Rule-based machine translation between Bulgarian and Macedonian

Tihomir Rangelov
University of Iceland

Second international workshop on free/open-source rule-based machine translation
Universitat Oberta de Catalunya, Barcelona

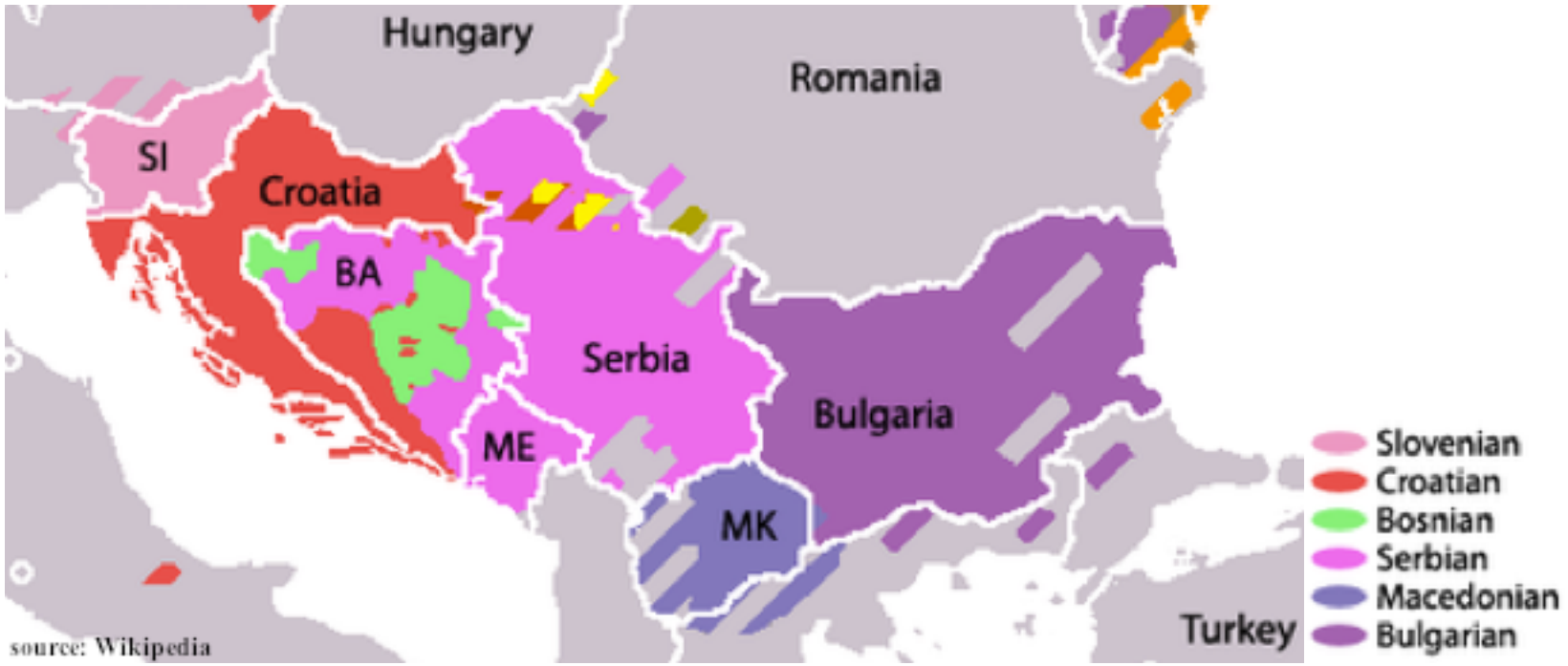
Contents

The Apertium platform

- An open-source platform for MT
- Uses linguistic information and rules
- Bilingual dictionaries for lexical transfer
- Morphological dictionaries for analysis/generation
- Rules for disambiguation and syntactic transfer
- 25 language pairs currently available

Bulgarian and Macedonian

- Comprise the eastern subgroup of the South Slavic languages
- Both new languages for Apertium



Available resources

- Bulgarian and Macedonian Wiktionaries
- SETimes Macedonian-Bulgarian parallel corpus
- Macedonian Reverse Dictionary
- Relatively few resources under open licences for Bulgarian and almost none for Macedonian

Lexical transfer

- The Macedonian-Bulgarian dictionary was created mostly manually
- Most frequent Macedonian words from the SETimes corpus
- Swadesh list
- Entries added semi-automatically include:
  - nouns and proper nouns from Wikipedia page titles
  - proper nouns from the frequency list, taking into account some spelling correspondences between Bulgarian and Macedonian (e.g. й = ј, ч = ќ)
  - loanwords, nouns derived with suffixes (-ност, -ство)

Analysis and generation

- A few paradigms already existed for Bulgarian in Wiktionary
- Most other work was done manually
- Some entries could be assigned paradigms semi-automatically:
  - nouns ending in -o (neuter), -a (feminine)
  - proper nouns
  - verbs with specific endings (e.g. -ува)

Disambiguation

- HMM POS tagger
- Fed the output from a Constraint Grammar module
  - very basic so far, only 41 rules for Macedonian and four rules for Bulgarian
  - priority was given to common homographs such as pronouns and forms of the auxiliary *to be*
  - `SELECT:r25 V-COP IF (0 (“<ce>")) (0 V-COP) (0 Pron) (1C A)` - selects the present-tense 3p. pl. form of the copular verb for the form *ce*, and ignores the tag for the homographic reflexive pronoun, when the token is followed by an adjective
  - `REMOVE:r10 Imprt IF (0 N) (0 Imprt) (-1C A)` - removes the imperative tag from a verb form when it coincides with a noun (usually derived from the same stem) if the preceding word is an adjective

Syntactic transfer

- Both Bulgarian and Macedonian have relatively free word order
- Small syntactic differences
- 33 rules for mk → bg
- 25 rules for bg → mk
- e.g. in adjective + noun concordances – a rule changes the gender of the adjective to correspond to that of the noun:

Macedonian		Bulgarian	
nov	aerodrom	novo	letište
new.MASC	airport.MASC	new.NEUT	airport.NEUT
‘new airport’			

Future in the past

- ‘I would arrive’
- MK:
  - ke pristignev
  - FUT.PART arrive.PAST.1P.SG
- BG:
  - štyah da pristigna
  - want.PAST.1P.SG to arrive.1P.SG

Present perfect tense

- ‘The cat has written’
- MK:
  - Mačkata ima pišuvano
  - Cat-THE have.PRES.3P.SG write.PAST.PASS.PART.NT.SG
- BG:
  - Kotkata e pisala
  - Cat-THE be.PRES.3P.SG write.PAST.ACT.PART.F.SG

Status

Module	No. Entries/Rules
Macedonian monolingual dictionary	8,693
Bulgarian monolingual dictionary	8,467
Bilingual dictionary	8,743
Transfer rules (mk→bg)	33
Transfer rules (bg→mk)	25
Macedonian CG rules	41
Bulgarian CG rules	4

Naïve vocabulary coverage

Corpus	Bulgarian	Macedonian
SETIMES	88.1%	92.1%
WIKIPEDIA	79.9%	-

Quantitative evaluation (mk → bg)

- Based on 57 sentences (1001 words) from twelve Wikipedia articles

Version	WER	PER
Apertium	25.31%	24.93%
Google	12.37%	12.17%

Comparative evaluation (mk → bg)

- Google Translate performs better than Apertium-mk-bg (for the time being)
- Google Translate sometimes inverts sentence meaning:
- [Macedonian original]:
  ... и слободно контактирајте со нас со свои реакции и сугестии.
  ... i slobodno kontaktirajte so nas so svoi reakciii sugestii.
  '... and feel free to contact us with your reactions and suggestions'
  [Google]:
  ... и колебайте да се свържете с нас чрез своите реакции и предложения.
  ... i kolebayte da se svăržete s nas črez svoite reakcii i predloženiya.
  '... and hesitate before contacting us with your reactions and suggestions'
  [Apertium]:
  ... и свободно *контактирајте с нас със свои реакции и *сугестии.
  ... i svobodno *kontaktirajte s nas săs svoi reakciii *sugestii.
  '... and feel free to *contact us with your reactions and *suggestions'

Qualitative evaluation (mk → bg)

Error type	number	percentage
dictionary coverage	131	49.81%
prepositions	14	5.32%
clitic pronouns	4	1.52%
definite article	7	2.66%
transfer (others)	38	14.45%
pronoun disambiguation	9	3.42%
disambiguation (others)	24	9.13%
multiword/idiom missing	4	1.52%
miscellaneous	32	12.17%

Qualitative evaluation (mk → bg)

- Clitic pronouns: same function, more widely used in Macedonian, differences in position (before or after the verb)
- Masculine definite article: in Bulgarian long form -ат/ят (-at/yat) for subjects or predicatives, short form -а/я (-a/ya) otherwise:
  - solved by a transfer rule
- Homographic pronouns in Macedonian:
  - personal and demonstrative: тоа (toa) “it/this”, тој (toj) “he/this one(masculine)”
  - CG rule (somewhat loose):
    - `REMOVE:r16 Pron + Dem IF (0 ("тој")) (NOT 1 N) (NOT -1 PREP) (NOT -1 V-COP) (NOT 1 CS)`

Future work

- Improve dictionary coverage
- Better and more CG rules
- Better syntactic rules (more thorough analysis of results)
- Adding frequent multi-word expressions and idioms, frequent prepositional phrases
- Evaluation for bg→ mk

Gràcies

- Благодаря!
- Благодарам!

- Въпроси? Прашања?