EMERGENCE OF LINGUISTIC FEATURES: INDEPENDENT COMPONENT ANALYSIS OF CONTEXTS

TIMO HONKELA[A] AAPO HYVÄRINEN[B] JAAKKO VÄYRYNEN[A]

[A] *Neural Networks Research Centre, Helsinki University of Technology*
P.O.Box 5400, FI-02015 TKK, Finland
timo.honkela@hut.fi

[B] *HIIT Basic Research Unit, Department of Computer Science, University of Helsinki*
P.O.Box 68, FI-00014 University of Helsinki, Finland
aapo.hyvarinen@helsinki.fi

We study how independent component analysis can be used to create automatically syntactic and semantic features based on analyzing words in contexts.

1. Introduction

We show that independent component analysis (ICA) (Hyvärinen et al. 2001) applied on word context data gives distinct features that reflect syntactic and semantic categories. The analysis gives features or categories that are both explicit and can easily be interpreted by humans. This result can be obtained without any human supervision or tagged corpora that would have some predetermined morphological, syntactic or semantic information. The results include both an emergence of clear distinctive categories or features and a distributed representation. This is based on the fact that a word may belong to several categories simultaneously in a graded manner. We wish that our model provides additional understanding on potential cognitive mechanisms in natural language learning and understanding. Our approach attempts to show that it is possible that much of the linguistic knowledge is emergent in nature and based on specific learning mechanisms.

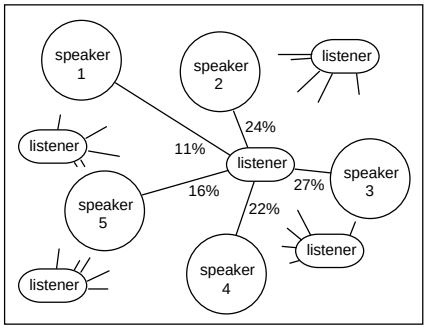
2. Independent Component Analysis

In the following, we consider the use of independent component analysis (ICA) (Comon 1994, Jutten & Hérault 1991, Hyvärinen et al. 2001) in the extraction of linguistic features for expression in due contexts. ICA learns features in an unsupervised manner. Several such features can be present in a word, and ICA gives the explicit values of each feature for each word. We expect the features

to coincide with known syntactic and semantic categories: for instance, we expect ICA to be able to find a feature that is shared by words such as “must”, “can” and “may”. In earlier studies, independent component analysis has been used for document level analysis of texts (see, e.g., Bingham et al. 2001, Bingham et al. 2002).

We first give a brief outline of the basic theory of independent component analysis (Hyvärinen et al. 2001). In the classic version of the ICA model, each observed random variable is represented as a weighted sum of independent random variables. The weights in the sum (which can be negative as well as positive) can be collected in a matrix, called the mixing matrix. The weights are assumed to be different for each observed variable, so that the mixing matrix can be inverted, and the values of the independent components can be computed as some linear functions of the observed variables.

A classical application for the ICA, blind signal separation (BSS) can be used as an example of the basic idea. In BSS, one studies signals that originate from a number of separate sources, e.g., discussants in a cocktail party who are speaking at the same time[a]. The signals are mixed in different proportions depending on the relative distance of a listener to each sound source. This kind of situation is illustrated in Figure 1 in which five listeners have the task of separating five signal sources from each other.



[a] The cocktail party should be considered merely as a simple illustration of blind source separation. A real cocktail-party problem is much more complicated than described here due to such factors as the slow propagation of sound in the air, echoes, nonlinearities in the microphones, and noise.

Figure 1. An illustration of a blind signal separation situation and the idea of a mixing matrix. Each listener receives a unique mixture of signals as a combination of five source signals.

Sources numbered from 1 to 5 refer to the original signals and each listener receives a unique combination of the signals. Based on the distance to each source, the vector that represents the mix of the five sources for one indicated listener in Figure 1 could be [0.11 0.24 0.27 0.22 0.16]. Putting together similar vectors for each listener we would gain a mixing matrix that, in general, would be unknown in a BSS task.

The goal in ICA is to learn the decomposition in an unsupervised manner, which means that we only observe the mixed signals and have no information about the mixing coefficients or the contents of the original signals. For the cocktail party problem, this would mean that ICA would provide an estimate of each original signal as well as of the mixing matrix. It may sound surprising that one could make an estimation of both the original signals and the mixing matrix provided only the observed signals. This challenge could be compared with a situation in which one is asked the sum of which two integers 63 is. The answer could, for instance, be 12+51 or 38+25, or, if negative values are included, there would be an infinite number of options.

The “trick” needed for the successful estimation of the original signals and the mixing matrix is based on is the simple assumption that the original signals are statistically independent. Two variables are independent if information on the value of one variable does not give any information on the value of the other. This does not need to hold for the observed variables. The independence assumption gives ground for finding the estimates the same way as if in summing up two integers to get 63 we would know that one of the numbers would need to be as closely as possible two times larger than the other. Then one could conclude that the sum would be 21+42. Two variables can be independent, for example, if they could be outcomes of two events that have nothing to do with each other, or random signals originating from two different physical processes. In case of two variables, the independence holds if and only if the joint probability of the variables is the same as the product of the two variables considered separately. This definition extends to any number of random variables. Thorough description of the ICA methodology and underlying mathematical and statistical principles is given in (Hyvärinen et al. 2001)

3. Analysis of Word Contexts

We can ask if a mixing structure introduced in the previous section is discernable in a higher level of abstraction than in the process of finding the original speech signals. For the BSS example presented above, the task was to find out what are the original signals and in which proportion each original signal is present in the perceived signals. In a more general level, ICA analysis has been used to find separate underlying components or variables. For linguistic data, the task could be to see whether different underlying categories could be found in an unsupervised manner. For instance, one can study if it is possible to find syntactic and semantic features based on an analysis of words in their contexts. The idea is that maybe words in contexts can be divided into a number of roles that they serve.

Contextual information has widely been used in statistical analysis of natural language corpora (consider, e.g., Church & Hanks 1990, Schütze 1992, Lund et al. 1995, Manning & Schütze 1999). One useful numerical representation for written corpora can be obtained by taking into account the sentential context in which the words occur. First, we represent each word by a vector in an n-dimensional space, and then code each context as an average of the vectors representing the words in that context. In the simplest case, the dimension can be taken equal to the number of different words, and each word is represented by a vector with one element equal to one and others equal to zero. Then the context vector simply gives the frequency of each word in the context. For computational reasons, however, the dimension may be reduced by different methods.

The ICA model in its basic form assumes that the components are independent, which may seem to be a very stringent assumption. However, this assumption need not be taken too seriously, because even if the components are not exactly independent, ICA algorithms can be interpreted as finding the most independent components. Furthermore, it can be shown that ICA estimation can often be interpreted as maximization of the sparseness of the components, i.e., the components should have only a few entries that are significantly different from zero. Sparseness is a very intuitive and useful property that has a long history in sensory coding, see, e.g., (Field 1994), so it seems likely that it it would be useful in this context as well.

For the ICA analyses on the word contexts we applied FastICA[a] software package for Matlab. We formed a context matrix in which the elements consisted of the number of occurrences of one word in the immediate context of another word, i.e, one word followed by another word with no words between them. We used this kind of short context but we are aware that one can use contexts of different sizes (consider, e.g., Redington et al. 1998, McDonald 2000).

We fed the word-context matrix to the FastICA algorithm so that each column was considered one data point, and each row one random variable. The dimension of the data was reduced to 10 by principal component analysis (PCA). The motivation for this operation can be considered, e.g., from the point of view that the ICA method can be divided into three consecutive steps: PCA, normalization of variances and ICA-rotation (Hyvärinen et al. 2001). The dimensionality reduction using PCA was motivated by computational efficiency.

The data used in the experiments consists of collection of e-mails sent to the connectionists mailing list. More details of the approach have been given in (Honkela et al. 2003, Honkela and Hyvärinen 2004).

The results of the ICA analysis corresponded in most cases very well or at least reasonably well with our preliminary intuitions. The system was able to automatically create distributed representations as a meaningful collection of emergent linguistic features; each independent component was one such feature.

In Figure 2, there are three examples of the analysis results[a]. In considering the feature distributions, it is good to keep in mind that the sign of the features is arbitrary. This is because of the ambiguity of the sign in the ICA model: one could multiply a component by -1 without affecting the model. Figure 2 shows how the third component is strong for nouns. For the words “systems” and “psychology” an additional component is strong. Namely, in more thorough analyses it became apparent that plural nouns shared the same pattern of peaked

[a] http://www.cis.hut.fi/projects/ica/fastica/

[a] More results can be found in (Honkela et al. 2003) and (Honkela & Hyvärinen 2004). In this paper we focus on the conceptual and modeling aspects of the ICA-based context analysis methodology.

third and fifth component. On the other hand, for nouns that refer to disciplines it became apparent that both third and fourth component is strong.

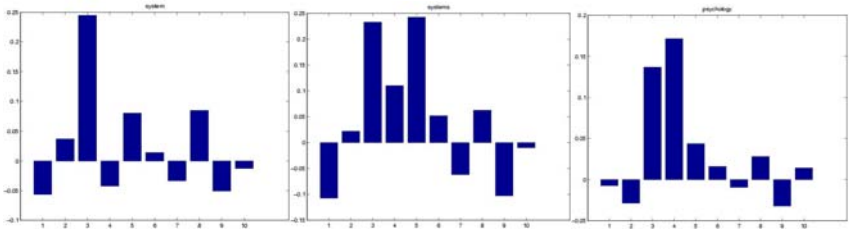


Figure 2. Ten independent components for three words (starting from the left): “system”, “systems” and “psychology”.

In this analysis we used ten as the number of ICA features which sets a limit on the complexity of the feature encoding. We used this limit in order to demonstrate the power and usefulness of the method in a simple manner. A higher number of features can be used in order to obtain more detailed feature distinctions.

In order to verify the match between ICA-based features and traditional linguistic categories systematically, we have conducted analyses in which the emergent features are compared with the categories provided in tagged corpora. In Figure 3, we show the results of one such analysis for nouns in plural form. Among the ICA-based features the one that has the closest match with the traditional category is first selected. In the study, 1000 most common words for which the traditional linguistic category was known were considered. Each of the points in the diagram corresponds to a word. The position of a point in the y-axis indicates the weight of a particular word in the mixing matrix for the particular feature. Among the words we have circled the ones that are known to belong to the category of plural nouns. Specifically, the comparison has been made using Brown corpus[a] and the corresponding tag is called NNS. The number of those words is 56 among the 1000 most common words. The successful match between the ICA-based feature and the traditional category is rather apparent because the circled points are located in the upper part of the diagram. The majority of the other words have a weight that is close to zero. For illustration, one can list the first 20 words that have the highest weight beginning from the best matching word: “parts”, “persons”, “places”,

[a] http://helmer.aksis.uib.no/icame/brown/bcm.html

"conditions", "circumstances", "animals", "times", "forms", "matters", "things", "words", "lines", "houses", "letters", "names", "days", "states", "officers", "books" and "works".

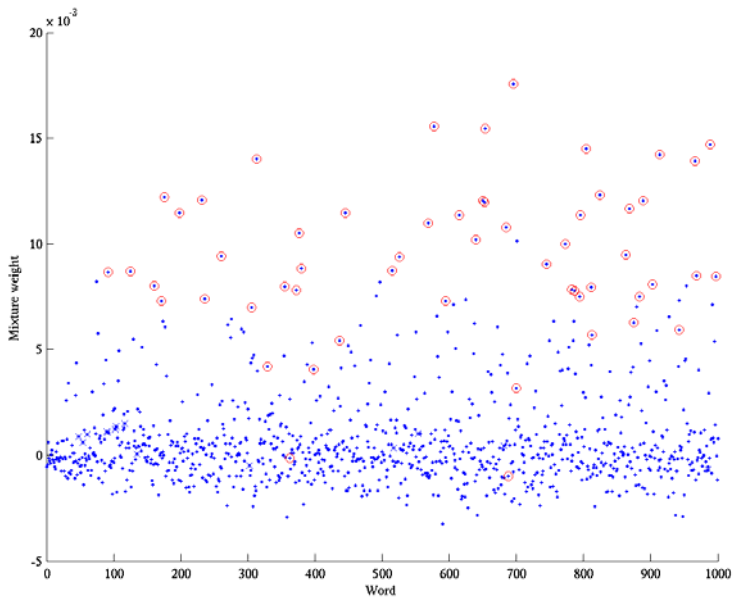


Figure 3. A comparison between an ICA-based emergent feature and a traditional linguistic category in a tagged corpus. Circled words belong to the plural noun (NNS) category. We can see that the words belonging to this particular category are clearly different from most of the other words when the mixture weight (y-axis) is considered.

4. Emergent Representations

In the following, we discuss the potential of independent component analysis for modeling some aspects of language learning and in adaptive conceptual modeling. We compare the merits of the ICA method with some earlier results gained by using the self-organizing maps (SOM). Earlier, the self-organizing map has been used in the analysis of word context data by, e.g., Ritter and Kohonen (1989) (artificially generated short sentences), and Honkela et al. (1995) (Grimm fairy tales). The result can be called a self-organizing map of words, or a word category map. Areas or local regions on a word category map can be considered as implicit categories or classes that have emerged during the learning process. The borderlines for the emergent implicit SOM-based

categories have to be determined separately. In the ICA analysis one is able to find the categories in an automated way.

In the SOM-based word category map each word appears in one location. This means, among other things, that one cannot have a map in which several characteristics or categories of one word would be represented unless the categories overlap and accordingly the corresponding areas of the map overlap. On the other hand, the ICA analysis provides a sparse encoding of the words in such a way that there can be a collection of features associated with each word. ICA can provide an arbitrary number of useful features per each word. The features can be syntactic as well as semantic like proposed already by Fillmore (1968). One advantage of using ICA is that the features emerge automatically.

One should also remember that the word context analysis based on written texts is only partially able to reveal the underlying syntactic and semantic feature structures. For instance, it is commonplace that opposites appear in similar contexts in texts. For more realistic language learning simulations it is necessary to include other kinds of contexts including visual perceptions, actions and activities associated with linguistic expressions.

5. Conclusions

We have shown how independent component analysis can find explicit features that characterize words in an intuitively appealing manner. There are various methods such as self-organizing maps (Kohonen 2001, Ritter & Kohonen 1989) and latent semantic analysis (Deerwester et al. 1990) that can be used for automatic statistical methods for linguistic analysis. However, independent component analysis appears to make possible a qualitatively new kind of results that have earlier been obtainable only through hand-made analysis. Moreover, the results indicate that autonomous agents would be able to learn linguistic feature systems in an unsupervised manner.

The analysis results show how the ICA analysis was able to reveal underlying linguistic features based solely on the contextual information. This means that no dictionary or thesaurus was used to guide the ICA analysis and the samples were not labeled. The results include both an emergence of clear distinctive categories or features as well as a distributed representation based on the fact that a word may belong to several categories simultaneously. For illustration

purposes in the first experiment we kept the number of features low, i.e., ten. However, similar approach scales well up to higher numbers of dimensions.

Future research directions include analysis of larger corpora for extracting larger number of independent components. On a qualitative level, polysemy will be considered as one particularly challenging research topic. Whether the component values can be applied as degrees of membership for each word in each category is a question of further analysis. The distributed representation can be used as a well-motivated low-dimensional encoding for words. The limited number of dimensions brings computational efficiency whereas the meaningful interpretation of each component provides basis for intelligent processing.

References

Bingham E., Kaban A. and Girolami, M. (2001). Finding topics in dynamical text: application to chat line discussions. *Poster proceedings of the 10th International World Wide Web Conference (WWW10)*, May 1-5, 2001, Hong Kong. (pp. 198-199)

Bingham E., Kuusisto J. and Lagus K. (2002). ICA and SOM in Text Document Analysis. *Proceedings of the 25th ACM SIGIR 2002 International Conference on Research and Development in Information Retrieval*, August 11-15, 2002, Tampere, Finland. (pp. 361-362)

Church K. and Hanks P. (1990). Word association norms, mutual information and lexicography. *Computational Linguistics*, 16, 22-29.

Comon, P. (1994). Independent component analysis - a new concept? *Signal Processing*, 36, 287-314.

Deerwester S., Dumais S., Landauer T., Furnas G. and Harshman R.A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41, 391-407.

Field, D.J. (1994). What is the goal of sensory coding? *Neural Computation*, 6, 559-601.

Fillmore C. (1968). *Universals in Linguistic Theory*, chapter The case for case, pp. 1-88. Holt, Rinehart and Winston, Inc.

Honkela T., Pulkki V. and Kohonen T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In *Proceedings of ICANN-95, International Conference on Artificial Neural Networks*, volume 2, EC2 et Cie. (pp. 3-7)

Honkela T., Hyvärinen A. and Väyrynen J. (2003). *Emergence of linguistic representation by independent component analysis*. Report A72, Espoo,

Finland: Helsinki University of Technology, Laboratory of Information and Computer Science.

Honkela T. and Hyvärinen A. (2004). Linguistic Feature Extraction using Independent Component Analysis. In *Proceedings of IJCNN04, International Joint Conference on Neural Networks*.

Hyvärinen A., Karhunen J. and Oja E. (2001). *Independent Component Analysis*. New York: John Wiley & Sons.

Jutten C. and Hérault J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24, 1-10.

Kohonen T. (2001). *Self-Organizing Maps*. Berlin, Heidelberg: Springer.

Lund, K., Burgess, C. and Atchley, R.A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, 660-665.

Manning C. and Schütze H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, MA: MIT Press.

McDonald, S. (2000). *Environmental determinants of lexical processing effort.* PhD dissertation, University of Edinburgh.

Redington, M., Chater, N. and Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425-469.

Ritter H. and Kohonen T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61 (4), 241-254.

Schütze H. (1992). Dimensions of meaning. *Proceedings of Supercomputing*. (pp. 787-796)