Large Corpora for Turkic Languages and Unsupervised Morphological Analysis

Vít Baisa, Vít Suchomel

Natural Language Processing Centre
Masaryk University, Brno, Czech Republic
{xbaisa, xsuchom2}@fi.muni.cz

Abstract

In this article we describe six new web corpora for Turkish, Azerbaijani, Kazakh, Turkmen, Kyrgyz and Uzbek languages. The data for these corpora was automatically crawled from the web by SpiderLing. Only minimal knowledge of these languages was required to obtain the data in raw form. Corpora are tokenized only since morphological analyzers and disambiguators for these languages are not available (except for Turkish). Subsequent experiment with unsupervised morphological segmentation was carried out on the Turkish corpus. In this experiment we achieved encouraging results. We used data provided for MorphoChallenge competition for the purpose of evaluation.

1. Introduction

Obtaining textual data from the web has become a popular way to build large corpora for linguistic research. All web data is in an electronic form, instantly accessible, in large volume and covering various topics in many languages.

On the other hand, the internet is quite wild: messy, unordered and much duplicate. Solutions to these problems are being developed by other researchers such as (Pomikálek, 2011) whose text cleaning software was used in this work.

Since the performance of NLP generally tends to improve with increasing amount of training data, our aim is to obtain as much grammatical sentences as possible. Many words occur sparsely (according to Zipf's law), so we need really huge text collections to be able to study rare words' behaviour on sufficient number of their utterances.

Turkic languages are interesting for their productive inflectional and derivational agglutinative morphology which causes that these languages have immense amount of various wordforms. Comparing two corpora of the same size: English and Turkish, the second will contain much more wordforms but with lower frequencies. Thus, for these languages, the need for large corpora is even more pronounced.

We chose Turkish, Azerbaijani, Uzbek, Kazakh, Turkmen and Kyrgyz for our work since these languages are more or less connected to the corresponding nations and countries. Unlike other Turkic speaking areas, there are internet top level domains associated with the selected countries. That is why we decided not to collect Uyghur and Tatar texts.

2. Related work

2.1. Building web corpora

Building web corpora has received much attention recently. Table 1 presents selected previous work showing that it is possible to create very large corpora from the web.

The successful techniques used in the former works are search engines querying, web crawling (traversing the internet and downloading documents) and thorough data post-processing. Also, (Baroni et al., 2006) present a web tool able to build a web corpus almost instantly. It performs all necessary steps to prepare the data for further studying, such as concordance queries or terms extraction. However, we argue building billions scale corpora using that tool would require massive search engine querying which could turn out problematic.

We took advice from the previous works and developed new crawler SpiderLing (Suchomel and Pomikálek, 2012). We used the crawler in cooperation with several tools developed by authors referenced in Table 1.

2.2. Corpora of Turkic languages

Probably the largest corpus for Turkish till now was *BOUN Corpus* (Sak et al., 2008) containing about 423M words and 491M tokens. Among others are *METU* corpus with 2M words whose part also forms Turkish METU–Sabanci Treebank (Say et al., 2002), 50M web corpus (Dalkiliç and Çebi, 2002), Turkish part of parallel corpus of Balkan languges containing about 34M tokens (Tyers and Alperen, 2010) and recently developed Turkish corpus with about 42M words containing also Turkish word sketches (Ambati et al., 2012). Still under development is Turkish National Corpus with target size 50M words (Aksan and Aksan, 2009).

Probably the largest corpus for Azerbaijani is described in (Mammadova et al., 2010), containing about 300M words but since there is no mention about boilerplate removing, cleaning and de-duplicating, it is hard to estimate actual size of the corpus.

As for Kazakh, Kyrgyz, Uzbek and Turkmen languages there are some corpora for these langauges but either very small or not accessible (only mentioned in papers, on web pages).

(Biemann et al., 2004) developed corpora of relatively small size for Kazakh, Kyrgyz, Azerbaijani, Turkish, Chuvash, Uzbek, and Tatar mostly from Wikipedia.

2.3. Unsupervised morphology segmentation

Developing corpora of Turkic languages with almost no language tools available, we are forced to use unsupervised methods. Fortunately, unsupervised morphology analysis

Table 1: Overview of selected previous work concerning building large web corpora

language	reference	corpus size
English	(Liu and Curran, 2006)	10 bn tokens
German, Italian	(Baroni and Kilgarriff, 2006)	3.6 bn tokens total
English	(Pomikálek et al., 2009)	5.5 bn words
Dutch, Hindi, Indonesian, Norwegian, Swedish, Telugu, Thai, Vietnamese	(Kilgarriff et al., 2010)	680 mil words total
American Spanish, Arabic, Czech, Japanese, Russian	(Suchomel and Pomikálek, 2012)	51.6 bn tokens total

and segmentation have been well studied since 2000's. Several methods were proposed: (Bernhard, 2006; Demberg, 2007; Snyder and Barzilay, 2008; Argamon et al., 2004) with *Morfessor* (Creutz et al., 2005) being probably the most significant representative of them.
For evaluative purpose, competition *MorphoChallenge* (Kurimo et al., 2006) has been organized several times since 2005 with focus on English, Finnish and Turkish languages. We used their evaluation method for our experiment described in 4.3..

3. Building six Turkic web corpora

There are many ethnic groups and language varieties mixed together in the six language–country pairs we selected. Moreover, some of the languages are somewhat spoken in other countries too. Since we do not understand Turkic languages, we had to carefully constrain crawling and post-processing of the data. Fortunately, most web sites offer documents in just one or two languages understood by major part of the population, but we could not rely on that. Crawling was limited to the respective internet top level domain.
Furthermore, five of the six languages currently use two or three writing systems. We decided to collect the scripts prevailing in the recent texts: Latin for Azerbaijani, Uzbek, Turkmen and Cyrillic (with extensions) for Kazakh and Kyrgyz.
Three language specific models for each selected language were trained using texts from the respective Wikipedia. Byte trigrams for character encoding detection (tool Chared[1]), character trigrams for language identification[2] and a wordlist for boilerplate removal. Filtering crawled documents through these tools/models further helps eliminating unwanted content. Thanks to the strict limits, we believe a good quality of the texts was achieved at the cost of the resulting corpora size.
A couple of seed URLs (the links to start the crawling with) is usually enough in a network of websites densely connected by many links. Since the Turkic presence on the internet is relatively scarce, we obtained more starting URLs to cover more websites (see Table 2) using Corpus Factory (Kilgarriff et al., 2010) and Wikipedia. To get more texts from scarce resources, we configured the crawler to visit websites with less text amount than usually expected.

[1] nlp.fi.muni.cz/projects/chared/
[2] code.activestate.com/recipes/326576

Table 3: Processing the Turkish web. Each line represents a crawling or post-processing step which prevented some data not to pass. Only the last part was put in the final corpus.

data processing phase	fraction of documents	fraction of data size
HTML not retrieved	22.0 %	—
wrong encoding detected	0.7 %	—
other language	17.0 %	13.5 %
boilerplate	14.4 %	49.0 %
exact duplicates	17.8 %	14.7 %
near duplicates	15.9 %	16.4 %
clean text	12.1 %	6.4 %

The texts were tokenized on spaces, punctuation was treated as a separate token. Boilerplate (HTML markup, very short paragraphs and non-grammatical sentences) was removed by Justext[3] (Pomikálek, 2011). Duplicate and near-duplicate paragraphs were removed by n-gram based deduplication tool Onion[4] (Pomikálek, 2011). Misspelling was not dealt with.
Table 2 contains information about data size during crawling and processing. A detailed view on processing the Turkish corpus is presented in Table 3. The corpora have been installed in SketchEngine[5] with enabled concordance querying and wordlist functionality. The final sizes of the corpora in SketchEngine are displayed in Table 4.

4. Unsupervised morphological analysis

4.1. Motivation

Despite there are some morphological analyzers for Turkic languages, namely *TRmorph* (Çöltekin, 2010) and two-level analyzer (Oflazer, 1994) for Turkish, *UZMORPP* for Uzbek (Matlatipov and Vetulani, 2009) and *Azmorph* for Azerbaijani developed within *Apertium* project (Forcada et al., 2009), we are interested in unsupervised methods since other Turkic languages are uncovered in this respect.
A morphological analysis and disambiguation should assign one lemma and one morphological tag to each token in a corpus. With this information one can search for more

[3] code.google.com/p/justext/
[4] code.google.com/p/onion/
[5] the.sketchengine.co.uk

Table 2: Size of crawled HTML data, filtered plaintext and deduplicated texts. $Crawler's\ yield\ rate = \frac{plaintext\ size}{raw\ data\ size}$. $Final\ yield\ rate = \frac{deduplicated\ plaintext\ size}{raw\ data\ size}$.

language	initial domains	raw data [MB]	plaintext [MB]	crawler's yield rate	deduplicated plaintext [MB]	final yield rate	crawling time [h]
Azerbaijani	727	61,479	4,644	7.55 %	834	1.36 %	168
Kazakh	431	68,817	9,425	13.70 %	1,935	2.81 %	168
Kyrgyz	277	13,646	787	5.77 %	271	1.99 %	151
Turkish	157	2,763,780	159,054	5.75 %	26,844	0.97 %	336
Turkmen	51	1,469	113	7.66 %	17	1.18 %	27
Uzbek	454	7,825	497	6.35 %	141	1.80 %	70

Table 4: Turkic corpora obtained using SpiderLing

language	tokens	words	raw wordlist	clean wordlist
Azerbaijani	115M	92M	1.7M	1.4M
Kazakh	175M	136M	2.4M	1.9M
Kyrgyz	24M	19M	684K	590K
Turkish	4,124M	3,370M	20.5M	16.1M
Turkmen	2M	2M	230K	200K
Uzbek	24M	18M	626K	320K

general concordances and e.g. discover grammatical collocates using queries with lemmata and morphological tags.
Although we do not have taggers for several Turkic languages we nevertheless want to provide users with more than just simple querying using regular expressions on wordforms.
As was mentioned, there are some unsupervised methods for morphological analysis (assigning of morphological tags) but we plan to exploit particularly morphological segmentation since this (sub)task of morphological analysis is believed to be much simpler with more reliable results (in the realm of unsupervised methods).
Morphological segmentation splits wordforms into smaller parts: stems, prefixes and suffixes. If we assigned appropriate segmentations to all wordforms in a corpus we would be able to find more general concordances based on queries using stems. In this respect, stems could partially compensate absence of lemmata and tags in a corpus.
The quality of the segmentation is crucial for this enhancement so we evaluated unsupervised segmentations obtained by tool Morfessor-MAP (Creutz et al., 2005). For unsupervised morphological segmentation, Morfessor needs only a wordlist and it was chosen because of its fine results comparing to its competitors and because of being purely unsupervised.

4.2. Evaluation of Morphological Segmentation

For evaluation we used a tool provided for competition *MorphoChallenge* 2005 (Kurimo et al., 2006). Within the competition, gold standards for English, Finnish and Turkish language were provided containing one or more possible morphological segmentations of selected wordforms.

Table 5: Quality of segmentation for various training data.

prec	recall	f-sc	source	WL size
71.15	72.55	71.84	100k	22,3k
77.37	**69.74**	**73.36**	500k	70,4k
72.11	**69.74**	70.90	500k	70,4k
73.83	68.10	70.85	1M	112,6k
73.75	65.09	69.15	5M	313,8k
76.20	65.53	70.46	10M	482,4k
79.90	65.20	65.30	WIN	582,9k
79.10	37.90	51.30	M1	582,9k
73.70	65.10	69.20	M2	582,9k
77.50	65.00	66.40	M3	582,9k

That is why we could evaluate only Turkish segmentations. Nevertheless, we suppose that for other Turkic languages, the quality would be similar.[6]
The evaluation is based on the placement of morpheme boundaries. For example Turkish word *taylanddaki* (in Thailand) should be segmented into two parts: *tayland* and *daki*.[7]
Every correctly placed morpheme boundary forms *hit* (H), missing morpheme boundary forms *insertion* (I) and redundant boundary forms *deletion* (D). Precision is then the number of hits divided by the sum of the number of hits and insertions: $\frac{H}{(H+I)}$, recall is the number of hits divided by the sum of the number of hits and deletions: $\frac{H}{(H+D)}$ and f-score is as usually the harmonic mean of precision and recall.

4.3. Unsupervised Segmentation Results

In Table 5 there are results for various training data (wordlists) extracted from our Turkish corpus.
First three columns stand for precision, recall and f-score as explained before. The fourth column indicates a source for training. The number (in the upper part of the table) means an amount of tokens in a subcorpus from which a wordlist was extracted. The last column contains number

[6] In general, results (f-measure) for English within MorphoChallenge are better than for Finnish and Turkish. Results for Finnish and Turkish are comparable.

[7] In this case, *daki* should be further segmented into two morphemes *da, ki* but for the purpose of querying corpora, the coarse-grained segmentation is good enough.

of wordforms in appropriate wordlist used for unsupervised training.

The lower part of Table 5 shows selected results from MorphoChallenge 2005 for purpose of comparison. WIN stands for highest precision, recall and f-score achieved by various participants. M1–3 stands for evaluation of three variants of Morfessor.

It is clear that we achieved best results in all three measures. Quite surprising is fact that the best score was achieved using relatively small wordlist with about 70,000 of Turkish wordforms. Lower scores for larger wordlists were probably caused by inappropriate setting of one parameter of Morfessor (perplexity threshold) which must be set according to training data size. We run the process with various thresholds and wordlists but did not achieve better results for any of them. Despite, even with larger wordlists, we achieved better results than any participant of MorphoChallenge 2005.

Among other things, we believe that these results support good quality of the Turkish corpus. Training of Morfessor with data provided for MorphoChallenge did not achieve such good results and we suppose it is caused by rather strict language filtering of text data and diversity of language data in our corpus.

5. Conclusion and future work

We have built corpora for six Turkic languages, Turkish with 3.37 bn words being the largest. We believe the corpora are relevant not only due to their size but also with regard to the easiness with which the texts were obtained.

The actual results for morphological segmentation are encouraging but usefulness of unsupervised segmentation for Turkic and other agglutinative languages must be further investigated.

6. Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

7. References

Y. Aksan and M. Aksan. 2009. Building a national corpus of turkish: Design and implementation. In *Working Papers in Corpus-based Linguistics and Language Education*, pages 299–310.

Bharat Ram Ambati, Siva Reddy, and Adam Kilgarriff. 2012. Word Sketches for Turkish. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Shlomo Argamon, Navot Akiva, Amihood Amir, and Oren Kapah. 2004. Efficient unsupervised recursive word segmentation using minimum description length. In *Proceedings of Coling 2004*, pages 1058–1064, Geneva, Switzerland, Aug 23–Aug 27. COLING.

M. Baroni and A. Kilgarriff. 2006. Large linguistically-processed web corpora for multiple languages. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations*, pages 87–90. Association for Computational Linguistics.

M. Baroni, A. Kilgarriff, J. Pomikálek, and P. Rychlý. 2006. Webbootcat: a web tool for instant corpora. In *Proceeding of the EuraLex Conference*, pages 123–132.

D. Bernhard. 2006. Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of 2nd Pascal Challenges Workshop*, pages 19–24.

C. Biemann, S. Bordag, G. Heyer, U. Quasthoff, and C. Wolff. 2004. Language-independent methods for compiling monolingual lexical data. *Computational linguistics and intelligent text processing*, pages 217–228.

Ç. Çöltekin. 2010. A freely available morphological analyzer for turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*.

M. Creutz, K. Lagus, K. Lindén, and S. Virpioja. 2005. Morfessor and hutmegs: Unsupervised morpheme segmentation for highly-inflecting and compounding languages. In *In Proceedings of the Second Baltic Conference on Human Language Technologies*.

G. Dalkiliç and Y. Çebi. 2002. A 300 mb turkish corpus and word analysis. *Advances in Information Systems*, pages 205–212.

V. Demberg. 2007. A language-independent unsupervised model for morphological segmentation. *Annual meeting of Association for Computational Linguistics*, 45(1):920.

M.L. Forcada, F.M. Tyers, and G. Ramírez-Sánchez. 2009. The apertium machine translation platform: five years on. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 3–10.

A. Kilgarriff, S. Reddy, J. Pomikálek, and A. Pvs. 2010. A corpus factory for many languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'10, Malta)*.

M. Kurimo, M. Creutz, M. Varjokallio, E. Arisoy, and M. Saraclar. 2006. Unsupervised segmentation of words into morphemes-morpho challenge 2005, an introduction and evaluation report. In *Proceedings of ICSLP*.

V. Liu and J.R. Curran. 2006. Web Text Corpus for Natural Language Processing. *EACL. The Association for Computer Linguistics*.

S. Mammadova, G. Azimova, and A. Fatullayev. 2010. Text corpora and its role in development of the linguistic technologies for the azerbaijani language. In *The Third International Conference Problems of Cybernetics and Informatics*.

G. Matlatipov and Z. Vetulani. 2009. Representation of uzbek morphology in prolog. *Aspects of Natural Language Processing*, pages 83–110.

K. Oflazer. 1994. Two-level description of turkish morphology. *Literary and Linguistic Computing*, 9(2):137.

J. Pomikálek, P. Rychlý, and A. Kilgarriff. 2009. Scaling to billion-plus word corpora. *Advances in Computational Linguistics*, 41:3–13.

J. Pomikálek. 2011. *Removing Boilerplate and Duplicate*

Content from Web Corpora. Ph.D. thesis, Masaryk University, Brno.

H. Sak, T. Güngör, and M. Saraçlar. 2008. Turkish language resources: Morphological parser, morphological disambiguator and web corpus. *Advances in natural language processing*, pages 417–427.

B. Say, D. Zeyrek, K. Oflazer, and U. Özge. 2002. Development of a corpus and a treebank for present-day written turkish. In *Proceedings of the eleventh international conference of Turkish linguistics*, pages 183–192.

B. Snyder and R. Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. *Proceedings of ACL-08: HLT*, pages 737–745.

V. Suchomel and J. Pomikálek. 2012. Efficient web crawling for large text corpora. In *Proceedings of the Seventh Web as Corpus Workshop*, Lyon, France, In print.

F. Tyers and M.S. Alperen. 2010. South-east european times: A parallel corpus of balkan languages. In *Forthcoming in the proceedings of the LREC workshop on Exploitation of multilingual resources and tools for Central and (South) Eastern European Languages*.