Building a 50M Corpus of Tajik Language

Gulshan Dovudov[1], Jan Pomikálek[2], Vít Suchomel[1], Pavel Šmerk[1]

[1] Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xdovudov,xsuchom2,xsmerk}@fi.muni.cz

[2] Lexical Computing Ltd.
73 Freshfield Road, Brighton, UK
jan.pomikalek@sketchengine.co.uk

Abstract. Paper presents 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. The paper brings a description of both of them, discusses their advantages and disadvantages and shows some statistics of the two respective partial corpora. Then the paper characterizes the resulting joined corpus and finally discusses some possible future improvements.

1 Introduction

Tajik language is a variant of the Persian language spoken mainly in Tajikistan and also in some few other parts of Central Asia. Unlike Iranian Persian, Tajik is written in the Cyrillic alphabet.

Since the Tajik language internet society (and consequently the potential market) is rather small and Tajikistan itself is ranked among developing countries, there are almost no resources for computational processing of Tajik at all. Namely the computer corpora of Tajik language are either small or even still only in the stage of planning or development. The University of Helsinky offers a very small corpus of 87 654 words.[1] Megerdoomian and Parvaz [3,2] mention a test corpus of approximately 500 000 words, and the biggest and the only freely available corpus is offered within the Leipzig Corpora Collection project [4] and consists of 100 000 Tajik sentences which equals to almost 1.8 million words.[2] Iranian linguist Hamid Hassani is said to be preparing a 1

[1] http://www.ling.helsinki.fi/uhlcs/readme-all/README-indo-european-lgs.html

[2] Unfortunately, the encoding and/or transliteration vary greatly: more than 5 % of sentences are in Latin script, almost 10 % of sentences seem to use Russian characters instead of Tajik specific characters (e.g. kh instead of Tajik h, which sound/letter does not exist in Russian) and more than 1 % of sentences uses other than Russian substitutes to Tajik specific characters (e.g. Belarussian short u instead of proper Cyrillic u with macron) — and only the last case is easy to repair automatically.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pp. 89–95, 2011.

million words Tajik corpus[3] and Tajik Academy of Sciences prepares a corpus of 10 million words[4]. Unfortunately, at least by now the latter is not a corpus of contemporary Tajik, but rather a collection of works—and moreover mainly a poetry—of a few notable Tajik writers (one of them is even from the 13th century).

In this paper we present a newly built corpus of contemporary Tajik language of more than 50 million words. All texts were taken from the internet. We used two different approaches to obtain the data for the corpus and we describe these methods and their results in the following two sections. In the Section 4 we present the resulting corpus and finally we discuss some possible future improvements.

2 Semi-automatically Crawled Part

The first part of the corpus was collected by crawling several news portals in Tajik language.[5] If the articles of a particular portal were numbered, then we tried to download all possible numbers, otherwise we get a list of articles either directly from the portal, or from the Google cache. Each single portal was processed separately to get the maximum of relevant (meta)information, i.e. correct headings, publication date, rubric etc.

Table 1. Statistics of the semi-automatically crawled part of the corpus.

source	docs	pars	words	w/doc	tokens	MB
ozodi.org	54301	343057	12037886	222	14073673	170
gazeta.tj archive	480	163572	5006650	10431	6032214	67
bbc.co.uk	8032	147073	3694769	460	4270087	51
jumhuriyat.tj	7598	98102	3485592	459	4134782	50
khovar.tj	15420	62280	2323445	151	2835757	36
tojnews.org	8377	64583	2254917	269	2740760	33
millat.tj	2373	44710	2004863	845	2361816	27
gazeta.tj	1608	31962	1130957	703	1357889	15
gazeta.tj library	130	98698	1053262	8102	1358584	15
pressa.tj	2165	20526	642394	297	782917	9
news.tj	1396	7759	269267	193	325164	4
muhabbatvaoila.tj	503	9957	241905	481	297964	3
all	**102383**	**1092279**	**34145907**	**334**	**40571607**	**480**

[3] http://www.tajikistan.orexca.com/tajik_language.shtml
http://en.wikipedia.org/wiki/Hamid_Hassani

[4] http://www.termcom.tj/index.php?menu=bases&page=index3&lang=eng (in Russian)

[5] Paradoxically, the two largest Tajik news portals are not located in Tajikistan, but in Czech Republic (ozodi.org, Tajik version of Radio Free Europe/Radio Liberty) and United Kingdom (bbc.co.uk, Tajik version of BBC).

In the Table 1 we present some statistics of the obtained data. Column docs contains a number of documents downloaded from the given source, pars is a number of paragraphs (including headings), words is a number of tokens which contain only characters from Tajik alphabet, w/doc is a words / document ratio (i.e. average length of possibly continuous texts), tokens is a number of all tokens (words, interpunction etc.) and MB is the size in megabytes of the data in vertical corpus format (i.e. plain text). The table is sorted by number of words. From the electronic library on `gazeta.tj` we choose only prose and omit all more structured texts as poetry, drama or e.g. computer manual. The articles in `gazeta.tj` archive are joined in one file on a weekly basis and that is why the words / document ratio is so high.

On allmost all websites, alongside the articles in Tajik there were also many articles in Russian. Because both alphabets, Tajik and Russian, contain characters which do not occur in the other alphabet, it is easy to distinguish between the two languages and discard the Russian articles even without any further language analysis.

3 Automatically Crawled Part

SpiderLing, a web crawler for text corpora [5], was used to automatically download documents in Tajik from the web. We started the crawl using 2570 seed URLs (from 475 distinct domains) collected with Corpus Factory [1].

The crawler downloaded 9.5 GB of HTML data in ca. 300,000 documents over three days. That is not much compared to crawling documents in other languages by the same tool. For example the newly built web corpus of Czech, which has roughly only two times more native speakers compared to Tajik, has more than 5 billion words, of course not in three days. We conclude the available online resources in Tajik are very scarce indeed. An overview of internet top level domains of URLs of the documents obtained can be found in Table 2.

Table 2. Number of documents by internet top level domain

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	2.6 %	1.4 %
ir	1.6 %	2.4 %
other	2.0 %	1.5 %

Since Russian is widely used in government and business in Tajikistan (and other language texts may appear), 33 % of the downloaded HTML pages were

removed by the SpiderLing's inbuilt language filter [5]. The obtained data was tokenized and deduplicated using Onion[6] with moderately strict settings[7]. Some statistics of the automatically crawled part of the corpus are in the Table 3 (only the ten most productive sources of data are detailed).

Table 3. Statistics of the automatically crawled part of the corpus.

source	docs	pars	words	w/doc	tokens	MB
*.wordpress.com	3385	68228	4007147	1184	4890924	55
bbc.co.uk	5238	97897	2556034	488	2939833	37
ozodi.org	6847	83955	2257307	330	2691184	33
khovar.tj	10716	30275	1801314	168	2202621	29
gazeta.tj	1655	8122	1110378	671	1334125	15
millat.tj	1381	20843	1108658	803	1299469	15
*.blogspot.com	800	18272	974842	1219	1208834	13
ruzgor.tj	1473	17575	943143	640	1084123	11
firdavsi.com	588	22433	857906	1459	1029023	11
pressa.tj	2312	13530	655840	284	792631	10
			...			
all	**61523**	**612178**	**28841537**	**469**	**34680994**	**405**

4 Corpus of Tajik Language

The two partial corpora were joined together and the result was deduplicated using Onion. We obtained a corpus of more than 50 million words, i.e. corpus positions which consists solely of Tajik characters, and more than 60 million tokens, i.e. words, interpunction, numbers etc. Detailed numbers follow in the Table 4.

It was rather surprising for us, that the fully automated crawling yielded even smaller data than the semi-automated approach. It has to be said, that at least 25 % of semi-automatically crawled data were inaccessible to the general crawler, as it cannot extract texts from RAR-compressed archives (gazeta.tj archive and library) and because it does not seem to exist any link to bigger part of older BBC articles although they remained on the server (we exploited Google cache to get the links). It is higly probable that also the other sites contain articles unreachable by any link chain and thus inaccessible for the general crawler. But even if we discount these data, the automated crawling did not outperform the semi-automated one in such an extent that we expected and which is common for many other languages. As we remarked in the Section 3, we attribute it to the scarceness of online texts in Tajik language. It means that we probably reach or almost reach the overall potential of internet resources,

[6] `http://code.google.com/p/onion/`

[7] removing paragraphs with more than 50 % of duplicate 7-tuples of words

Table 4. Statistics of the resulting corpus.

source	docs	pars	words	w/doc	tokens	MB
ozodi.org	56181	378790	12964584	231	15198491	183
gazeta.tj archive	480	163572	5006650	10431	6032214	67
bbc.co.uk	8188	147564	3706224	453	4283839	51
*.wordpress.com	3181	58604	3616710	1137	4416611	50
jumhuriyat.tj	7599	98141	3489592	459	4139709	50
khovar.tj	16554	66129	2502796	151	3048777	38
tojnews.org	8426	64770	2259311	268	2746100	33
millat.tj	2687	47823	2172612	809	2557686	29
gazeta.tj	1894	32537	1170516	618	1403929	16
gazeta.tj library	130	98698	1053262	8102	1358584	15
			...			
all	**134329**	**1430896**	**51768804**	**385**	**61943879**	**721**

i.e. even if we somehow get all Tajik online texts, the corpus might be bigger by half, but surely not for example ten times or even three times.

The new corpus is not freely available for a download at the moment, but eventual interested researchers can be given an access to it through the Sketch Engine[8].

5 Future Work

The Table 5 shows statistics of the texts which were new in the automatically crawled part compared to the semi-automatically crawled data. The numbers indicate that there is a room for an extension of the semi-automated part. We will prepare specialized scripts for the most productive portals to download the data in a some more controlled way.

It is worth clarifying the case of ozodi.org. The general crawler tries to get all reasonable texts on the page, which, on the news portals, may include the readers' comments. On the other hand, because the comments may contain a substandard language features, they were omitted during the semi-automated crawling. Thus the 1880 documents from ozodi.org were not some newly added ones, but they were results of the deduplication which discarded the article itself and leaved only the comments as it processes corpus by single paragraphs. This is also one of the reasons why we prefer the semi-automated crawling when it is possible: we want to tag these comments to allow a creation of subcorpora of a (presumably standard) language of articles as well as of a language of comments.

Another problem with the comments—but not only with them—is a common absence of Tajik-specific characters. The language model for the general

[8] http://www.sketchengine.co.uk/

Table 5. The contribution of automatically crawled part

source	docs	pars	words	w/doc	tokens	MB
*.wordpress.com	3181	58604	3616710	1137	4416611	50
ozodi.org	1880	35733	926698	493	1124818	13
*.blogspot.com	752	15462	841228	1119	1049364	12
ruzgor.tj	1399	13281	709124	507	814561	8
firdavsi.com	472	19624	684496	1450	820569	9
khatlonpress.tj	870	7983	541402	622	628142	6
kemyaesaadat.com	439	12008	523194	1192	631801	7
ucoz.ru	742	9068	494783	667	596656	7
nahzat.tj	2435	5478	465748	191	545995	7
ozodagon.com	1962	6920	450334	230	542911	7
			...			
all	**31946**	**338617**	**17622897**	**552**	**21372272**	**242**

crawler was trained using Tajik Wikipedia[9] of our corpus to make the crawler download texts in language which looks like the language of Tajik Wikipedia. Unfortunately, in many Wikipedia articles the Tajic-specific characters are replaced by some other characters. The unambiguous replacements were trivially repaired in the whole corpus, but e.g. Cyrillic kh can sometimes stand either for Tajic-specific h or also for kh itself. On the one hand we plan to tag such texts to allow a creation of subcorpora with or without them, on the other hand we want to develop a program which would be able to repair them. We will also train the language model with another sets of texts to see how it will affect the crawled data.

Acknowledgements This work has been partly supported by Erasmus Mundus Action II lot 9: Partnerships with Third Country higher education institutions and scholarships for mobility, and by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Kilgarriff, A., Reddy, S., Pomikálek, J., PVS, A.: A Corpus Factory for Many Languages. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010. Valleta, Malta (2010)
2. Megerdoomian, K.: Language Engineering for Lesser-Studied Languages, chap. Low-density Language Strategies for Persian and Armenian, pp. 291–312. IOS Press, Amsterdam (2009)

[9] The use of Wikipedia to train the language model is a part of default settings or a default scenario of the process of building corpora for new languages without any other utilizable resources. Of course we have better Tajik texts at hand, but the automatically crawled part of our corpus had also to act as a test of a general suitability of our technologies for the case of building corpora for low-density languages.

3. Megerdoomian, K., Parvaz, D.: Low-density Language Bootstrapping: The Case of Tajiki Persian. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008. Marrakech, Morocco (2008)
4. Quasthoff, U., Richter, M., Biemann, C.: Corpus Portal for Search in Monolingual Corpora. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006. Genoa (2006)
5. Suchomel, V., Pomikálek, J.: Practical Web Crawling for Text Corpora. In: Proceedings of the Fifth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2011. Masaryk University, Brno (2011)