggel megosszon. Egyelőre házon belül használjuk lexikális bővítésekhez, de célunk, hogy mielőbb nyilvánossá tegyük, és a felhasználók számára lehetővé váljon a rendszer on-line bővítése.

Elkészült egy cíamazonosító modul [3], amire nagy volt szükség mert a MoBiCAT-et elsősorban internetes böngészéshez ajánljuk, és hírportálokra a felhasználó először címeikkel találkozik. Az angol címek nyelvtana mint tudjuk, nagyon eltér a mondatszintaxisban leírtaktól.

Integráltuk a Brill-tagget. Statisztikai adatbázisait átvettük, de a szófaji egyértelműsítésre szolgáló eszköz kiértékelő függvényeit újraírtuk.

Kiemelendő még a jelentés-egyértelműsítő modul integrálása a programba [4]. A WEKA nevű Java alapú WSD program kiértékelő modulját C++ nyelven újraírtuk, és megkezdjük az egyértelműsítéshez szükséges minikorpuszok összeállítását. Ezekben a kérdéses szavak jelentése a WordNet rendszernek megfelelően meg van különböztetve, és a program a szó környezetében álló szavak gyakoriságából valószínűsíti az aktuális jelentést.

Minden modulunk (a MetaMorpho szerver, a MoBiCAT, a morphoWord, a morphoWeb és a RuleBuilder kliensek is) saját telepítőprogramot kapott, ami nemcsak a frissítéseket könnyíti meg, hanem a verziószám megjelenésével pontosabb felhasználói visszajelzésekre ad lehetőséget.

4 Webfejlesztések

Elkészültek weboldalaink: www.metamorpho.hu, és <http://www.mobicat.hu>, amelyek tájékoztatást nyújtanak a projektről általában és naprakész információval szolgálnak (letöltési oldal, hírek, GYIK). Ezekre az oldalakra érkeznek a felhasználói visszajelzések és itt van lehetőség a program minőségének véleményezésére is. Elkészültek az internetes értékesítéshez szükséges webes felületek, és a MoBiCAT visszajelzések kezelésére szolgáló internet oldal.

5 Tervek 2005-re

A jövő év nagy vállalkozása a magyarról angolra fordítás lesz. A 2005 elején induló NKFT által támogatott projekt a MTA Nyelvtudományi Intézetével és a Szegedi Tudományegyetem Informatikai Tanszékcsoportjával közös együttműködésben fog megvalósulni.

Jövő évi terveink között szerepel a hagyományos, szövegszerkesztőkben is használható fordítóprogram piacra bocsátása. Ugyancsak tovább folynak a MetaMorpho alapú fordítómémória fejlesztések, a szövegpárhuzamosítása [5] és a hasonlósági keresésre [6] irányuló kutatások.

Referenciák

1. Tihanyi László: A MetaMorpho projekt története. *I. MSzNy, Szeged* (2003)
2. Gröbler Tamás: Egyértelműsítés a MetaMorpho rendszerben. *II. MSzNy, Szeged* (2004)
3. Pohl Gábor, Ugray Gábor: Angol címek felismerése. *II. MSzNy, Szeged* (2004)
4. Miháltz: Gépi fordítórendszer támogatása jelentés-egyértelműsítő rendszerrel. *II. MSzNy, Szeged* (2004)
5. Pohl Gábor: Iteratív bekezdés- és mondatszinkronizáció. *II. MSzNy, Szeged* (2004)
6. Hodász: Nyelvi hasonlóságon alapuló keresés fordítómémóriában. *II. MSzNy, Szeged* (2004)