cgmorf2conllu.py

Programm morfoloogilise märgenduse teisendamiseks nn kitsenduste grammatika (CG) märgendusskeemist Universal Dependencies' märgendusskeemi

Sisukord

Skripti üldine kirjeldus

Otstarve

Skript *cgmorf2conllu.py* on mõeldud kitsenduste grammatika (CG) formaadis morfoloogilise märgendusega failide teisendamiseks *Universal Dependencies'* (*CoNLL-U)* formaati. Märgenduse teisendamise baasreeglid on skripti koodi sisse kirjutatud, kuid saab lisada reegleid ka tabeliformaadis (*tab-separated values*) etteantavate failide alusel.

Skript on kirjutatud keeles Python ning on testitud Python 3.5 keskkonnas.

Asukoht svn-s

svn+ssh://svnkrpvkp@imbi.at.mt.ut.ee/svnkrpvkp/trunk/rabauti/cgmorf2conllu

Kasutamine

cgmorf2conllu.py [-w lemmalistfolder] [-m] [-d] [-i inputfile] [-o outpufile]

Skripti käsurea parameetrid

Parameeter	**Kohustuslik**	**Kirjeldus**	**Kommentaar**
-w	ei	Kataloogi nimi (tsv formaadis failidega)	Vaikeväärtus on POS_LEMMA_RULES
-i	ei	Sisendfaili nimi	Vaikeväärtus STDIN
-o	ei	Väljundfaili nimi	Vaikeväärtus STDOUT
-d	ei	Väljundisse kirjutakse *debug*-info	Vaikimisi ei kirjutata
-m	ei	Lisatakse “ma”-tunnus verbi lõppu (*ole -> olema*)	Vaikimisi ei lisata

Kasutamise näited

$ python cgmorf2conllu.py < samples/geof.synt > geof.conllu
$ python cgmorf2conllu.py -i samples/geof.synt -o geof.conllu
$ python cgmorf2conllu.py –d -m -i samples/geof.synt -o geof.conllu

Skripti sisend ja väljund

Sisendiks on morfoloogilise märgendusega tekstifail, mille märgendus on kitsenduste grammatika (CG) kujul. Morfoloogiliste märgendite ja kategooriate kohta vt http://www.cl.ut.ee/korpused/morfliides/seletus.php?lang=et .

Süntaktiline märgendus (@ ja # algavad stringid) võib sisendis olla, aga ei pea olema.

Sisendfail peab olema UTF-8 kodeeringus.

Sisendi näide 1.

```
"<Esimestel>"
        "esimene" Ltel N ord pl ad l cap @AN> #1->2
"<kordadel>"
        "kord" Ldel S com pl ad @ADVL #2->10
"<oli>"
        "ole" Li V main indic impf ps3 sg ps af <FinV> <Intr> <0> @FMV #3->0
"<mul>"
        "mina" Ll P pers ps1 sg ad @ADVL #4->3
"<psüühiliselt>"
        "psüühiliselt" L0 D @ADVL #5->10
"<raske>"
        "raske" L0 A pos sg nom @PRD #6->3
"<talle>"
        "tema" Llle P pers ps3 sg all @ADVL #7->3
"<puurondiga>"
        "puu_ront" Lga S com sg kom @NN> @ADVL #8->3
"<pähe>"
        "pea" L0 S com sg adit @ADVL #9->10
"<virutada>"
        "viruta" Lda V main inf <NGP-P> @SUBJ #10->3
"<.>"
        "." Z Fst CLB #11->11
"</s>"
```

Sisendi näide 2

```
"<Tundub>"
        "tundu" Lb V main indic pres ps3 sg ps af cap <FinV> <Intr>
"<,>"
        "," Z Com CLB
"<et>"
        "et" L0 J sub
"<sellele>"
        "see" Lle P dem sg all
"<pärimisele>"
        "pärimine" Lle S com sg all <mine> "päri"
"<on>"
        "ole" L0 V main indic pres ps3 sg ps af <FinV> <Intr>
"<lihtne>"
        "lihtne" L0 A pos sg nom
"<vastata>"
        "vasta" Lta V main inf <NGP-P> <All>
"<.>"
        "." Z Fst CLB
"</s>"
```

Väljundi formaat

Väljundiks on tekstifail *CoNLL-U* formaadis.

CoNLL-U formaadi kohta saab lugeda siit:

Universal Dependencies' morfoloogiliste kategooriate kohta saab lugeda siit: http://universaldependencies.org/u/overview/morphology.html .

Tulp	Nimetus	Krjeldus
1	ID	Word index, integer starting at 1 for each new sentence; may be a range for multiword tokens; may be a decimal number for empty nodes
2	FORM	Word form or punctuation symbol
3	LEMMA	Lemma or stem of word form
4	UPOSTAG	Universal part-of-speech tag
5	XPOSTAG	Language-specific part-of-speech tag; underscore if not available
6	FEATS	List of morphological features from the universal feature inventory or from a defined language-specific extension; underscore if not available
7	HEAD	Head of the current word, which is either a value of ID or zero (0)
8	DEPREL	Universal dependency relation to the HEAD (root iff HEAD = 0) or a defined language-specific subtype of one
9	DEPS	Enhanced dependency graph in the form of a list of head-deprel pairs
10	MISC	Any other annotation

Väljundi näide

```
# sent_id = 45
# text = Toompea südames
1       Toompea Toom_pea        NOUN    S       Case=Gen|Number=Sing    0       root    _       _
2       südames süda    NOUN    S       Case=Ine|Number=Sing    1       dep     _       _

# sent_id = 46
# text = Riigiarhiiv asutati eeskätt ja ainult riigiasutuste dokumentide kogumiseks ja hoidmiseks .
1       Riigiarhiiv     riigi_arhiiv    NOUN    S       Case=Nom|Number=Sing    0       root    _       _
2       asutati asuta   VERB    V       Mood=Ind|Polarity=Pos|VerbForm=Fin|Voice=Pass   1       dep     _       _
3       eeskätt ees_kätt        ADV     D       _       1       dep     _       _
4       ja      ja      CCONJ   J       _       1       dep     _       _
5       ainult  ainult  ADV     D       _       1       dep     _       _
6       riigiasutuste   riigi_asutus    NOUN    S       Case=Gen|Number=Plur    1       dep     _       _
7       dokumentide     dokument        NOUN    S       Case=Gen|Number=Plur    1       dep     _       _
8       kogumiseks      kogumine        NOUN    S       Case=Tra|Number=Sing    1       dep     _       _
9       ja      ja      CCONJ   J       _       1       dep     _       _
10      hoidmiseks      hoidmine        NOUN    S       Case=Tra|Number=Sing    1       dep     _       _
11      .       .       PUNCT   Z       _       1       dep     _       _
```

Sõnastik ja reeglid

Süntaktiliste sõltuvuste teisendamine

Skript ei ole mõeldud süntaktiliste sõltuvuste teisendamiseks. Kõiki lähtefailis olevaid süntaktilisi sõltuvusi ja funktsioone ignoreeritakse. Väljundis, *CoNLL-U* formaati teisendamisel, muudetakse lause ***root***-elemendiks lause esimene sõne. Kõik ülejäänud sõned märgendatakse üldistatud sõltuvus-suhtega ***dep***.

Sõnaliigi märgendi (POS) teisendamise reeglid

Sõnaliikide märgendite teisendamise reeglid on skripti koodi sisse kirjutatud.

POS IN	MORF IN	POS UD OUT	FEATURES UD OUT
A	pos	ADJ	Degree=Pos
A	comp	ADJ	Degree=Cmp
A	super	ADJ	Degree=Sup
B		PART	
D		ADV	
E		SYM	
G		NOUN	Number=Sing Case=Gen
I		INTJ	
J	crd	CCONJ	
J	sub	SCONJ	
K	pre	ADP	AdpType=Prep
K	Post	ADP	AdpType=Post
N	card	NUM	NumType=Card
N	ord	ADJ	NumType=Ord
P		PRON	PronType=Dem,Int,Ind,Prs,Rcp,Rel,Tot
S	com	NOUN	
S	prop	PROPN	
T		X	
V	main	VERB	
V	aux	AUX	
V	mod	AUX	
X		ADV	
Y and CAPS in lemma		PROPN	Abbr=Yes
Y		SYM	
Z		PUNCT	

Morfoloogiliste kategooriate teisendamise reeglid

Sõnaliigi märgendi ”POS UD” ja morfoloogiliste kategooriate märgendite ”FEATURES UD” veergude väärtused väärtused tulevad “POS märgendite teisendamise reeglite” tabelist

Morfoloogiliste kategooriate teisendamise reeglid on skripti koodi sisse kirjutatud.

MORF IN	POS UD	FEATURES UD	COMMENT	FEATURES UD OUT
#Case=				
nom	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Nom
gen	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Gen
part	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Par
ill	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ill
ill	VERB, AUX	VerbForm=Sup		Case=Ill
adit	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Add
in	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ine
in	VERB, AUX	VerbForm=Sup		Case=Ine
el	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ela
el	VERB, AUX	VerbForm=Sup		Case=Ela
all	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=All
ad	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ade
abl	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Abl
tr	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Tra
tr	VERB, AUX	VerbForm=Sup		Case=Tra
term	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ter
es	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Ess
ab	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Abe
ab	VERB, AUX	VerbForm=Sup		Case=Abe
kom	NOUN, PROPN, ADJ, DET, PRON, NUM			Case=Com
#Degree=				
pos	ADJ			Degree=Pos
comp	ADJ			Degree=Cmp

#Mood= #VerbForm=				
indic	VERB, AUX			Mood=Ind VerbForm=Fin
imper	VERB, AUX			Mood=Imp
				VerbForm=Fin
cond	VERB, AUX			Mood=Cnd VerbForm=Fin
quot	VERB, AUX			Mood=Qot VerbForm=Fin
#Number=				
sg	NOUN, PROPN, ADJ, DET, PRON, NUM, VERB, AUX			Number=Sing
pl	NOUN, PROPN, ADJ, DET, PRON, NUM, VERB, AUX			Number=Plur
#NumForm=				
digit	ADJ	NumType=Ord		NumForm=Digit
digit	NUM			NumForm=Digit
l	ADJ	NumType=Ord		NumForm=Letter
l	NUM			NumForm=Letter
roman	ADJ	NumType=Ord		NumForm=Letter
#Person=				
ps1	VERB, AUX			Person=1
ps2	VERB, AUX			Person=2
ps3	VERB, AUX			Person=3
#Polarity=				
neg	AUX			Polarity=Neg
neg	VERB			Connegative=Yes
#Tense=				
impf	VERB, AUX			Tense=Past
past	VERB, AUX			Tense=Past
pres	VERB, AUX			Tense=Pres
#VerbForm=				
inf				VerbForm=Inf
ger				VerbForm=Conv
sup	VERB, AUX			VerbForm=Sup
partic	ADJ, VERB, AUX			VerbForm=Part
#Voice=				
ps	VERB, AUX			Voice=Act
imps	VERB, AUX			Voice=Pass
#Tense= #Voice=				
<v> partic				Tense=Pres Voice=Act

<tav> partic				Tense=Pres Voice=Pass

POS / Lemma -> POS teisendamisreeglid

Need reeglid ei ole skripti koodi sisse kirjutatud, neid saab skriptile ette anda TSV-formaadis failidena. Reegleid saab kirjutada ainult kindla lemma kohta.

Reeglite faili struktuur

Reeglite fail on TSV (*tab-separated values*) fail UTF-8 kodeeringus, väljade eraldajaks on tabulaator.

Reeglite faile loeb skript vaikimisi skripti kataloogist **POS_LEMMA_RULES**. Reeglite failide muud asukohta saab skriptile ette anda ka käsitsi käsurea parameetriga *–w*.

Reeglite faili nimi

Reeglite faili nimi peab algama sõnaliigimärgendi tähisega, sellele järgneb alakriips ja siis vabas vormis tekst. Faili laiend peab olema *tab*.

Reeglite faili nime näited:

A_PRON.tab
P_PRON.tab
Y_puudepanga_lyhendite_UD_morf.tab
Z_rules.tab

Reeglite faili struktuur

Reeglite fail koosneb kolmest tabulaatoriga eraldatud veerust:

- esimene veerg – lemma;
- teine veerg - UD sõnaliik;
- kolmas veerg – UD Features*.

*UD Features väljal on eraldajaks tühik

Näide P_PRON.tab reeglite failist

```
mina    PRON    PronType=Prs    Person=1
sina    PRON    PronType=Prs    Person=2
tema    PRON    PronType=Prs    Person=3

oma     PRON    PronType=Prs Poss=Yes
ise     PRON    PronType=Prs Reflex=Yes

iseenda PRON    PronType=Prs Reflex=Yes
omaenese        PRON    PronType=Prs Reflex=Yes
```

mina	PRON	PronType=Prs Person=1
sina	PRON	PronType=Prs Person=2
tema	PRON	PronType=Prs Person=3
iseenda	PRON	PronType=Prs

		Relflex=Yes
omaenese	PRON	PronType=Prs Relflex=Yes

Sõnavormi-põhised teisendamisreeglid

Sõnavormi-põhised teisendamisreeglid on skripti koodi sisse kirjutatud.

token	POS UD	MORF IN		FEATURES UD OUT
oma	PRON	sg gen		Poss=Yes

Lemma-põhised teisendamisreeglid

Lemma-põhised teisendamisreeglid on skripti loogikasse sisse kirjutatud.

lemma	POS UD		FEATURES UD OUT
ise	PRON, DET		Reflex=Yes
enda	PRON, DET		Reflex=Yes
enese	PRON, DET		Reflex=Yes
iseenda	PRON, DET		Reflex=Yes
iseenese	PRON, DET		Reflex=Yes
omaenda	PRON, DET		Reflex=Yes
omaenese	PRON, DET		Reflex=Yes

Skripti veateated ja logi

Skripti töö logi ja veateated kirjutatakse STDERR väljundisse.

Logi näide

```
$ python cgmorf2conllu.py < samples/martjatoivo.txt.cg3 > out
Read 'P' rules from : /Users/svnkrpsoft/cgmorf2conllu/POS_LEMMA_RULES/P_PRON.tab
Read 'Y' rules from : /Users/svnkrpsoft/cgmorf2conllu
/POS_LEMMA_RULES/Y_puudepanga_lyhendite_UD_morf.tab
Read 'A' rules from : /Users/svnkrpsoft/cgmorf2conllu /POS_LEMMA_RULES/A_PRON.tab
Read 'Z' rules from : /Users/svnkrpsoft/cgmorf2conllu /POS_LEMMA_RULES/Z_rules.tab

Done.
```

Veateated

```
Unable to parse morf info, sentence 243 line 14  "võrdu" lb v main indic pres ps3 sg ps af <finv> <intr>
<kom> fst clb

Unknown line format, sentence 573, line 7  "<>"
```

Veateade koosneb vea kirjeldusest, vigase lause numbrist sisendfailis ning vigase rea sisust.