Hogyan működnek a fordítóprogramok?

Megértőbb gépek

Pár éve még csak álom volt, ma már valóság: az angol nyelvű weboldalt tized másodpercek alatt magyarra fordítja egy hazai alkalmazás. Pár hét, és magyar szövegeinket is angolra fordíthatjuk géppel, de egyelőre várni kell a magyar diktálóprogramra, és a gépi szinkrontolmácsolás is csak a tudományos-fantasztikus filmekben létezik. Prószéky Gábor egyetemi tanárral a számítógépes nyelvészet világába kalandoztunk.

LUKÁCS

Négy ember összetanakodott – így kezdődik az 1991-ben alapított MorphoLogic krónikája. A cél a számítógépes nyelvészet kutatása, fejlesztése és alkalmazása volt. A vállalkozás azóta már a nemzetközi piacon is ismert, termékeit megtaláljuk a szövegszerkesztőkben, a keresőprogramokban és más számítógépes eszközökben is. A *Prószéky Gábor* vezetésével dolgozó cégnek jelenleg több mint húsz állandó munkatársa van, hazai tudományos kutatóközpontokkal dolgoznak együtt, és részt vesznek az Európai Unió kutatási projektjeiben. A „fordító egér" szoftver kapcsán (a MoBiMouse futtatásával elég a kurzorral a fordítani kívánt szót kijelölni, fölötte egy buborékban azonnal megjelenik az eredmény) 1999-ben Európai Információtechnológiai Díjat kaptak. Ennek továbbfejlesztése volt a teljes angol mondatot magyarító MoBiCAT, a www.webforditas.hu honlapon pedig ingyen használható weblap- és szövegfordító, keresőprogram és szótár működik.

Magyarországon egyedülálló módon, a Pázmány Péter Katolikus Egyetem (PPKE) Információs Technológiai Karán tanítanak nyelvtechnológiát. Prószéky Gábor 2006 óta a kar dékánhelyettese, a PPKE doktori iskolájában pedig egyre nő az ezen a szakirányon kutató doktoranduszok száma. Az itt tanított információtechnológia különlegessége, hogy a számítógép mellett az élő szervezet sajátosságait is vizsgálják, hiszen a gép és az ember által használt „technológia" között nagy a hasonlóság: az idegrendszert felfoghatjuk komplex hálózatként, az agyat memória- és irányítóközpontként, de még az immunrendszer működése és a genetika tudománya sem haszontalan a jövendő mérnökei számára.

Azért is fontos tudományág manapság a számítógépes nyelvészet, mert a huszonegyedik században szöveg szinte csak számítógépen születik: azon írjuk, szerkesztjük, továbbítjuk (publikáljuk) és fordítjuk a közlendőnket, és géppel keressük meg az ehhez szükséges információt is. Néha olyan nyelven, amelyet nem ismerünk – ilyenkor merül fel a gyors, hatékony fordítás szükségessége. De nem mindegy, hogyan: hosszú ideig – főleg a hidegháború idején – az amerikaiak arra esküdtek, hogy a gépi fordítását a jövő, és jó sok pénzt költöttek el a kidolgozására. A kívánt eredmények késtek, ezért leálltak a fejlesztéssel. Aztán a formálódó Európai Unióban alakult ki újra az igény a nyolcvanas évek elején, a soknyelvűség szorításában. Itt sem volt áttörés, ezért a kezdeti lendület alábbhagyott egészen addig, amíg meg nem jelentek a matematikusok a maguk statisztikai módszereivel. Ezek lényege, hogy meglévő szövegekből és fordításaikból végeznek statisztikai számításokat, és ezeket alkalmazzák az új, lefordítandó szövegekre – magyarul a gép megkeresi az adatbázisból az egyszer már lefordított mondatot. Végtelen számú minta esetén tökéletesen működne ez a módszer, hiszen a gép mindent megtanulna, de erre természetesen nincs lehetőség – a tapasztalat szerint a szótáralapú fordítók még mindig jobbak, de drágábbak is a statisztikai módszert használó alkalmazásoknál. Az ötlet kézenfekvő: össze kell fésülni a kettőt. EuroMatrix néven nagyszabású európai projektbe fogtak ez ügyben, a kutatásban részt vesz a MorphoLogic is. A hibrid rendszer kialakítása 2006 szeptemberében kezdődött, és két és fél évet szánnak rá. A folyamat végén elvileg minden európai nyelvről minden európai nyelvre fordítana az új szoftver.

A MorphoLogic igazgatója szerint az első tíz évben annyiféle nyelvi szoftvermodult hoztak létre (elemzőket, szótárakat és az ezeket kezelő technológiát), hogy az új évezred elején – *Tihanyi László* vezetésével – bátran nekifoghattak a gépi fordításnak. Elsősorban az angolra koncentrálnak, hiszen a weben fellelhető szövegek kétharmadát ezen a nyelven írják. A cél nem a tökéletes fordítás elkészítése, hanem inkább a megértés támogatása. A feladat nem egyszerű, mert a tükörfordítás gyakran tévútra visz, ezért a gépnek lehetőleg az adott szöveg egészét, kontextusát is meg kell értenie. Külön bonyolítja a dolgokat a szövegek, hírek címeinek fordítása, hiszen ezeknek más a grammatikájuk – a gépnek fel kell ismernie, hogy címmel van dolga, hogy az erre írt speciális nyelvtani programot indíthassa el.

FOTÓ: REUTERS – PETER MUELLER

UNESCO-konferencia Hamburgban. A legnagyobb hibaforrás még mindig az ember

Nehezíti a dolgot, hogy a világhálón fellelhető angol nyelvű szövegek csak látszólag íródtak angolul: nagy részüket nem anyanyelvi szinten írók-beszélők készítették. A kaotikus helyzet miatt a gépnek sokszor „ki kell találnia", hogy mit akart megfogalmazni, aki a szöveget elkészítette, és ezt lefordítania magyarra. Az emberi agy sajátossági miatt mi a legtöbb esetben megértjük a hibás szöveget is, a gép azonban nagyon könnyen tévedhet ebben a helyzetben: a rossz helyre tett vesszőkkel, elgépelt szavakkal, sok nyelvi hibával megírt szöveg esetében nagyságrendekkel nehezebb a dolga a számítógépnek. A MorphoLogic webfordítója a teljes szöveget lefordítja pár másodperc alatt – Prószéky Gábor szerint ezzel nagy kockázatot vállaltak, és nagy bátorság kellett a fejlesztőknek ahhoz, hogy megtanítsák a gépet „toleránsnak lenni", magyarul elfogadni a bizonytalan értelmezést. A formális nyelvészet szabályai szerint ugyanis csak az fordítható le, ami szabályos, de itt nem a tökéletesség, hanem a hatékonyság a cél.

A kétezerben Széchenyi-díjjal kitüntetett tudós szerint a huszadik század matematikája megmutatta, hogy az igen vagy nem logikája új szempontokkal bővült: azt is figyelni kell, mennyi időbe, energiába, ráfordításba kerül az igen. Bizonyos esetekben nem precíz, hanem gyors és használható fordítások kellenek – esetünkben, viccesen fogalmazva, „ilyen gyorsan, ilyen olcsón ilyen rosszat más nem tud produkálni".

Mindezt egy példán keresztül megvizsgálva: ha egy angolul nem beszélő magyar termelő az uniós weboldalakon fellelhető adattengerből próbál a maga számára hasznos információt kihalászni, akkor nem kell neki a több ezer oldalas dokumentumokat lefordíttatni, elég, ha a gép segítségével megérti, hogy azok nagyjából miről szólnak. A konkrét, őt érdeklő információt aztán már szakemberrel pontosan lefordíttathatja, de amíg ehhez eljut, sok órát spórolhat meg egy olyan szoftverrel, amely pillanatok alatt általa is megérthető állapotba hozza a szövegeket.

Érdekes a kutatók tapasztalata ez ügyben: miután a mintaszöveget a gép lefordította, a néha vicces, nem túl értelmesnek tűnő magyar fordításból kérdéseket tettek fel egy kísérleti csoport tagjainak. A válaszok alapján csaknem százszázalékos volt a megértés, miközben az emberek elutasították a magyar fordítást, és azt állították, hogy képtelenek megérteni. Tehát intoleránsak vagyunk a géppel, miközben embertársainktól sokkal inkább elfogadjuk a hibát. Pedig rászorulunk a gép segítségére, hiszen kevesen beszélünk nyelveket. A gépi eszköz elterjedésnek gátja az is, hogy akiknek a leginkább szükségük van rá (idősebb generáció), azok kevesebbet interneteznek, mint az újdonságra fogékony fiatalok.

A nyelvész szerint szűk szakmai területen jobb minőségű fordítást is el lehetne érni, hiszen radikálisan csökken a többjelentésű szavak mennyisége. Egy autó műszaki leírása vagy egy használati utasítás esetén a célszoftvernek nem kell más kontextusokban „gondolkodnia", így hatékonyan és jól tudna fordítani – a cég a jövőben ilyen alkalmazásokat akar piacra dobni.

Nagyon érdekes a másik irány: ha magyar szövegünket angolra fordíttatjuk, kritika nélkül elfogadjuk az eredményt, ha nem beszélünk angolul. A minőségi szűrőt ezért már a magyar nyelvnél beépítik – a hamarosan elkészülő magyar–angol MorphoWord fordítóprogramhoz készül egy kiegészítő modul, mely csak azt ülteti át a másik nyelvre, ami helyesen van megfogalmazva az eredetiben. Magyarul: ez az úgynevezett kontrollált nyelvi alkalmazás arra figyel, hogy ne írjunk le olyat, amit majd nem tud lefordítani a gép. Az új szoftver azoknak segít, akik nem beszélik ugyan a nyelvet, de üzleti vagy más okokból leveleznének angolul, netán lefordítanák weboldalukat.

A jövőbe mutató tendenciákról szólva Prószéky Gábor a fordító munkájának támogatását emelte ki: a hatékony környezet megteremtésével a gép a jövőben áldolgozik az embernek, elvégezve a kulimunkát. A jövőt kutatva fontos szerepük lehet a szűrőprogramoknak is. A kommunikáció ugyanis újra túl direktté vált – egy elfoglalt üzletembert telefonon nehéz elérni, de az SMS-ét és az e-mailjét nagy valószínűséggel még maga olvassa. És egyre többet kap belőlük, így egyre kevesebb ideje marad a munkájára – ha egy beépített nyelvtechnológiai eszköz automatikusan elemezné a tartalmakat, és osztályozná a leveleket, netán továbbítaná is őket az illetékeseknek, ezzel sok időt lehetne spórolni.

A beszélt nyelv információtechnológiával való támogatása esetében minket, magyarokat a kis nyelvek hátránya sújt: piaci szempontból nem vagyunk elegen ahhoz, hogy megérje például magyarul is értő diktálóprogramot fejleszteni. Ráadásul nagyon bonyolult a magyar, így kétszeresen nehéz a feladat. A szakember szerint baj van a saját igényességünkkel is. Nem akarunk, nem tudunk helyesen írni és pontosan fogalmazni. Arra már van példa, hogy a szövegszerkesztők helytelen beállításai miatt szokunk rá hibákra: sok helyen találkozunk nagybetűs hónapnévvel amiatt, mert van olyan szövegszerkesztő program, mely a pont után automatikusan nagybetűvel kezdi a mondatot. A felhasználó először csodálkozik, hogy ő még kis j-vel tudta a januárt, de aztán megszokja a látványt, és szabálynak fogadja el – pedig ki is kapcsolhatná ezt az opciót a programban, és ezzel minden hibás nagybetűsítés megoldódna.

– Ha a szövegünkben káosz van, ezt a gép megsokszorozza, míg a gondozott szöveget hatékonyan tudja kezelni – véli a szakember.

A számítógépes nyelvészeti alkalmazások legnagyobb hibaforrása tehát nem a gép, hanem még mindig az ember. A géppel az a baj, hogy a hibát – az emberrel ellentétben – nem nagyon képes értelmezni.