Élő vagy élettelen?

Sass Bálint
joker@nytud.hu

MTA Nyelvtudományi Intézet, Nyelvtechnológiai Osztály
PPKE, Információs Technológiai Kar, MMT Doktori Iskola

MSZNY2007
Szeged, 2007. december 6–7.

Hogyan fordítanánk angolra?

Alszik.

Elromlott.

Hogyan fordítanánk angolra?

Alszik. $\rightarrow$ *He/she is sleeping.*

Elromlott. $\rightarrow$ *It has gone wrong.*

Hogyan fordítanánk angolra?

Alszik. → *He/she is sleeping.*

Elromlott. → *It has gone wrong.*

Általános kérdés

Gépi fordításkor mit tehetünk, ha . . .

- a forrásnyelv nem specifikál bizonyos tulajdonságot;
- a célnyelv viszont ugyanazon a ponton elvárja a tulajdonság egy konkrétan megadott értékét.

Lehetőségek:

- a szövegkörnyezet alapján kitaláljuk;
- lexikonban rögzített alapértelmezett értékeket használunk.

Módszer:
becslés nagyméretű korpuszban mért gyakoriságok alapján

AZ ÉLŐSÉGI (ANIMACY) SKÁLA

- univerzális skála: *ember* $>$ *állat* $>$ *élettelen*
  - nyelvi prominenciaviszonyokat meghatározó egyik tényező
  - élő/élettelen $\rightarrow$ eltérő nyelvi forma
  - lehetővé teszi, hogy a dialógusban követni tudjuk, hogy éppen melyik szereplőről van szó
  - univerzális elv: élőség $\sim$ aktuális esemény befolyásolására való képesség
- *jelentősége:* gépi fordítás generálás fázisa
- *ember* | *állat*, *élettelen*

Konkrét kérdés

magyar *pro-drop* $\leftrightarrow$ angol nem

Probléma

Ha az egyes szám harmadik személyű magyar mondatban nincs kitéve a névmás

$\rightarrow$

az angol oldalon a „semmiből” kell élő vagy élettelen testes névmást generálni.

Baseline

- a prototipikus alany: élő és ágens →

Baseline

minden esetben élőnek vesszük az alanyt

nagyon jó eredményeket ad: 80-90 %-os accuracy

- a fordítórendszer alapértelmezés szerint *he/she*-t generál → *elvárás:* sose javasoljunk élő helyett élettelen alanyt ($P_l = 100\%$).

NYERSANYAG

- Magyar Nemzeti Szövegtár
  egyvonzatkeretes egységekre bontott változata.

  Egység („tagmondat”): ige + bővítmények
  $\longrightarrow$
  „összetett igék”
  (pl.: *kiderül vmiről vmi*, *rendben van vmi*),
  az igék kereteinek külön kezelése.

  Hiányosság: megy $\leftrightarrow$ *nyilvánosságra hoz vmit*
  az előbbi lekérdezés jóval zajosabb!

- tanuló- és tesztkorpusz: az 1500 leggyakoribb keretből

Alapötlet

Komlósy:

Bizonyos igék csak egyes szám 3. személyben használatosak, ezeknek az igéknek „az alanyi vonzata nem jelölhet személyt”.

1/2. személy: élő $\leftrightarrow$ 3. személy: élő vagy élettelen

Alapötlet

Harmadik-személy% (*3sz%*) módszer

Ha az ige túlnyomó többségében 3. személyben fordul elő, akkor alanya élettelen, különben élő.

ige	*élőség*	*3sz%*-érték
néz	élő	65,4%
alszik	élő	64,0%
megtörténik	élettelen	99,9%
tartalmaz	élettelen	99,9%

3sz%-módszer:

3. személy aránya > 90% $\Rightarrow$ élettelen az alany

KIÉRTÉKELÉS

$n = 68$

	A	P_l	R_l	P_A	R_A
3sz%	84%	57%	86%	96%	83%
baseline	79%				

- a baseline magas – a módszer kis mértékben jobb
- *hibák:* főleg a kellemetlenebb irányba (P_l)
  jellemző igék: *nyilatkozik*, *vélekedik*, *aláír*, *tárgyal vmiről*

Vannak igék, melyek lényegében
csak 3. személyben fordulnak elő mégis élő alanyúak.

Ötlet

- 1/2. személy: élő $\leftrightarrow$ 3. személy: élő vagy élettelen
- *Ötlet:* vannak olyan szópáraink,
  melyek funkciójukban azonosak,
  kizárólag abban különböznek,
  hogy az élő/élettelen jegyük értéke más: *aki/ami*.

aki/ami arány $\sim$ élő/élettelen arány

- nyilván: *ami* = { *ami*, *amely*, *mely* }

MÉRTÉK

az élettelen alanyok arányának becslését finomítjuk

KORRIGÁLT HARMADIK-SZEMÉLY% *(k3sz%)* MÓDSZER

a 3. személyű alanyok közül csak az *ami* összes alanyi pozícióban előforduló vonatkozó névmáshoz viszonyított arányának megfelelő számút tekintünk élettelennek.

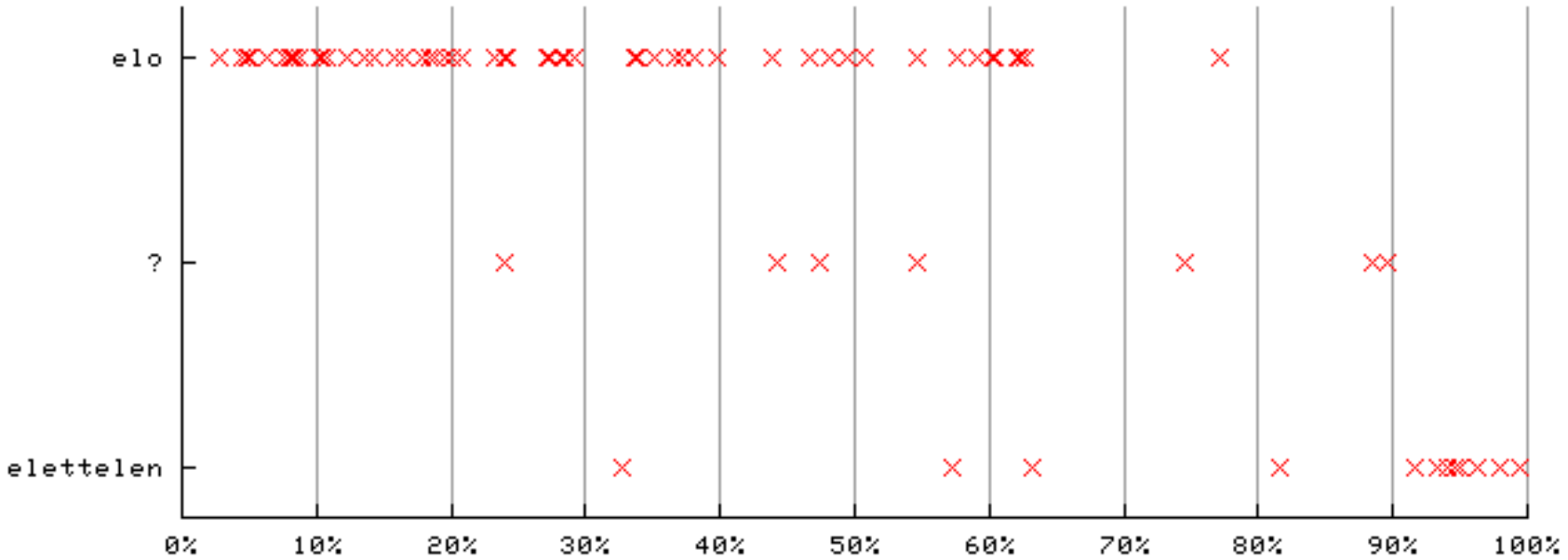
	1. sz.	2. sz.	3. sz. × *aki%*	3. sz. × *ami%*
3sz%	élő	élő	élettelen?	élettelen?
k3sz%	élő	élő	élő	élettelen

→ az alábbi mértéket alkalmazzuk:

3. személy aránya × *ami%*

DÖNTÉSI SZABÁLY

A *k3sz%* értékek eloszlása a tanulókorpuszon ($n = 68$):



Lényegében: 65% alatt élő $\leftrightarrow$ 90% fölött élettelen

Döntési szabály

A $P_I = 100\%$ követelménynek megfelelően a következő döntési szabályt rendeljük a mértékhez:

k3sz%-módszer:
3. személy aránya $\times$ *ami%* $> 90\% \Rightarrow$ élettelen az alany

KIÉRTÉKELÉS

- nagyobb, megbízhatóbb korpusz: $n = 383$, két annotátor
- kódolás: élő / élettelen / mindkettő
  egyetértés: 77% (296 ige), ebből egyértelmű: 278 ige

$n = 278$

	A	P_I	R_I	P_A	R_A
k3sz%	95%	95%	63%	95%	100%
baseline	88%				

- az eredmény jobb a korábbinál – baseline nagyon magas
- $P_I = 100\%$ követelmény: 1 hiba (tárgyatlan *jelent*)
- R_I: az élettelen keretek 63%-át találta meg

ILLUSZTRÁCIÓ

- A tesztkorpusz helyesen megtalált élettelen alanyú keretei:

  vezet vmihez, *kezdődik*, *kell vmihez*, *történik vkivel*, *következik vmiből*, *csökken*, *múlik vmin*, *megvalósul*, *létre jön vmi*, *véget ér vmi*, *épül vmire*, *kezdődik vmivel*, *szolgál vmire*, *irányul vmire*, *zajlik*, *keletkezik*, *kialakul vmiben*, *növekedik*, *fennmarad*, *zajlik vmiben*

- Igék különböző keretei (nagyobb igelistán futtatva):

  dicsér vmit élő $\leftrightarrow$ *dicsér munkáját* élettelen
  költözik élő $\leftrightarrow$ *költözik épületbe* élettelen

Alkalmazás

`http://www.webforditas.hu`

TOVÁBBI LEHETŐSÉGEK

- összevethető egy jóval erőforrásigényesebb módszerrel: a 3. személyű mondatok alanyi pozícióján megjelenő szavak gyűjtése és élő/élettelen kategóriákba sorolása.
- egyéb jegyek: felszólító mód $\rightarrow$ élő alany
- más nyelvekre is alkalmazható (pl. who/which?)
- egyéb „célpontok”: igék tárgya, igék egyéb bővítményei, predikatív melléknév alanya, birtok birtokosa
- nemek elkülönítése: *lány,nő/fiú,férfi* arány
  megnősül: 1/20, *férjhez megy*: 108/2
  megvéd: 4/20, *zokog*: 25/9.

TOVÁBBI LEHETŐSÉGEK

- összevethető egy jóval erőforrásigényesebb módszerrel: a 3. személyű mondatok alanyi pozícióján megjelenő szavak gyűjtése és élő/élettelen kategóriákba sorolása.
- egyéb jegyek: felszólító mód → élő alany
- más nyelvekre is alkalmazható (pl. who/which?)
- egyéb „célpontok”: igék tárgya, igék egyéb bővítményei, predikatív melléknév alanya, birtok birtokosa
- nemek elkülönítése: *lány,nő/fiú,férfi* arány
  megnősül: 1/20, *férjhez megy*: 108/2
  megvéd: 4/20, *zokog*: 25/9.

Köszönöm a figyelmet!