Proceedings of the International Conference on Engineering Research, Innovation and Education 2017 ICERIE 2017, 13 15 January, SUST, Sylhet, Bangladesh

Bangla Word Clustering Based on Tri-gram, 4-gram and 5-gram Language Model

Dipaloke Saha^{1,*}, Md Saddam Hossain¹, MD. Saiful Islam¹ and Sabir Ismail¹
Department Of Computer Science & Engineering,
Shahjalal University Of Science And Technology, Sylhet, Bangladesh.
dipsustcse12@gmail.com, mshossaincse@gmail.com, saiful-cse@sust.edu,
sabir.ismail01@gmail.com

Keywords:

• Word Cluster,

Natural

Language

Processing,

Machine

Learning,

N-gram Model,

Term

Frequency (tf).

• SUST, ICERIE.

Abstract: — In this paper, we describe a research method that generates Bangla word clusters on the basis of relating to meaning in language and contextual similarity. The importance of word clustering is in 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. In the history of word clustering, English and some other languages have already implemented some methods on word clustering efficiently. But due to lack of the resources, word clustering in Bangla has not been still implemented efficiently. Presently, it's implementation is in the beginning stage. In some research of word clustering in English based on preceding and next five words of a key word they found an efficient result. Now, we are trying to implement the tri-gram, 4-gram and 5-gram model of word clustering for Bangla to observe which one is the best among them. We have started our research with quite a large corpus of approximate 1 lakh Bangla words. We are using a machine learning technique in this research. We will generate word clusters and analyze the clusters by testing some different threshold values.

1. INTRODUCTION

Though Bangla is a widely spoken language, it has lack of resources in its research field. Recently, a new research dimension in Bangla is added called word clustering. In this paper, the research of word clustering for Bangla language is trying to be extended. For this, a large Bangla corpus containing 97,971 individual words is compiled to generate the word clusters. In this paper, an unsupervised machine learning technique and a method are proposed to cluster Bangla words on the basis of similarity in semantics and contexts.

In language processing word cluster has a wide range of applications. POS tag is one of them. Same clustered words usually contain the same POS tag. Word clustering can produce suggestions for an inaccurately typed word which is very much helpful for spell checker. Word sense disambiguation, sentence structure with grammatical mistakes can also be solvable using clustered words. In the case of recommender system if related products of the same category are clustered in the same group, more feasible suggestion can be produced. This type of work is also useful for Bangla search engine to find the appropriate content. So, there is a huge importance of word clustering in the field of natural language processing.

dipsustcse12@gmail.com

2. RELATED WORK

In Bangla the implementation of word clustering is in the neophyte stage. A previous work on Bangla word clustering exists in which an unsupervised machine learning technique is used to implement the bigram model by Sabir Ismail and M. Shahidur Rahman. In many other languages different types of techniques are used for word clustering. Finch and Chater (1992) implemented bigram model for the calculation of weight matrix of a neural network. N-gram language model is used on word clustering in a research proposed by Brown, Desouza, Mercer, Peitra, Lai (1992). Another effort using n-gram model is introduced by Korkmaz (1997) in which a similarity function and greedy algorithm is used to group the words into same cluster. However, with the use of delete interpolation method by Mori, Nishimura and Itoh (1998) they got the better result than the Brown, Desouza's method. This was done for Japanese and English language. Besides these, there exists quite a good number of researches of word clustering for some other languages like Russian, Arabic, Chinese etc.

3. PROBLEM DEFINITION

Clustering is an unsupervised machine learning technique that does not require any type of rules or predefined conditions. Items which are much similar either in semantically or contextually are grouped in the same cluster and which are dissimilar are in different clusters. The introduced method in this problem is concentrating on two types of similarity such as semantics and contextual similarity.

Consider the following four sentences:

- মা বাবার দেখাশুনা করা প্রত্যেক সন্তানের দায়িত্ব।
- মা বাবার দেখভাল করা প্রত্যেক সন্তানের দায়িত।
- 3. গতরাতে ভালুকায় কয়েকজন লোক নিহত হয়।
- গতরাতে ভালুকায় কয়েকজন লোক আহত হয়।

দেখাণ্ডনা and দেখভাল are similar in semantic meaning in sentence 1 and 2 and there is similarity in নিহত and আহত in sentence 3 and 4. Here, the theory of N-gram model is implemented . Probability distribution is used here to define n-th item in a sequence form previous or next (n-1) items. Tri-gram, 4th and 5th gram model is defined as size of 3, 4 and 5 of N-gram respectively. In this research, word clusters will be generated by implementing tri, 4th and 5th gram model. After finding the word clusters the most efficient model will be found out based on those clustering words.

4. METHODOLOGY

Firstly, quite a large corpus of 97,971 individual words Wi is used in this research. Next, a list of previous three words of a specific word for tri-gram, four words for 4-gram, five words for 5-gram are prepared. Similarly, a list of next three words of a specific word for tri-gram, four words for 4-gram, five words for 5-gram are prepared. Next, similarity between a pair of words to be included in the same cluster based on preceding three words, four words and five words are determined as follows:

In tri-gram for every pair of words Wi, Wj the number of matched preceding words from list list(Wi-3, Wi-2, Wi-1) and list(Wj-3, Wj-2, Wj-1)

```
P(Wi,Wj) = (Count(match(list(Wi-3,Wi-2,Wi-1),list(Wj-3,Wj-2,Wj-1)))/((Count(list(Wi-3,Wi-2,Wi-1)) + Count(list(Wj-3,Wj-2,Wj-1)))))
```

Similarly, calculation for the 4-gram model is:

and for 5-gram model is:

```
P(Wi,Wj) = (Count(match(list(Wi-5,Wi-4,Wi-3,Wi-2,Wi-1),list(Wj-5,Wj-4,Wj-3,Wj-2,Wj-1)))/((Count(list(Wi-5,Wi-4,Wi-3,Wi-2,Wj-1))) + (Count(list(Wj-5,Wj-4,Wj-3,Wj-2,Wj-1))))))
```

Again similarly, between a pair of words to be included in the same cluster based on following three, four and five words are determined as follows, For tri-gram,

```
P(Wi,Wj) = (Count(match(list(Wi+3,Wi+2,Wi+1),list(Wj+3,Wj+2,Wj+1)))/((Count(list(Wi+3,Wi+2,Wi+1)) + Count(list(Wj+3,Wj+2,Wj+1))))/((Count(list(Wi+3,Wi+2,Wi+1)))))
```

Similarly, calculation for the 4-gram model is:

```
P(Wi, Wj) = (Count(match(list(Wi+4, Wi+3, Wi+2, Wi+1), list(Wj+4, Wj+3, Wj+2, Wj+1)))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+1))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+2, Wi+1)))))/((Count(list(Wi+4, Wi+3, Wi+2, Wi+2, Wi+3, Wi+2, Wi+3, Wi+3, Wi+2, Wi+3, Wi+3,
```

and for 5-gram model is:

```
P(Wi,Wj) = (Count(match(list(Wi+5,Wi+4,Wi+3,Wi+2,Wi+1), list(Wj+5,Wj+4,Wj+3,Wj+2,Wj+1)))/((Count(list(Wi+5,Wi+4,Wi+3,Wi+2,Wi+1)) + Count(list(Wj+5,Wj+4,Wj+3,Wj+2,Wj+1)))))
```

If the above equations of a particular model yield values greater than a predefined threshold value they are grouped into the same cluster for that model.

For example, to implement the tri-gram model some of the following phrases are:

- 1. ভোরে সূর্য উঠার আগে।
- 2. আগে খাওয়া শেষ করি।
- 3. সকালে সূর্য উঠার পরে।
- 4. পরে কাজটি শেষ করি।

For word আগে preceding three words list:

```
For word আগৈ Following three words list: list(Wi+3, Wi+2, Wi+1) = \{ খাওয়া, শেষ , করি\} Count(list(Wi+3, Wi+2, Wi+1)) = 3
```

```
For word পরে preceding three words list:
    list(Wj-3, Wj-2, Wj-1) = { সকালে , সূর্য , উঠার }
    Count (list(Wj-3, Wj-2, Wj-1)) = 3
```

For word পরি following three words list:

$$list(Wj+3,\ Wj+2,\ Wj+1)=\left\{$$
 কাজটি , শেষ , করি $\right\}$ $Count(list(Wj+3,\ Wj+2,\ Wj+1))=3$

Number of matched words for word আগৈ with পরে based on preceding three words :

```
Count(match(list(Wi-3, Wi-2, Wi-1), list(Wj-3, Wj-2, Wj-1))) = 2

Count(list(Wi-3, Wi-2, Wi-1)) + Count(list(Wj-3, Wj-2, Wj-1)) = 6
```

Similarity between words with and $\Re G$ based on preceding three words: P(Wi, Wj) = 2/6 = 0.33

Number of matched words for word আগে and পরে based on following three words:

```
Count(match(list(Wi+3, Wi+2, Wi+1), list(Wj+3, Wj+2, Wj+1))) = 2
Count(list(Wi+3, Wi+2, Wi+1)) + Count(list(Wj+3, Wj+2, Wj+1)) = 6
```

Similarity between words with and has based on following three words: P(Wi, Wj) = 2/6 = 0.33

Similarly, 4th and 5th gram model can be implemented in the same way.

The value of similarity between words আগৈ with পরি when considering preceding three words is 0.33 and considering following three words it is also 0.33. Different types of threshold values are experimented and best result is earned with 0.20. Both words are grouped in the same cluster when all the probability scores are greater than this threshold value.

5. RESULT ANALYSIS

In the tri, 4th and 5th gram model we derive 2215, 3327 and 5730 word clusters in total respectively. Some clusters randomly from each of the model are represented here in the following tables:

Table 1 Word Cluster for tri-gram model

	or for an grain moder
<u>ফায়ার</u>	মেডিকেল
সার্ভিস	কলেজ
মুখোমুখি	দেখাশোনা
) जःपर्य	দেখভাল
উপপরিদর্শক	সহকারিসহ
এসআই	সঙ্গী
	· "
পিকআপকে	প্রথম
রিকশাকে	আলোকে
	THE HE T
বর্ডার	মহানগর
গার্ড	প্রথাণগর প্রভাতী
110	49101
(e) TOLZ	in the state of th
প্রমের	আশ্বাসের
সম্পর্ক	পূরণের
ফেরদৌস	বিচারিক
জুনায়েদ	হাকিম

Table 2 Word Cluster for 4th gram model

লাখ	নায়েক
টাকা	সুবেদার
সহকারিসহ	চালবোঝাই
সঙ্গী	বালুবোঝাই
মহানগর	ত্বমড়ে
প্রভাতী	মুচড়ে
কমিটি	মালামালসহ
গঠন	আসবাবপত্ৰসহ
ঘটনাস্থল	দেশলাইয়ের
পরিদর্শন	কাঠি
বার্ষিক	ফায়ার
ওয়াজ	সার্ভিস

Table 3 Word Cluster for 5th gram Model

ওরস	স্ট্রুডেন্ট
মাহফিল	ভিসায়
মিনার	আশঙ্কায়
মসজিদ	বিঘ্ন
ঘটনাস্থল	আশ্বাসের
পরিদর্শন	পূরণের
গোলা	মার্কেটিং
অবিস্ফোরিত	ম্যানেজার
বিপর্যয়	মেরে
চরম	পিটিয়ে
এনটিভি	গার্লস
আরটিভি	ক্যাডেট

After analyzing the word clusters of all the three models we find poor similarity in some word clusters such as 266 for tri-gram, 300 for 4th gram and 360 for 5th gram. So, we find 1949, 3027 and 5370 clusters in strong similarity for the tri, 4th and 5th gram model respectively. So, the accuracy for strong similarity in

```
Tri-gram :- 88%
4<sup>th</sup>gram :- 91%
```

5thgram :- 93%

So, it is observed that 4th gram is is better than tri-gram and 5th gram is the best in all of them.

6. CONCLUSION

Word clustering is important for various types of purpose for any language. For this reason in Bangla, trigram, 4th gram and 5th gram model is implemented here to proceed the previous work on word clustering. The analysis and result presented above on quite a large Bangla corpus has helped us to find the efficiency among the three mentioned models for word clustering. On the basis of the observation, it can be said that better efficiency is in the higher orders than the preceding orders of the N-gram model.

REFERENCES

Top 10 most spoken languages in the world, $\underline{\text{http://listverse.com/2008/06/26/top-10-most-spoken-languages-in-the-world/}}$

Unsupervised machine learning,

http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm

Y Goldberg. "Task-specific word-clustering for Part-of-Speechtagging".arXiv preprint arXiv:1205.4298, 2012.

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.

Sabir Ismail, M. Shahidur Rahman. https://www.researchgate.net/publication/261551758_Bangla_Word_Cl ustering Based on N-gram Language Model, in press.

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

P F Brown, P V Desouza, R L Mercer, V J D Pietra, V J Della. and J CLai. "Class-based N-gram Models of Natural Language". Computationalinguistics, 18 No: 4, 1992, P: 467-479.

EEKorkmaz. "A method for improving automatic wordcategorization". Doctoral dissertation, Middle East Technical University, 1997, in press.

S Mori, M Nishimura and N Itoh. "Word clustering for a word bi - gramModel". International Conference on Spoken Language Processing, 1998, in press.

Clustering – Introduction, http://home.deib.polimi.it/ matteucc/Clustering/tutorial h

Clustering-Introduction, ``http://www.stanford.edu/class/cs345a/slides/12-clustering.pdf ``.Stanford University-Clustering."

Similarity in semantics and contexts, http://www.ilc.cnr.it/EAGLES96/rep2/node37.html