17 Using Self-Organizing Maps for Natural Language Processing

Timo Honkela, Ville Pulkki, and Teuvo Kohonen

Development of large-scale natural language processing applications is restricted by quantitative and qualitative limitations. Quantitatively, a system for even a moderately narrow domain requires a substantial knowledge base. One suggested solution for this problem has been an approach where vast common repositories of knowledge items (frames, facts, rules) have been collected. Qualitatively problematic areas remain, e.g., graded phenomena, inherent ambiguity of natural language, and subjectivity and variation in natural language generation and interpretation. Gradual changes in the domain of the application make non-adaptive systems vulnerable.

The predominant approach among computerized models of language is based on predetermining and coding the linguistic categories and rules "by hand". The methodological basis is symbol manipulation. However, the fact that the expressions in natural language appear to be inherently symbolic and discrete does not imply that symbolic descriptions of linguistic phenomena are sufficient. This is expecially remarkable when semantic and pragmatic issues are considered, i.e., the ability of a system to interpret natural language expressions. To be able to model the gradually changing relation between continuous phenomena and discrete symbols, the building blocks of the theory must be sufficiently powerful.

The Self-Organizing Map [1] (SOM) algorithm can be used to automatically create implicit emergent categories from uncategorised linguistic input [2]. Our experiments have shown have unrestricted textual input can be analyzed by the SOM [3]. As the result a word category map is created. In the following, some of the details of the basic experiment are described.

The encoding of the input words was made using a 90-dimensional random real vector for each word. The codes were statistically independent so that there was no correlation between them. The code vectors of the words in the triplet, i.e., three subsequent words in the text, were then concatenated into a single input vector $\mathbf{x}(t)$, the dimensionality of which was thus 270. The 270-dimensional input vectors $\mathbf{x}(t)$ were used as inputs to the SOM algorithm. The SOM array itself was a planar, hexagonal lattice of 42 by 36 formal neurons. Our aim in this analysis was to study in what context the “keys” (middle parts in the triplets) occur. The mapping of the $\mathbf{x}(t)$ vectors to the SOM was determined by the whole vector $\mathbf{x}(t)$, but after learning the map units were labeled according to the middle parts of the $\mathbf{m}_i(t)$. In other words, when the “key” parts of the different $\mathbf{m}_i(t)$ were compared with a particular word in the list of the selected 150 words (the most frequent ones), the map unit that gave the best match in this comparison was labeled by the said word. It may then also be conceivable that in such a study one should also use only such inputs $\mathbf{x}(t)$ for training that have one of the 150 selected words as the “key” part.

In order to equalize the mapping for the selected 150 words statistically and to speed up computation, a solution used in [2] was to average the contexts relating to a particular “key”. In other words, if the input vector is expressed formally as $\mathbf{x} = [\mathbf{x}_i^T, \mathbf{x}_j^T, \mathbf{x}_k^T]^T$ where T signifies the transpose of a vector, and $\mathbf{x}_j$

is the "key" part, then the true inputs in the "accelerated" learning process were $[E\{\mathbf{x}_i^T|\mathbf{x}_j\}, 0.2\mathbf{x}_j^T, E\{\mathbf{x}_k^T|\mathbf{x}_j\}]^T$, where E now denotes the (computed) conditional average. (The factor 0.2 in front of $\mathbf{x}_j^T$ was used to balance the parts in the input vectors.) In this way there would only be 150 different input vectors that have to be recycled a sufficient number of times in the learning process. The information about all the 7624 words is anyway contained in the conditional averages. Although the above method already works reasonably well, a modification of "averaging" based on auxiliary SOMs was used. For each codebook vector a small, 2 by 2 SOM was assigned. It was trained with the input vectors made from the due word triplets. After training, each codebook vector in one small map described more specifically what context was used on the average with that "key" word.

The results of the computation are presented in Figure 27. The positions of the words on the map are solely based on the analysis of the contexts performed by the SOM. The general organization of the map reflects both syntactical and semantical categories. The most distinct large areas consist of verbs in the top third of the map, and nouns in the bottom right corner.

Word category maps can be used in practical large-scale natural language processing applications, like in intelligent information retrieval. This particular application area has been described in detail in the WEBSOM section of this report.

References

[1] Teuvo Kohonen. *Self-Organizing Maps.* Springer Series in Information Sciences, vol. 30. Springer, Heidelberg, 1995.

[2] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biol. Cyb.*, 61(4):241–254, 1989.

[3] T. Honkela, V. Pulkki, and T. Kohonen. Contextual relations of words in Grimm tales analyzed by self-organizing map. In F. Fogelman-Soulié and P. Gallinari, eds., *ICANN-95, Proc. of Int. Conf. on Artificial Neural Networks, Vol. 2*, pp. 3–7. EC2 et Cie, Paris, 1995.

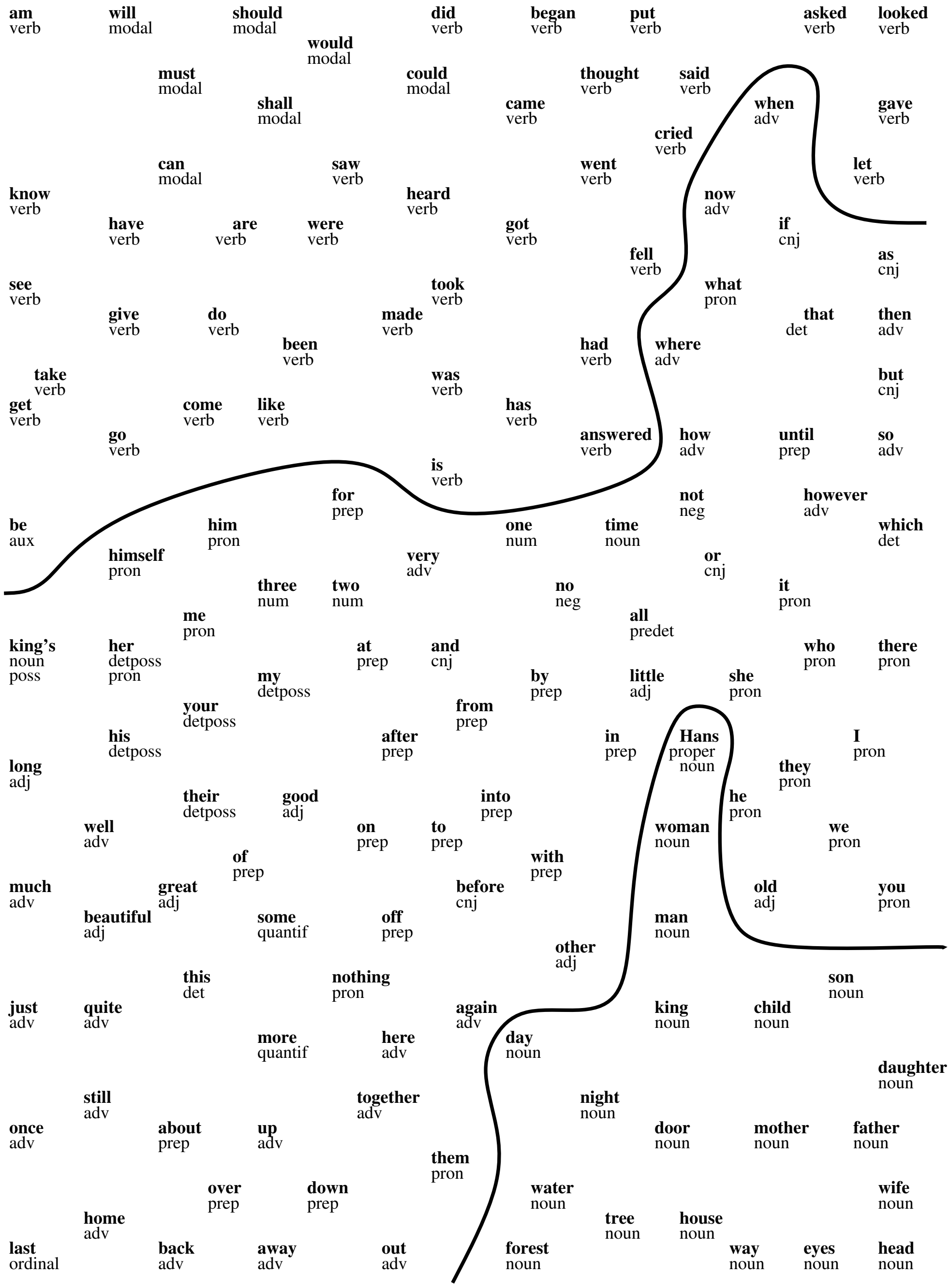

Figure 27: The 150 most frequent words of the Grimm tales, their statistical contextual relations being represented two-dimensionally by the SOM. The words are shown in their due position in the array; no symbols for "neurons" have been drawn. Many words are ambiguous but usually only the most common category relating to the tales is presented. All verbs can be found in the top section whereas the nouns are located in the lower right corner of the map. In the middle there are words of multiple categories: adverbs, pronouns, prepositions, conjunctions, etc. Modal verbs form a collection of their own among the verbs. Connected to the area of nouns are the pronouns. The three numerals in the material form a cluster. Among the verbs the past-tense forms are separated from the present-tense forms and located in the top right corner. Among the nouns, the inanimate and animate nouns forms separate areas of their own.