Investigating the semantic representation of Chinese emotion words with co-occurrence data and self-organizing maps neural networks

Yueh-Lin Tsai
Department of Psychology, National Cheng Kung University, Taiwan

Hsueh-Chih Chen
Department of Educational Psychology and Counseling, National Taiwan Normal University, Taiwan

Jon-Fan Hu
Department of Psychology, National Cheng Kung University, Taiwan

Keywords: Chinese emotion words, word co-occurrence, Latent Semantic Analysis, semantic representation, self-organizing maps.

Introduction

Regarding the investigation of the word representations, previous researchers often asked participants to rate the similarities between emotion words (Barrett, 2004; Cheng, Cheng, Cho, & Chen, 2013; Romney, Moore, & Rusch, 1997), or to give the scores upon certain psychological dimensions (e.g. valence, arousal, et. al) (Bradley & Lang, 1999; Cho, Chen, & Cheng, 2013; Morgan & Heise, 1988). These direct similarity-based or anchor-based ratings indeed emerge categorical properties of emotion words according to existing theoretical postulations. However, these methods are just based on subjective and retrospective report data. To date we might well grasp what the meanings of general concepts are, but what emotional concepts refer to is still not fully clear. Hence adopting more objective way and robust theories about how people learn and represent the semantic concepts of emotion words is crucial for leading us to in-depth investigation.

Analyzing general products of word use might shed light to answer the question. Latent Semantic Analysis (LSA, Landauer, Foltz, & Laham, 1998) is used to study the concepts behind words from the perspective of how human build-up the word meanings. Accordingly, the semantic representations of words are gradually shaped and learned through the multiple constraints of the input data from the environment since childhood. By calculating the co-occurrence matrix between words and documents, we could study the semantic knowledge from large-scale corpus, and hence explore the possible relationships between individual words.

Besides, previous success of neural networks showed the competence to study the inner representations of human mind. Particularly, Self-Organizing Maps (SOM) model can vectorize the representation relationships of categories of words and display the topological properties across the maps (P. Li, 2009; Ping Li, Burgess, & Lund, 2000; P. Li, Farkas, & MacWhinney, 2004). Parameters of SOM models are sensitive tools to extract subtle variations of word meanings, therefore grasping the common properties under an unsupervised learning manner to express similarity- or anchor-free semantic representations of the complex emotion word meanings. Instead of comparing pairs of emotion words or rating the semantic properties based on predefined dimensions, the present modeling study combined both corpus-based analysis (LSA) and connectionist model to delineate the complexity of semantic representations of Chinese emotion words.

Methods

The Academia Sinica Balanced Corpus of Modern Chinese 3.0 (Sinica Corpus 3.0) was used as the corpora to provide Chinese word uses. This corpus has nearly ten thousand documents, fifteen million words, which were collected from magazines, speeches, internet, and other media in Taiwan. Target emotion words were selected from Taiwan Corpora of Chinese Emotions and Psychophysiological Data on EmotioNet (http://ssnre.psy.ntu.edu.tw/). This data collected 218 Chinese emotion-describing words, with 353 participants rated each emotion words for valence, arousal, dominance, continuance, frequency, and typicality (Cho et al., 2013). We chose 161 two-character words from the database. For validate the data, 36 nouns similar to the stimulus used in Lund and Burgess (1996) were also selected. The words included 12 animal names, 12 words of body parts, and 12 words of nations.

The present study established the semantic representations in Sinica Corpus 3.0 using CTM_PAK (Zhao, Li, & Kohonen, 2011). Window size and list of words were adjusted to the data of emotional words as the parameters of interested. SOM was used to represent the semantic representations of the emotion words but not the nouns. 5 nearest neighbors of each emotional word were also generated and being rated by researcher if the nearest neighbors belong to the same group or not.

Results

The SOM model of 36 nouns showed that three categories of words were projected onto different map regions, with only few words locating at the wrong regions. The results here were at large consistent with the findings of Lund and Burgess (1996), although they used multi-dimensional

scaling as a plausible approach to probe semantic category in the corpus. Hence we might ascertain that the present data could at least distinguish different types of semantic categories under a coarse framework as Lund and Burgess (1996) reported.

However, the words were not form any relevant clusters in the SOM models, and the semantic-similar words were not projected into adjacent region even in the different map size, training length, and initial radius. For excluding the possibility that some low-frequency words would affect the model, SOM models with emotional words appeared more than fifty times in the whole corpus were built. The results also showed that no meaningful clusters were formed.

The average percentages in which five-nearest neighbor words were in the same category as the target word revealed that across data with different parameters, the accuracy of the five-nearest neighbors was all below to thirty percent. Although there were some variations between data, the patterns were not clear so that we treat these variations as random and had nothing to do with the size of moving window and the content of word list.

Discussion

As the result showed, although the data could categorize different types of noun, the semantic representations of emotional words were far from perfect. One of the possible reasons is that emotional words can't be well anchored by other co-occurring words because of the complex and subjective component. As the theory of LSA mentioned, the vectors of each word could just represent the semantic concept across all contexts. Because of the subjective characteristic, highly different emotional words might be able to apply to the same context. So maybe it's not sufficient to separate the emotional words and concepts with only co-occurrence data.

Despite there seems to have above possibility, when taking a closer look in the SOM model of 36 nouns, the arrangement within category were also showed no finer and meaningful categories. Hence although we couldn't reject that emotional words might involve some complex components, the way we extract concepts behind words might have its limitation. It's possible that we have to include some dimensions about subjective feelings to establish the semantic structure of emotional words better. Across the theories about knowledge structure, embodiment cognition highlights the importance of motor, perceptual, and introspective states while forming and retrieving concepts. Further research might get more insight with the inclusion of this kind of approach, and the semantic structure could be well established.

Conclusion

The present study strived to investigate the semantic representation of Chinese emotional words with corpus-based analysis. The results showed that although the data could separate different types of nouns, it is not sufficient to separate and categorized different emotional concept. Further studies have to take other theory into consideration to construct the mental representation more properly.

Reference

Barrett, L. F. (2004). Feelings or words? Understanding the content in self-report ratings of experienced emotion. Journal of personality and social psychology, 87(2), 266.

Bradley, M. M., & Lang, P. J. (1999). Affective norms for English words (ANEW): Instruction manual and affective ratings: Citeseer.

Cheng, C.-M., Cheng, J., Cho, S.-L., & Chen, H.-C. (2013). A Structure Analysis of Chinese Emotions. Chinese Journal of Psychology, 55(4), 417-438.

Cho, S.-L., Chen, H.-C., & Cheng, C.-M. (2013). Taiwan Corpora of Chinese Emotions and Relevant Psychophysiological Data -- A Study on the Norm of Chinese Emotional Words. Chinese Journal of Psychology, 55(4), 493-523.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological cybernetics, 43(1), 59-69.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, 25(2-3), 259-284.

Li, P. (2009). Lexical organization and competition in first and second languages: computational and neural mechanisms. Cogn Sci, 33(4), 629-664.

Li, P., Burgess, C., & Lund, K. (2000). The Acquisition of Word Meaning through Global Lexical Co-occurences. In E. V. Clark (Ed.), Proceedings of the thirtieth stanford child language research forum. Stanford: CA: Center for the Study of Language and Information.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. Neural Netw, 17(8-9), 1345-1362.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28(2), 203-208.

Morgan, R. L., & Heise, D. (1988). Structure of emotions. Social Psychology Quarterly, 19-31.

Romney, A. K., Moore, C. C., & Rusch, C. D. (1997). Cultural universals: Measuring the semantic structure of emotion terms in English and Japanese. Proceedings of the National Academy of Sciences, 94(10), 5489-5494.

Zhao, X., Li, P., & Kohonen, T. (2011). Contextual self-organizing map: software for constructing semantic representations. Behav Res Methods, 43(1), 77-88.