Journal of Computer and Communications, 2015, 3, 33-37
Published Online May 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.35004

Cluster Analysis Based on Contextual Features Extraction for Conversational Corpus

Qi Chen[1,3], Yue Chen[2,3], Minghu Jiang[3]

[1]College of Computer Science and Technology, Shandong University, Shandong, China
[2]Department of Chinese Language and Literature, School of Humanities, Tsinghua University, Beijing, China
[3]Lab of Computational Linguistics, School of Humanities, Tsinghua University, Beijing, China
Email: triplecq@gmail.com, yue-chen11@mails.tsinghua.edu.cn, jiang.mh@tsinghua.edu.cn

Received December 2014

Abstract

Cluster analysis related to computational linguistics seldom concerned with Pragmatics level. Features of corpus on Pragmatics level related to specific situations, including backgrounds, titles and habits. To improve the accuracy of clustering for conversations collected from international students in Tsinghua University, it required contextual features. Here, we collected four-hundred conversations as a corpus and built it to Vector Space Model. With the Oxford-Duden Dictionary and other methods we modified the model and concluded into three groups. We testified our hypothesis through self-organizing map neural network. The result suggested that the modified model had a better outcome.

Keywords

Conversational Corpus, Contextual Features, VSM, SOM

1. Introduction

Natural Language Processing (NLP) involves different levels, including Morphology, Syntax, Semantics and Pragmatics [1]. Different levels are applied to various methods and applications, such as Segmentation based on Morphology and Syntax. With advanced statistics models and algorithms these applications are able to perform a high accuracy. However, a few clustering algorithms for conversational corpus are not aware with levels in Semantics or Pragmatics, which results in ambiguity to categorize specific situations [2]. As an illustration, the traditional method of Vector Space Model (VSM) has some limitations. The method pays its entire attention to the frequency without concern about Semantics and Pragmatics, resulted in a "false positive" or "false negative" match. As a consequence, the algorithm based on VSM above lacks enough ability to represent the corpus and cluster them into right groups.

Some researchers have already applied Semantics into their studies. To overcome the limitations of VSM, there is a combination between VSM and some lexical databases such as WordNet [3]. Same work related to

How to cite this paper: Chen, Q., Chen, Y. and Jiang, M.H. (2015) Cluster Analysis Based on Contextual Features Extraction for Conversational Corpus. *Journal of Computer and Communications*, **3**, 33-37. http://dx.doi.org/10.4236/jcc.2015.35004

Chinese corpus was conducted based on HowNet [4]. The key is to modify the default VSM with Semantics information. Instead of the entire attention to word frequency, these studies combined similar words into a same concept. As a result, the corpus was built into a conceptual tree rather than VSM before. With these efforts the model had a quite low dimensionality and resulted in a good performance related to auto-summarization or categorization.

This paper concentrates more on Pragmatics, particularly on Contextual Features. The experiment is conducted on Conversational Corpus with eight different situations in **Table 1**, including "Hospital", "Restaurant", "Renting House", "Inside Class", "After Class", "Airport", "Baber's" and "Bank". Contextual features are the very representative ones in each situation, such as conversation backgrounds and titles. Features in different situations are distinct to others, thus they play a very important role in clustering. We use two methods to extract contextual features from corpus. The experiment applies these contextual features to modify the default VSM that is calculated based on word frequency and testify this hypothesis by Self-organizing map (SOM) neural network [5]. The experiment result suggests that the combination with contextual features in Pragmatics level bring a better outcome for clustering.

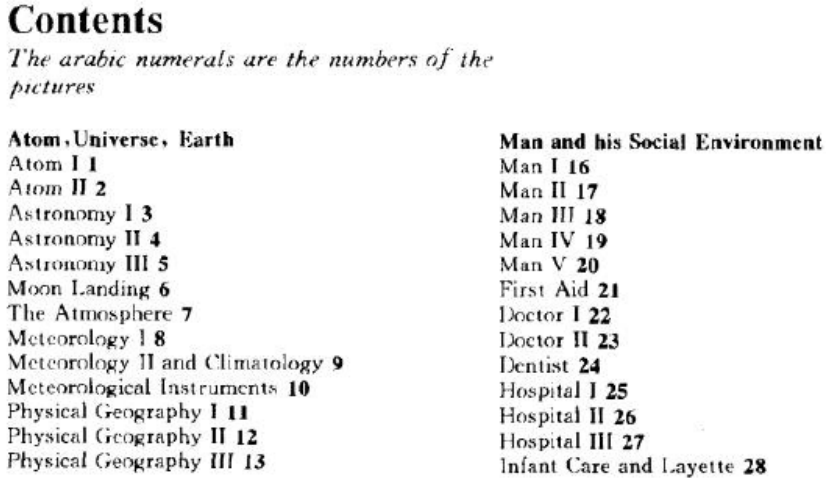
2. Materials and Methods

2.1. Conversational Corpus

The experiment was conducted on the conversational corpus collected from international students in Tsinghua University. The experiment was in eight categories and each of them represented a specific situation, including "Hospital", "Restaurant", "Renting House", "Inside Class", "After Class", "Airport", "Barber's" and "Bank". Each category consisted of fifty different conversations recorded of daily life from these international students. We got rid of titles in each conversation and reorganized them with word segmentation. As a result, the corpus had nearly 10,000 words and about 5000 word tokens after the stop list.

2.2. The Oxford-Duden Dictionary

The experiment used the Oxford-Duden dictionary to map the same category discussed above in order to extract specific contextual features on Pragmatics level. The dictionary was organized into several categories and illustrated with pictorial items within particular situation. This organization helped us to extract features like backgrounds easily. For example, there were several keywords related to "Hospital" illustrated on the dictionary, such as "Drag", "Blood" and "Alcohol", which were very common in the hospital and of great possibility to be referred in conversations. Besides, we were able to access to lots of features related to habits through the dictionary. It was significant to consider these features as integrity rather than separated words. With the Oxford-Duden dictionary, we selected several keywords and maintained a list for modification of VSM (**Figure 1**).



Contents

The arabic numerals are the numbers of the pictures

Atom, Universe, Earth
Atom I 1
Atom II 2
Astronomy I 3
Astronomy II 4
Astronomy III 5
Moon Landing 6
The Atmosphere 7
Meteorology I 8
Meteorology II and Climatology 9
Meteorological Instruments 10
Physical Geography I 11
Physical Geography II 12
Physical Geography III 13

Man and his Social Environment
Man I 16
Man II 17
Man III 18
Man IV 19
Man V 20
First Aid 21
Doctor I 22
Doctor II 23
Dentist 24
Hospital I 25
Hospital II 26
Hospital III 27
Infant Care and Layette 28

Figure 1. Categories in Oxford-Duden dictionary.

2.3. Procedures

The first step was to build VSM from collected corpus [6]. We used four hundred (50*8) conversations as input to construct VSM. The values in VSM were calculated by the frequency of each word. Without a clear-cut boundary for features in VSM, we also paid attention to low-frequency words, for some of them may contain important contextual information. Therefore, we kept all features in the model. The VSM in this state was called the "Default" group. General definition of VSM was illustrated as below:

$$\vec{d}_j = (t_{1,j}, t_{2,j}, \cdots, t_{N,j})$$

$$\vec{q}_k = (t_{1,k}, t_{2,k}, \cdots, t_{N,k})$$

$$sim(\vec{q}_k, \vec{d}_j) = \sum_{i=1}^{N} t_{i,k} \times t_{i,j} \quad sim(\vec{q}_k, \vec{d}_j) = \sum_{i=1}^{N} t_{i,k} \times t_{i,j} \tag{1}$$

$\vec{d}_j$ and $\vec{q}_k$ are two vectors while the Equation (1) is to make a comparison between them.

We selected eight keywords for each category which were often referred in each specific situation. For example, when we talked about words like "Food", "Waiter" and "Beverage", there was a great possibility that the conversation occurred in a restaurant. These words were very common and distinct because of their identities like backgrounds or titles. It was usual for us to hear words like beverage in a restaurant as well as various foods. Besides, the conversation in a restaurant was always occurred between customers and waiters. Therefore, the experiment maintained a self-selected keyword list of sixty-four words (8*8) which were representative for each situation (Details in **Table 1**). The default VSM was then modified by self-selected keyword list and was called the "Defined" group. For each word in VSM occurred in the list, we weighted it for twice than its original value.

Besides, we used the Oxford-Duden dictionary for the keywords selection. The experiment chose ten words for each specific situation from different categories in the dictionary. For example, words like "Teacher", "Student" and "Book" were selected from the dictionary. In the dictionary, these words above were all in a same category. Within this specific situation, keywords like above were considered as a key feature to represent the situation, because all of these keywords shared a same background or common relationship between people or even their habits. As a consequence, the experiment maintained a dictionary-selected keyword list of eighty (10*8) words which were representative for each situation (Details in **Table 1**). The default VSM was then

Table 1. Keyword list.

Group / Category	Defined	Dictionary
Hospital	Doctor, Patient, Nurse, Drug, Pain, Hospital	Alcohol, Blood, Drug, Gauze, Medicare Card, Operation, Surgery, Medicine, Illness, Scalpel
Restaurant	Meal, Eat, Vegetable, Rice, Soup, Menu, Beef, Beverage	Waiter, Beer, Glass, Meal, Dish, Combo, Utensil, Pepper, Dinner
Renting House	Landlord, Rent, House, Estate, Intermediary, Contract, Kitchen	Floor, Door, Window, Lock, Cabinet, TV, Heater, Tap, Gas, House
Inside Class	Teacher, Work, Homework, Content, Philosophy, History, Poetry	Class, Classroom, Student, Teacher, Library, Book, Boy, Girl, Experiment
After Class	Friend, Sports, Party, Activity, Play, Notes, Intern, Coffee	Teacher, Boy, Girl, Classroom, Library, University, Class, Book, Certificate
Airport	Baggage, Passenger, Airplane, Airport, Passport, Flight, Consignment	Baggage, Boarding, Terminal, Duty-free, Exit, Passport, Urgent, Control, Airplane, Passenger
Barber's	Barber, Hair, Haircut, Hairstyle, Perm, Hairdressing, Dye	Hair, Curly, Hairstyle, Barber, Bang, Towel, Blower, Shampoo, Perm, Haircut
Bank	Bank, Password, Employee, Account, Transaction, Rate, Deposit, Withdraw	Counter, Customer, Exchange, Bill, Receipt, Delivery, Bank, Date, Money, Employee

modified by dictionary-selected keyword list and was called the "Dictionary" group. For each word in the VSM occurred in the list, we also weighted it for twice than its original value. The method used to weight VSM this step and above paid its attention to all contextual features no matter what their frequency were. This solution overcame the limitation of VSM and applied Pragmatics information to modify VSM for clustering.

After the construction of three different groups of VSM, we used the SOM neural network to test our hypothesis. The neural network had one hundred cells and was trained in two hundred epochs. The equation to compute distance between each cell was:

$$d_j = \left| X - W_j \right| = \left(\sum_{i=1}^{m} (x_i - w_{ij})^2 \right)^{1/2} \tag{2}$$

While the weight was updated in:

$$\Delta w_{ij} = \eta (x_i(t) - w_{ij}(t)) \tag{3}$$

where η was a constant between 0 and 1.

3. Results

The results from different groups were illustrated clearly in the diagrams. Each group was run by SOM Neural Network in two hundred epochs. Figures were the distance between each cell. The left diagram was the space model of one hundred cells. Every red dot represented a cell while the blue line stood for the distance between each cell. The number in each axis was the weight of neural network. The right diagram illustrated one hundred cells in a 10*10 matrix. Every blue hexagon was a cell in the network which linked to others by the red short line. The color from light yellow to dark red represented the distance between each cell. The deepest color meant the furthest distance.

4. Discussion

As illustrated in diagrams above, the result of clustering in "Dictionary" group was the best. Firstly, the "Default" group in **Figure 2** was hardly to categorize conversations into right group while the distribution of cells in "Dictionary" group was much compact than other groups, which means that more conversations were clustered into the same group. It suggested that more conversations were concluded into a same category, for we collected eight categories of corpus artificially. Secondly, the distance between each cell in different category was further than others. In "Default" group, cells scattered in the model. The boundary between each clustered category was not clear, resulted in a difficulty for categorization. On the other hand, the "Defined" group in **Figure 3** was similar to the "Dictionary" group in **Figure 4**, because both VSMs were weighted. The different distribution was due to different keyword list. Some categories had a same outcome while others had a sharpen distribution.

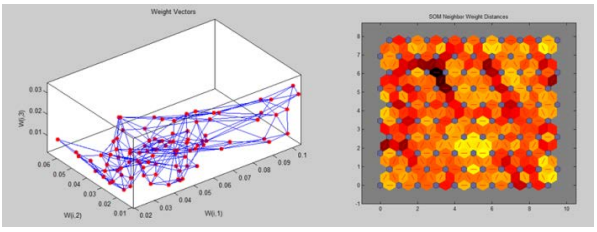


Figure 2. Default group.

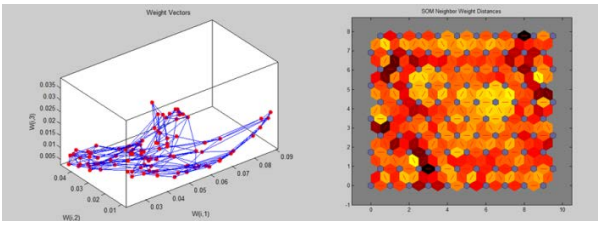


Figure 3. Defined group, twice weight.

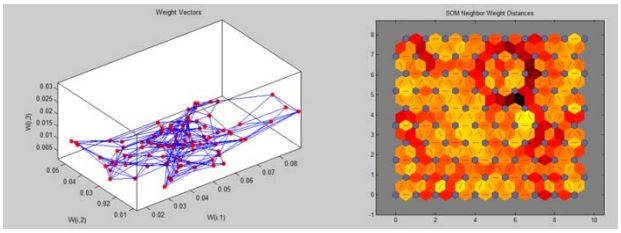


Figure 4. Dictionary group, twice weight.

Meanwhile, there were still some ambiguities in the clustering. Some cells in particular categories were distinct, shown in dark color in diagrams, while other categories were almost the same. For instance, the category "Restaurant" was much clear than others because of the distinction of its keywords, such as "Food", "Beverage" and "Waiter". These keywords were very uncommon in other situations, resulted in a higher accuracy. However, categories like "Inside Class" and "After Class" were hard to distinguish. Due to lots of shared keywords especially in a high frequency, such as "Teacher" and "Student", these two categories were clustered closer than others. Besides, some features which were the key to distinguish such categories were in a very low frequency or without a high weight. The difference between "Inside Class" and "After Class" was hard to define. Generally, we can differentiate these two categories from the main content in conversation, because there might be more contents related to academy in "Inside Class" rather than "After Class". However, conversation between students after class may also highly relate to academic contents while teachers may talk about some activities or jokes instead of class.

Moreover, the convergence was sharpened by the high weight on same keywords. For example, keywords like "Employee" and "Money" were of great possibility to be referred in situations like "Airport" and "Bank". Although these two categories had enough distinct features to be distinguished from each other, high weight on same words like "Money" may mislead the algorithm and come out with ambiguity. These two categories may be much closer if we largely enhance the weight. As a consequence, categories like "Hospital", "Restaurant" and "Barber's" got a more accurate clustering while few categories remained in ambiguity for above reasons.

The selection of keywords is significant to modified VSMs. Contextual features extraction and the weight on VSM will greatly influence the result of clustering. The self-defined keyword list was more flexible but resulted in a sharpen distribution while the dictionary-selected keyword list was much stable but remained some ambiguity. It is important to choose words which are representative in each situation with enough pragmatics information, such as background, title or habit, as well as distinct enough with other situations. With more accurate keyword list and proper weight, the result of clustering will be better.

Acknowledgements

This work was supported by the National Natural Science Fund (61171114) and Key Fund (61433015), and National Social Science Major Fund (14ZDB154 & 13ZD187) of China.

References

[1] Jurafsky and Martin (2000) Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall.

[2] Lewis, D.D. and Hayes, P.J. (1994) ACM Transactions on Information Systems: Special Issue on Text Categorization, Vol. 12. ACM Press.

[3] Ji, H., Luo, Z.S., Wang, M. and Gao, X.Y. (2002) Summarizing Based on Concept Counting and Hierarchy Analysis. *The Natural Language Processing and Knowledge Engineering* (*NLPKE*) *Mini Symposium of the* 2002 *IEEE International Conference on Systems, Man and Cybernetics* (*SMC*2002).

[4] Liao, S.S. and Jiang, M.H. (2005) An Improved Method of Feature Selection Based on Concept Attributes in Text Classification. *Advances in Natural Computation, Lecture Notes in Computer Science*, **3610**, 1140-1149. http://dx.doi.org/10.1007/11539087_152

[5] Kohonen, T. (1987) Self-Organization and Associative Memory. 2nd Edition, Springer-Verlag, Berlin.

[6] Salton, G., Singhal, A., Buckley, C., *et al.* (1994) Automatic Text Decomposition Using Text Segments and Text Themes. Text Retrieval Conference, Washington DC.