Towards 100 M Corpus of Tajik

Gulshan Dovudov, Vít Suchomel, Pavel Šmerk

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

8. 12. 2012

“History”

- a need for real world test data for Gulshan’s Tajik MA and other tools
- the first data taken mainly from ozodi.org and other news sources
  - ozodi.org is the Tajik version of RFE/RL broadcasted from Prague
- then we wanted to compare the “manually” crafted corpus with results of SpiderLing crawler & co.
- — and the result was and is by far the largest corpus of Tajik

Tajik Language

- a variant of Persian Language spoken mainly in Tajikistan
  - Indo-European language
  - ca. 5 M speakers
- unlike Iranian Persian, TP uses a bit extended Cyrillic alphabet
  - extra characters has a little software support
  - people use e.g. Belarussian Short U ў instead of proper Cyrillic U with macron ӯ
    - (the former is from cp1251)
    - some of these cases are easy to repair
  - or they write "without diacritics": they use the most simillar Russian character
    - e.g. х instead of ҳ, but х is also in Tajik
  - or they even write in Latin

Computer Corpora of Tajik Language

- the biggest planned: Tajik Academy of Sciences
  - 10 M words
  - collection of works (mainly poetry) of notable Tajik writers
    - even from the 13th century
- the biggest existing: within Leipzig Corpora Collection
  - 100 000 Tajik sentences, ca. 1.8 M words
  - source is ozodi.org
  - automatically crafted $\Rightarrow$ many problems with "encoding"
    - 5 % of sentences are in Latin script
    - 10 % seem to use Russian characters
    - 1 % uses non-Tajik Cyrillic characters

The Current State of the Corpus

- only Internet sources
  - we had to distinguish Russian and Tajik texts
- semi-automatically crafted part
  - set of Perl scripts, slightly modified for each site
    - news/media portals, books from gazeta.tj (prose only), ...
- automatically crawled part: CorpusFactory, SpiderLing, onion, ...
- now
  - 85 056 058 tokens, 58 M semi-automatically, 27 M automatically
  - 70 757 996 words (Cyrillic characters only), ca 92.5 % known to MA
  - growth of available data (automatically crawled part):

date	tokens	increment	per month
11/2011	34.6 M	—	—
03/2012	41.1 M	+18.6 %	4.7 %
05/2012	44.7 M	+8.6 %	4.3 %
11/2012	54.8 M	+22.6 %	3.8 %

Dealing with Texts of a Lower Quality

- six Tajik letters are missing in the most widespread encoding cp1251
- missing support for Tajik in Windows (and probably other major OSs?)
- people have to use some sets of replacements, "hand-made" fonts etc.:

Char	RS1	RS2	RS3	RS4
Ғғ	Ѓѓ	Љљ	Uu	Гг
Ӣӣ	Її	Њњ	Bb	Ии
Ҷҷ	Љљ	Її	Xx	Чч
Ҳҳ	Њњ	Ii	{[	Хх
Ққ	Ќќ	Ss	Rr	Кк
Ӯӯ	Ўў	Ўў	Ee	Уу

- Љљ, Њњ, and Її are shared by RS1/RS2 but have different meanings
  - we have to use MA and guess the whole RS for the document
- sometimes people use й instead of ӣ and х, к, or ч, instead of ҳқҷ

Dealing with Texts of a Lower Quality

- out of 192 664 documents 21,524 (12.2 %) need some changes
  - (document == web page, news article, . . .)
- 2391 documents use RS1, 1323 RS4, 778 RS3, and 113 RS2
  - ca 100 documents use some mix
- 859 641 words were modified, i. e. more than 1 % of all words
- types of changes and numbers of affected words are described in each document's metadata which allows users to create specific subcorpora
  - e. g. a subcorpus of texts without any changes
    - probably the most quality ones

(Near) Future Work

- 100 M words
- better tokenization ("some phrase"-po etc.) and RS detection
- better morphological analysis coverage (enlarging the lexicon)
- complex verbs and nominal phrases detection and annotation
  - corpus manager support?
- POS tagging, tagging
- word sketches