Handling Plurality in Bengali Noun Phrases

Biswanath Barik
Innovation Lab Kolkata
Tata Consultancy Services
1B, Ecospace, Rajarhat, Kolkata – 700156
biswanath.barik@tcs.com

Sudeshna Sarkar
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Kharagpur, India -721302
shudeshna@gmail.com

Abstract

Plurality of a Bengali Noun Phrase (NP) is not always determined by the plurality of its governing member (or the *head*). It is often seen that an NP is plural but the plurality is indicated through qualifiers or other means whereas the head noun has the singular form. In such scenarios, the plurality of the NP is determined by analyzing its non-head members or from other components (or context) of the sentence. Classification of Bengali NPs with respect to plurality is important for many applications including Machine Translation (MT) from Bengali to other language (say Hindi). The plurality of NPs in other languages like Hindi and English is always indicated by plurality of head irrespective of the plurality of the qualifiers. In this paper, we have investigated different sources from where the plurality information of head noun (or NP) can be collected and proposed an approach to automatically classify Bengali NPs by analyzing the identified sources.

1 Introduction

Identification of grammatical properties (or features) of different syntactic units of a sentence is a major task of Natural Language Understanding (NLU). Text processing tasks like context-sensitive spell checking, Named Entity Recognition (NER), Word Sense Disambiguation (WSD), parsing etc., which require detail grammatical description of the context, spend a considerable amount of processing time to identify such syntactico-semantic properties of the input sentence. Rule-based MT, on the other hand, not only explores such grammatical properties through multi-level analyses of a source language input sentence to decode the meaning, but also try to map them correctly to the target language so that a grammatically correct target sentence is generated. Therefore, identification of different grammatical properties of interest is a prerequisite for many text processing jobs and associated applications.

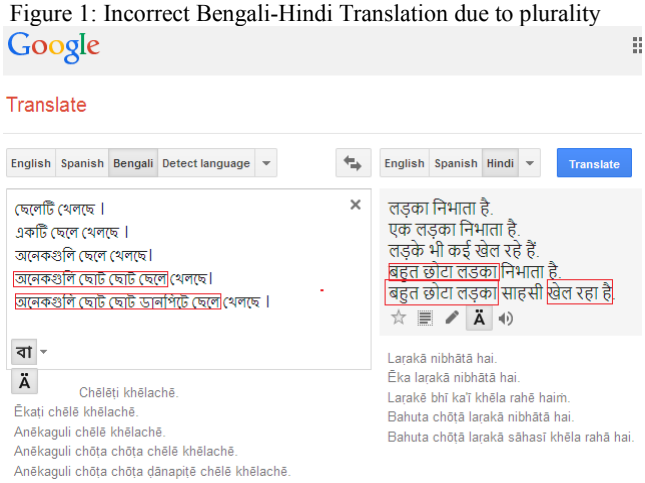
Unfortunately, the essential grammatical properties of different components of the sentence are not always directly expressed in a language. Some of them are hidden in the context or are missing. The hidden features can be identified by analyzing the context. For example, the *gender* of an anaphoric personal pronoun can be determined, in general, by the *gender* of the referring noun. In some languages (e. g., Hindi), the *gender* of such pronouns can be traced by analyzing the verb group (VG) as the inflection of the VG is based on the *gender* of the subject.

The problem of identifying hidden grammatical features becomes difficult if the context is very short or is syntactically ambiguous. In such a case, the most logical interpretation of the context has to be identified which requires domain knowledge and pragmatic interpretation of the context.

During the syntactic analysis of Bengali sentences (for developing a rule-based Bengali to Hindi Machine Translation or BHMT), we observe some mismatches on grammatical features between Bengali and Hindi constituents. Bengali and Hindi have strong agreements within/among constituents on some grammatical features like *person* and *case.* In Bengali, these features are identifiable by analyzing the corresponding surface words (nouns and verbs) and, therefore, are also available for target (Hindi) sentence generation. However, some

other features like *number* and *gender* which play an important role in Hindi sentence generation or agreement are either hidden in some cases or are missing in Bengali side. For example, the *gender* feature is usually missing in Bengali – neither the verb takes *gender* inflection not it can be traced from the morphological analysis of the subject noun phrase. The *number* or plurality information, on the other hand, is sometimes hidden in the context. Such missing/hidden features in source Bengali sentences have to be determined correctly so that they are used in correct target (Hindi) sentence generation which improves the translation accuracy of Bengali-Hindi MT. Figure 1 shows that the Google's Bengali-Hindi translation system fails to handle plurality issue statistically and thus produces incorrect translation.

Figure 1: Incorrect Bengali-Hindi Translation due to plurality



In this work, we discuss how to determine the *plurality*, an implicit morphological feature of Bengali Noun Phrases (NPs). In general, the head nouns of NPs hold the plurality and the NPs are accordingly classified as *singular* or *plural*. However, we have observed that the head nouns of Bengali NPs do not always take plural inflections. In such cases, the plurality of the NP is determined from the contextual information of the sentence. The automatic text processing and allied applications like Machine Translation thus require a computational approach to automatically identify the plurality of Bengali NPs. Sufficient resources like domain knowledge and infrastructures for analyzing large or inter-sentence context are required to handle such issues exhaustively. Using limited available resources, we propose a rule-based approach to classify Bengali NPs as *singular* or *plural* by analyzing different sources of plurality and have achieved an accuracy of 73.12%.

The paper is organized as follows: Section 2 describes the differences in plurality expression in Bengali and Hindi NPs. Section 3 illustrates a comparison on plural inflection of Bengali and Hindi lexical classes and shows how Bengali to Hindi machine translation is affected due to improper classification of Bengali NPs on plurality. Section 4 investigates different sources of plurality in Bengali sentences. Section 5 describes our approach in three steps to classify NPs by determining the number of the head noun (or equivalent). In the first step, the input sentence is analyzed at different syntactic level and produces some structural representation of the sentence which helps to identify different phrases and their heads, qualifiers of the heads and other necessary information. In the second step, the quantifiers are classified as *singular* or *plural* (as quantifiers are the major source for determining the number of head noun) and, in the last step, the NPs are classified based on the plurality of the quantifiers. Some syntactic patterns are also identified which helps to identify the plurality of NPs. Section 6 shows experimental result of our approach on 1000 sentences chosen randomly from ILMT Bengali corpus[1]. Section 7 categorizes the misclassification errors and justifies why proposed approach fails to classify them properly. Section 8 summarizes our work and concludes with future scope of improvements.

2 Related Work

The difference in plurality expressions in Bengali and Hindi NPs is an example of grammatical divergences between Bengali and Hindi. Some works are found in literature which address such divergence phenomena in specific language pairs. (Dave et al., 2001) addresses the divergence issues between English and Hindi from the perspective of computational linguistics which includes various aspects of syntactic and lexico-semantic divergences. Also, a considerable amount of work is done by the linguistic community on this issue in Indian and Western languages (Bholanath, 1987; Gopinathan, 1993). (Das, 2013) discussed different type of divergences observed in English to Bengali machine translation.

[1] ILMT Bengali corpus is created as a resource for Bengali-Hindi MT system under consortium project ILMT Phase I.

In Bengali, a considerable amount of work is done on different morpho-syntactic analyses like morphological analysis/synthesis (Dasgupta et al., 2004; Bhattacharya et al., 2005), Part-of-Speech (POS) tagging (Dandapat, 2009), chunking (Avinesh et al., 2007; Dandapat, 2007), parsing (Ghosh et al., 2009) etc.

The plurality issue in different foreign languages like Chinese (Bošković et al., 2012), Turkish (Walter, 2014) are studied and NP classification on plurality is also reported in (Li, 1999; Dryer, 2005). In Bengali, some work (Chacón. 2011; Biswas, 2012) are reported on the linguistic analysis of plurality.

3 Plurality in Bengali and Hindi NPs

Morphologically, nouns (NNs) and pronouns (PRPs) in Bengali, Hindi and English take null (ø) suffix for *singular* number. The *plurality* is indicated by the plural inflection or suffix attachment with the noun or pronoun. Generally, the plurality of the NP is determined by the plurality of the head noun as the head is the main (or dominating) member of the NP. However, unlike Hindi and English, the plurality of Bengali NP is not always determined by the plurality of the head. Instead, other non-head members of the NP like qualifiers, reduplicative adjectives etc. indicate the plurality of the NP. Examples (1a) and (1b) illustrate the divergences of plurality in Bengali, Hindi and English NPs. Bengali, Hindi and English examples are denoted by B: and H: and E:, respectively.

(1a) B: (সুন্দর/JJ ছেলেগুলো/NN) NP মাঠে খেলা করছে।
beautiful boys ground-in are playing

H: (सुंदर/JJ लड़के/NN)NP मैदान में खेल रहे हैं।
beautiful boys ground in are playing

E: (Beautiful/JJ boys/NN)NP are playing in the ground

(1b) B: আজকে (পাঁচজন/QF ছেলে/NN) NP স্কুলে এসেছে ।
today five-cl boy school-to came

H: आज (पांच/QF लड़के/NN)NP स्कूल में आये हैं।
today five boys school to came

E: (Five/ QF boys/NN)NP came to school today.

Example (1a) shows that the plurality of NP is kept in the head noun in Bengali, Hindi and English sentences. However, in example (1b), the head noun of Bengali NP does not hold plurality and it is indicated by its quantifier. On the other hand, the head of corresponding Hindi and English NP hold the plurality information although the plural quantifier is present in the NP.

In summary, it can be said that the plurality information is present in the head of a plural Hindi or English NP irrespective of the plurality of the dependents whereas, in Bengali, the head contains the plurality if the other non-head members of a plural NP has no plurality indicator.

4 Number Issues in Bengali-Hindi MT

In Bengali, the nouns and pronominal entities take inflection on *number*. In some cases, the quantifiers also take plural inflection like "গুলো (/gulo/)", "গুলি (/guli/)" etc. or classifiers like "জন (/jan/)" to indicate plurality. The verb forms are not inflected on plurality. Therefore, analyzing verb forms, it is not possible to identify the plurality of the subject (or object) NPs. The other grammatical classes like post-positions, adjectival and adverbial qualifiers etc. do not take plurality inflections. Therefore, the sources of plurality in lexical level are limited to nouns, pronouns or quantifiers. However, in Hindi, if the NP is plural, the plurality information is available in the head, its dependents and the verb group agreeing the subject NP. Table1 shows some representative Hindi and Bengali *singular* and *plural* word-forms taken from different lexical classes where all the plural forms are different than singular forms in Hindi but are not different in Bengali.

Examples (3a) and (3b) show that the plurality differences in lexical level cause higher difference in sentence level. It is observed from the examples that a small difference in Bengali sentences B1 and B2 (i.e., the noun "ছেলে /chhele/" is *singular* in B1 and is *plural* in B2) there has a significant difference in corresponding Hindi sentences H1 and H2 due to strong *number* agreements among/within constituents.

(3a) B1: ওই ছেলেটি স্কুলে পড়ছে ।
that boy-cl school-in studying

E: 'That boy is studying in the school.'

H1: वो लड़का स्कूल में पढ़ रहा है।

(3b) B2: ওই ছেলেগুলি স্কুলে পড়ছে ।
that boys school-in studying

E: 'Those boys are studying in the school.'

H2: वो लड़के स्कूल में पढ़ रहे हैं।

Table 1: Lexical Categories with Plural Inflection

Lexical Classes	Inflections with Number	
	Singular	Plural
Noun (Hindi)	लड़का (boy)	लड़के (boys)
	कुर्सी (chair)	कुसियाँ (chairs)
Noun (Bengali)	ছেলে (boy)	ছেলেগুলো (boys)
	চেয়ার (chair)	চেয়ারগুলো (chair)
Pronoun (Hindi)	मै (I)	हम (we)
	तू (You)	तुम (you)
Pronoun (Bengali)	আমি (I)	আমরা (we)
	তুমি (You)	তোমরা (you)
Adjective (Hindi)	काला (black)	काले (black)
Adjective (Bengali)	কালো (black)	
Post-Position(Hindi)	का ('s)	के ('s)
PostPosition(Bengali)	এর ('s)	দের ('s)
Verb (Hindi)	हूँ (be)	हैं (be)
	था (was)	थे (were)
Verb (Bengali)	হই (be)	
	ছিলে (was)	

Therefore, identifying the correct plurality of each Bengali NP helps in Hindi side agreement among the NP members (head and associated dependents) as well as the inter-constituents agreements like subject-verb, object-verb etc. during the translation from Bengali to Hindi.

5 Plurality Indicators in Bengali NPs

We have discussed in section 2 that the plurality of Bengali NP is not always indicated by the plurality of the head. Plurality can be collected from other sources of the sentence. In this section we investigate different sources (or contextual patterns) from where the plurality information can be collected in Bengali.

5.1 Existence of Quantifiers in NPs

The quantifiers are the major plurality indicators in Bengali noun phrases. NPs containing plural quantifiers are plural.

The quantifiers may appear at different distances from the quantifying nouns. With the increasing distance between a quantifier and a noun, it becomes difficult to relate them without having deep syntactic processing (parsing) of input sentence. To address this issue, we have divided quantifier-noun distance into three categories as described below and try to solve them separately with available resources and infrastructures.

Quantifiers in Short Range (SR) - quantifier appears just before the head noun (4a).

(4a) (কিছু/QC ওষুধ/NN) NP
Some medicine
'some medicines'

Quantifier in Medium Range (MR) - other noun modifiers (but not modifying phrases) occur in between the quantifier and the head (4b).

(4b) (কয়েকটা/QC ভালো/JJ ভালো/JJ কথা/NN) NP
few good good word
'few good words'

Quantifier in Long Range (LR) – other modifying phrases exists in between quantifier and its noun appear (4c).

(4c) (কয়েকজন/QC (আদিবাসী সম্প্রদায়ভুক্ত) JJP, (অতি সাধারণ মানের,) JJP অস্থায়ী/JJ কর্মচারী/NN) NP
some tribal community-belongs-to, very general quality, temporary worker
'Some temporary workers, belonging to tribal community and of very general quality'

5.2 Quantifiers outside NPs

In some typical cases, it is observed that the *numeral classifiers* are positioned after the head to show strong definiteness (Simpson, 2011) as shown in (5a) and (5b).

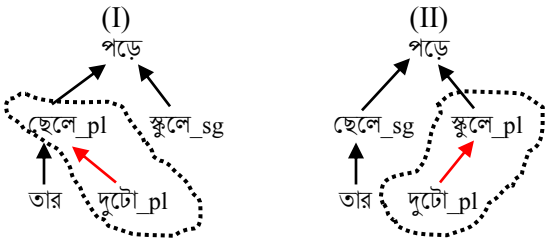
(5a) বছর/NN পাঁচেক/QC পরে সে আবার ফিরে এলো।
year five after he again returned back

'He again retuned back after five years'

(5b) রাজা নাটকের প্রথম সংস্করণে গান/NN ছিল মোট বাইশটি/QC।
King drama-gen first edition-in song was total twenty-two
'There were a total of twenty-two songs in the first edition of the drama "Raja"'

The major issue in handling such constructions is to dealing with the ambiguity. If a quantifier occurs in between two nouns, then the quantifier may be associated with the previous or the following noun. This causes semantic ambiguity as illustrated in 5(c).

(5c) তার ছেলে/NN দুটো/QC স্কুলে/NN পড়ে।
his boy two school-in reads



'His two children read in school,' 'His child reads in two schools'

In (I), the noun "ছেলে (/chhele/)" is plural (pl) as it is quantified by the following plural quantifier "দুটো (/duto/)" and the other noun "স্কুলে (/skule/)" is singular (sg) whereas in (II) it is reverse as the quantifier modifying the followed noun.

5.3 Plurality Collected from other Nouns

In some constructions, the plural nouns with (optional) modifiers are connected through coordinating conjunctions to form a larger NP unit. In such constructions, none but the last noun contains the plurality information of the NP.

(6) (রাজনৈতিক/JJ নেতা/NN ,/CC মন্ত্রী/NN ও/CC আমলাদের/NN) NP চাপে খুনের কেসটা চাপা পড়ে গেল।
political leader minister and bureaucrats pressure-due-to murder-gen case-cl suppresses-was

'The murder case was suppressed due to the pressure from political leaders, ministers and the bureaucrats'

In example (6), the plural NP consists of three nouns "নেতা (/netA/)", "মন্ত্রী (/mantrI/)" and "আমলা (/AmalA/)" whereas the plural inflection is visible in the last noun. Plurality of others can be collected from the last noun.

5.4 Reduplication & Compounding

Plurality of the phrase may be determined from the reduplication of the head noun (7a) or its modifiers (7b) and compounding (7c).

(7a) (পথে/NN পথে/NN) NP অবরোধ
road-in road-in obstruction
'obstructions in the roads'

(7b) (ছোট/JJ ছোট/JJ নদী/NN) NP
small small river
'small rivers'

(7c) (ভালো-মন্দ/JJ কথা/NN) NP
good-bad word
'good-bad words'

5.5 NP with Elliptic Phenomena

In some typical Bengali constructions (8), the noun is implicit and the inflection of the noun is attached with its modifiers. This phenomenon is an example of *ellipses*. In this type of construction, the resulting quantifier semantically behaves like a plural pronoun if the quantifier is plural.

(8) এই সিনেমাটি অনেকেই দেখেনি ।
This film-cl many-people seen-not
'Many people did not seen this film'

5.6 Spatio-Temporal Patterns (PAT)

Spatio-temporal entities (also referred as Nouns denoting Space and Time or NST) along with compatible post-position (PSP) holds the plurality of the noun without having any inflection.

(9) বছরের পর বছর ধরে কাজটা চলছে ।
year-of after-year work-cl is-in-porgress
'The working is in progress for many years'

In example (9), instead of taking the plural suffix, the temporal noun "বছর (/bachhar/)" replicate itself

with PSPs "পর (/par/)" and "ধরে (/dhare/)". In this type of patterns, the NST is plural.

6 NP Classification Approach

In this section, we will discuss how to identify the plurality of Bengali noun phrases. The NP classification process consists of three steps. In the first step we will analyze the input Bengali sentence and identify different phrases and necessary grammatical (morphological, syntactic) properties associated with the sentence. As we mentioned in section 4 that quantifiers are the major indicators of plurality, we will classify the quantifiers as *singular* or *plural* in the next step. Finally, we will classify NPs by applying some predefined rules on analyzed structure of the sentence. The following sub-sections elaborate the steps.

6.1 Syntactic Analysis of Input Sentence

To solve the plurality issue exhaustively, a detailed, multi-level syntactic analysis of input sentence is required which spans from word level to sentence level. Also, in some ambiguous cases, the analyzed structure needs to be represented in discourse or pragmatic level. In this experiment, we limit the analysis phase in the following steps due to resource constraints.

Morphological Analysis: Input sentences are tokenized and each word-token is analyzed morphologically with a rule-based Bengali Morphological Analyzer (MA) developed at IIT Kharagpur. MA produces all possible analyses of input word. Words having plural suffixes are identified as plural nouns in this step.

Part-of-Speech (POS) Tagging: A CRF based Bengali POS tagger is trained on ILMT POS tagged corpus and used to identify the grammatical functionalities of each word in the sentence.

Chunking: Local dependencies among consecutive words are identified using CRF-based chunking model trained on ILMT corpus with further post-processing (Chatterji et al., 2012).

Morphological Pruning: As a word-token may have more than one morphological interpretation, morphological analyses having context incompatibility are pruned out using the method specified in (Barik et al., 2014).

Head of the Chunk Identification: Chunking segments the sentence based on local dependencies and classify each segment. The *head* of each chunk is identified using a rule-based approach.

Dependency Parsing: To identify long range modifiers or modifying phrases, we have used a rule-based Bengali (partial) parser which is developed according to (Bharti et al., 2009).

6.2 Classification of Quantifiers

Quantifiers are the important sources of plurality. The plurality of the quantifier implies the plurality of its quantifying NP. Therefore, each quantifier has to be classified correctly.

The quantifiers are identified in the input sentence during POS tagging. The quantifiers are categorized as follows:

Numerals: Non-negative, integer quantifiers greater than one with (optional) classifiers like ১৮৫৭-জন (1857-cl), ২৫-টা (25-cl) are plural quantifiers.

Cardinal Number: Cardinal numbers without having decimal indicator like পাঁচ (five), দশ (ten), পাঁচশ (five hundreds) etc. are plural.

Ordinal Number: Ordinal quantifiers like প্রথম (first), চতুর্থ (fourth), একাদশ (eleventh) are not plural.

Indefinite Quantifier: For the case of indefinite quantifiers, classifiers play an important role for their plurality. Classifiers are used in Bengali to combine NPs with numerals and quantifiers (Biswas, 2012). For example, the indefinite quantifier "অনেক (/anek/)" with classifier "জন (/jan/)" is plural but with classifiers like "টা (/tA/)" or "খানা (/khana/)" is not plural.

Another observation is that some indefinite quantifiers take regular plural suffixes like "গুলো (/gulo/)", "গুলি (/guli/)" etc. and, thus, are plural. Quantifiers do not take the *associative plurals* (Chacón, 2011) like "রা (/ra/)" as they are special markers for animate nouns (Moravcšik et al., 2013).

Table 2: Plurality of Quantifiers w.r.t Classifiers

	ø	টা/টি	খানা/খানি	জন	গুলো/গুলি
কিছু	Y	×	NA	Y	NA
অনেক	Y	×	×	Y	Y
বহু	Y	NA	NA	Y	NA
কত/যত/এত	Y	×	×	Y	Y
সমস্ত	Y	×	NA	Y	NA
সব	Y	×	NA	×	Y
কতক	Y	×	NA	Y	Y

To determine the plurality of classifier attached indefinite quantifiers, we have prepared a quantifier set (Q) and a classifier set (C) and examined the plurality of each member of Q X C. Some examples are shown in Table 2. Here, "Y" and "×" denote that attachments are valid where the quantifiers are plural for the first case and not for the second. "NA" refers to invalid classifier attachment.

6.3 Rules for Classify NPs

With syntactic analysis of the input sentence, some information of each word-token (or different grammatical units) are found. Each quantifier in the sentence is identified and classified. Other information like measuring units for mass nouns, nouns denoting space or time etc. are listed. Therefore, if any NP is plural, the information is explicitly available in the analyzed structure of the sentence. We have developed sixteen rules to look for the plurality information within chunk members or context. Once the plurality is determined by matching a rule, the plurality information is copied to the head of the chunk so that plurality of the chunk is explicitly identified and it becomes available for grammatical functionalities like agreements. Some of the rules are explained below. Rules are denotes by 'R'.

Case I: *Existence of measuring units like "মিটার"* (meter)*, "পাউন্ড"* (pound) *or "ভোল্ট"* (volt) *in NP denotes mass noun and, thus, NP is not plural.*

Case II: *NP having numerals attachment with temporal nouns like "১৬ নভেম্বর" (16 November) or "১৭৯৯ সালে" (in 1799 year) are not plural.*

Case III: *NPs having plural quantifiers are plural.*

(10) এখানে (পাঁচজন/QC_pl লোকের/NN_sg)NP থাকার ব্যবস্থা আছে।
here five-cl people-gen accommodating-for arrangement has
'Here, there is an arrangement for accommodating five people'

R: (QC_pl (.*?) NN_**sg**) NP => (QC_pl (.*?) NN_**pl**) NP

Case IV: *If a NP consists of multiple quantifiers connected with coordinating conjunctions and the rightmost quantifier is plural then the NP is plural.*

(11) (এক/QC_sg, দুই/QC_pl বা পাঁচ/QC_pl মাস/NN_sg) NP বাদেও সে আসতে পারে।
one, two or five month after-also he come can
'He can also come after one, two or five years'

R: (QC* QC_pl NN_**sg**) NP => (QC* QC_pl NN_**pl**) NP

Case V: *Reduplications of temporal nouns with certain patterns (i.e., combination of temporal nouns with specific post-position(s)) are considered as plural nouns. If the pattern matches with the rule, the corresponding pattern is modified.*

(12) (বছরের/NN পর/PSP) NP (বছর/NN ধরে/PSP) NP এই রীতি চলে আসছে।
year-of after-year this ritual continued-over
'This ritual is continued over years'

R: (NNPSP%পর)NP (NN PSP%ধরে)NP
=> (NN_pl PSP%ধরে)NP

Case VI: *Reduplication of temporal nouns with no post-position attachments are not plural.*

(13) (বছর/NN) NP (বছর/NN) NP দূর্গপূজা হয়।
year year Durgapuja-celebration is
'Durgapuja is celebrated every year'

Case VII: *Reduplication of animate ('a') and inanimate ('i') nouns (identified during morphological analysis) without having post-position attachments are plural.*

(14) (গাছে/NN) NP (গাছে/NN) NP ফুল ফুটেছে।
tree-on tree-on flower blossomed

'Flowers have blossomed on the trees'

R: (NN_sg && case='এ' && type= 'a' or 'i') NP (NN_sg && case='এ' && type= 'a' or 'i') NP => (NN_pl && case='এ' && type= 'a' or 'i') NP

Case VIII: NPs having p*ronouns (PRPs) with reduplications are plurals*.

(15) (যাহা/PRP) NP (যাহা/PRP) NP বলি মন দিয়ে শুন।
what what saying carefully listen
'Listen carefully to what I am saying'

R: (X%PRP_sg) NP (X%PRP_sg) NP => (X%PRP_pl) NP

Case IX: *NPs having adjectival reduplications are plural.*

(16) এই পুকুরে (বড়/JJ বড়/JJ মাছ) NP আছে।
this pond-in big big fish have
'Big fishes are there in this pond'

R:((.*) X/JJ X/JJ (.*) NN_sg)NP=>((.*) X/JJ (.*) NN_pl) NP

7 Evaluation

To evaluate our rule-based NP classification method, we choose 1000 sentences randomly from ILMT corpus as test data. The sentences are analyzed by different steps as specified in section 5.1. The analyzed structures are manually examined and the errors are removed. The 1000 test sentences thus are annotated and are considered as ground truth (or gold standard data).

In the test 1000 sentences, we found 5817 noun chunk (NP) and a total of 5109 nouns which includes both singular (3572), plural (1271) and verbal nouns (39). Among these singular nouns identified by morphological analysis, 227 are actually plural which are detected and marked during manual annotation. The Bengali Morphological Analyzer (MA) fails to analyze them as plural because no plural suffix is attached with these nouns and MA does not consider the context to determine the number of any noun. As a result, 227 NP chunks are misclassified as singular NPs and they are marked in the test data.

Table 3: NP Classification Results

Target Plurals NPs	SR	MR	LR	PAT	SEM	MIS	Total
Ground Truth	141	29	23	17	6	11	**227**
Identified	157	33	7	15	0	0	212
Correctly Identified	129	19	4	15	0	0	**167**

Table 3 shows the distribution NPs having plurality kept in non-head members in ground truth, identified by our approach and correctly identified in different type of plurality sources.

From Table 3, we observe that quantifiers as plurality sources in short range (SR) is solved in most of the cases. The fixed pattern-based (PAT) number acquisition correctly captures all such patterns and classified NPs accordingly. However, plurality determinations of NPs based on the plurality of the quantifiers located in medium (MR) and long range (LR) are solved partially. Plurality determination of NPs where short or no contextual information is available or the context is ambiguous enough (SEM) is not worked. A details summary of the results on 1000 test sentences is given below.

Total NPs	Nouns by MA				After NP Classification	
5817	Sg	Pl	Verbal	Total	Sg -> Pl	Correct
GT	3572	1271	39	5109	-	227
Test	3587	1271	39	5109	212	167

The confusion matrix of the result is given below.

	TRUE	FALSE
Singular	3572	15
Plural	167	55

The overall accuracy is **73.12**%.

8 Errors Analysis of Our Method

8.1 NP Misclassification in Local Level

Classification of NPs on quantifier-noun number agreement depends on the type of the head noun. Some nouns are singular always and NPs with such

nouns are also singular. However, with the presence of plural quantifiers, such singular NPs are misclassified as plurals. For example, the specified NP in (17) is singular along with the the singular noun "পৃথিবী (/prithibI/)". The quantifier "সমস্ত (/samasta/)" has two aspects – one is quantitative which means 'all' and it is plural. The other aspect is qualitative which means 'whole'. In this example, the quantifier is realized as plural which is incorrect.

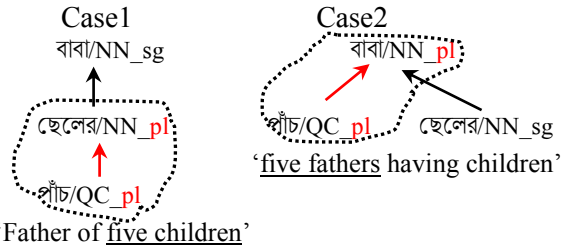
(17) তাজমহল (সমস্ত/QC পৃথিবীর/NN_pl)NP দর্শকদের আকৃষ্ট করে।
Tajmahal for-all earth-gen visitors attracts
'Tajmahal attracts the visitors of the entire world'

To handle this issue, deeper semantic representation of the context is needed.

8.2 Error due to Chunk Level Ambiguity

Quantifier followed by two nouns may quantify either one and, thus, resulting dependency ambiguity as described below.

(18) পাঁচ/QC ছেলের/NN বাবা /NN
five boy-gen father



'Father of five children'
'five fathers having children'

In case1, the noun "ছেলের (/chhelera/)" is plural due to the plural quantifier "পাঁচ (/panch/)" whereas in case2, the noun "বাবা (/bAbA/)" is plural.

8.3 Long Range Agreement Problems

With the unavailability of good accuracy deep parser in Bengali, the (partial) parser is used to identify long range dependencies which fails to identify some of the relations of long range and causes misclassification of NPs on numbers.

(19) দশটা/QC সুন্দর সুন্দর নক্সা করা নরম কাপড়ের তৈরি আসন/NN দেখিয়ে নন্দিতা বলল - "এখানে আপনারা বসুন"।
ten-cl beautiful beautiful designed soft cotton made seat pointing-to Nandita said – "here you-pl sit down"

'Pointing to ten, beautifully designed, soft, cotton-made seats, Nandita said "please sit down here".

Example (19) is an ambiguous sentence because the quantifier "দশটা(ten)" may attached with either "কাপড় (cloth)" or "আসন(seat)". But in this case by using pragmatics we see that since the sentence addresses a plural number of people are asked to sit, the plurality should be attached with "কাপড় (cloth)".

8.4 Unavailability of Contextual Information

Some cases are found where the given context is not sufficient to determine the plurality of the NP.

(20) জায়গাটা (ছেলেতে/NN_sg) NP ভর্তি হয়ে গেছে।
place-cl boy-with filled-up
'*The place is filled up with the boys*'

The NP with noun "ছেলেতে (/chhelete/)" is plural in (20). However, with available contextual information, our approach fails to identify it as plural.

9 Conclusion and Future Work

In this work, we showed that Bengali NPs cannot be classified on plurality based on the plurality of the head nouns. Also, we showed that the plurality issue in Bengali NPs should be resolved correct as it causes errors in different text processing applications (specifically Machine Translation from Bengali to other language like Hindi or English). Different sources and indicators of plurality are studied in this work and noticed that the quantifiers are the major sources of plurality indicator in Bengali. The quantifiers have been classified on plurality and a rule-based approach is proposed to examine noun phrases as singular or plural.

In error analysis phase, we have noticed that some NPs are not classified properly due to the unavailability of sufficient resources. With correct parsing information, domain knowledge and the facilities for analyzing large context in discourse or pragmatic level, the error rate of our approach can be reduced which is our future scope of work.

Acknowledgments

This work is financially supported by MCIT, Govt. of India through the consortium project "Indian Language to Indian language Machine Translation (IL-ILMT) Systems Phase II" vide order no. 11(1)/2010-HCC(TDIL) dated 27/12/2010. The problem addressed here is explored as a part of performance issues in Bengali-Hindi Machine Translation system. The work is done at the Department of Computer Science & Engineering, IIT Kharagpur during the MS course of the first author.

References

Akshar Bharati, Mridul Gupta, Vineet Yadav, Karthik Gali and Dipti Misra Sharma. 2009. *Simple Parser for Indian languages in a Dependency Framework,* In Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP, Singapore, pp. 162-165

A. Simpson, H. L. Soh and H. Nomoko. 2011. *Bare classifiers and definiteness: A cross-linguistic investigation,* Studies in Languages 35.1, John Benjamins Publishing Company, 168-194.

Biswanath Barik and Sudeshna Sarkar. 2014. *Pattern based Pruning of Morphological Alternatives of Bengali Wordforms,* In Proceedings of International Conference on Advances in Computing, Communications and Informatics (ICACCI), New Delhi.

D. A. Chacón. 2011. *Bangla and company: the distribution of associative plurals in Bangla, Japanese, and Mandarin Chinese,* Handout in Formal Approaches to South Asian Languages (FASAL I), University of Massachussets, Amherst

Aniruddha Ghosh, Pinaki Bhaskar,Amitava Das and Sivaji Bandyopadhyay, *Dependency Parser for Bengali: the JU System at ICON 2009*, Kharagpur, India.

Sandipan Dandapat, *Part of Speech Tagging and Chunking with Maximum Entropy Model,* In Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages (SPSAL-2007)

Sandipan Dandapat. 2009. *Part-of-Speech Tagging for Bengali,* MS Thesis submitted at IIT Kharagpur

Niladri Sekhar Dash. 2013. *Linguistic Divergences in English to Bengali Translation*, International Journal of English Linguistics; Vol. 3, No. 1

Edith Moravcšik and Michael Daniel. 2014. *The Associative Plural,* The World Atlas of language Structures Online, http://wals.info/chapter/36, accessed in July 2014

Omkar N. Koul. 2008. *Modern Hindi Grammar*, VA, Dunwoody Press.

Priyanka Biswas. 2012. *Plurality in Classifier Language: Two Types of Plural in Bangla,* In Proceedings of GLOW in Asia IX, Japan

Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya. "*Interlingua-based English–Hindi Machine Translation and Language Divergence.*" *Machine Translation* 16.4 (2001): 251-304

S. Gopinathan and S. Kandaswamy (eds): 1993, Anauvaad kI samasyaae [Problems of Translation], Lokbharti Prakashan, New Delhi, India.

Tiwari Bholanath and Naresh Kumar: 1987, *Problems of translation from various foreign languages*, Prabhat Prakashan, New Delhi, India

S. Dasgupta and M. Khan. 2004. *Morphological paring of Bangla words using PC-KIMMO*, Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT2004, Dhaka, Bangladesh)

S. Bhattacharya, M. Choudhury, S. Sarkar and A. Basu. 2005. *Inflectional morphology synthesis for Bengali noun, pronoun and verb systems*. In Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05), Dhaka, Bangladesh

Sanjay Chatterji, Nabanita Datta, Arnab Dhar, Biswanath Barik, Sudeshna Sarkar, Anupam Basu. 2012. *Repairing Bengali Verb Chunks for Improved Bengali to Hindi Machine Translation,* In Proceedings of the 10th Workshop on Asian Language Resources, Mumbai, pages 65–74

Avinesh. PVS, Karthik G. 2007. *Part-Of-Speech Tagging and Chunking using Conditional Random Fields and Transformation Based Learning,* In Proceedings of IJCAI Workshop on Shallow Parsing for South Asian Languages (SPSAL-2007)

Ž. Bošković and I. T. C. Hsieh, 2012, *On word order, binding relations, and plurality within Chinese NPs*. In Proceedings of the 13th International Symposium on Chinese Languages and Linguistics (pp. 19-47).

Walter, M. J. (2014). Morphosyntax and semantic type of noun phrases in Turkish

Y. H. A. Li, 1999, *Plurality in a classifier language*. Journal of East Asian Linguistics, 8(1), 75-99

Dryer, Matthew S. 2005. "*33. Coding of Nominal Plurality*."