Study on a weighting scheme of word vector using word topic specificity

Yuki Nakayama

School of Knowledge Science,
Japan Advanced Institute of Science and Technology
March 2016

Keywords: word topic specificity, weighting scheme, co-occurrence information, word space model, Latent Dirichlet Allocation

One of the most important procedures in natural language processing application such as information retrieval and text classification is putting word semantic relationship into the model. By discovering latent relation between words, we can enrich features for the model and achieve apparent improvement for tasks of natural language processing. When expressing the semantic relationship of words on the language processing model, many linguistic researchers use language resources such as thesaurus and ontology that are manually constructed by experts in specific area. This is because a good usability since various relationships of terms is structurally represented by hierarchical or network representation in such resources. However, construction of such resources costs a lot of money and time. In addition, because the increasing amount of knowledge required to construct language resources in specialized field such as medical fields, manually construction of resources became more difficult. Then, the method to automatically build language resources by an approach based on the distributional hypothesis has been studied from more than 20 years.

The distributional hypothesis is a hypothesis that indicates that semantic relationship of certain words can be estimated by the similarity of appearance patterns in context that each words was appeared. Based on this hypothesis, it would be able to express the characteristics of the meaning of a given word by counting the co-occurrence frequency of words that was appeared around the given word. To do it concretely, we can express the given word meaning by counting the co-occurrence frequency of the given word and

other words that appear before and after N words of the given word and recording it as a vector that exists in a high dimensional space whose dimension is expressed by contexts (generally co-occurrence words). Then, we can know whether a word has the similarity meaning with another word by calculating the closeness of the distance between each word semantic vectors.

Elements of word vectors gained from the above way are represented as co-occurrence frequencies between words and co-occur words. However, distinctiveness of the word semantic vectors whose elements are co-occurrence frequencies is bad because such vectors are strongly effected by the occurrence frequency of the context word. To deal with this problem, an weighting scheme is usually used. The weighting is to give greater weights to instances that only appear in certain situations and smaller weight to instances that are seen many situations. The most popular weighting methods that is used for the word semantic vector model are point-wise mutual information (PMI) and T-test. Those methods are based on co-occurrence information and determine the weight for elements by evaluating how much higher actual co-occurrence frequency between words compared to the expected co-occurrence frequency. Then, we can remove the frequency effect by using weighting scheme based on co-occurrence information. However, such weighting scheme tend to ignore the characteristic of word itself to emphasize the co-occurrence. For example, in these model, if there is necessity to co-occur with context word such as “recently” and “always”, such words are evaluated as elements to enforce the characteristic of word semantic vector unless it has no remarkable features itself.

Therefore, we proposed a new weighting scheme that consider the features of word itself by using word topic specificity (WTS). Word topic specificity defines how uniformly the word distributed over the topics. We assumed that the more concrete, the more unevenly distributed and the more abstract, the more evenly distributed over the topic. We determined whether words have specific meaning by observing the bias of the distribution of word over topics. In the first procedure to define WTS, we estimate word occurrence probability over the topic and topic occurrence probability by using topic model called latent dirichlet allocation (LDA). LDA does not work well if the number of topic is not proper, so we used Arun’s method to detect the natural number of the topics in the corpus. Then, we assumed that a word that has the most abstract meaning uniformly distributed over the topic and distribution of the word over topics is same as distribution over the topic itself. We compare this distribution with topic distribution of each words by using Jensen-Shanon divergence. Finally, we got WTS, the index that evaluates more concrete word closed to one and more abstract word closed to zero. We gave more weight to elements by considering features of word itself by using this weighting scheme. As

having mentioned above, the word vector space model estimates the semantic relationship of certain words by comparing the similarity of appearance pattern in context that each words appeared. In weighting scheme based only WTS, we cannot give weight that consider the context pattern. Therefore, we combine weight based on WTS with based on the co-occurrence information. We tried 2 approach to combine each weight. One is multiple approach, and another is addition approach. By doing so, we achieve an weighting scheme that considers not only co-occurrence information but also feature of word itself.

Finally, we do experiments in order to confirm that our proposed method enforce the discriminability of the word semantic vectors. After generating the word vector from Wikipedia using existing weighting scheme and proposed weighting scheme, we compare cosine similarity between the generated word vectors and evaluation dataset that consists word pairs and the similarity score of word pairs that rated by human by calculating Spearman's rank correlation coefficient. As a result, the proposed method generates semantic vector that has more discriminability than existing methods in most cases. Especially, in PMI, Spearman's rank correlation coefficient of the word vector that generated by the proposed method was higher than 5% of the coefficient by only PMI.