brat2inforem

Programm kitsenduste grammatika kujul sõltuvussüntaktiliselt märgendatud failide ja tarkvara brat abil viitesuhetega märgendatud failide ühendamiseks

Taust ja vajadus

Selleks, et kitsenduste grammatika kujul sõltuvussüntaktiliselt märgendatud failid sobiksid sisendiks tarkvarale Brat (http://brat.nlplab.org), eemaldati failidest morfoloogiline ning süntaktiline märgendus, kuid eelnevalt märgiti ära märgendamist vajavad asesõnad. Seejärel lisati failidesse tarkvara Brat kasutades uut informatsiooni (märgiti ära viitealused ning pronoomenid ning seosed nende vahel).

See skript on kirjutatud selleks, et viia uus informatsioon kokku selle informatsiooniga, mis enne Brat märgendite lisamist kõrvale heideti, st et ühes failis oleks morfoloogiline, süntaktiline ja viitealuste märgendus.

Skript on kirjutatud keeles Python, on testitud keele Python kahe versiooniga 2.7 ning 3.5.

Skript eeldab, et sisendfailid on utf-8 kodeeringus.

Skript kirjutab oma logi stderr väljundisse.

Probleemide korral võib mure kirja panna ja saata aadressile <rabauti@gmail.com>

Parameetrid

-i Brat sisendfailide asukoht + failinimi ilma laienditeta (prefiks)

Nt: -i samples/aja_EPL_2007_08_12.

Kui parameetri väärtuseks on "samples/aja_EPL_2007_08_12.", siis skript otsib sisendiks failid "samples/aja_EPL_2007_08_12.*.ann" ja "samples/aja_EPL_2007_08_12.*.txt"

-m originaalfail kitsenduste grammatika kujul, millele liidetakse Brati abil märgendatud failidest saadud informatsioon.

Nt: -m samples/aja_EPL_2007_08_12.tasak.inforem

Kitsenduste grammatika (CG) kuju on selline:

```
"<s>"
"<Üle>"
	"üle" L0 K pre @ADVL #1->4
"<poole>"
	"pool" L0 N card sg gen l @<P #2->1
"<neist>"
	"tema" Lst P pers ps3 pl el @<NN #3->2
"<lasti>"
	"lask" Lti V main indic impf imps af @FMV #4->0
"<maha>"
	"maha" L0 D @Vpart #5->4
"<pärast>"
	"pärast" L0 K pre @ADVL #6->4
"<Berliini>"
	"Berliin" L0 S prop sg gen @NN> #7->8
```

```
"<müüri>"
	"müür" L0 S com sg gen @NN> #8->9
"<püstitamist>"
	"püstitamine" Lt S com sg part @<P #9->6
"<.>"
	"." Z Fst #10->10
"</s>"
```

-o Tulemusfaili asukoht+nimi. Tulemusfailiks on kitsenduste grammatika (CG) formaadis fail, millele on liidetud Brat failidest saadud viidete informatsioon. Lisaks on lisatud lausetele ID-d, et viidata saaks ka teise lause sõnele.

Nt: -o aja_EPL_2007_08_12.tasak.inforem.ana

Tulemus salvestatakse jooksvasse kataloogi "aja_EPL_2007_08_12.tasak.inforem.ana" faili.

Kasutamise näide:

```
$ python brat2inforem.py -i samples/aja_EPL_2007_08_12. -m samples/aja_EPL_2007_08_12.tasak.inforem -o aja_EPL_2007_08_12.tasak.inforem.ana
```

Brati ekspordifailidest saadud viited salvestatakse väljundfailis analüüsirea lõppu loogeliste sulgude vahele.
Järgnevas näites on sõne "<neist>" märgitud Brat failis kui Pronoomen ning sellel on Coref seos tokeniga #18 lauses nr 18.

```
"<neist>"
	"tema" Lst P pers ps3 pl el @<NN #3->2 {Pronoomen} {Coref:18.16}
```

Tulemuse formaat on selline:

```
"<s id="19">"
"<Üle>"
	"üle" L0 K pre @ADVL #1->4
"<poole>"
	"pool" L0 N card sg gen l @<P #2->1
"<neist>"
	"tema" Lst P pers ps3 pl el @<NN #3->2 {Pronoomen} {Coref:18.16}
"<lasti>"
	"lask" Lti V main indic impf imps af @FMV #4->0
"<maha>"
	"maha" L0 D @Vpart #5->4
"<pärast>"
	"pärast" L0 K pre @ADVL #6->4
"<Berliini>"
	"Berliin" L0 S prop sg gen @NN> #7->8
"<müüri>"
	"müür" L0 S com sg gen @NN> #8->9
"<püstitamist>"
	"püstitamine" Lt S com sg part @<P #9->6
"<.>"
	"." Z Fst #10->10
"</s>"
```

Veateated ja logi

Skript ei tunnista viiteid, milles on üheks osapooleks märgitud mitmest sõnest koosnev osa:

Invalid line in ann file: T116 Viitealus 12201 12208;12209 12211;12217 12223 Boersma ja Ronnes

Erandiks on sõned, mis sisaldavad tühikuid, aga on juba algses kitsenduste grammatika kujul märgendatud failis lemmana analüüsitud kui üks sõne. Nt:

"<Brasiilias>"
"Brasiilia" Ls S prop sg in @<NN #13->12
"<Rio de Janeiros>"
"Rio de Janeiro" Ls S prop sg in @<NN #14->12

"Rio de Janeiro" on ka algses kitsenduste grammatika kujul olevas failis olemas ühe üksusena (lemmana), ning talle saab selliselt ka viidata.

Skript hetkel ignoreerib AnnotatorNotes kirjeid:

Ignored line: #3 AnnotatorNotes R53 kahtlane

Skripti töö logi näide

Searching for Brat files
2 files found
samples/aja_EPL_2007_08_12.tasak.ann
samples/aja_EPL_2007_08_12.tasak.txt

Parsing samples/aja_EPL_2007_08_12.tasak.txt

1..2..3..4..5..6..7..8..9..10..11..12..13..14..15..16..17..18..19..20..21..22..23..24..25..26..27..28..29..30..31..32..33..34..35..36..37..38..39..40..41..42..43..44..45..46..47..48..49..50..51..52..53..54..55..56..57..58..59..60..61..62..63..64..65..66..67..68..69..70..71..72..73..74..75..76..77..78..79..80..81..82..83..84..85..86..87..88..89..90..91..92..93..94..95..96..97..98..99..100..101..102..103..104..105..106..107..108..109..110..111..112..113..114..115..116..117..118..119..120..121..122..123..124..125..126..127..128..129..130..131..132..133..134..135..136..137..138..139..140..141..142..143..144..145..146..147..148..149..150..151..152..153..154..155..156..157..158..159..160..161..162..163..

Parsing samples/aja_EPL_2007_08_12.tasak.ann
.....
.....
Invalid tag: T75 Viitealus 351 354 lii
Invalid reference: R4 Coref Arg1:T4 Arg2:T75
Ignored line: #1 AnnotatorNotes T110 ilmselt

Ignored line: #2 AnnotatorNotes T115 kahtlane
Invalid line in ann file: T116 Viitealus 12201 12208;12209 12211;12217 12223Boersma ja Ronnes
Invalid reference: R47 Coref Arg1:T58 Arg2:T116
Invalid line in ann file: T119 Viitealus 11599 11606;11607 11609;11624 11629esemeid ja vorme
Invalid tag: T122 Viitealus 8342 8358 500 000 krooniga
Invalid reference: R52 Coref Arg1:T44 Arg2:T122
Invalid tag: T123 Viitealus 8991 9005 paarsada tuhat
Invalid reference: R53 Coref Arg1:T46 Arg2:T123
Ignored line: #3 AnnotatorNotes R53 kahtlane
Invalid tag: T127 Viitealus 10566 10579 Mercedes Benz
Invalid reference: R57 Coref Arg1:T50 Arg2:T127
Invalid reference: R59 Coref Arg1:T56 Arg2:T119

Merging Brat and inforem info

1..2..3..4..5..6..7..8..9..10..11..12..13..14..15..16..17..18..19..20..21..22..23..24..25..26..27..28..29..30..31..32..33..34..35..36..37..38..39..40..41..42..43..44..45..46..47..48..49..50..51..52..53..54..55..56..57..58..59..60..61..62..63..64..65..66..67..68..69..70..71..72..73..74..75..76..77..78..79..80..81..82..83..84..85..86..87..88..89..90..91..92..93..94..95..96..97..98..99..100..101..102..103..104..105..106..107..108..109..110..111..112..113..114..115..116..117..118..119..120..121..122..123..124..125..126..127..128..129..130..131..132..133..134..135..136..137..138..139..140..141..142..143..144..145..146..147..148..149..150..151..152..153..154..155..156..157..158..159..160..161..162..163..
Done.
Result saved to aja_EPL_2007_08_12.tasak.inforem.ana