Feature Extraction of Concepts by Independent Component Analysis

Altangerel Chagnaa*, Cheol-Young Ock*, Chang-Beom Lee, and Purev Jaimai*****

Abstract: Semantic clustering is important to various fields in the modern information society. In this work we applied the Independent Component Analysis method to the extraction of the features of latent concepts. We used verb and object noun information and formulated a concept as a linear combination of verbs. The proposed method is shown to be suitable for our framework and it performs better than a hierarchical clustering in latent semantic space for finding out invisible information from the data.

Keywords: *Independent Component Analysis, Clustering, Latent Concepts.*

1. Introduction

Our purpose in this work is to extract features of concepts from verb and noun contexts. These extracted verb concepts benefit many applications in natural language processing, such as word sense disambiguation and automatic thesaurus building, etc. In this scope, the application of an Independent Component Analysis (ICA) is proposed, which is a widely used method in signal processing, especially in blind source separation. However some contributions have been made in the text processing area for document analysis and linguistic feature extraction. As a brief word about these contributions, in document analysis studies (such as [2] and [3]) a term by document matrix is considered as the linear mixtures of a set of independent sources. The semantic spaces built in these works take the form of non-orthogonal term-occurrence histograms. For the linguistic feature extraction work in [4], the contextual information of words in a raw corpus is used to extract the linguistic features of the words.

Our contribution in this work differs from these previous studies, in the sense of (1) using a verb-noun pattern and (2) modeling from the "concept" view in the ICA, despite having different purposes. As mentioned above our model uses verb-noun information for a cognitive task concerning latent concepts and noun membership for those concepts. We formulate a concept as a linear combination of verbs/predicates. The ICA decomposes the verb-noun matrix into two parts, and thus the first part shows the weights of the verbs in the latent concepts and the second part gives the weights of the nouns in those concepts. See Fig. 2 for a more detailed illustration.

In the following section we give a brief introduction to the ICA, and then model its adaptation to extract the concept's feature verbs using a verb-noun distributional pattern. Section 4 describes the experimental results and is followed by the conclusion and suggestions for future works.

Manuscript received November 3, 2006; accepted January 20, 2007.
This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2004-211-420088) and in part by the MIC & the IITA through an IT leading R&D support project.
Corresponding Author: Cheol-Young Ock
* School of Computer Engineering and Information Technology, University of Ulsan, Korea(E-mail : goldenl@mail.ulsan.ac.kr, okcy@ulsan.ac.kr)
** Investigation & Analysis Team, Korean Transportation Safety Agency, Korea(E-mail : beom1009@kotsa.or.kr)
*** School of Information Technology, National University of Mongolia, Mongolia(E-mail : purev@num.edu.mn)

2. Independent Component Analysis

What follows is a brief description of the independent component analysis theory [1]. The classic version of the ICA model can be expressed as:

$$\mathbf{x} = \mathbf{As} \tag{1}$$

where $\mathbf{x} = (x_1,x_2,...,x_n)^T$ is the vector of the observed random variables, $\mathbf{s} = (s_1,s_2,...,s_n)^T$ is the vector of the independent latent variable (the "independent components"), and $\mathbf{A}$ is an unknown constant matrix, called the mixing matrix. If we denote the columns of matrix $\mathbf{A}$ by a_i the model can be written as:

$$\mathbf{x} = \sum_{i=1}^{n} a_i s_i \tag{2}$$

The goal in the ICA is to learn the decomposition in Eq. (1) in an unsupervised manner. That means that we only observe $\mathbf{x}$ and want to estimate both $\mathbf{A}$ and $\mathbf{s}$. The ICA can be seen as an extension to the principal component analysis and the factor analysis which underlie LSA. However, the ICA is a more powerful technique that is capable of

identifying the underlying factors in cases where the classic methods would fail.

The beginning point for the ICA is the simple assumption that the s_i components are statistically independent. The two variables y_1 and y_2 are independent if information on the value of y_1 does not give any information on the value of y_2, and vice versa. This assumption does not need to hold for the observed variables x_i. In the case of two variables, the independence holds if, and only if, $p(y_1, y_2)=p(y_1)p(y_2)$. This definition extends to any number of random variables.

There are three properties of the ICA that should be taken into account when considering the analysis results. First, one cannot determine the variances of the independent components s_i. The reason for this is that, with both **s** and **A** being unknown, any scalar multiplier in one of the sources s_i could always be canceled by dividing the corresponding column a_i of **A** by the same scalar. As a normalization step, one can assume that each component has a unit variance of $E\{ s_i^2\}=1$. The ambiguity of the sign still remains: a component could be multiplied by -1 without affecting the model.

The second property to be remembered is that one cannot determine the order of the components. While both **s** and **A** are unknown, the order of the terms in Eq. (2) can be freely changed and any one of the components can be called the first component.

The third important property of the ICA is that the independent components must be nongaussian for the ICA to be possible [1]. Then, the mixing matrix can be estimated up to the indeterminacies of order and sign discussed earlier. This is a main difference from such techniques as principal component analysis and factor analysis, which are only able to estimate the mixing matrix up to a rotation, which is quite insufficient for some purposes.

The ICA has many application areas such as speech processing, nueroimaging, face recognition, predicting stock market prices, mobile phone communication, and so forth.

In the application of the ICA for blind source separation, the observed random variables $(x_1(t),x_2(t),...,x_n(t))$ represent the signals recorded by microphones at different locations. Each microphone records the mixture of the speeches of the speakers. The ICA separates these individual speakers' speeches. The mixing matrix defines the location of the microphones. The following figure shows the scenario.

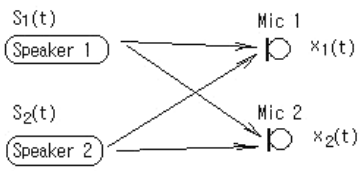


Fig. 1. Each microphone records the mixture of each individual speaker's speech in the case of two speakers and two microphones

For the ICA in our work we applied the FastICA [1] software package for Matlab. The FastICA uses a combination of two different approaches for a linear ICA (Comon's information theoretic approach and the projection pursuit approach) and is based on a fixed-point iteration scheme for finding the maximum of the nongaussianity (see [7] for more detail). It has many advantages as it is easy to use, is computationally simple, requires little memory space, and its independent components can be found one by one.

3. Modeling Concept with Verbs Using the ICA

For the concept's feature extraction model, we suppose that verbs indicate the latent concepts with their noun usage patterns. We assume that the influence of the concepts on the verb is a natural/language phenomenon, and that this information can be obtained using a source separation method like the ICA, as shown in Fig. 2. Verbs indicate the property of mixed independent concepts. That is why we aim to separate these latent concepts from the mixture. Verb-noun information (for example Verb-Object relation) is more detailed than just contextual information or a bag of words. The reason is that the contextual information is coarser than syntactically extracted verb-noun relations.

In the ICA analysis for this model, the observed random variables are verb-noun information, which consists of an individual verb's usage pattern with nouns. Then, the **x** variable in Eq.1 is a verb-noun matrix and its rows define the verbs' usage pattern with nouns (columns of the matrix). The independent components are the latent components that are expressed by nouns, and a mixing matrix defines the concepts with verbs (see Fig. 2). A column in the mixing matrix $(a_{1i},a_{2i},...,a_{ni})$ shows the verb weights in the concept.

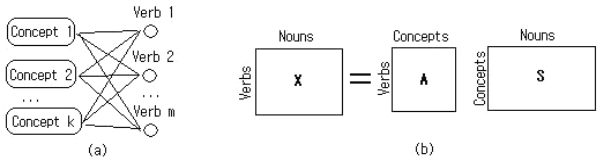


Fig. 2. The illustration of the ICA model from concept views: a) verbs act like microphones, indicating the mixture of concepts via their noun usage. b) The model in matrix notation.

By using this method we assume that we can find the hidden concepts expressed by verbs. In the following section, the experiment and results of the proposed approach are shown.

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

4. Experiment and Results

We used the Korean monolingual dictionary "Keum Sung" to extract verb noun information. This dictionary is PoS and is semantically tagged. There is a total of 24,932 distinct nouns and 8,172 distinct verbs (totaling 296,442 verb-noun pairs) in the explanations of the words given in the dictionary. Of these, 38,724 distinct verb-object noun pairs (3,820 distinct verbs and 11,904 distinct nouns) were found. From these pairs we selected nouns with frequencies of over 20 and verbs within a frequency of between 40 and 1000. The mutual information (MI) [5] was chosen for the measurement of the relationship between those verbs and nouns. After conducting the MI calculation, we set a threshold on the value of MI over 0.5. In our experiment we chose 15 common verbs and their 91 nouns from the verb and object noun data. The verbs are shown in the following table with their English meanings. The verbs shown in shaded rows have the same meanings; and from those 15 verbs we can see 12 distinct meanings.

Table 1. The verbs used in the experiment with their English meanings

1	그리_2 / to draw	9	싸_1 / to wrap
2	막 / to block	10	쓰_1 / to write
3	먹_3 / to eat	11	적_1/ to write, note
4	삼키 / to swallow	12	쓰_3 / to use
5	바꾸 / to change	13	알 / to know
6	뽑 / to pick out	14	잃 / to miss, to loose
7	사 / to buy	15	주_1 / to give
8	팔 / to sell		

The following picture shows the part of the verb and object noun data used in our experiment.

Object Nouns	그리_2 /to draw	막 /to block	먹_3 /to eat	삼키 /to swallow	바꾸 /to change	뽑 /to pick	사 /to buy	팔 /to sell	싸_1 /to wrap	쓰_1 /to write	적_1 /to note	쓰_3 /to use	알 /to know	잃 /to miss	주_1 /to
가치_2/worth													1,573		
값/value					1,194						0,777				
결과_2/result											2,353				
고기_1/meat			3,235												
곡식/grain							1,913	1,276							
행위/act													1,344		
효과/effect															0,563
힘/strength												1,294			1,972

Fig. 3. The excerpt from the verb and object noun data

As mentioned earlier we fed this verb-noun mutual information matrix to the FastICA algorithm and set the nonlinearity function *g* to the *tanh* function, using symmetric orthogonalization (see [7] for more details about the algorithm and parameters). The following paragraphs will show the analysis of the experimental results.

The verb-object noun matrix of the 15 rows by 91 columns is decomposed into 12 independent components. Each independent component corresponds to a concept. The resulting size of the mixing matrix is 15 rows by 12 columns, while the independent components matrix consists of 12 rows by 91 columns. The columns of the mixing matrix represent the latent concepts with which we are dealing.

For the extracted concepts, let us analyze the verb weights. Fig. 4 shows the verb weights in the concepts as bar graphs. The height of the bar indicates the weight of the verb in that concept. Note that because of the ambiguity of the ICA, the sign of the weight is arbitrary.

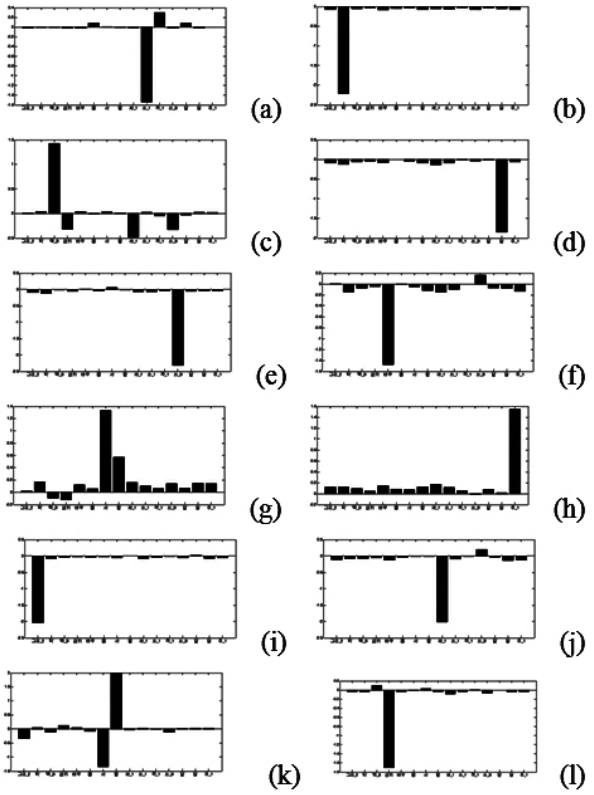


Fig. 4. Verb weights in 12 independent components shown as bar graphs

Similar verbs shows up with high weights in a concept, as verbs 쓰_1/*to write* and 적_1/*to write*, shown in the first graph (Fig. 4 (a)). Actually they have the same meaning of writing down. Also 먹_3/*to eat* and 삼키/*to swallow* (Fig. 4 (l)) have the same meaning of eating. We can say that these verbs are in the same concept. Another interesting point is that the verbs 사/*to buy* and 팔/*to sell* do not have same meaning. In fact, they have opposite meanings, but these verbs are nonetheless found in the same concept (Fig. 4 (k)). From a conceptual point of view, these are classed in the same concept of action as trading or exchanging products with money.

For the concept shown in Fig. 4 (c), its most weighted 4 verbs are 먹_3/*to eat*, 삼키/*to swallow*, 싸_1/*to wrap* and 쓰_3/*to use*. Actually 먹_3/*to eat* and 삼키/*to swallow* have the same meaning, but a further 2 words 싸_1/*to wrap* and 쓰_3/*to use* are found in this concept. This is because of the data. The verb 먹_3/*to eat* has three nouns which are also used by the verb 싸_1/*to wrap*, while only one noun is used by each of the two verbs 삼키/*to swallow* and 쓰_3/*to use* (see Fig. 5).

Nouns	먹_3/to eat	싸_1/to wrap	삼키/to swallow	쓰_3/to use
음식/food	3,466		3,259	
고기_1/meat	3,235			
밥_1/rice	3,187			
잎_1/leaf	2,945			
약_5/medicine	2,936	1,544		2,271
열매/fruit	1,558	2,283		
뿌리/root	1,269	1,739		
시체_3/corpse		2,354		

Fig. 5. Nouns used by the verbs in the concept shown in Fig. 4 (c)

Next, we compared the ICA results with the results of the traditional hierarchical clustering. Complete linkage-based clustering was carried out on the same data to discover the relations between these verbs. Here we preprocessed the actual data by singular value decomposition (SVD). This is because our model is a latent semantic model for extracting invisible information from the data. SVD is a method used in the latent semantic indexing model, and it is effective in information retrieval [8].

Euclidean distance was used as a similarity measure between objects in clusters. As seen from the result shown in Fig. 6, the complete linkage- based hierarchical clustering gives a neat figure of objects in the semantic space. At first, the most similar pairs of verbs are clustered. The verb pairs [쓰_1/*to write* - 적_1/*to write*], [사/*to buy* - 팔/*to sell*], [쓰_3 /*to use* – 잃/*to miss*], [삼키/*to swallow* - 뽑/*to pick out*] and [바꾸/*to change* - 알/*to know*] are clustered on the first level. Actually, the first two pairs are good clusters while the others are not.

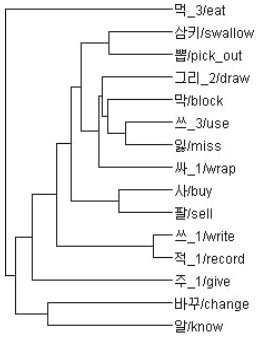


Fig. 6. Complete linkage-based hierarchical clustering of the verbs in the experiment

As compared with these hierarchical results, the ICA method is more effective in extracting hidden concepts from the data. For example, the verbs 먹_3/*to eat* and 삼키/*to swallow* are not found in the hierarchical clustering, but in the ICA model they are found in the same concept. Note that the hierarchical method is performed in the latent semantic space.

As a result, we observe that the ICA method is suitable for extracting the features of latent concepts and overcomes the traditional method based on a latent semantic space. This is because this method is a latent variable model that extracts the independent concepts based on the statistics of verb-noun usage patterns. The concepts are assumed to be independent from each other and not directly visible from the verb-noun information.

5. Conclusions and Future Works

This paper presented the implementation of a concept's feature extraction by a source separation method. We suppose that a verb shows the mixed characteristic of latent concepts. The concepts are assumed to be independent from each other. Thus we applied the Independent Component Analysis, a method of separating independent components from the mixture, in our task to extract the features of concepts.

The proposed approach is shown to be suitable for the task, and performs better than the traditional hierarchical clustering method in finding latent concepts from the experiment data. This method successfully finds similar verbs with high weights in a concept. Also, it finds verbs with opposite meanings in the same concept, which is true from the conceptual point of view. In this scope the comparison is not only made with the hierarchical method but also with the latent semantic model. For this, we transferred the original space into a latent semantic space by SVD decomposition.

This work is a preliminary work involving the clustering of words by the ICA. We gave here only the analysis of the verb and concept part (mixing matrix A in Fig. 2). Furthermore, we extend the work on noun clustering based on this work. If we look at the independent component part (matrix S in Fig. 2) the few best-fitting nouns in a concept show up with high similarities.

The work in this framework will be continued further and we expect good results. Future works will include the semantic clustering of words and the building of conceptual networks, and so forth.

Acknowledgements. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2004-211-420088) and in part by the MIC & the IITA through an IT leading R&D support project.

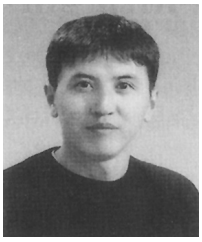
References

[1] Hyvarinen, A., Karhunen, J. and Oja, E., "Independent Component Analysis", John Wiley & Sons, 2001.

[2] Bigham, E., Kuusisto, J., and Lagus, K., "ICA and SOM in Text Document Analysis", Proceedings of the 25th ACM SIGIR 2002 International Conference on Research and Development in Information Retrieval, Tampere, Finland, pp. 361-362, 2002.

[3] Kolenda, T., Hansen, L., and Sigurdsan, S., "Independent Components in Text", In M. Girolami, editor,

Advances in Independent Component Analysis, Springer-Verlag, pp. 235-256, 2000.
[4] Honkela, T., Hyvarinen, A., and Vayrynen, J., "Emergence of Linguistic Features: Independent Analysis of Contexts", In Proc. of the Neural Computation and Psychology Workshop 9, Plymouth, UK, 2005.
[5] Church, K., and Hanks, P., "Word Association Norms, Mutual Information and Lexicography", Computational Linguistics, vol. 16, pp. 22-29, 1990.
[6] Manning, C., and Schutze, H., "Foundations of Statistical Natural Language Processing", Cambridge, MA: MIT Press, 1999.
[7] Hyvarinen, A. "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis", IEEE Tr. on Neural Networks, Vol. 10, No. 3, pp. 626-634, 1999.
[8] Scott C. Deerwester et al., "Indexing by Latent Semantic Analysis", Journal of the American Society of Information Science, no. 41, pp. 391-407, 1990.



Altangerel Chagnaa
BS, Dept. of Electronics, National University of Mongolia(2001)
MS, School of Information Technology, National University of Mongolia(2003)
Ph.D. candidate, School of Comp. Eng. and IT, University of Ulsan(present)
Research Field is Natural Language Processing and Machine Learning.

Cheol-Young Ock
BA, Dept. of Computer Engineering, Seoul National University(1982)
MA, Dept. of Computer Engineering, Seoul National University(1984)
Ph.D., Dept. of Computer Engineering Seoul National University(1993)
Visiting Professor, RUSSIA TOMSK Institute(1994), GLASGOW University(1996)
Professor, School of Comp. Eng. and IT, University of Ulsan(present)
Research Field is Natural Language Processing, Machine Learning, Knowledge Engineering, Ontology.

Chang-Beom Lee
BA, Dept. of Computer Science, Chonnam National University (1995)
MA, Dept. of Computer Science, Chonnam National University (2001)
Ph.D., Dept. of Computer Science, Chonnam National University (2005)
Research Professor, School of Comp. Eng. and IT, University of Ulsan (2005.3-2006.10)
Assistant Manager, Investigation & Analysis Team, KOTSA (present)
Research Field is Information Retrieval, Natural Language Processing, Text Summarization, Text Classification

Purev Jaimai
BS & MS, Academy of Economy, Poland(1979)
Ph.D., Mongolian University of Science and Technology(1994)
Visiting Professor, Sejong University, Korea(2002-2003)
Professor, School of Information Technology, National University of Mongolia(present)
Research Field is Natural Language Processing