Challenges in the development of annotated corpus of computer-mediated communication in Indian Languages: A Case of Hindi

Ritesh Kumar

Centre for Linguistics
Jawharlal Nehru University, New Delhi, India
riteshkrjnu@gmail.com

Abstract

The present paper describes an ongoing effort to compile and annotate a large corpus of computer-mediated communication (CMC) in Hindi. It describes the process of the compilation of the corpus, the basic structure of the corpus and the annotation of the corpus and the challenges faced in the creation of such a corpus. It also gives a description of the technologies developed for the processing of the data, addition of the metadata and annotation of the corpus. Since it is a corpus of written communication, it provides quite a distinctive challenge for the annotation process. Besides POS annotation, it will also be annotated at higher levels of representation. Once completely developed it will be a very useful resource of Hindi for research in the areas of linguistics, NLP and other social sciences research related to communication, particularly computer-mediated communication..Besides this the challenges discussed here and the way they are tackled could be taken as the model for developing the corpus of computer-mediated communication in other Indian languages. Furthermore the technologies developed for the construction of this corpus will also be made available publicly.

Keywords: Corpus, Computer-mediated communication, politeness analysis, sentiment analysis

1. Introduction

Computer-mediated communication (or, CMC) refers to all kinds of interaction carried out with the help of computer over the internet. Typically they are written communication (with the advent of voice/video chatting, video conferencing, etc. over the internet, CMC, technically, also includes spoken communication now). It is one of the dominant ways of communication in today's era. And it provides several unique challenges in natural language processing, which is quite distinct from the written as well as the spoken language. CMC is used for a multimodal activity (which is almost always coupled with tones, intonations, gestures, etc) in an unimodal way (i.e.,only through the written words and symbols), resulting in several innovations (like smileys, highly elliptical and scrambled structures, etc.) in the written text. As a result, using the standard tools (built with either written or spoken text in mind) on such texts do not return the desired results.

The corpus described in the present paper is especially designed and developed for the purpose of sentiment analysis (politeness analysis and recognition, in particular) of the texts available online. Consequently the compilation and design of the corpus is carried out keeping in mind this immediate purpose.

2. Compilation of the corpus

Since it is one of the first efforts in recent times to construct a corpus of the computer-mediated communication, the whole process was needed to be planned and executed from scratch (with inputs from (Beiwenger and Storrer, 2008), (King, 2009) and others). Some of the major challenges (some unique challenges specific to this corpus while others the general challenges of preparing any corpus compounded by the absence of any previous reference point) faced in the data collection process included -

2.1. What all to include in the corpus?

The most basic issue that crops up in the construction of a corpus is the contents of the corpus. Computer-mediated communication could include a plethora of things, ranging from such clear cases of communication like chatting, e-mails and e-forums to other fuzzy areas like wikipedia. Considering the fact that the corpus was being constructed for politeness analysis, the present corpus is compiled using data from 6 different sources of computer-mediated communication in Hindi. They include -

- *Blogs*: The data for Hindi blogs is mainly collected using http://chitthajagat.in (a blog aggregator, which stores, indexes and provides link for most of the blogs written in Hindi). Only those entries were saved which had received at least 3 comments.
- *Web Portals*: The data for web portals is collected from 3 different websites - Tehelka (http://www.tehelkahindi.com/), Josh18 (http://josh18.in.com/) and Web Duniya (http://hindi.webdunia.com/). As in blogs, only those entries which have received 3 or more comments have been included in the corpus.
- *E-forums*: The data for e-forums is collected from the Google groups using the Google group directory for Hindi language (http://groups.google.com/groups/dir?sel=lang%3Dhi&). However instead of three replies to the main entry in blogs and portals, here even if there is only one reply to the post that initiates a particular discussion on the forum then that is included in the corpus.
- *E-mails*: The data for e-mails have been taken from the personal e-mail account of 7 persons.

- *Public Chatting*: For public chats, the data is collected from the IRC (Internet Relay Chatting) channel using Mozilla's Chatzilla.

- *Private Chatting*: The data for private chatting is taken from Gmail chat logs of 7 persons (the same people who had given access to their personal e-mail accounts).

Table 1 gives the statistical summary of the raw corpora collected:

Data Source	**Number of Words (approx. in thousand)**	**Number of Sentences (approx. in thousand)**
Blogs	905	132
Portals	785	88
E-forums	164	38
E-mails	8731	1361
Public Chats	5033	623
Private Chats	404	66
Total	16023	2309

Table 1: The caption of the table

2.2. Availability of the data

Availability of data related to CMC in Hindi was one of the major challenges faced in the construction of the corpus. Except in case of blogs (which has seen an explosion in the last couple of years) and some web portals, there is hardly any data available in Hindi in other places where CMC takes place (like e-mails, chatting and e-forums). One of the most notable exclusions in the corpus is the data from social networking sites and it could be largely attributed to the unavailability of sufficient data in Hindi. The data from the review and user opinion sites are also not included in the corpus because they are generally not available in Hindi. Moreover the data which is included in the corpus is not completely in Hindi (particularly in chats and e-mails) since exclusion of English data would have rendered the rest of the data meaningless owing to the absence of the complete and proper discourse context. Furthermore some other languages (particularly in chats), besides English, are also present in the corpus for similar reasons.

It must be mentioned here that it is not a unique situation with Hindi; rather all the major languages of India present a similar picture where CMC involves English, along with the other languages. So considering the fact that corpus of this kind would be very helpful in understanding the phenomenon of code-mixing and code-switching in the Indian context and in also developing technologies to handle them more efficiently in the cyberspace, the data was kept as it occurred in the natural context.

2.3. Availability of technologies

Another significant issue which cropped up during the creation of the corpus was the lack of sufficient technologies for the acquisition and clean-up of the data. There were no crawlers available for language-specific crawling over the web, which could be used for getting Hindi data only; moreover there were lots of data in Hindi not written in the native Devanagari script, rather most of the time (except in blogs and web portals), it is written in Roman script (with highly variable and non-standard spelling patterns). Consequently accurate recognition of Hindi data automatically was very difficult. Furthermore besides a couple of rudimentary tools for cleaning-up the web data (which remove only the HTML tags), there is hardly any technology available for cleaning-up the web data.

In order to handle this situation, new technologies were required to be developed. It was extremely difficult to develop crawlers to recognise and index Hindi web pages written in Roman script (so most of the work was done manually). However for cleaning-up the data, a separate tool is being developed for each of the six sources of data since the structure of the noise present in each of these source is different and needs a specialised tool. Moreover tools are also developed for the addition of metadata (to a certain limited extent) while the word-level annotation is carried out using ILCI Annotation Tool (Kumar et al., 2012). A brief description of these tools is given below:

2.3.1. Web Data Processor (WeDPro)

WeDPro is a web-based application, developed using Java/JSP at the frontend (Figure 1), which extracts the relevant body text from any web page with its main body text in Devanagari Unicode (it is easily adaptable for other non-Roman Unicode text). It takes the raw/noisy files as well as the source of the files (like blogs, web portals, etc.) as input and gives the cleaned file as output. Along with this it is also able to automatically add some metadata information to the corpus. Currently it is able to correctly process the data from the following sources -

1. Blogs from 'Blogspot' and 'Wordpress': The data from these sites are cleaned using pattern matching and only the main body text is extracted from the complete web page (saved in 'text only' format using Mozilla Firefox browser). It cleans most of the nosiy data; however the text needs to be rechecked to ensure that the data is completely free of noise.

2. Web Portals: The data from the 3 web portals, mentioned above is correctly handled by the tool (and it is expected to handle data from other sources also). It works in a similar way as in the case of blogs.

3. E-forums: The data is cleaned in a similar way as in the case of blogs and web portals. Currently it works for the posts on Google groups.

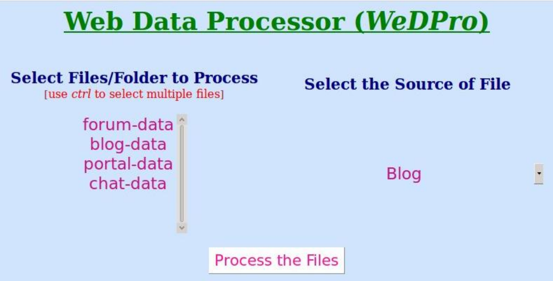
4. E-mails: The tool provides a facility to automatically extract e-mails from any' Gmail' account (other e-mail providers could be easily included with a little modification), given the username and the password of the account. The extracted emails are saved as plain text files on the local system. It uses Javamail API to extract the emails. These e-mails do not require much cleaning up as they are generally free of noise. However the redundancy in the mails (owing to the

conversation-like structure of the emails in Gmail) is automatically removed and other information like the date and time of receiving the mail, the sender of the mail and the subject of the mail are added as metadata.

This tool also anonymises the senders and receivers of the emails to protect their identity (and other relevant socio-economic information about them is later on added to the metadata). Furthermore it automatically creates the metadata with such information as the number of replies sent to the mail (or, the number of mails associated with this one), date and time when the mail was sent, subject of the mail etc.

5. Private Chats: Private chats are extracted, processed and saved in the corpus in the same way as the emails (since they are the transcripts of the chats saved in Gmail and so have a very similar structure as the Gmail emails). As in emails, the tool anonymises the participants in private chats and also prepare a similar kind of metadata.

6. Public Chats: The public chats from IRC also do not require much processing since Chatzilla provides the facility of saving the logs with minimal noise. However despite this some very minor noisy is completely removed automatically using the process of pattern matching.



Web Data Processor (WeDPro)

Select Files/Folder to Process
[use ctrl to select multiple files]

forum-data
blog-data
portal-data
chat-data

Select the Source of File

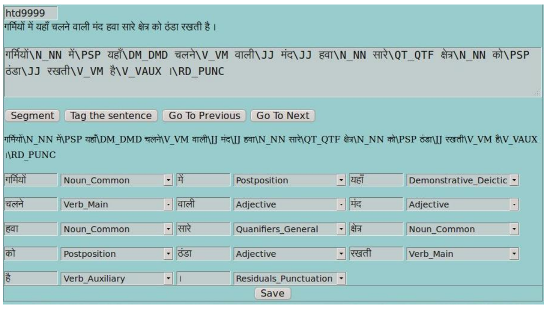
Blog

Process the Files

Figure 1: Home Page of the WeDPro

2.3.2. ILCI Annotation Tool

ILCI annotation tool (Figure 2) is also a web-based application which is developed as part of the Indian Languages Corpora Initiative for Part-of-speech (POS) annotation of the data. However it could be potentially used for any kind of word-level of annotation (given the proper tagset). Moreover with certain modifications it could also be adapted so as to carry out phrasal, sentential and discourse level annotation also. Along with providing a user interface for annotation, it also provides some kind of intelligence which helps in increasing the efficiency and reliability of the annotators (Kumar et al., 2012).



htd9999
गर्मियों में यहाँ चलने वाली मंद हवा सारे क्षेत्र को ठंडा रखती है ।

```
गर्मियों\N_NN में\PSP यहाँ\DM_DMD चलने\V_VM वाली\JJ मंद\JJ हवा\N_NN सारे\QT_QTF क्षेत्र\N_NN को\PSP
ठंडा\JJ रखती\V_VM है\V_VAUX ।\RD_PUNC
```

गर्मियों\N_NN में\PSP यहाँ\DM_DMD चलने\V_VM वाली\JJ मंद\JJ हवा\N_NN सारे\QT_QTF क्षेत्र\N_NN को\PSP ठंडा\JJ रखती\V_VM है\V_VAUX ।\RD_PUNC

गर्मियों	Noun_Common	में	Postposition	यहाँ	Demonstrative_Deictic
चलने	Verb_Main	वाली	Adjective	मंद	Adjective
हवा	Noun_Common	सारे	Quanifiers_General	क्षेत्र	Noun_Common
को	Postposition	ठंडा	Adjective	रखती	Verb_Main
है	Verb_Auxiliary	।	Residuals_Punctuation		

Figure 2: ILCI Annotation Tool Interface

2.4. Ethical issues

Handling the issues related to the ethics of data collection and the inclusion of data into the corpus is another very challenging issue in the creation of the corpus. In a country like India where the copyright laws (with respect to the data generated in the cyberspace) and its awareness are still in its infancy, it is very difficult to even know what may be the ethical practices. While, in general, the data generated in the cyberspace (and freely accessible by everyone) is considered to be something in the public domain and so free of copyright (at least for the academic purposes), those cases where copyright is clearly mentioned are excluded from the corpus. However there are several cases where the writers are anonymous, they use pseudonyms or no contact information is available and in such situations it becomes impossible to acquire or even know about the copyright holders.

3. Preparation of the Metadata

Two kinds of metadata have been maintained for the corpus-

3.1. General Metadata of the Corpus

The general metadata consists of the basic information about the corpus like the complete information about source from where data is taken, the statistics about the corpus and the information about the general structure of the corpus. This kind of metadata, on one hand, helps in arranging the data in a proper format, and on the other hand, it helps other people in making an optimum, informed and clear use of the corpus.

3.2. Content Metadata of the Corpus

The content metadata consists of the information regarding the kind of data included in a particular file, the kind of medium through which the data has been generated (i.e., whether it is taken from blog, web portals, e-mails, etc), the topic/domain of the data, the purpose for which the data was generated and other such similar information. This metadata is equivalent to the situational features described in the register-based annotation scheme (described in detail in (Kumar, 2011)) and will be used as semantic markups for the purpose of politeness analysis. One of the major challenges in the preparation of the content metadata is that in order to prepare the content metadata of the corpus, a detailed analysis of the corpus is required. Moreover at times the classification of the data into the purpose and domain fall into fuzzy areas and the decision to include it into one criteria is taken depending on the dominant purpose and domain of the data.

4. Annotation of the Corpora

Annotation of a corpus could be done at several levels of sophistication and complexity. The most basic level of annotation is done at the POS level and then depending on the purpose and the requirement, further deep annotation of the corpus is carried out at different levels of syntactic, semantic, pragmatic and dialog act information.

Since this corpus is being developed for the immediate purpose of politeness analysis, it was necessary to annotate the corpus with information which is relevant and necessary in the recognition of (im)politeness in the text. After a basic analysis of the ways in which (im)politeness is expressed in the language, an annotation scheme was developed for annotating the corpus. Since (im)politeness is a phenomenon which is expressed at all levels of linguistic representation (lexical, morphological, syntactic and pragmatic), in order to capture the phenomenon effectively and efficiently, the corpus was needed to be annotated at all these levels. However on the other hand, all kinds of structure and constructions at these levels are not relevant for (im)politeness analysis. So one of the first challenges in preparing the annotation scheme was to arrive at the categories for annotation in such a way that on one hand they could help in capturing the (im)politeness phenomenon in a fairly comprehensive way and on the other hand the scheme also do not become too large to be handled efficiently.

Finally a hierarchical annotation scheme, along with the annotation labels, was devised which does not annotate the corpus at some particular level of linguistic representation; rather it seeks to annotate the corpus at all levels of linguistic representation such that these markups help in (im)politeness analysis of the text but not so comprehensively in a complete linguistic analysis.

5. Summing Up

In this paper, the challenges faced and the ways in which they are dealt with in the preparation of a CMC corpus in Indian languages (with special reference to Hindi) are discussed. The structure of the corpus and the way it is constructed could be used as a reference point for the preparation of CMC corpus in other major as well as minor Indian languages also so that the unique challenges posed by the text generated as a result of CMC could be handled efficiently in the NLP.

6. Acknowledgements

I would like to express my sincerest thanks to my supervisors Dr. Ayesha Kidwai and Dr. Girish Nath Jha for their constant encouragement and support during the preparation of this corpus.

7. References

Michael Beiwenger and Angelika Storrer. 2008. Corpora of computer-mediated communication. In Anke Ldeling and Merja Kyt, editors, *Corpus Linguistics. An International Handbook (Vol. 2)*, pages 292–308. Mouton de Gruyter, Berlin.

Brian King. 2009. Building and analyzing coprpora of computer-mediated communication. In Paul Baker, editor, *Contemporary Corpus Linguistics*, pages 301–320. Continuum International, London/ New York.

Ritesh Kumar, Shiv Kaushik, Pinkey Nainwani, Esha Banerjee, Sumedh Hadke, and Girish nath Jha. 2012. Using the ilci annotation tool for pos annotation: A case of hindi. In *Presented at the 13th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2012), New Delhi, India, March 11th-17th, 2012*.

Ritesh Kumar. 2011. A register-based annotation scheme for co3h. In Rajendra Akerkar, editor, *Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS'11*. ACM, ICPS, ACM.