--- Page 1 ---

Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing

Article

December 1998

Source: CiteSeer

Citations and Reads

- **Citations:** 115

- **Reads:** 966

Authors

- **Timo Honkela**

  - University of Helsinki

  - 154 Publications

  - 3,108 Citations

  - [Profile](#)

- **Ville Pulkki**

  - Aalto University

  - 307 Publications

  - 6,407 Citations

Appeared in: Proceedings of the International Conference on Artificial Neural Networks, ICANN-95, F. Fogelman-Soulie and P. Gallinari (eds.), EC2 et Cie; Paris; 1995, pp. 3-7.

Contextual Relations of Words in Grimm Tales: Analyzed by Self-Organizing Map

Timo Honkela, Ville Pulkki, and Teuvo Kohonen

Helsinki University of Technology

Neural Networks Research Centre

Rakentajanaukio 2 C, FIN-02150, FINLAND

Tel: +358 0 451 3276, Fax: +358 0 451 3277

Email: Timo.Honkela@hut.fi

Abstract

Semantic roles of words in natural languages are reflected by the contexts in which they occur. These roles can be explicitly visualized by the Self-Organizing Map (SOM). In the experiments reported in this work, the source data consisted of the raw text of Grimm fairy tales without any prior syntactic or semantic categorization of the words. The algorithm was able to create diagrams that seem to comply reasonably well with traditional syntactical categorizations and human intuition about the semantics of the words.

---

1. Processing Natural Language with Self-Organizing Maps

It has earlier been shown that the Self-Organizing Map (SOM) can be applied to the visualization of contextual roles of words, i.e., similarities in their usage in short contexts formed of adjacent words [4]. This paper demonstrates that such relations or roles are also statistically reflected in unrestricted, even quaint natural expressions. The source material chosen for this experiment consisted of 200 Grimm tales (English translation).

In most practical applications of the SOM, the input to the map algorithm is derived from some measurements, usually after pre-processing. In such cases, the input vectors are supposed to have metric relations. Interpretation of languages, on the contrary, must be based on the processing of sequences of discrete symbols. If the words were encoded numerically, the ordered sets formed of them could also be compared mutually as well as with reference expressions. However, as no numerical value of the code should imply any order to the words themselves, it will be necessary to use uncorrelated vectors for encoding. The simplest method to introduce uncorrelated codes is to assign a unit vector for each word. When all different words in the input material are listed, a code vector can be defined to have as many components as there are words in the list. This method, however, is only practicable in very small experiments. If the vocabulary is large, as in the present experiments, we may then encode the words by quasi-orthogonal random vectors of a much smaller dimensionality [4].

To create a map of discrete symbols that occur within the sentences, each symbol must be presented in the due context. The context may consist of the immediate surroundings of the word in the text. Application of the self-organizing maps to natural language processing has been described earlier, e.g., [2], [3], [4], [5], and [6].

--- Page 3 ---

Experiments

2.1 Source Data

In the present experiments, the data consisted of a set of English translations of tales collected by the Grimm brothers. The number of words in the text was almost 250,000 in total, and the vocabulary size was over 7,000 words. Although the subject area of the tales is rather restricted, the language can be considered to be arbitrarily chosen and unrestricted. First, the language itself is not formal by any means. This choice of source data represents a significant generalization compared with the artificially produced sentences previously used in simpler experiments (e.g., in [4]). Second, the contents of the texts are very diverse.

2.2 Preprocessing

The texts of all tales were concatenated into one file. Punctuation marks were removed, and all uppercase letters were replaced by corresponding lowercase letters. Articles ("a," "an," "the") were also removed. Some tests were conducted with the articles included, separating nouns and personal pronouns to a greater extent.

Word triplets ("predecessor," "key," "successor") picked from the text file were chosen for the input vector \(x(t) \). The triplets were formed by taking the encoded representations of three subsequent words from the preprocessed text. All word triplets from the text were collected and stored as a source file.

The 150 most frequent words in the text file were chosen as the "key" words to be represented in the map. It's important to note that no words were ignored from the original text. Each word in the vocabulary was coded, and the predecessor and successor could be any word occurring in the text. Encoding was done using a 90-dimensional random real vector for each word. The codes were statistically independent with no correlation between them. The code vectors of the words in the triplet were then concatenated into a single input vector $x(t)$, resulting in a dimensionality of 270.

2.3 Learning Process

The 270-dimensional input vectors $x(t)$ were used as inputs to the SOM algorithm. The SOM array itself was a planar, hexagonal lattice of 42 by 36 formal neurons. To speed up computations, its codebook vectors were given ordered initial values, chosen from the signal space as follows:

First, a two-dimensional subspace, spanned by the two principal eigenvectors of the input data vectors, was defined. A hexagonal array corresponding to the size of the SOM array was then defined in the subspace, its centroid coinciding with the mean of $x(t)$, and the main dimensions of the array equaled the two largest eigenvalues of the covariance matrix of $x(t)$. The initial values of $m_i(0)$ were taken from these array points [1].

Our aim in this analysis was to study the context in which the "keys" (middle parts in the triplets) occur. The mapping of the $x(t)$ vectors to the SOM was determined by the whole vector $x(t)$, but after learning, we labeled the...
map units according to the middle parts of the mi(t). In other words, when we compared the "key" parts of the different mi(t) 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 that word. It may then be conceivable that in such a study we should also use only those inputs x(t) for training that have one of the 150 selected words as the "key" part.

In order to equalize the mapping for these 150 selected words statistically and to speed up computation, a solution used in [4] was to average the contexts relating to a particular "key."

In other words, if the input vector is expressed formally as \([x_T, y]\), where \(T\) signifies the transpose of a vector and \(x\) is the "key" part, then the true inputs in the "accelerated" learning process were \([E\{x^*|z_i\}, 0.2, E\{y|z_i\}]^T\), where \(E\) denotes the (computed) conditional average. The factor 0.2 in front of \(y\) was used to balance the parts in the input vectors. In this way, we only needed 150 different input vectors to be recycled a sufficient number of times in the learning process. The information about all the 7624 words is contained in the conditional averages.

Although the above method already works reasonably well, a modification of "averaging" based on auxiliary SOMs was used in this work. For each codebook vector, we assigned a small 2 by 2 SOM that was trained with input vectors made from the word triplets. After training, each codebook vector in one small map described more specifically what context was commonly used with that "key" word.

The computation of the map was conducted in two separate runs. The map was first pre-taught using CNAPS, a massively parallel neurocomputer with fixed-point arithmetic. Pre-teaching consisted of 600,000 learning cycles, during which we could use a large radius of neighborhood connections. After that, teaching continued at higher accuracy on a workstation with floating-point arithmetic. During this run, 400,000 learning cycles were used.

Results

The results of the computation are presented in Figure 1. The positions of the words on the map are solely based on the analysis of the contexts performed by the SOM. Explicit lexical categorization of the words was made following the Collins Cobuild Dictionary [7] to help the reader evaluate the results. 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.

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, including 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. The lexeme "one" is separated from "two" and "three" to some extent, which can be explained by its multiple meanings. Among the verbs, the past-tense forms are separated from the present-tense forms and located in the right corner. A distinct group of possessives can also be found: even the possessive form "king's" is among them, having rather similar contexts. The plain noun "king" is situated within the animate nouns.

The formation of syntactic categories on the map can be explained by sentential context. The context of a word is, quite naturally, dependent on the syntactical restrictions that govern word positions in the text.

The text discusses categories relating to the context of words, emphasizing that the presence of various categories is statistical. Within large, syntactically based groups, semantic relationships can be identified.

Consider the set of nouns as an example. Animate and inanimate nouns form their own groups. Sets of closely related words can be found, such as:

- "father-mother"
- "night-day"
- "child-son"
- "forest-tree"
- "head-eyes"
- "woman-man"

Some anomalies are apparent on the map, at least at first sight. The few adjectives are somewhat scattered, and the word "little" has an almost singular location. This may be due to specific uses of the word in phrases like "little by little" or "she walked a little farther."

--- Page 5 ---

Conclusion

The experimentally found linguistic categories, determined implicitly by the organizing map, seem to correlate with the categories of concepts occurring in actual cognitive processes. It may also be argued that, in human language learning, the naming of explicit syntactic relations is not necessary. Expressions heard in the proper context may be sufficient for creating a working model of language.

Contrary to this, the trend in theoretical linguistics has been to construct explicit symbolic models. Comparison of symbolic structures is easier for human linguists. However, the implicitness of "neural" models may be advantageous in practical applications where sharp-bordered symbolic models for semantic processing are too "rough" to meet the requirements of semantic mapping, especially when unrestricted natural language is considered.

Linguistic categories can thus be viewed as approximations (generalizations). This is reflected in the emergence of categories in a self-organizing process. The present study analyzed lexical items of a finite corpus based on a simple, statistically defined average context. Another approach would be to use complete expressions as input, which would allow for a more detailed analysis of phenomena like polysemy and impreciseness.

Contextual maps can be utilized as a central component in applications like information retrieval and machine translation. The present study concentrates on the analysis of a particular corpus. The method is generally applicable to the processing of any textual material, even a combination of text and other modalities.

References

1. Teuvo Kohonen: *Self-Organizing Maps,* Springer, 1995.

2. Risto Miikkulainen: *DISCERN: A Distributed Artificial Neural Network Model of Script Processing and Memory,* PhD thesis, Computer Science Department, University of California, Los Angeles, 1990. (Tech: UCLA-AL-90-05).

3. Risto Miikkulainen: *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory,* MIT Press, Cambridge, MA, 1993.

4. Helge Ritter and Teuvo Kohonen: "Self-organizing semantic maps," *Biological Cybernetics,* 61(4), 241-254, 1989.

5. J. C. Scholtes: "Kohonen feature maps in natural language processing," Technical report, Department of Computational Linguistics, University of Amsterdam, March 1991.

6. J. C. Scholtes: *Neural Networks in Natural Language Processing and Information Retrieval,* PhD thesis, Universiteit van Amsterdam, Amsterdam, the Netherlands, 1993.

7. John Sinclair, editor. *Collins Cobuild English Language Dictionary,* Collins, London and Glasgow, 1990.

--- Page 6 ---

Certainly! Here's the cleaned-up version of the text with improved structure and readability:

Verbs

- am
- will
- should
- did
- began
- put
- asked
- looked
- would
- must
- could
- thought
- said
- shall
- came
- gave
- cried
- can
- saw
- went
- let

- know
- heard
- have
- are
- were
- got
- fell
- see
- took
- give
- do
- made
- been
- had
- take
- was
- get
- come
- like
- has
- answered

Adverbs

- when
- now
- where
- then

- how
- so
- very
- quite
- just
- still
- once
- about
- together
- away
- often
- well
- up
- down
- back
- here
- over
- out
- last
- more
- much
- before
- again

Conjunctions

- if
- as

- but

- however

- or

- and

Nouns

- king's

- woman

- man

- son

- child

- daughter

- wife

- house

- tree

- night

- forest

- way

- eyes

- head

- door

- water

- home

- mother

- father

- Hans

Pronouns

- what
- whom
- who
- there
- it
- me
- all
- her
- they
- their
- he
- she
- you
- them

Prepositions

- until
- for
- to
- at
- by
- from
- in
- into
- on
- of

- with

- after

Adjectives

- little

- great

- old

- beautiful

- good

- long

- other

- some

- much

- many

- first

Quantifiers

- one

- two

- three

Negative & Negation

- not

- no

- nothing

Possessive

- his

- your

- their

- my

Miscellaneous

- time

- so

- very

- well

- this

- very

Note: The 150 most frequent words from the Grimm tales are organized to represent their statistical contextual relations. The words hold specific linguistic categories relevant to the tales.