Representing Words as Lymphocytes

Jinfeng Yang, Yi Guan, Xishuang Dong, Bin He
yangjinfeng2010@gmail.com, guanyi@hit.edu.cn,{dongxishuang, goohebingle}@gmail.com
School of Computer Science and Technology of Harbin Institute of Technology
Harbin, Heilongjiang, China 150001

Abstract

Similarity between words is becoming a generic problem for many applications of computational linguistics, and computing word similarities is determined by word representations. Inspired by the analogies between words and lymphocytes, a lymphocyte-style word representation is proposed. The word representation is built on the basis of dependency syntax of sentences and represent word context as head properties and dependent properties of the word. Lymphocyte-style word representations are evaluated by computing the similarities between words, and experiments are conducted on the Penn Chinese Treebank 5.1. Experimental results indicate that the proposed word representations are effective.

Introduction

The goal of word similarity is to compute the similarity degree between words. The study of similarity between words has been a part of natural language processing and information retrieval for many years. Word representations are key importance of computing similarities between words. A word representation is a mathematical object associated with each word, often a vector. Three types of word representations are mainly focused, named distributional word representations (Sahlgren 2006), clustering-based word representations (Brown et al. 1992) and distributed word representations (Bengio et al. 2003). It was shown that these word representations can be used to significantly improve and simplify many NLP applications (Collobert et al. 2011).

An inherent limitation of these word representations is their indifference to word order and their inability to represent idiomatic phrases (Mikolov et al. 2013), which mainly attributes to the absence of syntax information in these representations. From the perspective of sentence structure analysis, a (dependent) word syntactically or semantically depends on another (head) word in a sentence. Each word in a sentence both resides in a dominant context and a dependent context. So word properties may need to be grouped into head properties and dependent properties.

In this research, we propose a lymphocyte style word representation method, which represents words as B cells. Dependency features extracted from the context of head-dependent relations are used as word properties and are grouped into dependent properties and head properties, corresponding to idiotops and paratopes of B cells' receptors. The proposed lymphocyte style word representations are measured by computing the similarities between words. Experimental results on the Penn Chinese Treebank 5.1(CTB) (Xue et al. 2005) show that lymphocyte style word representation is an effective word representation.

Methods

Lymphocyte-style Word Representations

B cells, an important class of lymphocytes in the immune system, have a Y-shaped structure. At the tips of the Y, there exist paratopes and idiotopes. Paratopes of a B cell can specifically recognize idiotopes of another B cell. In this research, word representations are designed according to the Y-shaped structure of B cell receptors. Figure 1 shows the simplified design of B cell receptors and the word representation with head properties and dependent properties.

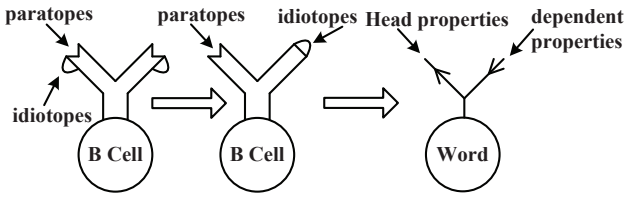

Figure 1: The design of B cell receptors and word properties

The proposed lymphocyte-style word representation is based on dependency syntactic information and is inspired by the analogies between words and lymphocytes. For learning of the representations, a multi-word-agent autonomous learning model (MWAALM) based on an artificial immune system is presented. The model is constructed by Cellular Automation, and words are modeled as B cell word agents (BWAs) and as antigen word agents (AgWAs). In this research, a Chinese dependency Treebank is used as a training set. Words in the training set are represented as B cells, and head properties and dependent properties of words are represented as paratopes and idiotopes on receptors of B

cells respectively. Dependency features extracted from head-dependent relations are used as properties of words. For a word w, $\{hf_1^w, hf_2^w, \cdots, hf_i^w, \cdots, hf_{N_{hf}}^w\}$ is the head feature set of w extracted from all head-dependent relations in which w is the head word and $\omega_{hf_i^w}$ is the weight of hf_i^w. Respectively, $\{df_1^w, df_2^w, \cdots, df_j^w, \cdots, df_{N_{df}}^w\}$ is the dependent feature set of w extracted from all head-dependent relations in which w is the dependent word and $\omega_{df_j^w}$ is the weight of df_j^w. The paratopes P^w and idiotopes I^w of a word w are formulated as Equation (1) and Equation (2). Just like affinities between B cells, strength of dependency relation between two words can be determined, please refer the section A2 of the Appendix[1] for detail. Inspirations from immune system and the learning of the representations is discussed in the section A1 and A2 of the Appendix.

$$P^w = \{(hf_1^w, \omega_{hf_1^w}), \cdots, (hf_{N_P}^w, \omega_{hf_{N_P}^w})\} \quad (1)$$

$$I^w = \{(df_1^w, \omega_{df_1^w}), \cdots, (df_{N_I}^w, \omega_{df_{N_I}^w})\} \quad (2)$$

Similarity Based on Word Representations

The Distributional Hypothesis states that words that occur in the similar contexts tend to have similar meanings (Harris 1954). We extend the Distributional Hypothesis: if two words both share dominant contexts and dependent contexts, the two words may be similar. Cosine similarity function $sim_{cosine}(x_1, x_2)$ is adopted to measure the similarity of both head properties and dependent properties. Based on the lymphocyte style word representation, word similarity can be computed according to equation (3). Distributional similarity between words can induce the paradigmatic relation between words (Sahlgren 2006). Paradigmatic relations are substitutable relations. So, the judgment about whether two words are similar is determined by the judgment about whether they share a paradigmatic relation.

$$similarity(w_1, w_2) = sim_{cosine}(P^{w_1}, P^{w_2}) \times sim_{cosine}(I^{w_1}, I^{w_2}) \quad (3)$$

Experimental Results

A dependency Treebank, built from the CTB, is employed as experimental data. For evaluation of the proposed word representations, words in the first 100 sentences of the CTB are considered. For each considered word, five words with most high similarities, according to equation (3), are chosen for evaluation. Two precision metrics are used to evaluate those mined similar words. The one is the precision of top one P_{Top1}, which means the percentage of those considered words whose top one candidate word is judged similar. The second is the precision of top five P_{Top5}, which means the percentage of those considered words for which one of the top five candidate words is judged similar. For the purpose of impartial evaluation, two persons evaluated the candidate similar words independently. Experimental results in detail can be found in the section A3 of the Appendix.

As shown in table 1, the evaluation results by two persons seem to be in high agreement. The results indicate that the proposed lymphocyte-style word representation can be successfully applied for word similarity computing and is proven to be an effective word representation.

Table 1: The evaluation results of similar words.

Evaluator 1		Evaluator 2	
P_{Top1}	P_{Top5}	P_{Top1}	P_{Top5}
0.6337	0.7864	0.6050	0.7840

Conclusions and Future Work

This research provides a completely new perspective on language and words. The important contribution is that a new lymphocyte-style word representation is presented. With this word representation, both similarities between words and strengths of dependency relations can be determined. Two aspects of future work will be mainly focused. One of the future works is to adapt the model for a large scale data set and make comparisons with other existing implementations of word representations, such as word2vec. Another future work is to try to apply the lymphocyte-style word representations to research tasks on NLP and to validate the performance of the word representations.

References

Bengio, Y.; Ducharme, R.; Vincent, P.; and C.Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* 3:1137–1155.

Brown, P.; DeSouza, P.; Mercer, R.; and et al. 1992. Class-based N-gram Models of Natural Language. *Comput. Linguist.* 18(4):467–479.

Collobert, R.; Weston, J.; Bottou, L.; and et al. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.* 12:2493–2537.

Harris, Z. 1954. Distributional structure. *Word* 10(23):146–162.

Mikolov, T.; Sutskever, I.; Chen, K.; and et al. 2013. Distributed representations of words and phrases and their compositionality. In Burges, C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc. 3111–3119.

Sahlgren, M. 2006. *The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Doctoral dissertation, Stockholm University.

Xue, N.; Xia, F.; Chiou, F.; and et al. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural Language Engineering* 11(2):207–238.

[1] https://github.com/yangjinfeng/wordrep/blob/master/aaai2014_appendix.pdf