Bangla Word Clustering Based on N-gram Language Model

Sabir Ismail, M. Shahidur Rahman
Department of Computer Science and Engineering
Shahjalal University of Science and Technology
sabir.ismail01@gmail.com, msrahman.bd@gmail.com

Abstract— **In this paper, we describe a method for producing Bangla word clusters based on semantic and contextual similarity. Word clustering is important for parts of speech (POS) tagging, word sense disambiguation, text classification, recommender system, spell checker, grammar checker, knowledge discover and for many others Natural Language Processing (NLP) applications. Computerization of Bangla language processing has been started a long ago, but still it is in neophyte stage and suffers from resource scarcity. We propose anunsupervised machine learning technique to develop Bangla word clusters based on their semantic and contextual similarity using N-gram language model. According to N-gram model, a word can be predictedbased on its previous and next words sequence. N-gram model is applied successfully for word clustering in English and some other languages. As word clustering in Bangla is a new dimension in Bangla language processing research, so we think this process is good way to start and our assumption is true as our result is quite decent. We produced 456 clusters using a locally available large Bangla corpus. Subjective score derived from the clusters reveal strong similarity of the words in the same cluster.**

Keywords—word cluster; information retrival; natural language processing; machine learning; n-gram model

I. INTRODUCTION

Bangla language is the 7th most spoken language [1] in the world. However, due to resource scarcity, the effort of Bangla computerization could not reachup to satisfactory level compared to other widely spoken languages. To our best knowledge, word clustering for Bangla is a new research dimension which is not explored yet.Bangla language is inherently so vast that it is not possible to develop manual or rule based word clusters. Recently, we compiled a large Bangla corpus containing 2,51,89,733 individual words that facilitates effective employment of unsupervised machine learning[2] based clustering. In this paper, we proposed a method to develop Bangla word clusters based on similarity in semantics and contexts.

Word cluster has many fold applications in language processing.For example, POS tag of an unknown word can be determined from the word cluster [3]. Every word in the same cluster generally belongs tothe same POS tag. In Bangla language a word can have multiple forms based on tense, person, number and gender. In some cases, a word can have more than 200 forms. To overcome the problem of word-sense disambiguation [4] word clustering can also suggest the most appropriate form of a word. Another application of word clustering is spell checker. Word cluster can generate suggestions for an incorrectly typed word. In recommender system, for example if several books of the same author or several movies of the same actor are clustered in the same group, more effective suggestion can be generated. Sentence structure with grammatical mistakes is also solvable using clustered words. In Bangla search engine, word cluster can make efficient suggestion. When performing machine translations, word cluster can suggest more appropriateform of words. Thus word cluster plays a vital rule in many natural language processing fields.

As mentioned earlier, word clustering for Bangla is not explored yet. However, various techniques are reported for word clustering for many other languages. There exists previous work in which the unigram and the bigram models are used for cord clustering. Finch and Chater [5] used bigram model to determinethe weight matrix of a neural network. A pioneer work on word clustering is proposed by Brown, Desouza, Mercer, Pietra, Lai [6] where they used n-gram language model. They reported that, class based n-gram model has higher perplexity than the word based n-gram model. Another attempt using n-gram model is reported by Korkmaz [7]. They use a similarity function and a greedy algorithm to put the words into clusters.Delete interpolation method is used by Mori, Nishimura and Itoh [8] for Japanese and English language and they achieve betterperformance than the Brown's method [6]. Ding, Al-Mubaid and Kotagiri [9] used Naive Bayes method effectively for English in classifying words using surrounding context words as features. McMahon and Smith [10] used annealing based algorithm and compared results with some other classification systems. Bellegarda [11] presented latent semantic analysis based clustering. His experiments indicate that, the clusters are intuitively satisfactory. Further, many other approaches have been reported in literature for other languages like Russian, Arabic, Chinese and Japanese.

In this paper, we propose an unsupervised machine learning technique for clustering Bangla words based on their semantic and contextual similarity. For example, consider the following two sentences:

1. আকাশে বিভিন্ন রঙের আলোর খেলা
2. আকাশে নানা রঙের আলোর খেলা

Here, বিভিন্ন and নানা are used to express the same meaning. In this case, the words preceding and following them have

similarity in context. This similarity is utilized to produce word clusters using tri-gram language model where two lists are generated for every word. One list contains two preceding words and another list contains two followingwords of the tri-gram model. Then for a particular pair of words, similarity is measured based on a score which corresponds the total number of matchesin the preceding and following word-list. If the similarity score is greater than a predefined threshold value, both words are clustered in the same group.

The remainder of this paper is organized as follows. SectionII is concerned with problem definition. In section III, we describe our methodology for producing word clusters. Section IVdescribesthe experimental results.Finally, Section V summarizesthe paper.

II. PROBLEM DEFINITION

Clustering[11-12] is the task of grouping a set of objects in such a way that objects in the same group are more similarto each other than to those in other groups. Clustering is an unsupervised machine learning technique where no rule or predefined conditions are required.

Consider the following fourpartial sentences:

1. তারা সবাই সভায় উপস্থিত থাকবে
2. তারাসবাই অনুষ্ঠানে উপস্থিত থাকবে
3. ব্যবসায় লাভ হচ্ছে না
4. ব্যবসায় ক্ষতি হচ্ছে না

In sentence-1 and sentence-2 সভায় and অনুষ্ঠানে are similar in semantic meaning. Again, in sentence-3 and sentence-4 লাভ and ক্ষতি are similar in context. The proposed method focuses on these two kinds of similarity. Principles of n-gram model [13] is employed here. An n-gram model is a subsequence of n consecutive items in any given sequence. It is used to predicate n-th item in a sequence from previous or next (n-1) items by means of probability distribution. An n-gram of size three is referred as tri-gram model. We use tri-gram model with some modification for our purpose. We make list of every three consecutive words. Then, from this list we make two separate lists for every word, one containing two previous words and another containing two following words. For example, Table 1and Table 2 shows a typical phrase list for অনুরোধ and for আবেদন respectively.

TABLE I. A TYPICAL PHRASE LIST FOR "অনুরোধ"

Words before**অনুরোধ**	Words after **অনুরোধ**
করার জন্য **অনুরোধ**	**অনুরোধ** করব বলার
করে দিতে **অনুরোধ**	**অনুরোধ** করা হয়
কর্তৃপক্ষের কাছে **অনুরোধ**	**অনুরোধ** করা হয়েছে
দেওয়ার জন্য **অনুরোধ**	**অনুরোধ** করেন তিনি
দেয়ার জন্য **অনুরোধ**	**অনুরোধ** জানানো হলে
না করার **অনুরোধ**	**অনুরোধ** জানানো হয়
বলার জন্য**অনুরোধ**	
বার বার **অনুরোধ**	
সরকারের কাছে **অনুরোধ**	
প্রধানমন্ত্রীর কাছে **অনুরোধ**	
ভিসার জন্য **অনুরোধ**	

TABLE II. A TYPICAL PHRASE LIST FOR "আবেদন"

Words before **আবেদন**	Words after **আবেদন**
কর্তৃপক্ষের কাছে **আবেদন**	**আবেদন** করেন তিনি
কাছে আমাদের **আবেদন**	**আবেদন** জানানো হয়
কাছে আমার **আবেদন**	**আবেদন** করতে পারবেন
কাছে সাহায্যের **আবেদন**	**আবেদন** করতে হবে
প্রধানমন্ত্রীর কাছে **আবেদন**	**আবেদন** করতে হয়
ভিসার জন্য **আবেদন**	**আবেদন** করা হয়
সরকারের কাছে **আবেদন**	**আবেদন** করা হয়েছে
বলার জন্য **আবেদন**	**আবেদন** খারিজ করে
	আবেদন জানানো হয়েছে

From Table 1 and Table 2, based on the number of similar words preceding অনুরোধ and আবেদন and following অনুরোধ and আবেদন Table III is obtained.

TABLE III. SIMIAR WORD LIST BETWEEN "অনুরোধ" & "আবেদন"

Similar previous two words	**Similar next two words**
সরকারের কাছে **অনুরোধ**	**অনুরোধ** করেন তিনি
সরকারের কাছে **আবেদন**	**আবেদন** করেন তিনি
ভিসার জন্য **অনুরোধ**	**অনুরোধ** করা হয়েছে
ভিসার জন্য **আবেদন**	**আবেদন** করা হয়েছে
প্রধানমন্ত্রীর কাছে **অনুরোধ**	**অনুরোধ** করা হয়
প্রধানমন্ত্রীর কাছে **আবেদন**	**আবেদন** করা হয়
কর্তৃপক্ষের কাছে **অনুরোধ**	**অনুরোধ** জানানো হয়
কর্তৃপক্ষের কাছে **আবেদন**	**আবেদন** জানানো হয়
বলার জন্য **অনুরোধ**	
বলার জন্য **আবেদন**	

Finally,similarity between these two words অনুরোধ and আবেদন is measured using a probability-based method described in the next section.

III. PROPOSED METHOD

First, we use a large corpus of 2,51,89,733 individual words containing 5,21,391 unique words w_i. Next, a list of three-consecutive words w_i,w_{i+1},w_{i+2} is prepared. Then, from this list we make two separate lists for every word w_i, one contains previous two words w_{i-2},w_{i-1} and the other contains next two words w_{i+2},w_{i+1}.Then for every pair of words w_i,w_j we calculate number of matched preceding words from list $list(w_{i-2},w_{i-1})$and $list(w_{j-2},w_{j-1})$, number of following matched words form $list(w_{i+2},w_{i+1})$ and $list(w_{j+2},w_{j+1})$.

We use tri-gram model equations for calculating similarity. In tri-gram model, we can predict 3rd term of a sequence from previous or next two terms.

$$P(w_n|w_{n-1},w_{n-2}) = count(w_{n-2}w_{n-1}w_n)/count(w_{n-2}w_{n-1})$$

Or

$$P(w_n|w_{n+1},w_{n+2}) = count(w_nw_{n+1}w_{n+2})/count(w_{n+1}w_{n+2})$$

Similarity between a pair of words to be included in the same cluster is measured with respect to both the preceding and following two words. For every case, again the similarity of the first word with the second word and the second word with the first word in the pair is measured to make cluster more precise.

Similarity between a pair of words to be included in the same cluster based on preceding two words is determined as follows.

- Similarity of the first word with the second word in the pair:

$$P(w_i,w_j) = \frac{Count(match(list(w_{i-2},w_{i-1}),list(w_{j-2},w_{j-1})))}{Count(list(w_{i-2},w_{i-1}))} \quad (1)$$

- Similarity of the second word with the first word in the pair:

$$P(w_j,w_i) = \frac{Count(match(list(w_{j-2},w_{j-1}),list(w_{i-2},w_{i-1})))}{Count(list(w_{j-2},w_{j-1}))} \quad (2)$$

Again, similarity between a pair of words to be included in the same cluster based on following two words is determined as follows.

- Similarity of the first word with the second word in the pair:

$$P(w_i,w_j) = \frac{Count(match(list(w_{i+1},w_{i+2}),list(w_{j+1},w_{j+2})))}{Count(list(w_{i+1},w_{i+2}))} \quad (3)$$

- Similarity of the second word with the first word in the pair:

$$P(w_j,w_i) = \frac{Count(match(list(w_{j+1},w_{j+2}),list(w_{i+1},w_{i+2})))}{Count(list(w_{j+1},w_{j+2}))} \quad (4)$$

If all four equations yield values greater than a predefined threshold value, we group them into the same cluster.

For example, we have the following phrases in three consecutive word lists.

অনুরোধ করেন তিনি
অনুরোধ জানানো হয়
দেওয়ার জন্য **অনুরোধ**
ভিসার জন্য **অনুরোধ**
সরকারের কাছে **অনুরোধ**
আবেদন করেন তিনি
আবেদন জানানো হয়
সরকারের কাছে **আবেদন**
প্রধানমন্ত্রীর কাছে **আবেদন**
ভিসার জন্য **আবেদন**

For Word অনুরোধ:

- Preceding two words list:

list(w_{i-2},w_{i-1})={দেওয়ার জন্য,ভিসার জন্য, সরকারের কাছে}
Coun(list(w_{i-2},w_{i-1})) = 3

- Following two words list:

list(w_{i+1},w_{i+2})={করেন তিনি,জানানো হয়}.
Count(list(w_{i+1},w_{i+2})) = 2

For Word আবেদন:

- Preceding two words list:

list(w_{j-2},w_{j-1})={সরকারের কাছে,প্রধানমন্ত্রীর কাছে, ভিসার জন্য}
Count(list(w_{j-2},w_{j-1}))= 3

- Following two words list:

list(w_{j+1},w_{j+2})={করেন তিনি,জানানো হয়}.
Count(list(w_{j+1},w_{j+2})) = 2

Number of matched words for Word অনুরোধ with আবেদন based on preceding two words:

Count(match(list(w_{i-2},w_{i-1})), list(w_{j-2},w_{j-1})))=2

Number of matched words for Word আবেদন with অনুরোধ based on preceding two words:

Count(match(list(w_{j-2},w_{j-1})), list(w_{i-2},w_{i-1}))) =2

Similarity between words আবেদন and অনুরোধ based on preceding two words (Eqs. (i) and (ii)):

$P(w_i,w_j) = 2/3 = 0.667$
$P(w_j,w_i) = 2/3 = 0.667$

Number of matched words for Word অনুরোধ with আবেদন based on following two words:

Count(match(list(w_{i+1},w_{i+2})), list(w_{j+1},w_{j+2}))) =2

Number of matched words for Word আবেদন with অনুরোধ based on following two words:

Count(match(list(w_{j+1},w_{j+2})), list(w_{i+1},w_{i+2}))) =2

Similarity between words অনুরোধ and আবেদন based on following two words (Eqs. (iii) and (iv)):

$P(w_i,w_j) = 2/2 = 1$
$P(w_j,w_i) = 2/2 = 1$

Similarity between words অনুরোধ with আবেদন when considering previous two words is 0.667 and 1.000 when considering following two words. Again, similarity of আবেদন with অনুরোধ when considering previous two words is 0.667 and 1.000 when considering next two words. We try with several threshold values and best result we achieve with 0.200. When, all the probability scores are greater than this threshold value, both words are clustered in the same group.

IV. RESULT ANALYSIS

To evaluate the performance of the method, we started with a three- consecutiveword list containing 64690 number of entries. Then we derive 456 word clusters in total with maximum 12 words in a cluster. Table IV shows results for 81 clusters randomly taken from the 456 clusters. Though the total number of clusters is small, strong similarity is observed in the words in the same cluster.

TABLE IV. WORD CLUSTERS

কোণ বাহু ----------	উদ্যোগ নির্মাণ ----------

লাভবান	যথাযথ
ক্ষতিগ্রস্ত	প্রয়োজনীয়
---	---
অস্বীকার	ঝিনাইদহর
স্বীকার	নরসিংদীর
উল্লেখ	নেত্রকোনার
---	---
তৈরি	সহযোগিতা
সৃষ্টি	সহায়তা
---	---
রক্ষা	সংক্রান্ত
বর্ণনা	সম্পর্কিত
---	---
উদ্ধার	পারছেন
গণ্য	পারলাম
বিবেচিত	পারল
---	পারছে
তারা	পারছি
তিনি	---
---	তোমার
যুদ্ধের	ওর
সংগ্রামের	আপনার
---	আমরা
প্রতিষ্ঠানের	---
সংস্থার	বুধবার
---	শুক্রবার
আগ্রহী	মঙ্গলবার
সক্ষম	শনিবার
সচেষ্ট	রোববার
---	সোমবার
গণ্য	---
বিবেচিত	পারবেন
---	চান
সাথে	চায়
সঙ্গে	চাই
---	পারি
সভায়	পারেন
অনুষ্ঠানে	পারছে
---	পারছি
বিশ্বাস	পারবে
অস্বীকার	---
আশা	দেখার
---	বোঝার
বিশ্বের	---
পৃথিবীর	বিপর্যস্ত
---	দুর্বিষহ
আর	---
কিন্তু	বিকল্প
---	উপায়
কঠোর	---
শান্তিমূলক	ইচ্ছা
---	ইচ্ছে
নিষিদ্ধ	পছন্দ
বহিষ্কার	---
সংশ্লিষ্টরা	সন্দেহ
বিশেষজ্ঞরা	আগ্রহ
---	আপত্তি
নির্দ্বিধায়	সংশয়
নিঃসন্দেহে	---
---	তেলের
গুরুত্বপূর্ণ	জ্বালানির
উল্লেখযোগ্য	---
---	রংপুর
চাঁদা	রাজশাহী
ঘুষ	---

---	জবাবে
ইন্টারনেট	উত্তরে
কম্পিউটার	---
---	বিশেষজ্ঞরা
নির্মাণে	সংশ্লিষ্টরা
প্রকল্পে	বিশ্লেষকেরা
---	---
সিদ্ধান্ত	মুগ্ধ
উদ্যোগ	অবাক
---	---
করবে	উচিত
করছে	দরকার
করবেন	---
করব	অথবা
---	কিংবা
সুষ্ঠু	---
নিরপেক্ষ	অবদান
---	ভূমিকা
লিখতে	---
ভাবতে	যুবলীগ
বাঁচতে	ছাত্রলীগ
---	---
চুরি	পাস
লুট	প্রণয়ন
পাচার	---
---	সংগীত
আইসিটি	সঙ্গীত
তথ্যপ্রযুক্তি	---
---	জনগণের
এরপর	সবার
তখনই	---
এছাড়া	নাগালের
তারপর	ক্রয়ক্ষমতার
---	---
জবাব	দেরি
উত্তর	দূর
---	---
ঘটনার	দেশে
কর্মকাণ্ডের	বাংলাদেশে
কাজের	---
---	তৃতীয়
কর্মী	চতুর্থ
সহযোগী	দ্বিতীয়
---	---
প্রণয়ন	বেগম
পাস	খালেদা
---	---
বিষয়ে	যুগ্ম
সম্পর্ক	ভারপ্রাপ্ত
---	সাংগঠনিক
প্রমাণ	---
ব্যাখ্যা	এছাড়া
---	এরপর
নেতারা	---
কর্মীরা	সম্পর্কে
সদস্যরা	বিষয়ে
---	---
মন্তব্য	কিন্তু
দাবি	আর
উল্লেখ	এবং
---	তবে
বাড়ার	---
বৃদ্ধির	যদিও
---	উল্লেখ্য

পরিষ্কার স্পষ্ট ----------	এমনকি বর্তমানে অবশ্য অথচ ----------

Further, a subjective evaluation of the method is attempted.A group of five users were asked to score the 456-word clustersin a scale of 5 where 5 means word clusters have very strong similarity and 1 means very poor similarity. Table:Vshow the user ratings for 456 clusters. In Table: V, User 1 gives rating 5 to 433 clusters, rating 4 to 19 clusters and so on. Table: VI shows average rating and percentage of similarity based on user ratings.

TABLE V. USER EVALUATION

Rating	1	2	3	4	5
User 1	0	1	3	19	433
User 2	0	2	4	38	412
User 3	0	0	8	20	428
User 4	0	4	4	26	422
User 5	0	2	7	16	431

Table: VI shows that in an average 425.2 cluster scores rating 5, and only 1.8 cluster scores rating 2. In percentage, 93.245 percent clustershave very strong similarity, and only 0.384 percent clustershave poor similarity. No clusters is scored rating zero.

TABLE VI. USER EVALUATION

Rating	1	2	3	4	5
Average	0	1.8	5.2	23.8	425.2
Percentage	0	0.394	1.140	5.219	93.245

V. CONCLUSION

Clustered word corpus for Bangla language is not available. In this paper, we proposed an unsupervised machine learning method to develop such a preliminary corpus in Bangla. Considering the prospect of word clustering, this paper may prove to be a starting point in an endeavor to conduct a large-scale analysis in various applications of language processing. The results presented above are produced using a tri-gram model, use of higher order n-gram model may results in better performance.

REFERENCES

[1] Top 10 most spoken languages in the world, "http://listverse.com /2008/06/26/top-10-most-spoken-languages-in-the-world/". Last accessed on 20th January, 2014.

[2] Unsupervised machuine Llearning,"http://www.aihorizon.com/essays /generalai/supervised_unsupervised_machine_learning.htm".Last accessed on 20th January, 2014.

[3] Y Goldberg."Task-specific word-clustering for Part-of-Speech tagging".arXiv preprint arXiv:1205.4298, 2012.

[4] H A Sánchez, A P Porrata and R B Llavori. "Word sense disambiguation based on word sense clustering". Advances in Artificial Intelligence, Springer Berlin Heidelberg, 2006. P: 472-481.

[5] S Finch and N Chater. "Automatic methods for finding linguistic categories". In Igor Alexander and John Taylor, editors, Artificial Neural Networks, Volume 2. Elsevier Science Publishers, 1992.

[6] P F Brown, P V Desouza, R L Mercer, V J D Pietra, V J Della. and J C Lai. "Class-based N-gram Models of Natural Language". Computational linguistics, 18 No: 4, 1992, P: 467-479.

[7] E E Korkmaz. "A method for improving automatic word categorization". Doctoral dissertation, Middle East Technical University, 1997.

[8] S Mori, M Nishimura and N Itoh. "Word clustering for a word bi-gram Model". International Conference on Spoken Language Processing, 1998.

[9] W Ding, H Al-Mubaid and S Kotagiri."Word classification: An experimental approach with Naïve Bayes". Conference on Computers and Their Applications, 2009.

[10] J McMahon and F J Smith. "Structural tags, annealing and automatic word Clsssfication". In Artificial Intelligence and the Simulation of Behaviour Quarterly, 1994.

[11] J R Bellegarda. "Latent semantic mapping information retrieval". Signal Processing Magazine, IEEE 22.5, 2005, P: 70-80.

[12] Clustering – Introduction, "http://home.deib.polimi.it/ matteucc/ Clustering/tutorial_html". Last accessed on 20th January, 2014.

[13] Clustering – Introduction, "http://www.stanford.edu/class/cs345a/ slides/12-clustering.pdf ". Stanford University-Clustering, Last accessed on 20th January, 2014.

[14] D Jurafsky and J H Martin. "An introduction to natural language processing, computational lingustics and speech processing". 2nd Edition, Chapter: 4.