Digitális terminológus

Prószéky Gábor

MorphoLogic
http://www.morphologic.hu

A terminológia modellezése

- Terminus technicus: szakszövegben olyan szó vagy kifejezés, amelyeket *konzisztensen kell fordítani*
- … tehát rögzült kifejezések (fixed expressions)
- Számítógépen: felszíni (morfoszintaktikai) jegyekkel

Módszerek (1)

- Szabályalapú módszerek
  - Az alapszótártól eltérő szókincs keresése
    - Statisztikai vizsgálattal: ismeretlen szavak kollokációit is keressük
  - A terminus technicusok belső morfoszintaktikai szerkezetének vizsgálata
  - A terminus technicusok környezetének vizsgálata

Módszerek (2)

- Statisztikai módszerek
  - Szokatlan gyakoriságú elemek keresése
    - Szavakra
    - Bigráfokra
  - Asszociációs mértékek alkalmazása
    - Klasszikus korpuszstatisztika

A módszerek összekapcsolása

- Fedés és pontosság: dichotómiát alkotnak
- Nagy fedés, alacsony pontosság:
  - Belső morfoszintaktikai szerkezet vizsgálata
  - Asszociációs mértékek alkalmazása
- Potenciálisan nagy pontosság:
  - A környezet vizsgálata
  - Szokatlan gyakoriságú elemek keresése

Induktív terminológiakeresés

- Ha van kiinduló szószedet, ne dobjuk el!
- Kétféle indukció:
  - Szószedetbeli kifejezések kollokációi
  - Szószedet kifejezéseihez hasonlók keresése

1. kísérlet

- Meglevő (informatikai) szószedet címszavainak összes előfordulását megkeressük egy (informatikai) korpuszban egy programmal(1)
- Megnézzük, milyen morfoszintaktikai minták formájában jelennek meg
- Kiértékeljük, megszűrjük a mintákat
- Egy másik programmal(2) a minták alapján új terminusjelölteket gyűjtünk ismeretlen szövegből

1. kísérlet – eredmények

- Absztrakt minták kigyűjtése
  - Korpusz: angol – 1,2 M, magyar – 1,6 M
  - Szószedet: angol – 30 765, magyar – 23 186
  - Minták: angol – 2 225, magyar – 2 520
- Új terminusok gyűjtése új szövegből
  - Angol: szöveg – 338 215, jelöltek – 25 702 (13 869)
  - Magyar: szöveg – 230 389, jelöltek – 15 141 (14 398)

1. kísérlet - példák

Angol

6 terminal service	Terminal services	[N]+[N][PL]
6 warning element	Warning element	[N]+[N]
6 worker process isolation mode	worker process isolation mode	[N]+[N]+[N]+[N]
6 XML parser	XML parser	[UNKNOWN]+[N]
4 server role	server roles	[N]+[N][PL]
4 shadow copy client	Shadow Copy Client	[ADJ]+[N]+[N]

Magyar

3 automatikus rendszer-helyreállítás	automatikus rendszer-helyreállítás	[ADJ][NOM]+[N]
3 hozzáférési jog	hozzáférési jogok	[ADJ][NOM]+[N][PL][NOM]
3 tartomány-nyilvántartó központ	tartomány-nyilvántartó központ	[N] +[N]
2 aktív tartalom	aktív tartalom	[N][NOM]+[N][NOM]
2 biztonsági házirend	biztonsági házirend	[N][NOM]+[N][NOM]
2 elérési út	elérési út	[N][NOM]+[N][NOM]

2. kísérlet – gyakorlati alkalmazás

- Spekulatív morfoszintaktikai minták
- Valódi terminológiai előkészítést végzünk
- Két angol nyelvű könyv szövege:
  - egyik: 100 963 szövegszó
  - másik: 74 626 szövegszó
- A kiemelt minták mellett keressük az alapszótárban nem szereplő szavakat is
- A két listát egyesítjük

Eredmények

- 1. kísérlet:
  - véletlenszerűen választott minták kézi kiértékelése
  - Angol minták
    - összes: 3 743, helyes: 2 412, pontosság: 64,44%
  - Magyar minták
    - összes: 2 107, helyes: 968, pontosság: 45,94%
  - Korrekció után (szűrtük a produktív szavakat):
    - angol: 77,08% (szűrve: *new*, *all*, *other*, *same*, *such* stb.)
    - magyar: 67,08% (szűrve: *elérésű*, *adott*, *alábbi*)

Eredmények

- 2. kísérlet:

	Tömeg (szövegszó)	Előfordulások	Különböző minták	Elfogadott minták
1. könyv	100 963	25 860	5 523	1 595 (28,88%)
2. könyv	74 626	14 330	6 535	2 275 (34,81%)

- Azonban:
  - 2 perc számítógépes feldolgozás, 3 óra utómunka
  - 400-500 oldal végigolvasása helyett!

További fejlesztések

- Módszerek integrálása
- Fordítások automatikus előállítása (amikor lehetséges)
- Végfelhasználói alkalmazás fejlesztése, integráció fordítástámogató programokba

Kapcsolódó projektek

- A pályázatról
  - IKTA-00181/2003: Digitális terminológus
  - MorphoLogic, PPKE ITK, SZAK Kiadó
  - kezdete: 2004, befezése: 2006
- Inspiráció
  - OTKA-NWO 048.011.040: Finding and Processing Multi-word Lexemes
  - Rijksuniversiteit Groningen – MorphoLogic
  - Befejeződött: 2004. december
- Alkalmazás
  - MetaMorpho projekt
  - …