Igekötő-kapcsolás

MSZNY2022
2022. január 27-28.

Pethő Gergely, Sass Bálint, Kalivoda Ágnes,
Simon László, Lipp Veronika

Nyelvtudományi Kutatóközpont, Lexikológiai Intézet
sass.balint@nytud.hu

vázlat

- mit?
- hogyan?
- kiértékelés
- eredmény
- korpusz
- annotáció
- keresés
- demó
- miért?
- reprodukálhatóság

vázlat

1. mit? *el lehetne nem költői módon is* ***mondani*** $\rightarrow$ ***elmondani***
2. hogyan? szabályalapon, kétlépéses módszer, *„nehéz”* korpusz
3. kiértékelés módszere
4. eredmény SOTA, $F_1 = 0{,}9889$
5. korpusz hozzáférhető, *„általános”* és *„nehéz”*
6. annotáció +3 mező
7. keresés új lehetőségek
8. demó 3 példa
9. miért? (korpusz)nyelvészet, lexikográfia
10. reprodukálhatóság „gombnyomásra”

mit?

meg kell ***próbálni***
szivárogtatta *ki*
tudtam *csak meg*
meg van ***győződve***
be nem ***tartásának***
meg lehetett volna ***küldeni***
miért ***szólna*** *ő ebbe bele*
fel kell, hogy ***vállalja***

hogyan?

– szabályalapú megoldás

– iteratívan, két körben:

1. adatokból ránézésre – train/dev/test!
2. hibaelemzés – *„nehéz”* korpuszon

– megfontolások Kalivoda Ági disszertációja alapján

- végig az igei természetű szavakon
- az eleve igekötős alakok megjelölése
- `-3..+3` ablak
- közbeeső szavak jellege
- igei természetű szavak kezelése
- részletek a kódban

kiértékelés

szigorúan csak ezt a modult értékeljük ki
= az `emTag` kimenetét hibátlannak vesszük
= az az igekötő, amit az `emTag` annak határoz meg

az igekötőt és az igét indexáljuk$_3$ össze$_3$
igekötő-annotálási feladatként tekintjük
= az igekötők indexét vizsgáljuk

gold		*pred*	$\rightarrow$ osztály
$\emptyset$	=	$\emptyset$	`TN`
?	$\neq$	j	`FP`
i	$\neq$	$\emptyset$	`FN`
i	=	j	`TP`

eredmény

számok

„általános” minta

	pontosság (%)	fedés (%)	F_1	accuracy (%)
`emPreverb`	99,80	98,00	0,9889	97,80
`nearest-verb`	96,56	98,76	0,9765	95,40
`max2`	97,94	96,94	0,9744	95,00
`emStanza`	96,04	90,66	0,9328	87,40

„nehéz” minta

	pontosság (%)	fedés (%)	F_1	accuracy (%)
`emPreverb`	99,14	94,28	0,9665	93,62
`nearest-verb`	76,62	93,79	0,8434	73,14
`max2`	92,28	79,48	0,8540	75,00
`emStanza`	90,83	89,24	0,9003	81,91

túl nehéz

1. nagy távolság

1. *El tudná már végre valaki nekem* ***árulni****, hogy*
2. *olyankor ő is külhonokba volt* ***távozni*** *kénytelen, trónjáról el, pár jótanácsért dörzsölt pártfogóihoz*
3. *hogy meg tuggyon szegénykém valamit* ***nyomni*** *a* Press any key to continue.. *üzenet megjelenésekor*

4. *át- meg át**járták***
5. *de azért el el **járogatok***
6. *A TV3-at nem le, hanem **meg-tiltani** kellene.*
7. *a távolság is csak egy probléma, amit meg kell és meg is lehet **oldani**.*
8. *Ha nem ki, hanem be akar **kapcsolódni** a világba, átkapcsol és kész.*
9. *Józsi tudta, hogy az öregje addig senkinek cipőt, csizmát nem szab a lábára, amíg meg nem **nézette**, **szagoltatta**, **tapintatta** velük a bőröket, hogy melyik lenne igazán a kedvükre való*

korpusz

„általános” korpusz 503 mondat 500 igekötő

→ `[msd!=".*IGE.*"] [msd="IK"]` →

„nehéz” korpusz 1477 mondat 1668 igekötő

hivatalos dev/test split a *„nehéz”* korpuszon

„nehéz” dev	1150 mondat	1292 igekötő
„nehéz” test	327 mondat	376 igekötő

elérhető:

`https://github.com/ril-lexknowrep/hungarian-preverb-corpus`

annotáció

form	lemma	xpostag	compound	prev	previd	prevpos
kapaszkodik	kapaszkodik	[/V][...]	kapaszkodik			
→						
kapaszkodik	kapaszkodik	[/V][...]	kapaszkodik			
eloldozódott	eloldozódik	[/V][...]	el#oldozódik			
→						
eloldozódott	eloldozódik	**[/Prev]**[/V][...]	el#oldozódik	**pfx**		
tér	tér	[/V][...]	tér			
vissza	vissza	[/Prev]	vissza			
→						
tér	**visszatér**	**[/Prev]**[/V][...]	**vissza#tér**	**sep**	**7**	**+1**
vissza	∅	[/Prev]	∅	**conn**	**7**	
haza	haza	[/Prev]	haza			
akarok	akar	[/V][...]	akar			
menni	megy	[/V][Inf]	megy			
→						
haza	∅	[/Prev]	∅	**conn**	**26**	
akarok	akar	[/V][...]	akar			
menni	**hazamegy**	**[/Prev]**[/V][Inf]	**haza#megy**	**sep**	**26**	**-2**

keresési lehetőségek 1/2

1. igekötős igék ***összes* korpuszbeli találata** elválástól függetlenül:
   CQL: `[lemma="előjön"]`
2. összes igekötőtlen ige:
   CQL: `[xpostag="\[/V\].*"]`
   összes igekötős ige:
   CQL: `[xpostag="\[/Prev\]\[/V\].*"]`
3. összes *meg* igekötős ige:
   CQL: `[compound="meg#.*" & xpostag="\[/Prev\].+"]`
4. összes tapadó alak:
   CQL: `[lemma="előjön" & prev="pfx"]`
   összes elvált alak:
   CQL: `[lemma="előjön" & prev="sep"]`

keresési lehetőségek 2/2

5. odakapcsolt igekötők:
   CQL: `[prev="conn" & xpostag="\[/Prev\]"]`
   árván maradt igekötők:
   CQL: `[prev="" & xpostag="\[/Prev\]"]`
6. `prevpos` mezőből → gyakorisági lista
   az igekötő eltávolodásának eloszlása
   adott/összes ige és adott/összes igekötő vonatkozásában
7. adott igekötő (*meg*) mennyire szeret elválni:
   CQL: `[prev="(sep|pfx)" & compound="meg#.*"]`
   általában mennyire szeretnek elválni az igekötők:
   CQL: `[prev="(sep|pfx)"]`

explicit annotáció → könnyű lekérdezhetőség

demó

1. igekötős ige összes találata:
   CQL: `[lemma="előjön"]`

2. árván maradt igekötők:
   CQL: `[prev="" & xpostag="\[/Prev\]"]`

3. *meg nem értés*:
   CQL: `[xpostag="\[/Prev\]\[/N\].*" & prevpos="-2"]`

miért?

- hasznos és triviálisnak *tűnő* feladat,
  mégsem volt eddig alapos megoldás rá

- az igekötős ige *egy egység*
  - korpuszalapú nyelvészeti vizsgálatok
  - lexikográfia

reprodukálhatóság

- minden elérhető a `github`-on
  a `https://github.com/ril-lexknowrep` alatt

- megfelelő liszenszekkel:
  - `emPreverb` – LGPL v3.0
  - `hungarian-preverb-corpus` – CC BY-NC-ND 4.0
  - `emCompound` – LGPL v3.0

- kiértékelés reprodukálása „gombnyomásra”

```
git clone https://github.com/ril-lexknowrep/emPreverb
make evaluate
```

  → előáll a közölt kiértékelési táblázat

összefoglalás

1. mit?	*el lehetne nem költői módon is* ***mondani*** $\rightarrow$ ***elmondani***
2. hogyan?	szabályalapon, kétlépéses módszer, *„nehéz”* korpusz
3. kiértékelés	módszere
4. eredmény	SOTA, $F_1 = 0{,}9889$
5. korpusz	hozzáférhető, *„általános”* és *„nehéz”*
6. annotáció	+3 mező
7. keresés	új lehetőségek
8. demó	3 példa
9. miért?	(korpusz)nyelvészet, lexikográfia
10. reprodukálhatóság	„gombnyomásra”