A CLUSTERING TECHNIQUE FOR THE VIETNAMESE WORD CATEGORIZATION

Nguyen Minh Hiep[a*], Nguyen Thi Minh Huyen[b],

Ngo The Quyen[b], Tran Thi Phuong Linh[a]

[a]*The Faculty of Information Technology, Dalat University, Lamdong, Vietnam*
[b]*The Faculty of Informatics, VNU University of Science, Hanoi, Vietnam*

Article history
Received: January 04th, 2016
Received in revised form: March 10th, 2016
Accepted: March 16th, 2016

Abstract

In natural language processing, part-of-speech (POS) tagging plays an important role, as its output is the input of many other tasks (syntax analysis, semantic analysis. . .). One of the problems related to POS tagging is to define the POS set. This could be solved using unsupervised machine learning methods. This paper presents an application of the DBSCAN clustering algorithm to classify Vietnamese words from a large corpus. The features used to characterize each word are naturally defined by the context of that word in a sentence. We use a large corpus containing sentences automatically extracted from the online Nhan Dan newspaper.

Keywords: Clustering; Corpus; DBSCAN; POS; POS tagging; Tag set.

1. INTRODUCTION

The question of Vietnamese word classification has been discussed in several linguistic studies [1]. This problem can be solved by the method called unsupervised machine learning method. We present technique that clusters Vietnamese words from a store of documents in the order to identify a tagged lexical class. The feature which is used to cluster words is the context of this word in the sentence. The algorithm DBSCAN is used to cluster words. Data training are automatically clustered big size Vietnamese document store from Nhan Dan online and Thanh Nien online newspapers.

* Corresponding author: Email: hiepnm@dlu.edu.vn

This article comprises three parts. Part 1 introduces the research motivation of authors, some existing methods have been studied and published, the approaches and methods that we use. In part 2, we introduce the similar information probability measure of two words, and DBSCAN clustering algorithm was improved conforming to features of clustering Vietnamese part-of-speech problems. In the next section, we present the results of clustering and evaluate these results. The last part is the conclusion of the article.

2. UNSUPERVISED APPROACHES FOR POS TAGGING

A group of approaches considers POS tagging as a clustering problem, where the words are clustered into syntax categories that each represents a POS tag.

Brown et al. (1992) employs an information theoretic approach where the word clusters yielding the greatest average mutual information between adjacent classes are discovered. To this end, initially each word is assigned to a separate cluster. Then the cluster pair which yields the minimum loss in the average mutual information is merged. The process is repeated until a set of clusters is found. Finally, each word is replaced into another cluster, if the resulting cluster is greater average mutual information. The algorithm would be terminated if no more moves are possible, which leads to greater average mutual information. Some of the earlier work represents the words in terms of their context vectors, where the adjacent words are used to measure the similarity among words. At the end, vector space models are widely used to represent statistics regarding the contexts of the words.

Finch & Chater (1992) consider the two preceding and the two following words that are in the most frequent 150 words as the context. To measure the linguistic similarity among context vectors, a Spearman Rank Coefficient of Correlate is used. Using the similarity measure, hierarchical agglomerative clustering is performed to capture the linguistic categories in a hierarchical structure.

Schutze (1993) uses context vectors that keep the counts of the context words in a variable size of window. Because of the unfeasibility of such large vectors, Singular Value Decomposition (SVD) (in Deerwester et al. (1990)) is used to reduce the dimensionality in the concatenated context vectors. In the reduced space, nearest neighbors are induced to form individual clusters by Buckshot clustering (Cutting et al. 1992). Schutze (1993) also uses neural networks to cluster ambiguous words which are poorly clustered by the Buckshot clustering.

Schutze (1995) improves the previous work also using the contexts of the context words, in addition to the context words itself. Another difference in this approach is that the context vectors are used separately instead of being combined in to a single context vector.

Clark (2000) follows the same distributional hypothesis within a distributional clustering algorithm. On the other hand, he defines the contexts probabilistically where a word is defined by radio between probability distribution and possible contexts. Instead of using context words, the clusters of the context words are used to eliminate the sparseness problem. Kullback-Leibler (KL) divergence is used to measure the divergence among the clusters, to decide which merges will be appropriate in each step.

Freitag (2004) employs an information theoretic co-clustering algorithm (Dhillon et al. 2003) to induce the POS tags of the words. The algorithm makes use of both words and their contexts in a similar fashion to the other approaches given in this section. Words and their contexts are replaced in the clusters finding the clusters which will maximize the mutual information between the words and the contexts in a particular cluster. He also develops a Hidden Markov Model (HMM) tagger to tag low frequency words.

Biemann (2006) employs a graph based clustering algorithm to induce POS tags. One advantage of the graph based on clustering algorithms is that the number of clusters does not need to be initially defined. In a graph clustering algorithm, the number of clusters is discovered while the graph is formed. Biemann uses two graphs; one for high frequency words where there is sufficient contextual information and one for medium and low frequency words where only likelihood statistics are being used. In his approach, to assign the classes, he uses the Chinese Whispers (CW) graph-clustering algorithm (see Biemann et al. (2007), with more detailed definition of the algorithm and its application to natural language). A graph is constructed for the high frequency words by using the context similarity of the words to draw an edge between two words. A threshold is used employing the cosine similarity of the words. Another graph is constructed by using the log-likelihoods and the number of common neighbors shared among the words. Both graphs are partitioned by the CW algorithm which produces some syntactic categories. However, Biemann defines a trigram model in which the joint probability of the tags and the words are maximized in a corpus to enlarge the dataset for tagging.

3. CLUSTERING ALGORITHM AND EVALUATION

3.1. DBSCAN Clustering Algorithm

The algorithm is based on DBSCAN clustering algorithm (Density-Based Spatial Clustering of Applications with Noise) is a data clustering algorithm proposed by Martin Ester, Hans-Peter Kriegel, Jorg Sander and Xiaowei Xu in 1996. It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN is one of the most common clustering algorithms and also most cited in scientific literature [2].

Basic idea: DBSCAN's definition of a cluster is based on the notion of density-reachable. Basically, a word *q* is directly density-reachable from a word *p* if it is not farther away than a given distance Eps (i.e., is part of its Eps-neighborhood) and if *p* is surrounded by sufficiently many words such that one may consider *p* and *q* to be part of a cluster. *q* is called density-reachable (note the distinction from “directly density-reachable”) from p if there is a sequence $p_1 \ldots p_n$ of words with $p_1 = p$ and $p_n = q$ where each p_{i+1} is directly density-reachable from *p*.

Note that the relation of density-reachable is not symmetric. *q* might lie on the edge of a cluster, having insufficiently many neighbors to count as dense itself. This would halt the process of finding a path that stops with the first non-dense point. By contrast, starting the process with *q* would lead to *p* (though the process would halt there, *p* being the first non-dense word). Due to this asymmetry, the notion of density-connected is introduced: two words *p* and *q* are density-connected if there is a word o such that both p and q are density-reachable from o. Density-connectedness is symmetric. A cluster, which is a subset of the words of the database, satisfies two properties:

All words within the cluster are mutually density-connected.

If a word is density-connected to any point of the cluster, it is part of the cluster as well.

DBSCAN requires two parameters: Eps and the minimum number of words required to form a cluster (MinWords). It starts with an arbitrary starting word that has not been visited. This word's neighborhood is retrieved, and if it contains sufficiently many words, a cluster is started. Otherwise, the word is labeled as noise. Note that this point might later be found in a sufficiently sized Eps-environment of a different word and hence be made part of a cluster.

If a word is found to be a dense part of a cluster, its Eps-neighborhood is also part of that cluster. Hence, all points that are found within the Eps-neighborhood are added, as is their own Eps-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited word is retrieved and processed, leading to the discovery of a further cluster or noise.

The clustering algorithm is divided into two phases. Phase 1: Apply algorithm DBSCAN clustering to clustering. Differently, the algorithm is improved so that each word after clustering has been considered in other clusters corresponding to contexts set since each word can be in many different clusters corresponding to their different contexts. For example: the word "rock" is both verb in context of "The horse kicked" which is same as the word in context of "Bricks are made of stone" (Figure 1). The second stage can cluster the same context cluster together. The context of each cluster c is defined as follows:

$$V_c = \cup_i v_i : v_i \; is \; context \; set \; of \; the \; word \; w_i \in c \qquad (1)$$

Then $d(w_i; c)$ is given by equation (7), this means the distance from word to cluster, this is sufficient condition to determining whether the word wi is in a cluster or not. Necessary condition is $d(w_i, w_j) \geq Eps$ and $(w_j, w_i) \geq Eps$, where $w_j \in c$ is core word.

Input: W = (w_1, w2, …,w_N), Eps, MinWords

Output: Cluster set: C = (c_1, c_2, …, c_K)

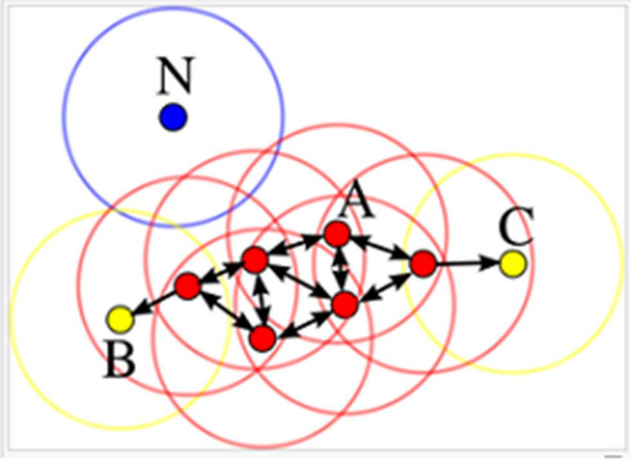


Figure 1. Words of A point are core words

The others of B and C point are density-reachable from A and thus density-connected and belong to the same cluster. Word N is a noise word that is neither a core word nor density-reachable (MinWords = 3 or MinWords = 4)

3.2. Evaluation

We use the V-measure [3] for evaluating clustering results. V-measure is an external entropy-based cluster evaluation measure.

Suppose the data set with N data points and has two partitions of this data set: a set of classes, $C = \{c_i \,|i = 1, 2, …,n\}$ and a set of clusters, $K = \{k_i|i = 1, 2, …m\}$.We build a table $A = \{a_{ij}\}$ where a_{ij} is the number of data points that are members of classes ci and elements of cluster k_j. V-measure introduces two criteria for a clustering solution: homogeneity and completeness. A clustering result satisfies homogeneity if all of its clusters contain only data points which are members of a single class. A clustering result satisfies completeness if all the data points that are members of a given class are elements of the same cluster. The homogenity and completeness of a clustering solution run roughly in opposition: Increasing the homogeneity of a clustering solution often results in decreasing its completeness [3].

Calculation Homogeneity:

$$h = \begin{cases} 1 & \text{if } H(C|K) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & \text{else} \end{cases} \tag{2}$$

Where

$$H(C|K) = -\sum_{k=1}^{|K|}\sum_{C=1}^{|C|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}} \tag{3}$$

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{N} \log \frac{\sum_{k=1}^{|K|} a_{ck}}{N} \tag{4}$$

Calculation Completeness:

$$c = \begin{cases} 1 & \text{if } H(K|C) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & \text{else} \end{cases} \tag{5}$$

Where

$$H(K|C) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|} \frac{a_{ck}}{N} \log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}} \tag{6}$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{N} \log \frac{\sum_{c=1}^{|C|} a_{ck}}{N} \tag{7}$$

Calculation V-measure:

We calculate a clustering solution's V-measure by computing the weighted harmonic mean of homogeneity and completeness [3].

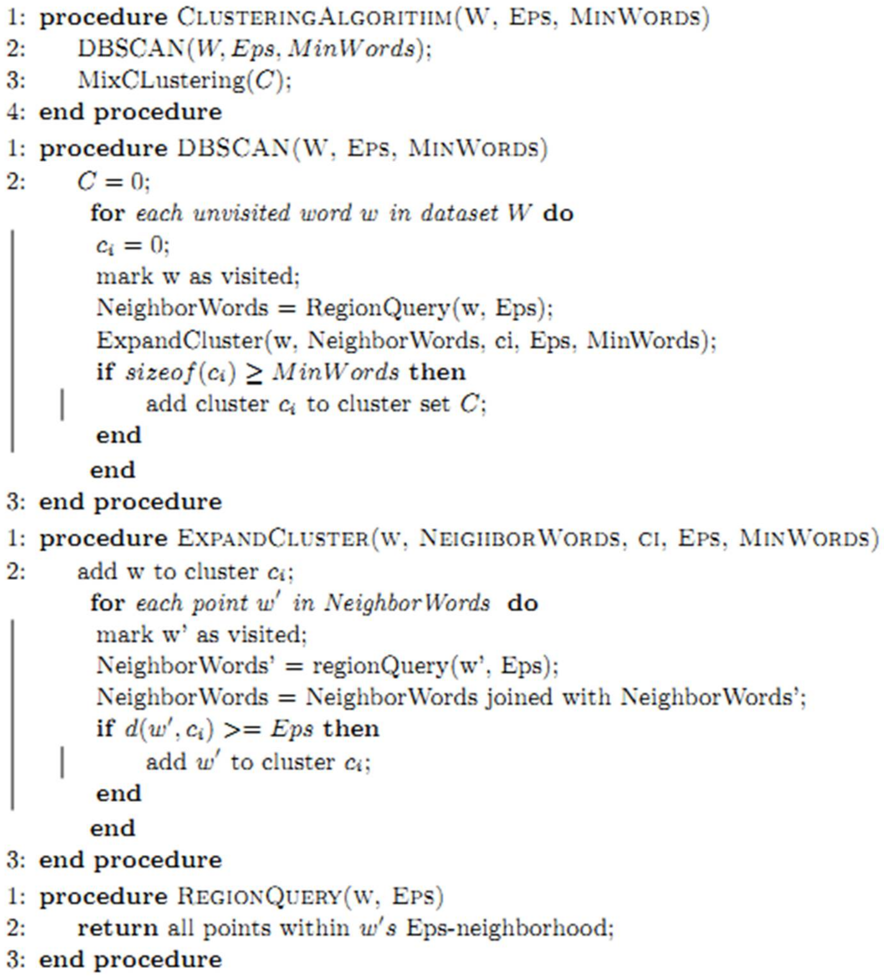
$$V_\beta = \frac{(1+\beta) \times h \times c}{(\beta \times h) + c} \tag{8}$$

If β is greater than 1, completeness is strongly weighted in the calculation, if β is less than 1, homogeneity is strongly weighted [3].

4. A CLUSTERING METHOD FOR VIETNAMESE WORD CATEGORIZATION

4.1. Word Clustering Method

Vietnamese clustering problem is stated as follows: Give $W = (w_1, w_2, \ldots, w_N)$: The words set of the Vietnamese. We need to determine that $C = (c_1, c_2, \ldots, c_K)$ is a catalog set. So each word from w_t corresponding to ci or c_j is certain. This means that a word can belong to many different catalogs, depending on the context of words. For example, in the sentence of "Tôi đang đá bóng" (playing/kicking) the word "đá" is a verb, but in onother context it's a noun, e.g. "Gạch được làm từ đá" (stone) (Figure 2).



```
procedure ClusteringAlgorithm(W, Eps, MinWords)
    DBSCAN($W, Eps, MinWords$);
    MixCLustering($C$);
end procedure
procedure DBSCAN(W, Eps, MinWords)
    $C = 0$;
    for each unvisited word $w$ in dataset $W$ do
        $c_i = 0$;
        mark w as visited;
        NeighborWords = RegionQuery(w, Eps);
        ExpandCluster(w, NeighborWords, ci, Eps, MinWords);
        if $sizeof(c_i) \geq MinWords$ then
            add cluster $c_i$ to cluster set $C$;
        end
    end
end procedure
procedure ExpandCluster(w, NeighborWords, ci, Eps, MinWords)
    add w to cluster $c_i$;
    for each point $w'$ in NeighborWords do
        mark w' as visited;
        NeighborWords' = regionQuery(w', Eps);
        NeighborWords = NeighborWords joined with NeighborWords';
        if $d(w', c_i) >= Eps$ then
            add $w'$ to cluster $c_i$;
        end
    end
end procedure
procedure RegionQuery(w, Eps)
    return all points within $w's$ Eps-neighborhood;
end procedure
```

Figure 2. DBSCAN Algorithm

In this section, we introduce a clustering technique for Vietnamese based on the context. Each word corresponds to a set of context, which indicates its neighbor relationships. We are giving a similarity information measure of two words based on context set and algorithms including Vietnamese phrases based on this measure. Here are some of the concepts to be applied in Vietnamese clustering.

Let a training text set *D = (d₁, d₂, …)*: Vietnamese files. Each di is a text file. Includes two types of data: The first type consists of 2610 files with more than 2 million syllables have been split word and checked manually by linguists. The second type, we collect the data from online newspapers: Nhan Dan and Thanh Nien on the internet that have been pre-processed and separated by vntokenizer of Le Hong Phuong author in which each $d_i = (s_1, s_2, \ldots)$: A set of sentences in the text. Each sentence sj is a series of words which are denoted as follows:

$$v_i = (w^1_{i-2} w^1_{i-1} w^1_{i+1} w^1_{i+2} w^1_{i-2} \ldots w^{||vi||}_{i-2} w^{||vi||}_{i-1} w^{||vi||}_{i+1} w^{||vi||}_{i+2} w^{||vi||}_{i-2}) \tag{9}$$

Let $I_L = v_i \cap v_j$ be an intersection of two left context sets of v_i and v_j. Let $I_L = v_i \cap v_j$ be an intersection of two right contexts sets of v_i and v_j . Then, we define the similarity information distance measure of two words as following: **Definition**: Let w_i , w_j in *W*, the similarity information distance measure of two words: w_i and w_j, denoted: $d(w_i, w_j)$ is defined as follows:

$$d(w_i, w_j) = \frac{\sum_{w_{i-2}w_{i-1} \in I_L} P(w_i|w_{i-2}w_{i-1}) + \sum_{w_{i+1}w_{i+2} \in I_R} P(w_i|w_{i+1}w_{i+2})}{2} \tag{10}$$

$$d(w_j, w_i) = \frac{\sum_{w_{j-2}w_{j-1} \in I_L} P(w_j|w_{j-2}w_{j-1}) + \sum_{w_{j+1}w_{j+2} \in I_R} P(w_j|w_{j+1}w_{j+2})}{2} \tag{11}$$

Features:

- $0 \le d(w_i, w_j) \le 1$
- $d(w_i, w_j) \ne d(w_j, w_i)$

4.2. Results

In this paper, we used training data of about 2 million syllables which were splitted and checked by linguists. Result of clustering with 2889 words is clustered. Numbers of clusters $K = 279$.

Numbers of cllowpses C = 27 (27 Category of words). Value of V-measure V = 0:32. (Table 1) The value of homogeneity h = 0:53. This value shows the ability of words in cluster K_i that are members of class C_j. If this value is high, the cluster K_i which is labeled C_j is more accurate.

Table 1. Result of Clustering

Eps	Number of words	Number of Clusterings	V-Measure
0.3	2289	279	0.32
0.3	2888	253	0.3
0.3	3002	359	0.31
0.3	3546	122	0.25
0.2	3515	88	0.24
0.2	3504	68	0.24

Because the numbers of K are more than the numbers of C (10 times more), so the value of completeness is low c = 0:23, which reduces the value of V-measure. However, we do not pay much attention to the completeness regarding the problem of POS tagging.

Table 2. Example of Clustering

1	công_ty sở cơ_quan_chức_năng nhà_khoa_học bộ_ngành bộ khu_vực vùng cụm trung_tâm …
2	xây_dựng sử_dụng hạ_tầng bắt_nguồn thực_hiện thi_công hoàn_tất xây rao khai_thác làm_quen quản_lý bảo_vệ duy_tu sửa_chữa …
3	bảo_trợ hiện_diện lây_lan tiến_bộ tôn_trọng tàn_phá tò_mò vận_dụng say_mê xâm_phạm tín_nhiệm nhiêu_khê lạm_dụng truy ngăn_cản …
4	hòa_bình bà_rịa kiên_giang tây_ninh bình_dương quảng_bình bình_thuận ninh_thuận br long_an sóc-trăng tiền_giang bạc_liêu an_giang
5	bán mua thuê có_mặt nhện lấy bắt_nguồn công tự_túc góp tiếp_sức ôm ôm_chầm vớt năn_nỉ…
6	giao_thông y_tế tài_chính tài_nguyên nội_vụ nn thương_mại văn_hóa khoa_học nông_nghiệp giáo_dục gtvt gtcc kinh_tế công_nghiệp …
…	

To develop the POS tagging problem, clustering results (Table 2) will be the linguistic evaluation and to determine the POS by hand from results based on machine learning to better POS current set and apply to problems in the next phase: POS tagging and parsing.

5. CONCLUSIONS

This paper built Vietnamese part-of-speech clustering system based on similarity information measure of context. We rely on DBSCAN clustering algorithm and improved to suit the features of the problem. Our training data are based on two sets of data. The first set of about 2 million syllables have been separated and checked manually by linguists. This result will be developed to solve POS tagging problem by unsupervised machine learning method in Vietnamese. Furthermore, we can use it for determining POS problem in Vietnamese that is still weary of linguists.

REFERENCES

[1] Hong, C.N.: “Vấn đề phân định từ loại trong tiếng Việt”. T/c Ngôn ngữ số 2(1) (2003)

[2] Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: in Proceedings of the 2nd International Conference on Knowledge Discovery and Data mining, Germany (1996).

[3] Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, June (2007).

MỘT KỸ THUẬT PHÂN CỤM CHO TỪ LOẠI TIẾNG VIỆT

Nguyễn Minh Hiệp[a*], **Nguyễn Thị Minh Huyền**[b],
Ngô Thế Quyền[b], **Trần Thị Phương Linh**[a]

[a]*Khoa Công nghệ Thông tin, Trường Đại học Đà Lạt, Lâm Đồng, Việt Nam*
[b]*Khoa Toán – Cơ – Tin học, Trường Đại học Khoa học Tự nhiên Hà Nội, Hà Nội, Việt Nam*
[*]*Tác giả liên hệ: Email: hiepnm@dlu.edu.vn*

Nhận ngày 04 tháng 01 năm 2016
Chỉnh sửa ngày 10 tháng 03 năm 2016 | Chấp nhận đăng ngày 16 tháng 03 năm 2016

Tóm tắt

Trong xử lý ngôn ngữ tự nhiên, gán nhãn từ loại (POS tagging) đóng một vai trò quan trọng, là đầu ra, đầu vào của nhiều nhiệm vụ khác (phân tích cú pháp, phân tích ngữ nghĩa...). Một trong những vấn đề liên quan đến việc gán nhãn từ loại là xác định tập từ loại (POS). Điều này có thể được giải quyết bằng các phương pháp học máy không giám sát. Bài viết này trình bày một ứng dụng của thuật toán phân cụm DBSCAN để phân loại từ tiếng Việt từ kho ngữ liệu lớn. Các đặt trưng được sử dụng để mô tả từng từ được định nghĩa một cách tự nhiên bởi ngữ cảnh của từ đó trong câu. Chúng tôi sử dụng một kho ngữ liệu lớn chứa câu được trích tự động từ báo Nhân Dân.

Từ khóa: Corpus; DBSCAN; Gán nhãn từ loại; Phân cụm; Từ loại; Tập từ loại.