Building a 50M Corpus of Tajik Language

Gulshan Dovudov, Jan Pomikálek, Vít Suchomel, Pavel Šmerk

Natural Language Processing Centre
Faculty of Informatics
Masaryk University

http://nlp.fi.muni.cz/en

3. 12. 2011

“History”

- a need of real world test data for Gulshan’s (prepared) Tajik morphological analyser
- the first data were taken mainly from ozodi.org and some 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 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)
    - 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

New 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
    - ozodi.org 12 M words, gazeta.tj 6 M, bbc.co.uk 4 M, ...
  - some books from gazeta.tj archive (prose only)
- automatically crawled part: CorpusFactory, SpiderLing, onion, ...

TLD	docs downloaded	docs accepted in the corpus
tj	55.0 %	51.7 %
com	23.0 %	28.1 %
uk	8.9 %	7.2 %
org	6.8 %	7.7 %
ru	6.2 %	5.3 %

Some Numbers

	docs	words	w/doc	tokens	MB
semi-aut.	102383	34145907	334	40571607	480
automatic	61523	28841537	469	34680994	405
aut. contr.	31946	17622897	552	21372272	242
all	**134329**	**51768804**	**385**	**61943879**	**721**

- surprise: fully automated crawling yielded smaller data than semi-automated one
  - ca. 25 % were inaccessible for SpiderLing crawler
    - RAR compressed .docs, BBC articles not linked from anywhere
  - even after discounting these data, results are incomparable with e.g Czech — more than 5 billion tokens
  - $\Rightarrow$ we (some of us :-) believe that we almost reach the overall potential of internet resources