QUANTIFYING SEMANTIC RELATEDNESS

Ing. Viliam Ročkai
supervisor: doc. Ing. Marián Mach, CSc.

Department of Cybernetics and Artificial Intelligence,
Technical University of Košice, Letná 9, 041 20 Košice,
Slovak Republic.
E-mail: viliam.rockai@gmail.com

ABSTRACT

Humans find it extremely easy to say if two words are related or if one word is more related to a given word than another one. For example, if we come across two words - 'car' and 'bicycle', we know they are related since both are means of transport. Also, we easily observe that 'bicycle' is more related to 'car' than 'fork' is. In the paper we describe our approach on quantifying the semantic relatedness of concepts based on the theory of associative learning of concepts.

1 INTRODUCTION

The theory [1] is based on the neuro-physiological model of thalamic-cortical information processing by R.Hecht–Nielsen [2]. In [2] the model assumes the existence of a fixed lexicon of ”symbols” in the human brain thalamus, which is created during an early period of the growth of the individual. Learning occurs by establishing associations between the symbols mimicking neural connections between cortical regions. The dynamics of learning in the network is modelled by Hebbian learning method. The associative learning of concepts is an unsupervised learning processed over the stream of symbols. The goal of the process is to learn a knowledge representation structure resembling semantic networks. The method is based on the induction of associations between symbols observing their co-occurrences in a context window which the learning is processed through. The condition necessary to create an association between a particular token pair is given by their statistically ”non-random” occurrences. After it has been determined that an association exists and therefore has to be created, only its weight is altered during the learning process [1].

2 THEORY OF ASSOCIATIVE LEARNING OF CONCEPTS

2.1 Representation

In [1] the *token* is given by a set of excited neurons in some cortical region. This *token* represents an invariant sensory input and can be seen as an *attribute* of the perceived object. In the theory *Symbols* are defined as the representation of these *attributes*. A *concept* is defined as a symbol with it‘s associated neighbourhood. Thus, the same concept can be activated by different sets of symbols. Associations anchor the symbols to each other. The establishment of an association between two symbols depends on whether the co-occurrences are significant or not. The significance can be determined by utilising the following formula:

$$S(i,j) = \frac{p(i,j)}{p(i).p(j)}$$

where *i* and *j* are symbol occurrences. *S(i,j)* stands for mutual significance and is based on the information theoretic formula of the mutual information. The mutual significance is defined as the ratio of mutual probability of symbols *i, j* to their prior probabilities. If i and j were independent then:

$$p(i, j) = p(i).p(j)$$

So if two symbols coexist in joint context non-randomly, it will be true that:

$$S(i,j) = \frac{p(i,j)}{p(i).p(j)} >> 1$$

If S(i,j) > 1, the tokens will be considered as associated tokens. This will be the main parameter during the learning process. We call this parameter the threshold and it can be altered for filtering more non-random occurrences. The weight of an association of two symbols x and y is given by the formula:

$$w(x,y) = \frac{p(i,j)}{p(j)}$$

The weight of an association is the only knowledge we learn in the process. These weights are represented in *fascicles* – our knowledge bases ***Fx***= ***w(i,j)***, symmetrical matrices. We learn several of these ***Fx*** knowledge bases each standing for different context distances (features) of the two symbols learned [1]. The calculation of weights w(i,j) is given by an approximation of the above given formulas:

$$S(i,j) = \frac{\frac{C(i,j)}{T}}{\frac{p(j)}{T}} \qquad w(i,j) = \frac{\frac{C(i,j)}{T}}{\frac{C(i)}{T}.\frac{C(j)}{T}}$$

Where C(i,j) is joint occurrence of tokens i and j and T is the total token count incremented during the learning process.

2.1.1 Semantic relatedness

We observe that humans find it extremely easy to say if two words are related, For example if we come across two words -- 'car' and 'bicycle', we know they

are related as both are means of transport. Also, we easily observe that one word is more related to a given word than another word - 'bicycle' is more related to 'car' than 'fork' is. But is there some way to assign a quantitative value to this relatedness? By utilising the knowledge acquired by learning the fascicles, we constructed the following measure:

$$R = \frac{|O(x) \bigcap O(y)|}{|O(x) \bigcup O(y)|}$$

where *O(x)* stands for the semantic neighbourhood of the symbol *x*. *O(x)* is such a set of symbols, where for every *l* from the lexicon *L* exists an association between *l* and the symbol *x*. Lexicon *L* is the set of all known tokens. The formula is the proportion of cardinalities of the sets of the commonly shared properties between symbols *x* and *y* to the set of all the properties of *x* and *y*. Since several fascicles can contribute with different features of symbols, we need to create an integral measure by a combination of all the contributing features:

$$R_{int} = \sum v_i \cdot \frac{|O_i(x) \bigcap O_i(y)|}{|O_i(x) \bigcup O_i(y)|}$$

R_{int} then represents weighted value of the relatedness between symbol x and y. v_i stands for the weight assigned to the relevant i-th fascicle.

3 EXPERIMENTS

To test our method we have generated 10 000 sentences using an artificial grammar able to form some fairly complex English sentences. An example of our generated training set can be seen here:

...
John walks.
Mangy dogs bite.
A mangy cat walks.
The nasty dog bites the girls who walk.
The boy who eats walks.
The dogs who Mary walks bark.
...

We carried out learning based on our training set and performed clustering on the learned knowledge bases for all symbols existing in the Lexicon (terminals of our grammar). The resulting dendrogram is depicted on the figure one.

From the figure one can see that the algorithm clustered tokens to morphological language classes with nearly 99% success rate. The only difference is ”a” and ”the” tokens, which are very frequent in the text and thus they have many associations. This suggests that our formula for computing mutual significance should be adjusted for very frequent symbols.

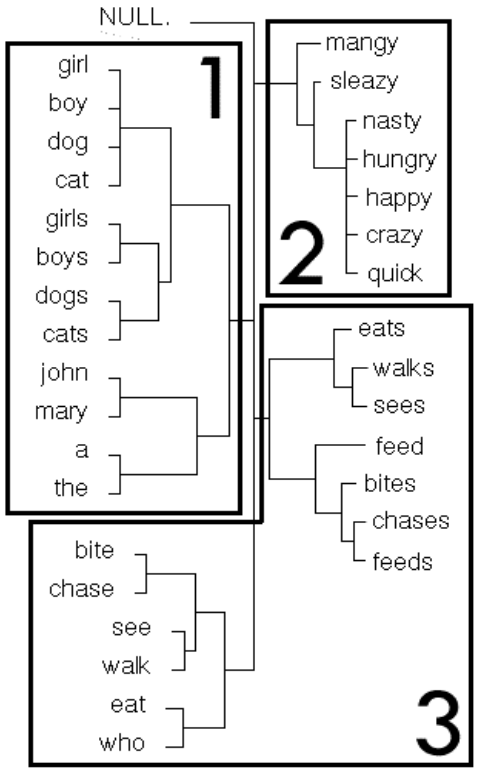

Figure 1. Dendogram

4 CONCLUSION

Our method achieved very good results in clustering tokens into morphological subsets upon artificial grammar. Thus, the next goal would be to realise experiments and further develop our method for the requirements of natural languages.

Acknowledgement

The work presented in the paper was supported by the Slovak Grant Agency of Ministry of Education and Academy of Science of the Slovak Republic within the 1/1060/04 project ”Document classification and annotation for the Semantic web”.

REFERENCES

[1] ROČKAI, V..: Mining of Concepts and Semantic Relations from Texts in Natural Language. Košice: FEI TU, 2005. Diploma thesis.

[2] R. Hecht-Nielsen & T. McKenna (Eds) : Computational Models for Neuroscience: human cortical information processing, Springer-Verlag, 2003.