POS Annotated 50M Corpus of Tajik Language

Gulshan Dovudov, Vít Suchomel, Pavel Šmerk

Natural Language Processing Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

zarif_dovudov@mail.ru, xsuchom2@fi.muni.cz, smerk@mail.muni.cz

Abstract

We present by far the largest available computer corpus of Tajik language of the size of more than 50 million words. To obtain the texts for the corpus two different approaches were used. We also developed morphological analyzer of Tajik and here we offer some statistics of its application on the corpus.

1. Tajik Language and Corpora

Tajik language:

- variant of Persian spoken mainly in Tajikistan and written mostly in the Cyrillic alphabet;
- since the Tajik language internet society (and the potential market) is small, available NLP tools and resources for Tajik as well as publications in the field are rather scarce.

Computer corpora of Tajik:

- either small or still only in development;
- Leipzig Corpora Collection [2] offers the biggest and the only freely available corpus with 100 k sentences (almost 1.8 M words)[1];
- The Tajik Academy of Sciences prepares a corpus of 10 M words (`www.cit.tj`), but at least by now it is not a corpus of contemporary Tajik, but a collection of works—moreover mainly a poetry—of a few notable Tajik writers (even from the 13th century).

2. New Corpus of Tajik

A new corpus of contemporary Tajik:

- more than 50 million words;
- all texts were taken from the internet;
- two different approaches to obtain the data.

Semi-automatically crawled part:

- we crawled around a dozen Tajik news portals;
- each portal was processed separately to get maximum of relevant (meta)information and the utmost clean text data;
- this part of the corpus is supposed to contain data of a higher quality.

Automatically crawled part:

- we used SpiderLing crawler which combines
  - collecting seed URLs with Corpus Factory,
  - character encoding detection tool chared,
  - general language detection based on a trigram model trained on Wikipedia articles,
  - boilerplate removal tool jusText,
  - deduplication on the paragraph level;
- possibly contains data of a lower quality.

The two parts were joined and deduplicated. As a result we obtained a corpus of more than 50 M words, i.e. corpus positions which consists solely of Tajik characters, and more than 60 M tokens, i.e. all words, interpunction, numbers etc. Detailed numbers follow:

source	docs	words	tokens
ozodi.org	59943	13426445	15738683
gazeta.tj archive	480	5006432	6031951
bbc.co.uk	9288	4129179	4772807
jumhuriyat.tj	8106	3703685	4397650
*.wordpress.com	3080	3235436	3946319
tojnews.org	9653	2532572	3077917
khovar.tj	17079	2512232	3082293
millat.tj	2803	2268000	2673004
gazeta.tj	2209	1389318	1665672
kemyaesaadat.com	1863	1182353	1404072
	...		
all	138701	**51722009**	61837585

3. Morphological Analysis of Tajik

A new morphological analyzer of Tajik:

- an analyzer of Tajik already exists [3], but is not usable for corpus annotation for it is too slow, non-portable and cannot offer a lemma;
- we extracted (partially) the information about Tajik morphs from the existing analyzer;
- we used Jan Daciuk's approach [1]:
  - data are triplets *word:lemma:tag*,
  - lemma is encoded: `kardem:Can:tag` says "delete 2 (A=0,B=1,...) chars and add `an`"[2],
  - such data form a finite formal language,
  - Daciuk's tools create a minimal automaton,
  - analysis is trivial (the C++ code can have <400 lines) passing through the automaton,
  - analysis is thus also very fast;
- data are generated from a dictionary of base forms according to a simple (80 lines) description of Tajik morphology — both the dictionary and the description are to be enriched, as this work is still at the beginning;
- some statistics of the new analyzer data follow in the tables:

count of word+lemma+tag triplets	8,476,108
size of the input data in bytes	175,845,264
size of the automaton in bytes	1,138,480
bytes per line of the input data	0.13
count of lemmata in the dictionary	14934
average number of triplets per lemma	568

Meaning	Tag	# of forms
nouns	01	6267182
adjectives	02	941209
numerals	03	25572
pronouns	04	52
verbs	05	372778
infinitives	06	646500
adjectival participles	07	217273
adverbial participles	08	5253
adverbs	09	86
prepositions	10	44
	...	

4. Annotation of the Corpus

For now, we annotate only lemma and POS, as the rest of the information is currently not in a fully consistent state. The analyzer recognizes 87.2 % of words and 25.6 % of them are ambiguous. Table shows the top 10 most frequent word forms, their analyses and frequency:

dar	dar:01;dar:05;dar:10	1626855
ba	ba:10	1572867
va	va:12	1417227
ki	ki:04;ki:12	1226404
az	az:10	1173474
in	in:04;in:14	773985
bo	bo:10	513154
ast	ast:05	347578
on	on:04	301493
Tojikiston	Tojikiston:01	281627

The corpus is accessible through the Sketch Engine on `http://ske.fi.muni.cz/open/`.

References

[1] J. Daciuk. *Incremental Construction of Finite-State Automata and Transducers, and their Use in the Natural Language Processing*. PhD thesis, Technical University of Gdańsk, 1998.

[2] U. Quasthoff, M. Richter, and C. Biemann. Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the Fifth International Conference LREC 2006*, Genoa, 2006.

[3] Z. D. Usmanov, G. M. Dovudov, and O. M. Soliev. Таджикский компьютерный морфоанализатор. National patent ZI-03.2.220, National Patent Information Centre, Tajikistan, 2011.

[1] The encoding/transliteration varies greatly: more than 5 % of sentences are in Latin script, almost 10 % seem to use Russian characters instead of Tajik specific characters (e.g. х instead of Tajik ҳ, which sound/letter does not exist in Russian) and more than 1 % uses non-Russian substitutes for Tajik specific characters (e.g. Belarussian ў instead of proper Tajik ӯ) — and only the last case is easy to repair automatically.

[2] This example would work only for suffixes. To handle also the prefixes, it is possible to employ the same principle again, for example: `namekardem:ECan:tag`, where the `E` and `Can` denotes that to get the correct lemma kardan the first four and the last two letters are to be deleted and `an` is to be added to the end (and nothing is to be added to the beginning).