Extracting Collocations from Bengali Text Corpus

Bidyut Das

Abstract

Automatic collocation extraction is very important in various applications in the field of natural language processing such as machine translation, word sense disambiguation, information retrieval, and language modelling in speech processing, lexicography and many more. The success of extracting collocations depends on the technique of pre-processing. A systematic pre-processing technique is described in this paper. Then the pre-processed data is used to extract collocation by using two methods: Point-wise Mutual Information and Fuzzy Bi-gram Index. The paper mainly focuses on bi-gram extraction from a Bengali news corpus. Collocations of higher length i.e., n-grams (n>2) are then obtained when the extracted collocations of lower lengths are treated as individual words.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of C3IT

Keywords: Collocation extraction; Text pre-processing; Fuzzy Bi-gram Index; Natural language processing; Bengali text corpus;

1. Introduction

A collocation is just a set of words occurring together more often than by chance in a corpus. In [1] defines a collocation is a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components. According to [2] collocation can be of two types i.e. rigid or flexible. Rigid collocation are those n-grams that always occur side by side and appear in same order whereas flexible collocation are n-grams that can have intervening words placed between them or can occur in different order. This paper concentrates on first types of collocation i.e. Rigid.

Most methods for collocation extraction are based on verification of typical collocation properties. These properties are formally described by mathematical formulas that determine the degree of association between components of collocation. Such formulas are called association measures which compute an association score for each collocation candidate extracted from a corpus. The scores indicate a chance of a candidate to be a collocation. They can be used for ranking or for classification – by setting a threshold. Finding such a threshold depends on the intended application. In the last thirty years a number of different methods employing various association measures have been proposed [3-6]. A comprehensive study of statistical aspects of word co-occurrences can be found in [7]. As all the association measures have their shortcomings and certain benefits, a combination of different measures can produce better results [8]. The paper is organized as follows. Pre-processing technique is portrayed in Section 2. Existing collocation extraction methodologies are illustrated in section 3. The proposed methodology is described in Section 4. Dataset and Experimental results are given in section 5. Section 6 gives the conclusions.
2. Pre-Processing

In this paper, a systematic approach of pre-processing is proposed for finding collocations from a Bengali text corpus. The corpus is pre-processed in such a way that the frequency of co-occurrence word-pair is easily counted from the corpus. The pre-processing approach consists of the following steps. Initially, the HTML tags are removed from the corpus if it is collected from the web. Then, special characters including numeric digits are removed. It is noticed that if any special character is present in between any two consecutive words then they are not considered as a co-occurrence word-pair to extract collocation. Therefore, after removal of any special character the line is broken to extract collocation properly. After that the stopwords are eliminated from the corpus and the line is again broken where the stopword is present. After removal of stopwords the line that contains only a single word is removed from the corpus because single word can not form collocation. Finally, stemming [9] is performed manually and the pre-processed data is ready. Then the pre-processed data is used to count the co-occurrence frequency.

3. Existing Methods

The 84 association measures are proposed in [10] to extract collocations. It includes some popular methods such as point-wise mutual information, Pearson’s test, z-score, odds ratio, or log likelihood ratio [11] that gives good result to extract bi-gram from the corpus. Among all standard collocation extraction indices, Point-wise Mutual Information (PMI) yields the best results in general language collocation extraction studies [10]. Therefore, proposed method Fuzzy Bi-gram Index (FBI) is compared with Point-wise Mutual Information (PMI) for extracting collocations from Bengali text corpus.

3.1. Point-wise Mutual Information (PMI)

The PMI has been utilized to find the closeness between word pairs [12]. PMI for two events $x$ and $y$ is defined as $I(x, y)=\log_2 \left[\frac{P(x, y)}{P(x)P(y)}\right]$. If $w_1$ and $w_2$ are written for the first and second word respectively, instead of $x$ and $y$, then the PMI for the two words $w_1$ and $w_2$ is given by $I(w_1, w_2)=\log_2 \left[\frac{P(w_1, w_2)}{P(w_1)P(w_2)}\right]$ where $P(w_1, w_2)$ is the probability of two words $w_1$ and $w_2$ coming together in a certain text and $P(w_1)$ and $P(w_2)$ are the probabilities of $w_1$ and $w_2$ appearing separately in the text, respectively. If $P(w_1, w_2)=P(w_1)P(w_2)$ that is, the two words are independent to each other, then $I(w_1, w_2)=0$ which indicates that these two words are not good candidates for collocation. A high PMI score signifies the presence of a collocation.

4. Proposed Methodology

Based on fuzzy sets, a new approach is proposed for finding collocations from a text corpus. As mentioned earlier, a collocation is just a set of words occurring together more often than by chance in a corpus. Collocations are extracted based on the frequency of the joint occurrence of the words as well as that of the individual occurrences of each of the words in the whole corpus. Intuitively, when a set of words is extracted as a collocation, then the joint occurrence of the words must be high in comparison to that of the constituent individual words. Researchers modeled this intuition as the ratio of the joint word appearance value to the individual word appearance values, either explicitly or implicitly.

It is to be realized that the concept of high occurrence is imprecise/vague in nature as it depends on many factors such as the corpus size, occurrences of other words or word combinations in the corpus. Accordingly, the concept of fuzzy set theory is brought here to manage such imprecision in representing the notion of high occurrences. Here two fuzzy sets, namely, ‘High Word Occurrence’ (HWO) and ‘High Word Pair Occurrence’ (HWPO) are considered where HWO corresponds to an individual word and HWPO corresponds to an adjacent word combination. Membership functions for such fuzzy sets are decided based on the word appearance statistics in the corpus. Finally, the said fuzzy membership values are combined to define the Fuzzy Bi-gram Index (FBI). FBI assigns a value in $[0, 1]$ characterizing the degree of an adjacent word-pair to be a bi-gram (collocation of length 2). Adjacent bi-grams are analyzed...
later to determine collocation of higher length i.e., n-grams (n>2).

4.1. High Word Occurrence(HWO)

This is a fuzzy set which corresponds to an individual word and its membership value represents the degree of being high appearance of the word in the corpus. Based on the occurrences of all individual words in the corpus, the membership function of HWO is decided as \( \mu_{HWO}(w) = \frac{C_{iw}(n)}{T_{iw}} \) where \( n \) is the number of occurrence of the word \( w \) in the corpus, \( T_{iw} \) is the total number of words in the corpus, \( C_{iw}(n) \) is the number of words occurred up to \( n \) times (cumulative frequency) in the corpus i.e.,

\[
C_{iw}(n) = \sum_{i=1}^{n} f_{iw}(i)
\]

where \( f_{iw}(i) \) is the number of words appeared \( i \) times in the corpus. Obviously \( 0 \leq \mu_{HWO}(w) \leq 1 \). As expected, \( \mu_{HWO} \) is 0 for all those words which are absent in the corpus and it is 1 for the words which occurred maximum number of times in the corpus.

Usually the value of \( f_{iw}(i) \) decreases with the increase of \( i \). It is observed that the value of \( f_{iw}(i) \) is considerably high for \( i=1 \) with respect to other values of \( i \). That is, the value of \( \mu_{HWO}(w) \) is high if \( w \) is appeared only once in the corpus and it conflicts our intuition. In order to overcome this, the actual value of \( f_{iw}(1) \) is replaced by the average of other appearance values i.e.,

\[
f_{iw}(1) = \frac{1}{\text{IW max}} \sum_{i=1}^{\text{IW max}} f_{iw}(i)
\]

where \( \text{IW max} \) is the maximum number of occurrence of any word in the corpus. A typical membership function of the fuzzy set HWO is shown in Fig.1.(a)

4.2. High Word Pair Occurrence(HWPO)

This is the second fuzzy set which corresponds to a word pair and its membership value represents the degree of being high occurrence of the word-pair. Based on the occurrence of all word pairs in the corpus, its membership function is decided as \( \mu_{HWPO}(w1,w2) = \frac{C_{wp}(m)}{T_{wp}} \) where \( m \) is the number of appearance of the word-pair \( (w1,w2) \) in the corpus, \( T_{wp} \) is the number of word-pairs in the corpus , \( C_{wp}(m) \) is the number of word-pairs occurred up to \( m \) times (cumulative frequency) in the corpus i.e.,

\[
C_{wp}(m) = \sum_{i=1}^{m} f_{wp}(i)
\]

where \( \text{WP max} \) is the maximum number of occurrence of any word-pair in the corpus. A typical membership function of the fuzzy set HWPO is shown in Fig.1.(b)

4.3. Fuzzy Bi-gram Index(FBI)
The membership values of above two fuzzy sets HWO and HWPO are combined to define a collocation measure, named as Fuzzy Bi-gram Index (FBI). It is to be noted that the degree of a word pair of \((w_1,w_2)\) to be a bi-gram is directly proportional to \(\mu_{HWPO}(w_1,w_2)\) and inversely proportional to the values of \(\mu_{HWO}(w_1)\) and \(\mu_{HWO}(w_2)\). Accordingly, \(FBI(w_1,w_2) = \mu_{HWPO}(w_1,w_2)\left[1-\alpha(\mu_{HWO}(w_1)+\mu_{HWO}(w_2))\right]\) where \(\alpha\in[0,0.25]\), and occurrence of \(w_1\) and \(w_2\) exclude the occurrence as word pair \((w_1,w_2)\) in calculation of \(\mu_{HWO}(w_1)\) and \(\mu_{HWO}(w_2)\). Here \(FBI(w_1,w_2)\in[0,1]\), and it provides a measure to be a bi-gram. That is, the more is the value of \(FBI(w_1,w_2)\), the more is the possibility of \((w_1,w_2)\) to be a bi-gram. A word pair of \((w_1,w_2)\) is identified as a bi-gram if the value of \(FBI(w_1,w_2)\) is greater than a threshold value which is taken 0.5 in our experiments. One may decide any other threshold value depending on requirements.

5. Experimental Results

The Bengali News Corpus (BNC) developed by Computer Vision and Pattern Recognition (CVPR) unit of ISI-Kolkata is used for the experiment. The untagged corpus consists of around 250 documents from different domains. The Bengali stop-word list provided by IIT-Kharagpur is used for eliminating the stopwords from the corpus. The list consists of around 275 Bengali stopwords that are used most frequently in Bengali text corpus. The output of each stage of pre-processing is shown below. Initially, all HTML tags are wiped out from the corpus (Fig.2.(a)). Then, the Bengali special characters including numeric characters and the most frequently coming character ‘danda’ are eliminated. Every line of the corpus is broken when any special character appears in the line (Fig.2.(b)). After that, a stopword list is used to eradicate stopwords. Every line is again broken when any stopword appears in the line (Fig.2.(c)). Finally, the line that contains single word is removed from the text corpus (Fig.2.(d)).

![Fig.2.(a) Selected portions of a text after removal of HTML tags; (b) Line break for appearance of any special character; (c) Line break for presence of any stopword; (d) Removal of single-word line](image)

The pre-processed data is ready. Then the data set is used for extracting collocations by using PMI and FBI methods. Among various evaluation measures, precision and recall are the ones used widely for collocation extraction and so, the results are evaluated by these two measures. Precision can be seen as a measure of exactness or fidelity, whereas recall is a measure of completeness. In collocation extraction, the precision is the proportion of true positive (i.e., collocations of the desired type) among the n collocations and how many of all suitable collocations that could have been extracted from the corpus are actually found in the n-best list are called recall. Precision and recall scores for extracting collocations (bi-grams) are shown in tables 1 and 2 respectively. The portions of ranked output lists, returned by each of the two measures, are considered to calculate precision and recall. These results demonstrate better performance of the FBI than the most popular PMI measure, especially when the considered rank list is
large.

**Table 1. Precision Score**

| % of Ranked List Consider | FBI | PMI |
|---------------------------|-----|-----|
| 10                        | 73.02 | 70.92 |
| 20                        | 69.12 | 61.40 |
| 40                        | 68.86 | 44.59 |
| 60                        | 67.50 | 42.89 |
| 80                        | 58.56 | 41.95 |
| 100                       | 56.32 | 41.12 |

**Table 2. Recall Score**

| % of Ranked List Consider | FBI | PMI |
|---------------------------|-----|-----|
| 10                        | 7.16 | 7.09 |
| 20                        | 13.50 | 12.25 |
| 40                        | 27.54 | 17.41 |
| 60                        | 38.86 | 24.19 |
| 80                        | 46.03 | 35.64 |
| 100                       | 56.60 | 41.93 |

6. Conclusions

In this article Fuzzy Bi-gram Index is compared with Point-wise Mutual Information for extracting collocations from a Bengali News Corpus. It is difficult to say that a single method is best for collocation extraction. It depends heavily on data, language, and notion of collocation itself. In [13] FBI is compared with the t-score and log-likelihood ratio to extract collocation from a biomedical corpus. In this paper FBI is compared with point-wise mutual information. The performance of the FBI scheme is verified on Bengali text and is found perform better than the PMI. Therefore, FBI is much effective than the other statistical association technique to extraction collocation from the Bengali text corpus.

References

1. Choueka, Yaacov (1988). Looking for needles in a haystack. In Proceedings of RIAO’88, pages 609–623.
2. Goldstein, Ira. Collocations in Machine Translations [Internet]. Version 4. Knol. 2008 Jul 27.
3. C. Manning and H. Schutze “Foundations of statistical natural language processing”, MIT Press, Cambridge, 1999.
4. D. Pearce, “A comparative evaluation of collocation extraction techniques”, In Proc. 3rd International Conf. on language Resources and Evaluation, Las Palmas, Spain, 2002.
5. K. Kita, Y. Kato, T. Omoto, and Y. Yano “A comparative study of automatic extraction of collocations from corpora: Mutual information vs. cost criteria”, Journal of Natural Language Processing, Vol. 1, 1994, pp.21–33.
6. S. Evert and B. Krenn, “Methods for the qualitative evaluation of lexical association measures”, In Proceedings 39th Annual Meeting of the Association for Computational Linguistics, 2001, pp.188–195.
7. S. Evert “The Statistics of Word Co-occurrences: Word Pairs and Collocations”, Ph.D. thesis, University of Stuttgart, 2004.
8. J. Lin, S. Li, Y. Cai, “A New Collocation Extraction Method Combining Multiple Association Measures” Proc. of the 7th International Conference on Machine Learning and Cybernetics, Kunming, 2008.
9. M. F. Porter, “An Algorithm for Suffix Stripping”, Program: electronic library and information systems, Vol. 14 (3), pp. 130-137, July 1980.
10. Pavel Pecina, “An Extensive Empirical Study of Collocation Extraction Methods”, Proceedings of the Association for Computational Linguistics Student Research Workshop, 2005, pages 13–18.
11. T. E. Dunning, “Accurate methods for the statistics of surprise and coincidence”, Computational Linguistics, Vol. 19, 1993, pp.61–74.
12. W. K. Church, P. Hanks, Word association norms, mutual information and lexicography, In Proceedings of the 27th meeting of the Association of Computational Linguistics (1989) 76-83.
13. Bidyut Das, Deba Prasad Mandal and Kaushik Roy, “Extraction of Collocations from a Text Corpus: A Fuzzy Measure”, In Proc. IEEE International Conference on Advances in Communication, Network, and Computing, 2010, pp 134-136.