Developing Politeness Annotated Corpus of Hindi Blogs

Ritesh Kumar

Dept. of Linguistics, Dr. Bhim Rao Ambedkar University
Agra, India
Email: riteshkrjnu@gmail.com

Abstract

In this paper I discuss the creation and annotation of a corpus of Hindi blogs. The corpus consists of a total of over 479,000 blog posts and blog comments. It is annotated with the information about the politeness level of each blog post and blog comment. The annotation is carried out using four levels of politeness – neutral, appropriate, polite and impolite. For the annotation, three classifiers – were trained and tested maximum entropy (MaxEnt), Support Vector Machines (SVM) and C4.5 - using around 30,000 manually annotated texts. Among these, C4.5 gave the best accuracy. It achieved an accuracy of around 78% which is within 2% of the human accuracy during annotation. Consequently this classifier is used to annotate the rest of the corpus.

Keywords: Hindi politeness, Hindi blogs, politeness annotation

1. Introduction

Politeness is one of the most important aspects of human communication which could single-handedly decide the success of any human interaction. Over the last few decades, different theoreticians have tried to understand politeness in different ways. Starting from the seminal works of Brown & Levinson (1978, 1987) who tried to understand politeness in terms of face-saving strategies and Leech (1983, 2007) who explains politeness in terms of conversational maxims to recent discursive approaches (Watts 1989, 2003) which understands politeness in terms of relational work which is an emergent property of the conversation and is subject to discursive struggle, politeness has remained a matter of great theoretical debate. While each theory comes from a different philosophical perspective and different set of assumptions these theories are not contradictory in nature. And at the same time none of these could be taken as a comprehensive theory of politeness which could explain all its aspects. The way the earlier theories of politeness were conceptualised, linguistic structures played a central and almost exclusive role in the evaluation of politeness. In fact these theories hardly talk about the "evaluation" or hearer-related part of politeness; rather they focus on "production" or speaker-related part of politeness. In these theories, it was assumed that politeness is inherent in linguistic structures or speech acts and these particular structures and acts will always produce same politeness effects. Even though the role of context in production of politeness was acknowledged in principle, it was not incorporated in the theoretical framework itself. On the other hand the later discursive theories completely ruled out the possibility of any kind of inherent semantics related to politeness in the linguistic structures. Rather it is argued that any structure is evaluated as polite or impolite only within a socio-cultural context and outside of that context they do not carry any politeness effects. Politeness is argued to be an emergent property of human interaction in these theories.

However a closer look at the phenomenon of politeness reveals that it could be best explained by taking it as a generalised conversational implicature generated out of its normative aspects (Terkourafi 2001, 2003, 2005). This implicature is generated because of the regular co-occurrence of certain linguistic structures with certain politeness effects in specific contexts. However it is argued that this implicature only creates potential for politeness but this structure may not be finally evaluated as polite in an interaction because of some of the local contextual factors. Thus even though linguistic structures are not inherently polite, they have the potential to be polite.

Following this view (also known as the interactional approaches to politeness), politeness could be understood as a loose mapping from syntactic structures to the semantic evaluations of these structures as polite which is based on the prior experience of the speakers. This could also explain the possible idiolectal nature of politeness whereby individual differences in politeness evaluation is found among the speakers. At the same time this understanding of politeness also gives a principled and theoretically valid ground for automatic annotation of texts as polite.

In this paper I discuss the construction of a corpus of Hindi blogs which is annotated for its politeness value. I discuss the annotation scheme (which is inspired by the literature on politeness) and how the corpus is annotated using this scheme. The complete corpus is annotated using supervised machine learning techniques where 30,000 manually annotated texts are used to annotate the complete corpus of over 479 thousand texts.

2. The Corpus

The corpus of Hindi blogs is automatically collected using the links aggregated by Chitthajagat, one of the most popular aggregators of blogs in Hindi. The link of the Hindi blogs thus obtained are crawled through using the Google API for bloggers. The data thus obtained was saved in XML format, along with the metadata information like the identity of the authors and the basic statistics like the number of words and sentences.

Two kinds of metadata information are stored for the corpus.

a. Information about the data collection: These are the information about the collection of the data and are maintained separately. A sample of this kind of metadata (taken from the blogs dataset) is given in Table 1.

Blog Link	File Name	No. of Comments	Date of Retrieval	Time of Retrieval
http://blog4varta.blogspot.com/2013/01/4_18.html	blog_corpus_1	6	20/01/13	00:21:55
http://blog4varta.blogspot.com/2013/01/4_16.html	blog_corpus_2	9	20/01/13	00:21:55
http://blog4varta.blogspot.com/2013/01/4_14.html	blog_corpus_3	15	20/01/13	00:21:55

Table 1: A sample of metadata about the data collection

This kind of information not only keeps an exact record of location from where the data is retrieved and the time when the data is retrieved but also it gives an added advantage for the future. By giving information about the links which have already been visited, it makes the task of extending the corpus easy and non-repetitive.

b. Information about the data itself: The information about the data is included in each XML file itself. A sample is given in table 2.

```
<async_info>
  <author>सगत पर</author>
  <blog_name>बग 4 वर</blog_name>
 <post_title>झटक कजए न हलल क ... बग 4 वर .. सगत
पर</post_title>
  <date>2013-01-18</date>
  <time>04:00:00</time>
  <words>804</words>
  <sentences>33</sentences>
</async_info>

<comment_info>
  <commentator>सध शर</commentator>
  <date>2013-01-18</date>
  <time>14:24:00</time>
  <words>4</words>
  <sentences>1</sentences>
</comment_info>
```

Table 2: Sample metadata of blog post and comments

In its current form the corpus consists of data from a total of 41,553 main blog posts and 437,952 comments on these blog posts. Overall the corpus is composed of approximately 905,000 words in 479,505 texts.

3. Annotation Scheme

I have used four tags for marking the level of politeness that a text exhibits. These four tags are – neutral, appropriate, polite and impolite.

3.1. Neutral Text

Neutral texts contain neither elements of politeness nor impoliteness and include objective description of some place, object, technology, etc. A text may be neutral only if it is a pure description and does not include any kind of instructions. For example. if the author is giving instructions on how to use a technology or why and how to visit some places then they cannot be neutral. Neutral would be just plain description and nothing else. For example, the following text would be marked as neutral

ऱाम के चार बेटे हैं । पहला बेटा दिल्ली मे इ़ंजीनियर है । दुसरा यहीं पर प्रोफेसर है । तीसरा और चौथा अभी पढाई कर रहे हैं । किसी की भी अभी शादी नही हुई है। [Devanagari]

rɑm ke cɑr bete hɛ. pəhlɑ betɑ dilli me ĩnɟinɪər hɛ. dusrɑ jəhĩ pər prophesər hɛ. tisrɑ ɒr cɒthɑ ɜbhi pərhɑi kər rəhe hɛ. kisi ki bhi ɜbhi ʃɑdi nəhi hui hɛ [IPA]

rɑm ke cɑr bete hɛ

[Ram has four sons. The first son is engineer in Delhi. The second one is professor over here. The third and fourth are still studying. None of the them are married till now]

Since there is no judgment or any thing else involved in this text except plain statement of facts, it is a neutral text. However if there is a slightest hint of any kind of judgment or instruction is observed then that text cannot be neutral. So the following text is not neutral since now there is a subjective evaluation of the facts which may differ according to different persons and situations.

ऱाम के चार **अच्छे** बेटे हैं । पहला बेटा दिल्ली मे इ़ंजीनियर है । दुसरा यहीं पर प्रोफेसर है । तीसरा और चौथा अभी पढाई कर रहे हैं । किसी की भी अभी शादी नही हुई है। [Devanagari]

rɑm ke cɑr **ɜcche** bete hɛ. pəhlɑ betɑ dilli me ĩnɟinɪər hɛ. dusrɑ jəhĩ pər prophesər hɛ. tisrɑ ɒr cɒthɑ ɜbhi pərhɑi kər rəhe hɛ. kisi ki bhi ɜbhi ʃɑdi nəhi hui hɛ [IPA]

[Ram has four **good** sons. The first son is engineer in Delhi. The second one is professor over here. The third and fourth are still studying. None of the them are married till now]

The neutral texts are not refutable by different persons and in different situations.

3.2. Appropriate Text

An appropriate text contains as much elements of politeness as is required. This tag is equivalent to the Wattsonian concept of 'politic' text. Thus most of the instances where the speakers/authors use a language which cannot be termed impolite, the text could be termed appropriate. So for example in Hindi if someone is talking to an elder or a stranger or someone whom one respects then the use of honorific pronoun and honorific verb form or the use of subjunctive form of verb with someone who shares a very formal relationship will be appropriate. It is to be noted that the 'appropriate' usage is the unmarked usage in the language in the sense that ***not***

using these kinds of markers will make the utterance impolite but using them will be taken as 'normal' and will not be marked as 'polite'. These are the expected ways of interaction in the language and thus their use goes unnoticed but not using them is considered bad.

3.3. Polite Text

Polite text contains elements of politeness more than is required for it to be appropriate. These are neither the expected norms of interaction nor are they generally used in the interaction. But sometimes the speakers/authors, out of enthusiasm or a desire to show extra respect/ give extra attention use some of the politeness markers which are not required. These are the marked forms of polite behaviour in any language. So if they are not used then the utterance remains unmarked but if they are used then the utterance becomes positively marked and others may comment on the extra polite behaviour of the speakers. However if it goes really overboard then there is also the danger of it slipping into the domain of 'too polite' which is not considered good and so such sentences should be marked 'impolite' by the annotators. Let us take an example of comments on the blog. Comments like 'बहुत अच्छा' bəhut ɜcchɑ [very good] or 'बहुत बढिया' bəhut bərhij ɑ [very nice] or 'बधाई' bədhɑi [congrats] etc are considered 'appropriate since it is expected that when you read someone's blog you acknowledge that in a good way. However when the comments become more than this customary greeting and takes a form like 'बहुत बहुत अच्छा' bəhut bəhut ɜcchɑ [very very good] or 'इतनी सुन्दर कविता मैने आज तक नही पढी' itni sundər kəwitɑ mɛne ɑɟ tək nəhi pərhi [I have not read such a beautiful poem till today] then it becomes an instance of 'polite' text since the commentator is using the intensifiers in congratulating more than it is sufficient.

3.4. Impolite Text

The kind of text which contains elements of impoliteness. It includes all instances over-politeness, use of inappropriate lexical items like slang, not using proper mitigation strategies while attacking someone, etc. In short all that is none of the above three could be classified as 'impolite' text.

4. Annotation of the Corpus

A total of 30,000 texts from the corpus are manually annotated by four annotators. Each blog post or blog comment is taken as a text. Thus a complete post\comment is annotated with the politeness information.

Inter-annotator agreement among the three annotators is calculated using 150 texts to see how far the annotators agree on their judgment of texts as far as politeness level is concerned. Since calculation of inter-annotator requires that the same text is annotated by all the annotators and the number of texts in this case is very large, it was not possible to calculate the agreement value on all the texts. Consequently only a tiny subset of the corpus was used to calculate the inter-annotator agreement and taken as a proxy for the whole corpus.

I have calculated both the percentage and the Fleiss' Kappa so that the agreement measure of both kinds (taking chance into account and without taking chance into account) is calculated.

4.1. Calculating percentage agreement

The simple percentage of agreements among the four annotators is summarised in Table 2. It is calculated using the simple formula of percentage, (Sum of agreed instances x 100/Total Number of Instances)

Annotators	Percentage Agreement (Exp 2)	
	4 tags	**3 tags**
A and B	79.0	90.0
A and C	57.0	86.0
A and D	84.0	91.0
B and C	61.0	88.0
B and D	84.0	94.0
C and D	60.0	90.0
A, B, C and D	48.0	81.0

Table 2: Percentage agreement among the annotators

4.2. Calculating Fleiss' Kappa

Fleiss' Kappa is a generalization over Scott's pi to calculate the inter-annotator agreement among more than 2 annotators. Since the present experiment involved four annotators, Fleiss' Kappa, generally considered more reliable and accurate than percentage calculation was also calculated. In order to arrive at a better picture vis-a-vis the percentage agreement as well as see if the overall agreement is affected by one annotator, both the inter-annotator agreement in between each pair of annotators as well as the overall agreement is also estimated. The values of Fleiss' Kappa for each pair of annotator is summarised in Table 3.

Annotators	Fleiss' Kappa (Exp 2)	
	4 tags	3 tags
A and B	0.66827416	0.8024178
A and C	0.32230315	0.7367966
A and D	0.7464186	0.8280469
B and C	0.38450804	0.77630615
B and D	0.7397632	0.8942793
C and D	0.35568196	0.8022122
A, B, C and D	0.53590107	0.80671656

Table 3: Fleiss' Kappa for inter-annotator agreement among the annotators

It is observed that as the number of classes for classifying the text increased, the agreement decreased significantly. If the annotators were asked to annotate using all the four classes then kappa was approximately 0.53. However if the distinction between 'appropriate' and 'polite' was removed and the annotators were asked to classify the text in one out of only three categories viz., 'neutral', 'polite' and 'impolite' then the agreement dramatically increased to 0.80.

4.3. Automatic annotation of the corpus

The manually annotated text is used to train and test a C4.5 classifier. I used the C4.5 implementation included in the Mallet package. The training file was given in SVM light format. The feature set consisted of three kinds of features – unigrams, bigrams and polite linguistic structures of Hindi which were manually identified.

4.3.1. Linguistic Structures

The specific linguistic structures used as features for training the classifier are discussed below.

<u>Subjunctives:</u> The subjunctive form of the verb is formed by adding -ẽ suffix to the last element of the verbal complex (leaving the copula) in Hindi. These are used as a very prominent politeness marker in Hindi, especially in formal contexts.

1	अगर	मुनासिब	समझे	तो
IPA	əgər	munɑsib	səmɟ^{h}e	t̪o
Gloss	if	proper	think	then
	मुझे	भी	अपने	समाज
	muɟ^{h}e	b^{h}i	ɜpne	səmɑɟ
	i.ACC	also	own	society
	में	शामिल	करें	
IPA	mẽ	ʃɑmil	kərẽ	
Gloss	in	include	do.SUBJ	
Free Translation	If you think it to be proper then please include me also in your society.			

<u>Honorifics:</u> In Hindi, verbs agree with nouns and pronouns with respect to their honorificity. The +honorific forms of the verbs are generally formed by adding -ije suffix to the TAM bearing element(s) of the verbal complex. This form of the verb is generally used to show respect to the elders. However it could also be used with the strangers, irrespective of their age, and in some cases with the acquaintances also as a mark of respect. While in such situations this +honorific marker is not required, its use sends positive signals to the hearer about his/her face considerations by the speaker. An example is given below

2.	उम्मीद	है	मैं
IPA	ʊmmid̪	hɛ	mɛ̃
Gloss	expected	is	i
	भी	कभी	ऐसा
	b^{h}i	kəb^{h}i	ɛsɑ
	also	sometime	like this
	लिख	पाऊंगा...	अगर
	lik^{h}	pɑʊ̃ŋgɑ...	ɜgər
	write	ECV…	if

	कोई	स्पेशल	टिप्स
IPA	koi	əspesəl	tips
Gloss	any	special	tips
	हो	तो	ज़रूर बताईएगा.
IPA	ho	t̪o	zərur bət̪aiegɑ
Gloss	be	then	necessarilytell.HON
Free Translation	It is expected that I shall also be able to write like this some day… if there is some special tips then do tell me.		

Suggestion Marker: Suggestion markers (deontics) are one of the prominent ways of marking politeness in Hindi, especially in informal contexts. An example is given below

3	अनाम	भाई	को	ज़्यादा
IPA	ɜnɑm	bʰɑi	ko	zjɑd̪ɑ
Gloss	anonymous	brother	ACC	excessive
	क्रोध	नहीं	करना	चाहिये।
	krod̪ʰ	nəhĩ	kərna	cɑhije
	anger	NEG	do	ECV.DEO
Free Translation	Anonymous brother should not carry excessive anger.			

Ability Marker: Just like the suggestion markers, ability markers (or, epistemic modals) may also used to mark politeness in Hindi. While suggestion is indicated by the light verb 'cɑhnɑ', ability is indicated by the light verb 'səknɑ' in Hindi. An example is given below

38	स्वास्थ्य	से	सम्बंधित
IPA	swɑstʰjə	se	səmbəndʰit̪
Gloss	health	about	related
	भी	जानकारी	के लिए
IPA	bʰi	ɟɑnkɑri	ke lije
Gloss	also	information	for
	कभी	भी	किसी
	kəbʰi	bʰi	kisi
	anytime	also	any
	आप	फोन	भी
IPA	ɑp	pʰon	bʰi
Gloss	you.HON	phone	also
Free Translation	For any kind of information related to the health anytime you could call give a call.		

Conditional Sentences: These structures are one of the most common ways of face-threat mitigation in blog comments. The commentator begins with a canonical praise of the blog post and then goes on to point out the mistakes or fallacies in the post. An example is given below

5	बेहतरीन	अभिव्यक्ति!	परन्तु
IPA	beht̪ərin	ɜbʰivjəkt̪i!	pərənt̪u
Gloss	marvellous	expression!	but
	हर	रचना	में
	hər	rəcnɑ	mẽ
	every	composition	in
	इतनी	उदासी	क्यों?
IPA	it̪ni	ʊd̪ɑsi	kjõ?
Gloss	so much	sadness	why
Free Translation	Marvellous expression! But why is there so much of sadness in every composition?		

4.3.2. Feature Selection and training

Only about 25% of the total available features were included in the feature vector for training. Those features which had a frequency of less than 10 were excluded from the training feature vector. Moreover only those features which had a Gini Index of more than 0.35 were included in the training feature vector. For each feature its Gini Index was used as its weight for training.

For training and testing 10-fold cross-validation was used and an average of the 10 trials are taken as the average accuracy. The classifier gave an accuracy of around 78% on the test set, which is within 2% of human accuracy. This result is almost at par with the current state-of-the-art reported for English texts (Danescu-Niculescu-Mizil et al (2013)) Considering the high accuracy given by the classifier, I used it for training the complete corpus with the politeness value of each text.

5. Possible applications

The resource thus created could prove to be very useful for both theoretical studies as well as in different computational applications ranging from machine translation to language pedagogy. It could also be used for assisting the the non-native speakers during the inter-cultural communication with the native speakers and helping them avoid miscommunication because of the lack of pragmatic competence related to politeness evaluations.

6. References

Brown, Penelope, and Stephen Levinson. 1978. Universals in Language Usage: Politeness Phenomena. In *Questions and Politeness*, ed. Esther Goody, 56–289. Cambridge: Cambridge University Press.

———. 1987. *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL 2013*

Leech, Geoffrey. 1983. *Principles of Pragmatics*. London: Longman.

———. 2007. Politeness: is there an East-West divide? *Journal of Politeness Research*. 3:2, pp. 167-206

Terkourafi, Marina. 2001. Politeness in Cypriot Greek : A Frame-based Approach. Ph.D. Thesis, University of Cambridge.

———. 2003. Generalised and Particularised Implicatures of Linguistic Politeness. In *Perspectives on Dialogue in the New Millennium.*, ed. Peter Kühnlein, Hannes Rieser, and Henk Zeevat, 149–164. Amsterdam: John Benjamins Publishing Company.

———. 2005. Beyond the Micro-level in Politeness Research. *Journal of Politeness Research. Language, Behaviour, Culture* 1 (2) (July): 237–262. doi:10.1515/jplr.2005.1.2.237. http://www.degruyter.com/view/j/jplr.2005.1.issue-2/jplr.2005.1.2.237/jplr.2005.1.2.237.xml.

Watts, Richard J. 1989. Relevance and Relational Work : Linguistic Politeness as Politic Behavior. *Multilingua* 8 (2/3): 131–166.

———. 2003. *Politeness*. Cambridge: Cambridge University Press.