One-Expression Classification in Bengali and its role in Bengali-English Machine Translation

Apurbalal Senapati
CVPR Unit, Indian Statistical Institute
Kolkata, India
apurbalal.senapati@gmail.com

Utpal Garain
CVPR Unit, Indian Statistical Institute
Kolkata, India
utpal.garain@gmail.com

Abstract— **This paper attempts to analyze one-expressions in Bengali and shows its effectiveness for machine translation. The characteristics of one-expressions are studied in 177 million word corpus. A classification scheme has been proposed for the grouping the one-expressions. The features contributing towards the classification are identified and a CRF-based classifier is trained on an authors' generated annotated dataset containing 2006 instances of one-expressions. The classifier's performance is tested on a test set (containing 300 instances of Bengali one-expressions) which is different from the training data. Evaluation shows that the classifier can correctly classify the one-expressions in 75% cases. Finally, the utility of this classification task is investigated for machine translation (Bengali-English). The translation accuracy is improved from 39% (by Google translator) to 60% (by the proposed approach) and this improvement is found to be statistically significant. All the annotated datasets (there was none before) are made free to facilitate further research on this topic.**

Keywords- ***one-expressions; Bengali; corpus, machine translation***

I. Introduction

One-expressions play important role in many areas of NLP, for instance, anaphora resolution, question-answering, machine translation, etc. Consider the following sentences in Bengali: **S1**: এক সময় সেখানে এক রাজা ছিলেন| [ek samay sekhAne ek rAjA chilen/Once upon a time, there was a king.] There are two one-expressions (both are এক/ek) in the Bengali sentence, **S1.** While translating this sentence into English, the first one-expression is translated to "once" and the second one-expression is translated to "a". There are instances when the same one-expression (e.g. এক/ek) is used in an inflected form and is translated to the number "one". For example, **S2**: বাজারে একটাও লোক নেই| [bAzAre ektAo lok nei/there is no one in the market.] In this sentence, the one-expression, একটাও (an inflected form of এক/ek) is translated to "one". Sometimes, the one-expression is not translated at all. For example, consider this sentence, **S3**: রাম ও শ্যাম কে এক করে দেখা ঠিক নয়| (rAm o shyAm ke ek kore dekha thik noi/It is not right to treat Ram and Shyam similarly). In this sentence the one-expression এক/ek(one) has not been translated at all.

The above discussion shows that the same one-expression behaves differently in different context. Therefore, its translation in the target language varies depending upon its particular type (or class). Hence, the classification of one-expressions is an important task to understand their behavior and subsequently to determine their translation. This work aims at doing this task for Bengali.

II. Previous Studies

The computational analysis of one-expression in Bengali has not been explored before. Even there is hardly any linguistic study on classification of one-expressions in Bengali. Computational hardship refers to unavailability of annotated data sets (making one-expressions in sentences and then tagging them with their respective class). Research on machine translation of Indic languages into other language(s) is also gaining importance in the recent times[1] and therefore, one-expressions still have not got a chance to be looked upon. However, statistics show that in Indic languages like Bengali, one-expressions are used often. A study on 177-million-word FIRE Bengali Corpus[2] shows that about 1.34-million words refer to one-expressions. Obviously, they demand additional processing effort for machine translation. This finding conforms to the observation of Hwee Tou Ng et al. [1] who showed the statistical measure of the word *one* in the 100-million-word British National Corpus (BNC) and claimed that one cannot just ignore *one* in any NLP application.

In case of English, one-expressions have been studied while dealing with one-anaphora [2, 3, 4, 5]. Halliday and Hasan [2], Dahl [4] and Luperfoy [5] identify major criteria that distinguish the non-anaphoric uses of *one* from each other. Hwee Tou Ng et al. [1] classified the uses of *one* into six classes: Numeric (John has *one* blue T-shirt), Partitive (A special exhibition of books for children forms *one* of the centrepieces), Anaphoric (Would you like *this book*? Yes, I would like that *one*), Generic (*One* must think a little deeper to discover the underlying social roots of the problem), Idiomatic (It would be perfect to have a loved *one* accompany me in the whole trip), and Unclassifiable (Cursed be *one* who curses you). Out of these classes, they concentrate on the *anaphoric* class and used a machine learning approach to identification and resolve of *one*-anaphora. In our study, we follow the classification scheme of Hwee Tou Ng et al. [1] for classifying the Bengali one-expressions with some extension.

III. Our Contribution

The distinct contributions of this work refer to (i) an exhaustive study of Bengali one-expressions (from a 177

[1] A nation-wide consortium for machine translation of Indic languages is being funded by the Ministry of Information Technology, Govt. of India, http://www.tdil-dc.in.

[2] FIRE: Forum for Information Retrieval Evaluation; http://www.isical.ac.in/~fire/data.html

978-1-4799-5330-1/14/$31.00 ©2014 IEEE

million-word corpus) and their classification, (ii) preparation of two annotated datasets (details is in section V): the first one containing 1806 sentences consisting of 2006 instances of one-expressions and the second one containing 296 sentences consisting of 300 instances of one-expressions. Each one-expression in these datasets is tagged with their respective class, (iii) study of the features contributing significantly for classification of one-expressions and then developing a CRF-based classifier for automatic classification of the one-expressions. One (the bigger one) of the annotated datasets is used to train the CRF-based classifier which is tested on the second dataset; (iv) demonstration of the utility of this work in the context of Bengali-English machine translation.

IV. Frequency Of One-expressions And Their Classifications In Bengali

One expression in Bengali is more complex compared to English like language. Since Bengali is highly agglutinative language, most of the words are highly inflected. In our experiment, we have identified twenty one commonly used such forms of *one*-expressions [এক/*ek*, একটা/*ekta* (*ek with -ta classifier*), একটি/*ekti* (*ek with -ti classifier*), একটাই/*ektai* (*ek with -tai inflection*), একটার/*ektar* (*ek with -tar inflection*), …].

The use of one-expression is quite frequently in Bengali. We investigated the frequency of one-expressions in the 177-million-word FIRE Bengali Corpus [6] and found that about 0.76% words (about 1.34 million words) in the corpus are one-expressions. This counts all morphological variations of one-expressions. Frequency of *one*-expression clearly shows the dominant presence of *one*-expressions in Bengali. As a reference one may note that the words "না"/na (no) and "করে"/kare (do) are the two most frequent words in FIRE corpus and their occurrence frequencies are 0.66% and 0.60%, respectively.

Classification of one-expressions is based on the instances found in the FIRE corpus. We follow the classification scheme of Hwee Tou Ng et al. [1] with one exception. Instead of six classes we found seven dominant classes among the Bengali one-expressions. The *Equality* class (explained next) which is not relevant for English it found to be quite in use for Bengali. The seven classes are explained as follows:

A. Idiomatic one (IDO)

In Bengali it acts like a particle and generally associated with definite or indefinite singularity of any entity. Functionally it is very similar to the indefinite/definite article "*a/an, the*" in English.

Example: একদা অযোোদ্ধায় [এক] রাজা ছিল (*ekoda ajyodhay* [*ek*] *rAjA chhilo*)/once upon a time there was [a] king at Ajyodha.

B. Numeric one (NUM)

It indicates the numeric (cardinal) value *one*.

Example: আমার কাছে মাত্র [এক] টাকা আছে (*AmAr kAchhe mAtro* [*ek*] *tAkA Achhe*)/I have only [one] rupee.

C. Partitive one (PAT)

Selects an individual from a group of object.

Example: কোনও [এক] ঝুপড়িতে রান্নার সময়েই ওই আগুন লাগে বলে সন্দেহ করা হচ্ছে ([*kono ek*] *jhuprite rAnnAr samayei oi Agun lAge bale sandeha karA hachhe*)/It is suspected that, the fire broke out in any [one] of the huts from cooking oven.

D. Anaphoric one (ANA)

The *one* having a referent.

Example: ওর দুটো ওয়াকম্যান আছে , [একটা] আমি নিয়ে নেবো (*or duto walkman Achhe* , [*ektA*] *aami niye nebo*)/He has two walkmans, I will take [*one*].

E. Equality one (EQU)

Used of this *one* is for equality for two or more entities.

Example: সন্ত্রাসবাদীদের সঙ্গে গোটা ইসলামি দুনিয়াকে [এক] করে দেখা ঠিক নয় (*santrashbAdider sange gotA islami duniyAke* [*ek*] *kore dekhA thik noi*) / It is injustice to treat the terrorism and the entire Islamic community [equally]. Note: One interesting property in Bengali the frequent use of the word একই/*ek-i*/*same* whose root form is এক/*ek*/*one*. But the expression একই (*ek-i*) does not come under *one*-expression.

F. Generic one (GEN)

A pronominal use that refers to a generic entity.

Example: প্রাথমিক ভাবে পুলিশের অনুমান, সভায় হাজির কেউ [এক] জন বোমাটি সঙ্গে নিয়ে এসেছিল (*prAthamik bhAbe pulicer onumAn, sabhAy hAjir keo* [*ek*] *jan bomAti sange niye esechhilo*)/Primarily the police suspects that some [one] attending the meeting carried the bomb.

G. Other one (OTH)

The *one*-expression other than above six classes.

Example: [এক] কথায়, রাজনীতির ঘূর্ণাবর্তে পড়িয়া বাঙ্গলা আজ নানা দিকেই পর্যুদস্ত ([*ek kothai*], *rAjnitir ghurnAbarte pariA bAnglA Aj nAnA dikei paryudasta*)/[*In brief*], Bengal, in many aspects, is now in a disastrous condition due to its political practices.

V. Preparation Of Annotated Data

From the FIRE corpus, we randomly selected 1806 sentences containing 2006 *one*-expression and manually annotated these with one of the seven classes described above. The distribution of each class in the annotated corpus is shown in the TABLE I. We call this annotated dataset $\Im_r$ as this has been used to train a CRF-based classifier as explained in the next section.

It is noted that this distribution *one*-expression is differs from other languages. For example, the experiment was conducted in [1] found Numeric class (46.9%) as the most frequent one followed by Partitive (25.3%). Idiomatic (1.6%) was seen to be very less frequent in their dataset of 1,577 one-expressions randomly selected from the BNC corpus.

TABLE I. Distribution Of One-expression In The Annotated Data Set

Class	Frequency	Percentage %
Idiomatic	544	27.12
Partitive	415	20.69
Numeric	362	18.05
Generic	266	13.26
Equality	114	5.68
Anaphoric	98	4.88
Other	207	10.32
Total	2006	100

VI. Automatic Classifications Of Bengali One-expression

We configured a CRF-based classifier for automatic classification of Bengali one-expressions. In our experiment, we have configured the Java-based an open-source package MAchine Learning for LanguagE Toolkit (MALLET)[3]. A set of seven features that contribute significantly for classifying the one-expressions is identified with the help of linguists. Description of these seven features is given below:

- POS tag of *one* (W_0): We have considered the POS of the *one* as a feature. In our experiment we have found the POS of *one*-expression is either QC (cardinal) or NN (common noun). [For POS tagging we have used a previously developed Bengali POS tagger which is basically obtained by retraining the Stanford tagger on about 10K tagged Bengali sentences. The Bengali tagger is found to be 92% accurate].
- Inflections (classifier) of *one*: The inflections (classifiers) of *one* are considered as feature. We have twenty-one such inflections (and classifiers) {*-ta, -ti, -tai, -tir, -tite, …*}.
- Previous word (W_{-1}) of *one*: The immediate previous word of *one*.
- Next word (W_{+1}) of *one*: The immediate next word of *one*.
- Sentence starts with *one*: Whether the *one* is the starting word of the sentence.
- Sentence ends with *one*: Whether the *one* is the ending word of the sentence.
- Measuring unit followed by *one*: Whether the next word of one is measuring unit (*hAzAr*/thousand, *keji*/kilogram, …).

VII. Training Data

The annotated dataset $\Im_r$ is used for training the CRF. The sentences in $\Im_r$ are POS tagged and the one-expressions are tagged with their respective class labels. The annotated dataset is presented in a column format as shown in TABLE II and TABLE III shows the detailed description of data format. The CRF-based classifier uses maximum likelihood for training, for feature expectations uses the forward backward algorithm and uses a Gaussian prior on parameter optimization.

TABLE II. The Training Data Format

```
………………………………………………………………
txt1.txt   0    ১১                 QC      o
txt1.txt   1    সেপ্টেম্বরের          NN      o
txt1.txt   2    এক                 QC      NUM
txt1.txt   3    সপ্তাহ                NN     o
txt1.txt   4    আগেই     NST      o
txt1.txt   5    মার্কিন               NN      o
txt1.txt   6    প্রশাসন   NN       o
………………………………………………………………
```

TABLE III. Description Of Training Data

Column	Type	Description
1	Document Id	Contains the filename
2	Word number	Word index in the sentence
3	Word	Word itself
4	POS	POS of the word
5	Classification	Classification tag

VIII. Evaluation

The classification system has been evaluated by the publicly available ICON 2011 data set [7] which was prepared primarily for Bengali anaphora resolution. This dataset consists of nine text pieces and we have extended this dataset by adding four more texts. This combined dataset ($\Im_e$) has been previously used for the evaluation of Bengali anaphora resolution systems [8, 9]. Choice of this dataset is somewhat intentional as this dataset has been annotated for anaphora resolution. As one-anaphor is one of the one-expression classes, annotation with one-expression information would help subsequent research on resolution of one-anaphors. The data in $\Im_e$ is presented in the same format as shown in TABLE II. TABLE IV shows the coverage of $\Im_e$ in terms of number of text pieces, words and one-expressions.

TABLE IV. Coverage Of Test Data Set, $\Im_E$

Data	Test data
#texts	13
#words	27454
#one-expressions	300

TABLE V. Evaluation Of One Expression Classification

Class label	#Instances	# Correct classification	# Incorrect classification	Recall	Precision	F1-score
IDO	103	85	25	.83	.77	.80
NUM	84	77	37	.92	.68	.78
OTH	43	29	3	.67	.91	.77
PAT	32	12	0	.38	1.0	.55
GEN	17	7	1	.41	.88	.56
ANA	16	13	8	.81	.62	.70
EQU	5	3	0	.60	1.0	.75
Total	300	226	74	.75	.75	.75

TABLE V gives the results for the *one*-expression classification for each of the seven classes. The average accuracy of *one*-expression classification is about 75% where as the accuracy of Idiomatic (80%), Numeric (78%) and Partitive (77%) are relatively better. As far as recall and precisions are concerned, the NUM class shows the highest recall and the PAT class shows the highest precision. The most dominant class, i.e. IDO shows the highest F1-score.

IX. Error Analysis

Most of the errors occur due to inter-class confusion. TABLE VI shows the confusion matrix that shows that IDO and NUM are two classes which create major confusions. For all other classes, a dominant tendency is to be confused with either IDO or NUM classes. This is because, some features (classifier/inflections, e.g., *-ta*/*-ti*; POS tags QC/cardinal, etc.) are strongly favourable for Idiomatic and Numeric classes. Many instances of PAT

[3] http://mallet.cs.umass.edu/sequences.php

class are also confused but such confusions are spread over three different classes, confusion with ANA being the most significant. This is because the features for Partitive (PAT) class are much closed to the features of Anaphoric (ANA) class. In fact, some instances of Partitive class are special kind of anaphoric *one*-expressions.

TABLE VI. CONFUSION MATRIX FOR CLASSIFICATION OF ONE-EXPRESSION

	IDO	NUM	PAT	ANA	EQU	GEN	OTH
IDO	×	17	0	0	0	0	1
NUM	6	×	0	0	0	0	1
PAT	5	6	×	8	0	1	0
ANA	1	1	0	×	0	0	1
EQU	2	0	0	0	×	0	0
GEN	6	4	0	0	0	×	0
OTH	5	9	0	0	0	0	×

X. EFFECT ON MACHINE TRANSLATION

In the beginning of our discussion we show that as the one-expressions behave differently in different context, production of their right translation (for example, in English the Bengali one-expressions can be translated to a / an / the / one / only one/ someone/ once / equally / similarly /etc.) is a challenging task. We hypothesize that classification of one-expressions would help in producing the right translation. We tested our hypothesis in the following way.

The dataset $\Im_e$ contains 300 instances of one-expressions. These instances occur 296 sentences. These 296 sentences are translated to English using Google Bengali-English translator[4] and the translations of 300 one-expressions are marked. The proper English translations of these 300 one-expressions (in context of their containing sentences) are done manually. It is found that the Google translator could produce correct translations for 117 (39%) one-expressions. Note that we are concerned about the translation of the one-expressions only.

Next, we associate the most dominant English translation to each of the six classes of one-expressions (for the OTH class, we could not associate any translation). For example, the most dominant translation of the IDO class is *a/an*, of the NUM class is *one*, etc. Note that the one-expressions under a particular class can have several English translations but for the sake of simplicity we consider just one of them, the most dominant translation.

Once a Bengali one-expression is classified, its English translation is the word associated with the respective class. By doing so, we could produce translation of 257 one-expressions (for 43 OTH class instances, we could not produce any translation). It is found that out of these 257 translations, 179 (60%) are correct. The accuracies originate from two sources: (i) classification errors (note that we can classify with about 75% accuracy) and (ii) even if the classification is correct, the most dominant translation (a static one which does not consider any context) does not always give the correct translation. Even this simple framework improves the translation accuracies from 39% (by Google translator that considers context) to 60% (by simply classifying the one-expressions and replacing them by their class-specific dominant (static) translation). This improvement is found to be statistically significant (p-value < 0.01 in a 2-tailed paired t-test).

XI. CONCLUSIONS

This work strongly supports to have an additional effort for processing one-expressions for Bengali NLP. Future works refer to further investigation of the features used for classification and class-wise better translation strategy. We have generated annotated datasets of 2246 sentences (combining $\Im$r and $\Im$e) containing 2306 instances of one-expressions. We make this datasets free for facilitating further research on this area and at the same time the datasets need to be enlarged in order to design more robust machine learning based classification scheme. The present CRF-based classifier gives about 75% accuracy and one of the major reasons behind its inaccuracies is the small size of the training data. We also plan to extend the present research for resolution of one-anaphors. For this work, we plan to classify the one-expressions as one-anaphor or not (two-class problem) and then resolve the one-anaphors by finding their correct antecedents.

REFERENCES

[1] H. Tou Ng, Yu Zhou, Robert Dale and Mary Gardiner (2005). Machine Learning Approach to Identification and Resolution of One-Anaphora.

[2] M. A. K. Halliday and R. Hasan (1976). Cohesion in English. Longman

[3] B. Webber (1979). A Formal Approach to Discourse Anaphora. Garland Publishing Inc.

[4] D. A. Dahl (1985). The structure and function of one anaphora in English. PhD thesis, Univ. of Minnesota.

[5] S. Luperfoy (1991). Discourse Pegs: A Computational Analysis of Context-Dependent Referring Expressions. PhD thesis, Univ. of Texas at Austin

[6] Forum for Information Retrieval Evaluation; http://www.isical.ac.in/~fire/data.html

[7] ICON NLP Tools Contest (2011). Anaphora Resolution in Indian Languages," In 9th Int. Conf. on Natural Language Processing (ICON), Chennai, India.

[8] A. Senapati and U.Garain (2012). Anaphora Resolution in Bangla using global discourse knowledge. In Int. Conf. of Asian Language Processing (IALP), 49-52, Hanoi, Vietnam.

[9] A. Senapati and U. Garain (2013). GuiTAR-based Pronominal Anaphora Resolution in Bengali, in ACL, 126-130, Sofia, Bulgari

[4] Google translator: http://translate.google.co.in/#en/bn/