Spatial models of Croatian, Bosnian and Serbian

`hbs-twitter-space/`

Documentation

May 1, 2016

Contents

1 Synopsis

The goal of `hbs-twitter-space` is to prepare an annotated corpus of specific linguistic features contained in Bosnian, Croatian, Montenegrinian and Serbian (BKMS) tweets. The main question this package helps to solve is: does the distribution of linguistic variables follow the BKMS' administrative pattern?

2 Architecture

`hbs-twitter-space/`

- `../lexicons/apertium-hbs.hbs_HR_purist.mte.gz`
- `../lexicons/apertium-hbs.hbs_SR_purist.mte.gz`
- `hrsrTweets.gz`
- `hrsrTweets.var.gz`
- `extract_variables.py`
- `custom-lexicons/`
  - `diftong-v-lexicon.gz`
  - `hdrop-lexicon.gz`
  - `kh-lexicon.gz`
  - `rdrop-lexicon.gz`
  - `st-c-lexicon.gz`
  - `yat-lexicon.gz`
  - `ch-lexicon.gz`
  - `genitiv-og-eg-lexicon.gz`
  - `hr-months.gz`
  - `int-months.gz`
  - `ir-is-lexicon.gz`
  - `ir-ov-lexicon.gz`
  - `verbs-inf-lexicon.gz`
  - `verbs-lexicon.gz`
  - `verbs-pres-lexicon.gz`
- scripts/
  - `extract_discr_lexicons.py`
  - `extract_not_discr_lexicons.py`
- `lang-id/`
- `evaluation/`

3 Step-by-step instructions

Create customized lexicons:

- $ `python extract_discr_lexicons.py`
- $ `python extract_not_discr_lexicons.py`

Annotate the Twitter corpus with linguistic variables

- $ `python extract_variables.py`

4 Components

4.1 hrsrTweets.gz

- Twitter corpus
- size: 1'350'101 tweets
- tweets with "guessed language" Serbian, Croatian or Bosnian: 738'589
- tweets with "guessed language" English: 226'706
- The corpus contains (tab-separated):
  1. tweet id (ex.: 463380928270962688)
  2. user (ex.: ugromir)
  3. time (ex.: 2014-05-05T18:13:17)
  4. "guessed" tweet language + certainty (ex.: sl:0.851)
  5. longitude (ex.: 21.9306105)
  6. latitude (ex.: 43.335324)
  7. tweet (ex.: @malibambi znas kakav sam, sta drugo da ti kazem)

4.2 apertium-hbs.hbs_HR_purist.mte.gz

- (currently work in progress)
- Morphologic lexicon of Croatian language

4.3 apertium-hbs.hbs_SR_purist.mte.gz

- Morphologic lexicon of Serbian language

4.4 scripts/

The python scripts in scripts/ produce customized lexicons which are later used in the extraction of linguistic variables from the Twitter corpus. Except for ch-lexicon.gz, the words in all the customized lexicons are stripped of diacritic marks.

- `extract_discr_lexicons.py`
  - extracts customized lexicons for discriminative features (either HR or SR)
  - idea: iterate the HR Apertium lexicon and when a specific feature is encountered (ex. the yat-reflex -ije- in "lijep") replace it with its opposition (ex. -e-: lijep -> lep). If the replaced version belongs to the same pos-tag and if it is contained in SR Apertium but not in HR one, add both words to the output list and mark their variables in the same file (ex. lijep (tabulator) je (newline) lep (tabulator) e)
  - outputs:
    - yat-lexicon.gz
    - diftong-v-lexicon.gz
    - hdrop-lexicon.gz
    - kh-lexicon.gz
    - st-c-lexicon.gz
    - ir-is-lexicon.gz
    - ir-ov-lexicon.gz
- `extract_not_discr_lexicons.py`
  - extracts customized lexicons for non-discriminative features (lexicons may contain same items for SR and HR)
  - idea: extract from Apertium lexicon all SR and HR words having the specific pos-tag (ex. verbs) or/and a specific string sequence (ex. ending -og in "starog") and save each list of words in a separated lexicon
  - outputs:
    - ch-lexicon.gz
    - verbs-lexicon.gz
    - verbs-inf-lexicon.gz
    - verbs-pres-lexicon.gz
    - verbs-vmf-lexicon.gz
    - genitiv-og-eg-lexicon.gz

4.5 custom-lexicons/

TABLE
custom lex
seize
structure

Table 1: Metainformation about lexicons

Lexicon	Size (tokens)	Manually added	Structure
ch-lexicon.gz	314915		token
diftong-v-lexicon.gz	932		token+var
genitiv-og-eg-lexicon.gz	64811		token
hdrop-lexicon.gz	322		token+var
hr-months.gz	84	x	token
int-months.gz	79	x	token
ir-is-lexicon.gz	6066		token+var
ir-ov-lexicon.gz	11846		token+var
kh-lexicon.gz	1026		token+var
rdrop-lexicon.gz	16	x	token+var
st-c-lexicon.gz	686		token+var
verbs-inf-lexicon.gz	18123		token
verbs-lexicon.gz	243538		token
verbs-pres-lexicon.gz	59354		token
verbs-vmf-lexicon.gz	47355		token
yat-lexicon.gz	105726		token+var

4.6 extract_variables.py

The functions in extract_variables.py take the tokenized tweet text as input, check if it contains the specific linguistic variable and output the related variable. The output variable is scalar. If a certain tweet contains two oppositions, the function returns "both". If it does not contain the specific variable, the respective function returns "NA".

- remove_diacritics(text)
- tokenize(text)
- clean(text): function for assigning the meta-information for each tweet. It returns:
  - automatic (I'm at Komercijalna Banka - @kombank (Novi Sad, Serbia))
  - English tweet (RT @RZual: RT if you would bend her over!)
  - noise website (‸̂http://t.co/FbXagAKvPd)
  - noise user (@OZKARBM :))
  - website (http://t.co/rBoVPHZ2rA)
  - is not alpha (!!!!!!1)
  - NA
- yat(text)
  It returns:
  - je (A sta bi sa onim" l**ije**pa njihova"?)
  - e (kolko mi se malopr**e** spavalo sad mi se uopste ne spava)

  - both (Bolje ovako poned**e**ljkom odradit prvu sm**e**nu zam**e**ne i vozdra c**ije**lu ned**e**lju spavam do sutra.)
  - NA

- kh(text)
  It returns:
  - h (Desiće se, valjda, da i ja jednom izađem iz **h**aosa.)
  - k (He, he, he! Dobro da nije s njim u suradnji, vijest bi bila u Crnoj **k**ronici! :D)
  - both (none found)
  - NA

- hdrop(text)
  It returns:
  - h (RT @AJBalkans: Tema: Nastanak i **h**istorijski razvoj selefizma)
  - h_drop (Škotlanđani čuje li se **i**storijsko DA , il je zapelo negde oko Dambartona .)
  - both (Viđet' mene u **h**aljinu je **i**storijski trenutak)
  - NA

- rdrop(text)
  It returns:
  - r (@MarijaZG jeo ih juče**r** :D)
  - r_drop (Svi hoće sve za juč**e**.)
  - both (none found)
  - NA

- st_c(text)
  It returns:
  - št (@cpcp89 a ko je to uop**št**e?)
  - ć (Dragi @Jutarnji u Rijeci ne pada kisa uop**c**e[...])
  - both (none found)
  - NA

- c_ch(text)
  It returns:
  - ć dev (Pij, imamo gdje povračati.)
  - č dev (Jedi,pij jogurt,žvaći burek)
  - both (ČAO MAĆKO KOJE SI GOD JESI LI ZA HOT SMS MMS GPS PGP RTS ILI NEŠTO VIŠE?)
  - NA

- diftong(text)
  It returns:

  - eu/au (Danas mi je **eu**ropski dan :-)[...])
  - ev/av (@stonexman na **ev**ropskom putu gasa.)
  - both (none found)
  - NA

- sa_s(text)
  It returns:
  - s dev (Vazda mi je san bio ono kad se nedje ide s autobusom **s s**kolom kad svi udju ja izadjem i prevrnem autobus)
  - sa dev (Licis mi na nekoga ko ima seks samo **sa u**gasenim svetlom.)
  - both
  - NA

- tko_ko(text)
  It returns:
  - tko (Kuca Mujo na vrata vidovnjaka. - **Tko** je? Šta **tko** je?! I ti si mi neki vidovnjak, pih!)
  - ko (**Ko** je rek'o kafa?)
  - both (neka me **neko** vodi 14og u Bl na parni valjak, moze ?" Evo i mi objavimo pa se mozda **netko** nadje :))
  - NA

  The function returns "tko" (resp. "ko") not only if the tweet contains "tko" but also "netko", "svatko", etc.

- sta_sto(text)
  It returns:
  - što
  - šta (Pa me sutra pitaj **šta** radim...ccc šta)
  - both (Na sva pitanja odgovaram sa 'eo'. Navika ili ne znam **sto** jee.)
  - NA

  "sto" is considered without diacritics: it means that tweets containing "sto" intended as number (100) may also be falsly marked

- da_je_li(text)
  It returns:
  - da li (O **da li** sam se ja to upravo odljubila)
  - je li (**jel** jos neko izvaljuje koliko sam ja nepismen na ovom tviteru ili samo ja?)
  - both (**Jel** zna neko **da li** radi sutra studentska ambulanta?)
  - NA

- usprkos(text)
  It returns:

  - usprkos (@MRenic a svi su nešto bolesni **usprkost** suncu i prekrasnom danu!)
  - uprkos (Dobro jutro,**uprkos** svim sranjima)
  - unatoč (nikad lošije izdanje na tenisu, **unatoč** pobjedi. ccc.)
  - NA

- treba_da(text)
  It returns:
  - treba da (Pijem kafu i kuliram a **treba** da ucim eee)
  - trebaX da (**Trebas** da prodjes ono najgore da dobijes ono najbolje)
  - both (Djeca **trebaju** da budu pioniri i **treba** da polažu zakletvu. Ima u tome nešto.)
  - NA

- bre(text)
  It returns:
  - bre (Kakav **bre** ses, sta mi tu glumite ludila i finocu!)
  - bolan (idi **bolan** umij se) (same for "bona")
  - ba (zadnje fizicko.. de **ba** nemoguce **ba**)
  - NA

- mnogo(text)
  It returns:
  - mnogo (Mada se ne razlikuje **mnogo** od stare)
  - puno (**Puno** bolje razumijem njemački nego ženski.)
  - vrlo (Evo slusam, **vrlo** pazljivo, i da definitivno glupost nema granice...)
  - jako (Jako sam lepa veceras. Vodi me na pivo i luk.)
  - NA

- months(text)
  It returns:
  - mnogo
  - HR months (Ljeto u **studenom**)
  - international months (Je li prosao 8. **mart** ?)
  - both (prvoga **maja** hladio sam jaja. sada prvoga **svibnja** muci nas krivnja ;) ak me razmete)
  - NA

- tjedan(text)
  It returns:
  - tjedan (U ova dva **tjedna** godisnjeg, danas se prvi put lijepo naspavala:))
  - sedmica (Ponedeljak je samo još jedan dan u **sedmici** ... do oktobra)

  - nedelja (I onda sta mi je pre ciniti kada odem posle 3 **nedelje**?!?)
  - nedjelja (Kao slag na tortu, white chicks na kraju ove dosadne **nedjelje**...)

  Possible problem: ambiguity ned(j)elja (sunday/week)

- drug(text)
  It returns:
  - drug (Hajde, Boze, budi **drug**, pa okreni jedan krug unazad planetu...)
  - prijatelj (Nina nedostajes mi, Nina vrati se.... Nina.... Moj jedini **prijatelju**...)
  - both (AnaMarija ti je **drugarica**? Mnogo fina i lepa **prijateljica**.)
  - NA

  All the declinated forms are considered

- inf_without_i(text)
  It returns:
  - inf with i (Plasim se da mu pisem, da ce me to jos vise povredit**i**)
  - inf without i (može a i ne mora **bit**')
  - both (Nikad necu moc**i** presta**t** da izgovaram to 'ah' na kraju rijeci)
  - NA

- synt_future(text)
  It returns:
  - synt inf (ne lajkaj mi sliku od prije 17 godina **ranicu** te)
  - nosynt inf (**Poludet cu** opet me napala upala mjehura :()
  - both (none found)
  - NA

- da(text)
  It returns:
  - da (presence of "da" in the tweet)
  - NA

- da_present(text)
  It returns:
  - da pres (Volela bih **da se udam** za @ademljajic Pazila bih ga i mazila...)
  - NA

  Returns "da pres" if "da" is followed by verb in present tense in the +2 window, and NA if there is no "da" in the tweet
  To be decided whether to return "da pres" only if precedeed by a modal verb (and with which context window).

- genitiva(text)
  It returns:

  - oga (Divan pocetak radn**oga** dana)
  - og (LadyAnjanas cijelu kitu gorsk**og** cvijeca a kamo li jednu ružu)
  - both
  - NA

  Endings -ega/eg are also returned as oga/og

- ir_is(text)
  It returns:
  - irati (pogledajte kako je defin**irana** pobkeda u 2. krugu)
  - isati (inspir**isan** si večeras)
  - both (U BiH malo šta **funkcioniše** kako treba I sve to bez obzira na enorman trud SNSD-a da BiH **profunkcionira**)
  - NA

- ir_ov(text)
  It returns:
  - irati (Ni prvu kavu nisam popio a vec kombiniram. Dobro jutro tl.)
  - ovati (Ako vas interesuje sta radim evo gledam bumbu sa sestrom od sest meseci)
  - both ("Šta jadni narod, jadni narod, narod treba da se organi**zira** kao što smo se organiz**ovali** prije pedeset godina i da se bori" - Džoni Štulić)
  - NA

- inf_verb_ratio(text) It returns:
  - verbs in infinitiv/all verbs in tweet (float) (*Onokad buraz krene postavljati sav beštek po špagi. :D [...]*: 0.5))

- cyrillic(text)
  It returns:
  - mix cyrillic (shožu s uma, mne malo malo malo teb)
  - sr cyrillic (none found (!))
  - latin

  Mix because different Cyrillic alphabets (not only Serbian) are present

4.7 lang-id/

- Contains manually annotated tweets according to their language
  - hr (Croatian)
  - sr (Serbian)
  - ba (Bosnian)
  - cg (Montenegrinian)

4.8 evaluation/

TBA

5 Further work

Other features we could consider:

- phonetic
  - Opposition -u/e (burza vs. berza)
  - Opposition -u/i (tanjur vs. tanjir)
  - Opposition -o/u (baron vs. barun)
  - Opposition -io/iju (milion vs. milijun)
  - Opposition -i/je after l/t (stjecaj vs. sticaj)
  - Opposition -s/z (inzistirati vs. insistirati)
  - Opposition -s/c (financije vs. finansije)
  - Opposition -t/ć (sretan vs. srećan)
  - Opposition -l/o after o (sol vs. so)
  - Opposition -v/h (muva vs. muha)
  - Opposition -cija/tija (diplomacija vs. diplomatija)
- morphologic
  - Opposition -kinja/ica (studentkinja vs. studentica)
  - Opposition -ka/ica (profesorka vs. profesorica)
  - Opposition -ac/telj (gledalac vs. gledatelj)
  - Suffix -a vs. no suffix (planeta vs. planet)
- lexical
  - toponomy (Španjolska vs. Španija)
  - kamo/kuda
  - hiljada/tisuća