Rule-based machine translation between Bulgarian and Macedonian

Tihomir Rangelov
Department of Humanities
Háskóli Íslands
101 Reykjavík, Iceland
trr2@hi.is

Abstract

This paper describes the development of a two-way shallow-transfer rule-based machine translation system between Bulgarian and Macedonian. It gives an account of the resources and the methods used for constructing the system, including the development of monolingual and bilingual dictionaries, syntactic transfer rules and constraint grammars. An evaluation of the system's performance was carried out and compared to another commercially available MT system for the two languages. Some future work was suggested.

1 Introduction

Bulgarian and Macedonian comprise the Eastern group of the South Slavic languages and are therefore synchronically and diachronically very closely related. The Modern Bulgarian literary language was standardised in the 19th century based on dialects in the central and eastern part of northern Bulgaria. The Macedonian language was standardised in 1944 and recognised as the official language of the Republic of Macedonia. There is a high mutual intelligibility between Bulgarian[1] and Macedonian.[2] There are relatively limited linguistic resources under free licences for Bulgarian and almost none for Macedonian. Besides creating a machine translation solution, the purpose of the present project is to create open-source linguistic resources for the two languages. This paper gives an account of the creation of a shallow-transfer rule-based machine translation (MT) system for Bulgarian and Macedonian in the Apertium platform.[3] I will first describe the Apertium platform and Constraint Grammar (CG). Section 3 gives an overview of how the problems of lexical and syntactic transfer were handled. An evaluation of the system is made in the fourth section, and the final section contains a discussion of the results. It is important to point out that the project was initially intended as a unidirectional MT system for translation from Macedonian to Bulgarian. It was later expanded to a bidirectional system. However, the quality of translation from Macedonian to Bulgarian was given some priority.

2 Design

2.1 The Apertium platform

The Bulgarian-Macedonian language pair is integrated in the Apertium platform, which follows a shallow-transfer MT model. The system comprises three dictionaries written in XML. First of all, a bilingual dictionary is used for lexical transfer in both directions. One unique lexical item in one of the languages normally corresponds to another unique item in the other language. The two other dictionaries are monolingual (one for each language), and they are used for both morphological analysis and generation. These dictionaries

[1] Especially the western dialects

[2] http://www.lmp.ucla.edu/Profile.aspx?menu=004&LangID=42

[3] http://www.apertium.org

F. Sánchez-Martínez, J.A. Pérez-Ortiz (eds.)
Proceedings of the Second International Workshop on Free/Open-Source Rule-Based Machine Translation, p. 53–59
Barcelona, Spain, January 2011. http://hdl.handle.net/10609/5642

contain each lemma from the bilingual dictionary, along with morphological information in the form of paradigms. It is possible to constrain certain lexical items to solely generation or analysis.

After a word in the source language (SL) is analysed, it is passed through a Constraint Grammar module for partial disambiguation (cf. section 2.2). Then a Hidden Markov model (HMM) part-of-speech (POS) tagger performs further disambiguation before lexical and syntactic transfer and morphological generation to the target language (TL) is performed. Syntactic rules for both directions of transfer were also written in XML.

2.2 Constraint Grammar

The Bulgarian-Macedonian language pair makes use of a Constraint Grammar (CG) module for partial disambiguation. CG (Karlsson et al., 1995) is a method, which has proven very efficient for rule-based taggers for various languages. The module is fed a POS-tagged input, where each token can have more than one possible tag, and uses hand-written rules to remove or select a certain tag given some morphological information in the preceding or following parts of the sentence. CG rules can also be written to add syntactic tags.

3 Development

3.1 Resources

I have been able to use a number of free resources for constructing the three dictionaries, which comprise the language pair. However, a lot of manual work was needed to ensure good accuracy and completeness. Among these resources are the Bulgarian[4] and Macedonian[5] Wiktionaries, the SETimes Macedonian-Bulgarian parallel corpus (Tyers and Alperen, 2010) and the Macedonian Reverse Dictionary.[6]

3.2 Analysis and generation

Two monolingual dictionaries were constructed for the purposes of analysis and generation. Some paradigms already existed in the Bulgarian Wiktionary but they needed to be revised and many new ones needed to be added. Most of the work was done manually, but some semi-automatic methods were also used. For example in both languages nouns ending in -a are usually feminine nouns, those ending in -o are neuter, etc. Most of the proper nouns in the dictionaries were generated automatically from the corpus and assigned paradigms.

3.3 Disambiguation

The SL text is first tagged by the HMM POS tagger and it is passed together with all possible tags to the CG model. Currently the CG for Macedonian contains 41 disambiguation rules and a few rules, which add syntactic labels. I have tried to give priority to some very common homographs, such as pronouns and forms of the verb e "to be". For example the following rule:

```
  SELECT:r25 V-COP IF (0
("<ce>")) (0 V-COP) (0 Pron) (1C
A);
```

selects the present-tense 3-person plural form of the copula verb for the form ce, and ignores the tag which corresponds to the homographic reflexive pronoun, when the token is followed by an adjective. I also wrote rules for open class words such as the following:

```
  REMOVE:r10 Imprt IF (0 N) (0
Imprt) (-1C A);
```

which 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.

The Bulgarian CG currently has four rules.

The output from the CG module is fed to the HMM POS tagger, which was trained unsupervised on the SETimes corpus for each of the languages.

3.4 Lexical transfer

Lexical transfer is carried out with the help of the Macedonian-Bulgarian dictionary. It was built mostly manually. A frequency list for all tokens was made from the Macedonian SETimes corpus and entries were added to the dictionary accord-

[4] http://bg.wiktionary.org

[5] http://mk.wiktionary.org

[6] http://macedonia.auburn.edu/revdict/index.html

ing to it. Besides, all words from the Swadesh list[7] were also added. In some cases, where automatic generation of entries was possible, some less frequent entries were added. Automatic generation of entries was possible in the following cases:

- proper nouns and some terms from Wikipedia page titles;
- proper nouns from the frequency lists made from the Bulgarian and Macedonian SETimes corpora. I used information about an entry's frequency and took into account some spelling rules in the two languages, e.g. Bulgarian Николай corresponds to Macedonian Николај, Macedonian Чавиќ corresponds to Bulgarian Чавич, etc.;
- many pairs of loanwords and technical terms, as well as nouns derived from adjectives or verbs with the help of certain suffixes (-ност, -ство, -ние/ње etc.) were derived automatically from the corpora, as they tend to have a total correspondence in spelling or vary solely by a few regular spelling changes.

All of the entries, which were added automatically have been manually checked and modified or discarded where needed. Besides, in some cases so many pairs were generated that checking them manually would have been too time consuming for this stage of the project. I always gave priority to more frequent words.

Verb pairs were as a rule added manually and therefore only verbs at the top of the frequency list are included. However, as verbs in both languages were tagged for aspect, in many cases I have added verbs, which are not as high on the frequency list, so that pairs of perfective/imperfective verbs are formed.

3.5 Syntactic transfer

Bulgarian and Macedonian are very closely related languages and differences in syntactic structures on the phrase level are generally small. Although they are the only Slavic languages without extensive case marking of nouns, both languages generally have very free word order. To make up for that, both languages make extensive use of clitic pronouns (Koneski, 1967; Pashov, 2005). However, the rules for the use of these pronouns vary. Generally, their usage is much more frequent in Macedonian. Besides, their position within the sentence is more constrained in Bulgarian than in Macedonian. Some of these differences have been solved by syntactic rules for both directions of transfer.

Among other issues, which were dealt with through syntactic transfer rules are the following:

- Bulgarian masculine nouns, adjectives and pronouns have two different articles depending on the syntactic function of the word, the longer form -ат/ят (-at/yat) being used when the word is a subject or predicative, and the shorter form -а/я (-a/ya) otherwise (Pashov, 2005). Macedonian does not make this distinction and I have chosen the shorter form as a default form when Bulgarian is the TL simply because its usage is more frequent. A rule was written to change that into the longer form when following the verb съм (sam) "to be" (i.e. when it is a predicate) as in (1).

  (1) Tova e tvoyat glas.
  This is your.DEF.LONG voice.
  'This is your voice'

- In noun phrases consisting of an adjective/pronoun and a noun, the former are transferred with the wrong gender or number if the gender/number of the nouns does not correspond in Bulgarian and Macedonian. A simple rule was written, which transfers the gender and number of the TL noun to a preceding adjective or pronoun, see Macedonian in (2-a) and Bulgarian in (2-b).

  (2) a. nov aerodrom
  new.MASC airport.MASC
  'new airport'
  b. novo letište
  new.NEUT airport.NEUT
  'new airport'

[7] http://en.wiktionary.org/wiki/Appendix:Macedonian_Swadesh_list

- In compound future tense constructions (such as past future, future perfect and past future perfect tenses) Bulgarian uses different forms of the verb ща (shta) “want”, which behaves like an auxiliary and therefore carries information about the tense, whereas Macedonian uses the (indeclinable) particle ќе (ke) and the main verb carries information about the tense, e.g. in the past future construction in Maceodnian (3-a) and Bulgarian (3-b).

  (3) a. ke pristignev
  FUT.PART arrive.PAST.1P.SG
  ‘I would arrive’
  b. štyah da
  want.PAST.1P.SG to
  pristigna
  arrive.1P.SG
  ‘I would arrive’

- In Bulgarian the present perfect tense is always formed by a present-tense form of the auxiliary съм (sam) “to be” followed by a past active participle (so-called l-participle) in the appropriate gender and number. Besides this same form, Macedonian also uses a construction comprising a present-tense form of the auxiliary има (ima) “to have” followed by a neuter singular form of the past passive participle as in Macedonian (4-a) and Bulgarian (4-b).

  (4) a. toj ima
  he have.PRES.3P.SG
  pišuvano
  write.PAST.PASS.PART.NT.SG
  ‘He has written’
  b. toy e
  he be.PRES.1P.SG
  pisal
  write.PAST.ACT.PART.M.SG
  ‘He has written’

- One of the major differences between the conjugation of Bulgarian and Macedonian verbs is that Macedonian lacks a present active participle. This presents a problem when translating from Bulgarian to Macedonian. I have solved this problem for one of the most frequent usages of the present active participle - when it is used with a form of the auxiliary съм (sam) “to be” preceding it to mark a continuous action.

  (5) a. toy e
  he be.PRES.1P.SG
  hodešt
  walk.PRES.ACT.PART.M.SG
  ‘He is walking’
  b. toj odi
  he walk.PRES.3P.SG
  ‘He walks’

3.6 Status

The current number of entries in all the dictionaries and the number of syntactic rules and CG grammar rules are shown in table 1.

4 Evaluation

This section presents an evaluation of the performance of the system, including its vocabulary coverage against some available corpora, a quantitative evaluation of errors, and a comparison of its performance with another commercially available system for MT between Bulgarian and Macedonian. A quantitative evaluation has so far been made only for translation from Macedonian to Bulgarian. I also give a short account of the most common errors made by the system at this stage.

4.1 Coverage

Vocabulary coverage numbers are given in Table 2. Coverage here means that for any given form in the SL, at least one analysis is returned by the system (naïve coverage). The system was tested on two corpora for Bulgarian (SETimes and Wikipedia) and on the SETimes corpus for Macedonian. Unsurprisingly, the numbers are highest for the coverage of the Macedonian SETimes corpus, as a frequency list based on it was used to construct the dictionaries.

4.2 Quantitative evaluation

The quantitative evaluation of the system was performed by selecting 57 sentences (1001 words) from twelve articles on diverse topics in the Macedonian Wikipedia, translating them with the

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

Table 1: Status of the Apertium-mk-bg pair as of November 21, 2010, version 0.2.0.

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

Table 2: Naïve coverage of the Apertium-mk-bg pair as of November 2010, version 0.2.0.

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

Table 3: Quantitative evaluation of the Apertium-mk-bg pair as of December 19, 2010, version 0.2.0, as opposed to Google Translate as of December 20, 2010. The evaluation has been performed only for translation from Macedonian to Bulgarian on 57 sentences from various Wikipedia articles.

system and then editing the translations manually. The word error rate (WER) and the position-independent error rate (PER) were calculated by the number of changes, which needed to be made to the translated text by a human post editor. Results are shown in table 3.

4.3 Comparative evaluation

As far as I know the only other available system for MT between Bulgarian and Macedonian is Google Translate,[8] which is a statistical MT system. To compare the results of the Apertium system against those of Google Translate, I made a quantitative evaluation of Google Translate's module for translation from Macedonian to Bulgarian on the same 57 sentences from the Macedonian Wikipedia, using the same method. The results are shown in table 3.

The results from the evaluation show that Google Translate performs better than Apertium.

[8] `http://translate.google.com`

Before using the Wikipedia corpus for quantitative evaluation, an evaluation of both Apertium-mk-bg and Google Translate was made against a similarly large set of random sentences from the SETimes corpus. Those results were slightly better for both systems.[9] I found out, however, that Google Translate seems to have used at least a part of the SETimes corpus for training because a few sentences were translated exactly as the corresponding Bulgarian sentence in the parallel SETimes corpus, despite some idioms or non-straightforward word order or choice of words. Besides, the SETimes corpus was also used for creating the Apertium-mk-bg system, so I decided to use the Wikipedia corpus as it was more independent. However, one important observation from the evaluation involving the SETimes corpus is that there was at least one case where Google Translate inverted sentence meaning (6-b), while Apertium did not (same meaning as original, despite the two unknown words, marked with asterisks) (6-c).

4.4 Qualitative evaluation

The evaluation in the previous subsection returned 263 errors for the Apertium-mk-bg system out of the 1001 words in the test corpus. These errors were of different types and table 4 shows the number of errors for a few different categories. Half of all errors were due to words missing in the dictionaries.

The second most common type of errors were those, which can most often be attributed to syntactic transfer. They include errors related to prepositions, clitic pronouns and definite articles, among others (cf. table 4). The use of the clitic

[9] All numbers were by 2-5 percentage points lower than the ones in table 3.

(6) a. ... и слободно контактирајте со нас со свои реакции и сугестии.
... i slobodno kontaktirajte so nas so svoi reakcii i sugestii.
'... and feel free to contact us with your reactions and suggestions' [Macedonian original]

b. ... 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' [Google]

c. ... и свободно *контактирајте с нас със свои реакции и *сугестии.
... i svobodno *kontaktirajte s nas săs svoi reakcii i *sugestii.
'... and feel free to contact us with your reactions and suggestions' [Apertium]

pronouns differs in the two languages (cf. section 3.5) and I have written some transfer rules for changing their position or simply deleting them where needed. Prepositions in some idioms or collocations, where the preposition differs in the two languages, are not always correctly translated. These errors can be fixed by adding and improving transfer rules and possibly multiword expressions. Besides, the transfer rule for the definite article in Bulgarian (cf. section 3.5) needs to be revised and new rules can be added to fix some related errors.

The third most common error type were disambiguation errors, which can be fixed with more CG rules. Homographic pesonal and demonstrative pronouns (тоа (toa) "it/this", тој (toj) "he/this one(masculine)"), and interrogative and relative pronouns (кој (koj) "who/whose") in Macedonian, which have seperate counterparts in Bulgarian also need special attention.

A few errors could also be fixed by adding more multiword expressions and idioms to the dictionaries.

The analysis of the evaluation also showed that there were 69 free rides (6.89% of all tokens), i.e. word forms not covered by the dictionaries, but considered correct translations as they have exact correspondence in Bulgarian and Macedonian.

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%

Table 4: Quantitative analysis of the 263 errors in the evaluation of the Apertium-mk-bg pair against a corpus of 1001 words from twelve Wikipedia articles.

5 Discussion

This paper presented the design of an MT system for Bulgarian and Macedonian in the Apertium platform. The results are generally good and require a realistic amount of post-editing. An evaluation of the translation from Bulgarian to Macedonian will also need to be made. Currently the system performs worse than its competitor, but I believe that its performance can be improved with future work, especially on dictionary coverage and transfer rules. A deeper analysis of the current results will be needed in order to expand the CG modules, which I believe could contribute significantly to the accuracy. Besides, the CG's for both languages, as well as the other resources created for this system could be of great value for other open-source language-technology related solutions for Bulgarian and Macedonian.

Acknowledgements

This work would not have been possible without the assistance and devotion of Francis Tyers. Many thanks to Igor Kalkov for answering endless questions on Macedonian grammar, as well as to Rosen Vladimirov for responding to questions and to Prof. George Mitrevski at Auburn University for providing his Macedonian Reverse Dictionary. Thanks also to the publishers of the newspaper "Narodna volya" in Blagoevgrad, Bulgaria, for providing a copy of their Macedonian-Bulgarian dictionary, and especially to Mr. Hristov.

References

Karlsson, F., Voutilainen, A., Heikkilä, J., and Anttila, A. (1995). *Constraint Grammar: A language independent system for parsing unrestricted text*. Mouton de Gruyter.

Koneski, B. (1967). *Gramatika na makedonskiot literaturen jazik*. Kultura.

Pashov, P. (2005). *Balgarska gramatika*. Hermes.

Tyers, F. M. and Alperen, M. S. (2010). South-east european times: A parallel corpus of the balkan languages.