Large Web Corpora of High Quality for Indian Languages

Uwe Quasthoff[1], Ritwik Mitra[2], Sunny Mitra[2], Thomas Eckart[1], Dirk Goldhahn[1], Pawan Goyal[2], Animesh Mukherjee[2]

[1] Natural Language Processing Group, University of Leipzig, Germany
[2] Indian Institute of Technology Kharagpur, Kharagpur, India
Email: [1] {quasthoff, teckart, dgoldhahn}@informatik.uni-leipzig.de
[2] {ritwik.mitra, sunny.mitra, pawang, animeshm}@cse.iitkgp.ernet.in

Abstract

Large textual resources are the basis for a variety of applications in the field of corpus linguistics. For most languages spoken by large user groups a comprehensive set of these corpora are constantly generated and exploited. Unfortunately for modern Indian languages there are still shortcomings that interfere with systematic text analysis. This paper describes the Indian part of the Leipzig Corpora Collection which is a provider of freely available resources for more than 200 languages. This project focuses on providing modern text corpora and wordlists via web-based interfaces for the academic community. As an example for the exploitation of these resources it will be shown that they can be used for the visualization of semantic contexts of terms and for language comparison.

Keywords: corpus generation, text acquisition, comparative corpus analysis

1. Availability of Indian text resources

Many applications in the field of corpus linguistics require the availability of large corpora. Well-known examples for corpora based on Indian languages are the web corpora built by Kilgariff and Duvuru (2011) or the large Sanskrit corpus by GRETIL that is provided for free. The Technology Development for Indian Languages (TDIL) Programme also provides some corpora for Indian languages for download and further resources are available from the Central Institute of Indian Languages (CIIL).

Unfortunately many of the existing corpora or resources lack features that are strongly desirable for their use in the scientific context. These shortcomings include problems with availability (in some cases the use of very specific interfaces is required), lack of currentness (a problem especially when dealing with ongoing political developments), high costs or strict licences that permit reuse and data aggregation. As some of these problems can't be eliminated in general (like in the context of copyright and personality rights) it would be desirable to have more resources that can be used with as less restrictions as possible and that can be useful for further progress in the exploitation of Indian language corpora and other text-based resources .

2. Indian Resources

The Leipzig Corpora Collection (LCC)[1] has been collecting digital text material for more than 20 years. Starting with a focus on European languages it became apparent that a lot of the developed strategies and tools could be reused for other languages as well. Over the last years the established tool chain for text acquisition and text processing was adopted to deal with non-Latin scripts and used to create and improve resources based on Indian text material.

2.1. Text Acquisition Strategies

Different strategies are combined for collecting textual data from the WWW. The main goal is to ensure that corpora of large extent and high diversity concerning topics or genres can be created for specific languages. Especially, Indian languages that are spoken in many countries require a variety of approaches to achieve this objective.

2.1.1. Generic Web Crawling

A framework for massively parallel Web crawling is applied that utilizes the standard Web crawler and archiver *Heritrix*[2] of the Internet Archive. Among other enhancements, it was enriched with means for the automatic generation of crawling jobs.

Heritrix is used in several ways. On one hand whole Top Level Domains are crawled. In this case a small list of domains of a country of interest is used as an input. Heritrix is then configured to follow links within this TLD. This has been conducted for several countries.

On the other hand News sources are downloaded using the Heritrix based Web crawler. Basis is a list of more than 32,000 news sources in about 120 languages provided by *ABYZ News Links*[3]. This service offers URLs and information regarding country and language. This way, news texts for several Indian languages were collected. This includes text data excluded in the TLD crawling because of non-country TLDs used such as ".com".

2.1.2. Distributed Web Crawling

FindLinks (Heyer and Quasthoff, 2004) is a distributed Web crawler using a client-server architecture. The

1 http://corpora.uni-leipzig.de

2 http://webarchive.jira.com/wiki/display/Heritrix

3 http://www.abyznewslinks.com

Java-based client runs on standard PCs and processes a list of URLs, which it receives from the *FindLinks*-server. FindLinks has been used with community support for several years and allowed us to crawl the WWW to a large extent.

2.1.3. Bootstrapping Corpora

In addition, an approach similar to Baroni (2004) and Sharoff (2006) was applied. Frequent terms of any specific language are combined to form Google search queries and the resulting URLs are retrieved as basis for the default crawling system.

A small set of frequent terms is needed for languages in question. Therefore existing corpora of the LCC or other sources such as the *Universal Declaration of Human Rights* (*UDHR*)[4] were utilized as a resource.

Based on these lists, tuples of three to five high frequent words are generated. These tuples are then used to query Google and to collect the retrieved URLs, which are then downloaded.

2.1.4. Crawling of Special Domains

Certain domains are beneficial sources for Web corpora since they contain a large amount of text in predefined languages.

One example is the free Internet encyclopedia Wikipedia, which is available in more than 200 languages and of course also contains entries for Indian languages.

Wikipedia dumps for these Indian languages, were downloaded. *Wikipedia Preprocessor*[5] was used for further processing and text extraction.

2.2. Corpus Creation Toolchain

Necessary steps for the creation of dictionaries are text extraction (mostly based on HTML as input material), language identification (Pollmächer, 2011), sentence segmentation, cleaning, sentence scrambling, conversion into a text database and statistical evaluation.

An automatic and mainly language independent tool chain has been implemented. It is easily configurable and only few language-dependent adjustments, concerning e.g. abbreviations or sentence boundaries, have to be made.

In a final step statistics-based quality assurance is applied to achieve a satisfying quality of the resulting dictionaries (Quasthoff, 2006b) (Eckart, 2012). Using features such as character statistics, typical length distributions, typical character or n-gram distributions, or tests for conformity to well-known empirical language laws during corpora creation can be detected and corrected.

The processing of Indian language text required several changes to the existing toolchain. Most of the developed tools could be reused but specific configurations had to be changed. This includes changes to components like sentence segmentation or quality assurance procedures. Besides some minor problems the general system again proved to be stable enough.

2.3. Sentence Scrambling

For all corpora the sentences had to be "scrambled" to destroy the original structure of the documents due to copyright restrictions. This inhibits the reconstruction of the original documents. With respect to German copyright legislation this approach is considered safe.

2.4. Available Resources

Corpora of this collection are typically grouped regarding the dimensions language, country of origin, text type (newspaper text, governmental text, generic Web material, religious texts etc.) and time of acquisition. Table 1 gives an introduction to currently available resources. It contains the number of sentences for different languages and genres. For comparison, the size of the corresponding Emille[6] corpora is given. As the crawling is an ongoing process new corpora are added at least every year.

To do a sanity check over the corpora, 1000 high frequency words from Bengali as well as Hindi corpora were taken. These words were then manually checked by native speakers of Bengali and Hindi. For both Bengali and Hindi, more than 94% of the words turned out to be valid. The remaining 6% contained spelling errors, mathematical symbols etc.

2.5. Available Interfaces

The corpora are available via Web-based interfaces. There is a freely available web portal where a variety of information can be accessed based on a word level (like sample sentences, word co-occurrences, co-occurrence graphs etc.). Furthermore many corpora can be downloaded for free in different formats. These include plain text versions of the textual material and also MySQL databases. For the later a platform-independent browsing tool is provided which allows examining the corpus locally.

3. Applications

3.1. POS-Tagging

For many Indian languages (like Hindi, Telugu, Tamil, Kannada, Punjabi, Urdu, Bengali and Marathi) POS taggers are available. They all use Latin transliteration and most of them use the WX transliteration scheme. Therefore the used tag sets are comparable.

For testing purposes some of the Hindi sentences were tagged using the Hindi shallow parser[7]. In addition to morphological analysis and chunking, this tool also gives the POS tagging analysis of a sentence.

4 http://www.ohchr.org

5 http://sourceforge.net/projects/wikiprep/

6 http://www.emille.lancs.ac.uk/

7 http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php

As an example the Hindi sentence (in Devanagari)

इसका उत्तर राजू ने अपने पत्र में लिखा कि सत्यम को तंगी से उबारने के लिए उन्होंने ऐसा किया ।
(English translation: *Raju answered this in his letter that he did it to rescue Satyam from scarcity.*)
is transliterated (using roman transliteration) to:

isakā uttara rājū ne apane patra meṃ diyā ki satyama ko taṃgī se ubārane ke lie unhoṃne aisā kiyā .

The output from the POS tagger is:

isakā/PRP uttara/NNPC rājū/NNP ne/PSP apane/PRP patra/NN meṃ/PSP diyā/VM ki/CC satyama/NNP ko/PSP taṃgī/NN se/PSP ubārane/VM ke/PSP lie/PSP unhoṃne/PRP aisā/PRP kiyā/VM ./SYM

3.2. Diachronic Comparisons

Newspaper corpora collected on yearly basis can be used to investigate changes in the frequency of words. These changes may reflect different types of modern developments.

3.3. Lexicography

Frequency ordered word lists help identifying lemmas for dictionaries, especially for the enlargement of existing dictionaries. Especially neologisms (i.e. words found in the corpus of the current year with a certain frequency and not seen before) are interesting for the study of language change.

3.4. Semantic Relations given by Word Co-occurrences

Significant word co-occurrences often show semantic relations between those words. The investigation of word co-occurrence graphs also helps identifying the most prominent topics in the corpus. Figure 1 shows two clusters of words that significantly often occur in the context of "Mumbai": one cluster corresponding to sports events and the other corresponding to the terror attacks in Mumbai 2008.

3.5. Comparisons between Countries and Regions

For languages spoken in more than one country, the corpora will reflect their differences. For the languages spoken in India, the corpora can be used for:

- Comparison of Urdu, Sindhi and Punjabi in India and Pakistan,
- Comparison of Tamil in India and Sri Lanka
- Comparison of Bengali in India and Bangladesh

4. Outlook

The Leipzig Corpora Collection will continue in aggregating Web-based text material to extend the amount and quality of available resources. Currently 220 GB of additional raw material crawled from the Indian TLD are processed and will be available in spring 2014. Furthermore, the result of these efforts will be provided to all the interested users.

Until mid of 2014 a new Web portal will be deployed that provides extended functionality and a more user-friendly interface. The underlying REST-based web services are also freely available and can be used for external applications as well. As a next step in exploiting word lists as a valuable resource in information extraction and language comparison, it is planned to publish a volume in the series of frequency dictionaries focusing on word frequency information in specific Indian languages.

5. References

Baroni, M.; Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

Eckart, T.; Quasthoff, U.; Goldhahn, D. (2012). Language Statistics-Based Quality Assurance for Large Corpora.Proceedings of Asia Pacific Corpus Linguistics Conference 2012, Auckland, New Zealand.

Heyer, G.; Quasthoff, U. (2004). Calculating Communities by Link Analysis of URLs. Proceedings of IICS-04, Guadalajara, Mexico and Springer LNCS 3473.

Kilgariff, A.; Duvuru, G. (2011). Large Web Corpora for Indian Languages. Information Systems for Indian Languages. Communications in Computer and Information Science. Volume 139, pp 312-313.

Pollmächer, J. (2011). Separierung mit FindLinks gecrawlter Texte nach Sprachen. Bachelor Thesis, University of Leipzig.

Quasthoff, U.; Biemann, C. (2006). Measuring Monolinguality. Proceedings of LREC-06 workshop on Quality assurance and quality measurement for language and speech resources.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini, editors, WaCky! Working papers on the Web as Corpus.Gedit, Bologna.

Language (ISO 639-3)	*News*	*Wikipedia*	*For comparison: Emille*
asm	-	-	100,095
ben	109,855	240,128	153,948
guj	848,723	-	601,947
hin	5,162,167	727,882	469,395
kan	-	389,395	76,445
kas	-	-	11,858
mal	216,788	185,928	75,645
mar	774,201	149,420	96,296
ori	-	-	80,262
pan	507,059	-	429,948
pnb	-	39,606	8,587
sin	-	-	287,554
tam	1,341,954	-	1,298,802
tel	326,233	430,723	198,669
urd	1,733,995	144,312	60,903

Table 1: Amount of available text resources in number of sentences

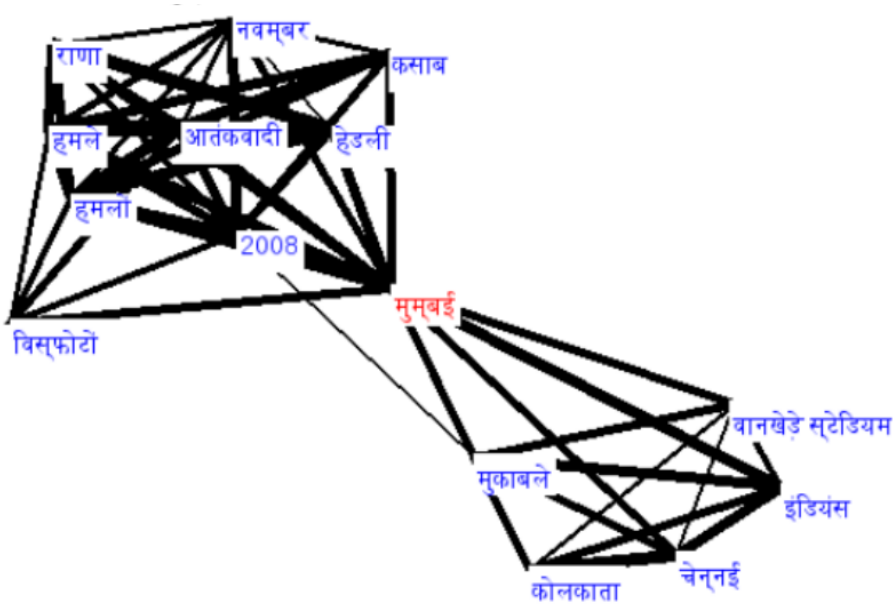

Figure 1: Word co-occurrence graph