Intelligent Semantic-Based System for Corpus Analysis through Hybrid Probabilistic Neural Networks

Keith Douglas Stuart[1], Maciej Majewski[2], and Ana Botella Trelis[1]

[1] Polytechnic University of Valencia, Department of Applied Linguistics
Camino de Vera, s/n, 46022 Valencia, Spain
{kstuart,apbotell}@idm.upv.es

[2] Koszalin University of Technology, Faculty of Mechanical Engineering
Raclawicka 15-17, 75-620 Koszalin, Poland
maciej.majewski@tu.koszalin.pl

Abstract. The paper describes the application of hybrid probabilistic neural networks for corpus analysis which consists of intelligent semantic-based methods of analysis and recognition of word clusters and their meaning. The task of analyzing a corpus of academic articles was resolved with hybrid probabilistic neural networks and developed word clusters. The created prototypes of word clusters provide the probabilistic neural networks with possibilities of recognizing corpus clusters. The established corpus comprises 1376 articles, from specialist leading SCI-indexed journals, and provides representative samples of the language of science and technology. In this paper, a review of selected issues is carried out with regards to computational approaches to language modelling as well as semantic patterns of language. The paper features semantic-based recognition algorithms of word clusters of similar meanings but different lexico-grammatical patterns from the established corpus using multilayer neural networks. The paper also presents experimental results of word cluster semantic-based recognition in the context of phrase meaning analysis.

Keywords: corpus analysis, artificial intelligence, probabilistic neural networks, semantic networks, phrase meaning analysis, natural language processing, applied computational linguistics.

1 Introduction

The hypothesis that modelling a language involves probabilistic representations of allowable sequences, determines two areas of knowledge that might be applied to text analysis. One is word clusters: it is often the case that strings of words are repeated or tend to cluster together for semantic and/or syntactic reasons. The other is the fact that given a sequence of words one might want to try and predict the next word based on what restrictions exist on the choice of next word. Another way of putting this might be that given a sequence of possible words,

D. Liu et al. (Eds.): ISNN 2011, Part I, LNCS 6675, pp. 83–92, 2011.

estimate the probability of that sequence. In a corpus of size N, the assumption is that any combination of n words is a potential n-gram. Each n-gram in our model is a parameter used to estimate probability of the next possible word. Low frequency n-grams are the most frequent. In other words, it is very common to find strings that have low frequency. In the same way, it is very common to find words that only occur once in a corpus (hapax legomena).

A corpus is a collection of linguistic data, consisting of a large and structured set of texts, which can be used for linguistic description or as a means of verifying hypotheses about a language [7,10,12]. Text corpus is nowadays usually electronically stored, distributed and processed for statistical analysis and used for checking occurrences or validating linguistic rules on a specific universe [16]. We have been doing research into word clusters in a corpus of 1376 academic articles and we have found that repetition is constant across many word sequences.

Our corpus comprises 1,376 articles, from specialist leading journals (a total of 6,104,323 tokens, 71,516 types, and 1.17 type/token ratio). The articles have all been published in journals cited in the Science Citation Index (SCI). They have been distributed in 23 knowledge areas, each of which constitutes per se a sub-corpus. They are representative samples of the language of science and technology. The corpus has been tagged with meta-textual information and transferred to an Access database by means of an application in Visual Basic.

Once the corpus had been designed and implemented, we proceeded to analyse the data by creating wordlists of technical and semi-technical terms through frequency counts and keyword identification. This process involved initially comparing a general English wordlist (from the 100 million BNC corpus) with a wordlist from our corpus. Frequencies were compared and a keyword list was created from our corpus. Analysis was conducted by processing both the corpus as a whole and each of the subject areas separately. Then, we proceeded to generate 3 to 8 word clusters (n-grams), which were transferred to a database specially designed to carry out queries related to the clusters. Furthermore, we carried out research into collocational structures which are obtained by calculating the total number of times a word is found in the neighbourhood of the node word using as the default collocation horizon 5 words to the left and 5 words to the right of the node word (although it is possible to calculate collocations using much larger horizons). Both clusters and collocational structures provide clues to lexico-grammatical patterns. For this paper, we have mainly used the data from the 3 to 8 word clusters.

The aim of the research is intelligent corpus analysis through meaning recognition of word clusters using artificial intelligence methods. We have developed a method which allows for development of possible word cluster components in a corpus for training hybrid probabilistic neural networks. The networks are capable of recognizing word clusters with similar meaning but different lexico-grammatical patterns. In other words, we are working with the idea that there is a strong tendency for sense and syntax to be associated [17]. Corpus

Linguistics needs computational tools to be able to map the close association between pattern and meaning and neural networks are ideal for pattern recognition and, consequently, semantic meaning.

2 The State of the Art

Automated tools are used by researchers for analyzing corpus frequency data. They have analyzed the differences between various registers of corpora such as fiction and academic writing and have found that many features of corpora differ between registers [1,2,3]. The features they discuss range from syntactic, through lexical, to discourse.

Parsed corpora have made it possible to generate more reliable syntactic frequency information. Much of the work with that data has looked at the frequencies of specific structures occurring with specific verbs [6,10].

Previous work on corpus analysis faced several limitations: the number of words covered, the number of structures covered, and limits on the amount of data available for low frequency items imposed by the size of the corpora [5,7,12]. While some work [11,15] has used data from larger corpora, it is an important goal to develop new reliable and efficient automatic extraction methods. Towards this goal, various automated tools have been developed during the last few years. However, most of them use old-fashioned methods, lacking functionalities such as sophisticated capabilities which could be delivered with use of artificial intelligence methods [8,9,22,23,24,25,26]. This paper proposes an approach to deal with the above mentioned problem.

3 Description of the Method

The proposed intelligent semantic-based system for corpus analysis shown in abbreviated form on Fig. 1a, consists of two subsystems: statistical corpus processing and intelligent corpus processing [26].

In the corpus processing subsystem, words are isolated from text extracted from the corpus, which are developed into various combinations of word clusters based on the statistical models of word sequences. The developed word clusters representing appropriate N-gram models are processed further for training hybrid probabilistic neural networks with learning patterns of words and clusters.

In the intelligent corpus processing subsystem, text is retrieved from the corpus using a parser. In the next step, word clusters are extracted by the parser using lexical and grammar patterns. The separated words are processed for letter strings isolated in segments as possible cluster word components. This analysis has been carried out using Hamming neural networks. The output data of the analysis consists of processed word segments. Individual word segments treated here as isolated possible components of the cluster words are inputs (Fig. 1b) of hybrid probabilistic neural networks for recognizing words. The networks use learning files containing words and are trained to recognize words as word cluster components, with words represented by output neurons.

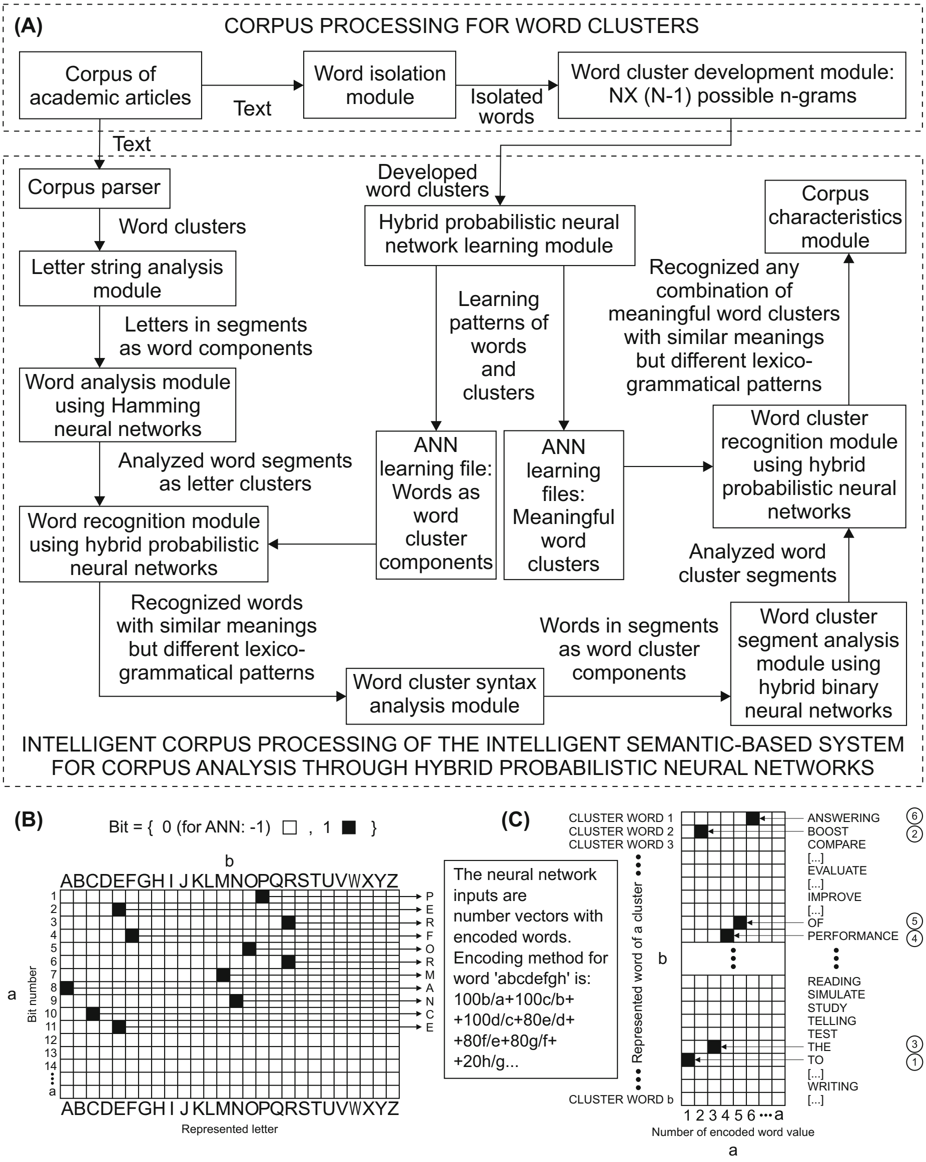
(A)
CORPUS PROCESSING FOR WORD CLUSTERS
Corpus of academic articles
Text
Word isolation module
Isolated words
Word cluster development module: NX (N-1) possible n-grams
Text
Corpus parser
Word clusters
Letter string analysis module
Letters in segments as word components
Word analysis module using Hamming neural networks
Analyzed word segments as letter clusters
Word recognition module using hybrid probabilistic neural networks
Recognized words with similar meanings but different lexico-grammatical patterns
Developed word clusters
Hybrid probabilistic neural network learning module
Learning patterns of words and clusters
ANN learning file: Words as word cluster components
ANN learning files: Meaningful word clusters
Word cluster syntax analysis module
Words in segments as word cluster components
Word cluster segment analysis module using hybrid binary neural networks
Analyzed word cluster segments
Word cluster recognition module using hybrid probabilistic neural networks
Recognized any combination of meaningful word clusters with similar meanings but different lexico-grammatical patterns
Corpus characteristics module
INTELLIGENT CORPUS PROCESSING OF THE INTELLIGENT SEMANTIC-BASED SYSTEM FOR CORPUS ANALYSIS THROUGH HYBRID PROBABILISTIC NEURAL NETWORKS
(B)
Bit = { 0 (for ANN: -1) □ , 1 ■ }
b
ABCDEFGHIJKLMNOPQRSTUVWXYZ
a
Bit number
1 2 3 4 5 6 7 8 9 10 11 12 13 14 ⋮ a
P E R F O R M A N C E
ABCDEFGHIJKLMNOPQRSTUVWXYZ
Represented letter
(C)
The neural network inputs are number vectors with encoded words. Encoding method for word 'abcdefgh' is: 100b/a+100c/b+ +100d/c+80e/d+ +80f/e+80g/f+ +20h/g...
CLUSTER WORD 1
CLUSTER WORD 2
CLUSTER WORD 3
Represented word of a cluster
b
CLUSTER WORD b
ANSWERING 6
BOOST 2
COMPARE
[...]
EVALUATE
[...]
IMPROVE
[...]
OF 5
PERFORMANCE 4
READING
SIMULATE
STUDY
TELLING
TEST
THE 3
TO 1
[...]
WRITING
[...]
1 2 3 4 5 6 ••• a
Number of encoded word value
a

Fig. 1. (A) Diagram of the proposed intelligent semantic-based system for corpus analysis, (B) inputs of the word recognition module, (C) inputs of the word cluster recognition module

The intelligent cluster word recognition method allows for recognition of words with similar meanings but different lexico-grammatical patterns. In the next stage, the words are transferred to the word cluster syntax analysis module. The module creates words in segments as word cluster components properly, which are coded as vectors. Then they are processed by the module for word cluster segment analysis using hybrid binary neural networks. The analyzed word cluster segments become inputs of the word cluster recognition module using hybrid probabilistic neural networks (Fig. 1c). The module uses multilayer probabilistic neural networks, either to recognize the cluster and find its meaning or else it fails to recognize it. The neural networks of this module use learning files containing patterns of possible meaningful word clusters. The intelligent analysis and processing allow for recognition of any combination of meaningful word clusters with similar meanings but different lexico-grammatical patterns. The overall detailed results of the intelligent analysis are subject to processing for corpus characteristics and its linguistic description including: statistical analysis, checking occurrences, and validating linguistic rules.

The proposed intelligent semantic-based system for corpus analysis contains hybrid probabilistic neural networks which are pattern classifiers. They can become effective tools for solving classification problems of lexico-grammatical structures in corpus linguistics, where the objective is to assign cases of clusters of letters or words to one of a number of discrete cluster classes. Pattern classifiers place each observed vector of cluster data x in one of the predefined cluster classes k_i, i=1, 2, ..., K where K is the number of possible classes in which x can belong. The effectiveness of the cluster classifier is limited by the number of data elements that vector x can have and the number of possible cluster classes K. The Bayes pattern classifier implements the Bayes conditional probability rule that the probability $P(k_i\,|x\,)$ of x being in class k_i is given by (1):

$$P(k_i\,|x\,) = \frac{P\,(x\,|k_i\,)\;P\,(k_i)}{\sum_{j=1}^{K} P\,(x\,|k_j\,)\;P\,(k_j)} \tag{1}$$

where $P(x\,|k_i\,)$ is the conditioned probability density function of x given set k_i, $P(k_j)$ is the probability of drawing data from class k_j. Vector x is said to belong to a particular class k_i if $P(k_i\,|x\,) > P(k_j\,|x\,)$, $\forall j = 1,\ 2,\ \ldots,\ K,\ j \neq i$. This classifier assumes that the probability density function of the population from which the data was drawn is known a priori. This assumption is one of the major limitations of implementing Bayes classifier.

The probabilistic neural network was first introduced by Specht [18, 19, 20, 21], who was inspired by the work of Parzen [14]. The network is interesting, because it is possible to implement and develop numerous enhancements, extensions, and generalizations of the original model. It offers a way to interpret the network's structure in the form of a probabilistic density function. The probabilistic neural network simplifies the Bayes classification procedure by using a training set of clusters for which the desired statistical information for implementing Bayes classifier can be drawn. The desired probability density function of the cluster class is approximated by using the Parzen windows approach

[4,13,14]. The probabilistic neural network learns to approximate the probability density function of the cluster training samples. It should be interpreted as a function that approximates the probability density of the underlying cluster samples distribution, rather than fitting the cluster samples directly. It approximates the probability that vector x belongs to a particular class k_i as a sum of weighted Gaussian distributions centred at each cluster training sample. The output of the model is an estimate of the cluster class membership probabilities.

The architecture of the hybrid probabilistic neural networks in the proposed system is shown in Fig. 2. The network is composed of many interconnected processing units or neurons organized in successive layers. The hybrid probabilistic neural network for recognition of clusters of letters or words consists of five layers: cluster processing, cluster input, cluster pattern, summation and output layers. The cluster processing layer performs input value normalization of each value in the input vector. The cluster input layer unit does not perform any computation and simply distributes the input to the neurons in the next layer. In the pattern layer, there is one pattern neuron for each cluster training sample. Each pattern neuron forms a product of the weight vector w_j^i and the given cluster sample, where the weights entering a neuron are from a particular cluster sample. This product is then passed through the exponential activation function (2):

$$\exp\left(-\frac{\left(w_j^i - x\right)^T \left(w_j^i - x\right)}{2\sigma^2}\right) \tag{2}$$

It is necessary that both vectors x and w are normalized to unity. On receiving a pattern x from the cluster input layer, the neuron x_j^i of the cluster pattern layer computes its output (3):

$$\phi_j^i\left(x\right) = \frac{1}{\left(2\pi\right)^{s/2}\sigma^s}\exp\left(-\frac{1}{2\sigma^2}\left(x - x_j^i\right)^T\left(x - x_j^i\right)\right) \tag{3}$$

where s is the dimension of the cluster input pattern x, σ is a smoothing factor and x_j^i is the j-th training vector for the cluster patterns in class k_i. The superscript T denotes the transpose of the vector, and *exp* stands for the exponential function. The total number of the cluster pattern layer nodes is given as a sum of the cluster pattern units for all classes. The summation layer neurons compute the maximum likelihood of cluster pattern x being classified into k_i by summarizing and averaging the output of all neurons that belong to the same cluster class (4):

$$P_i\left(k_i\left|x\right.\right) = \frac{1}{\left(2\pi\right)^{s/2}\sigma^s}\frac{1}{N_i}\sum_{j=1}^{N_i}\exp\left(\frac{-\left(x - x_j^i\right)^T\left(x - x_j^i\right)}{2\sigma^2}\right) \tag{4}$$

where N_i is the number of cluster training patterns in class k_i. Eq. (4) is a sum of small multivariate Gaussian probability distributions that are centred at each cluster training sample. This function is used to generalize the classification

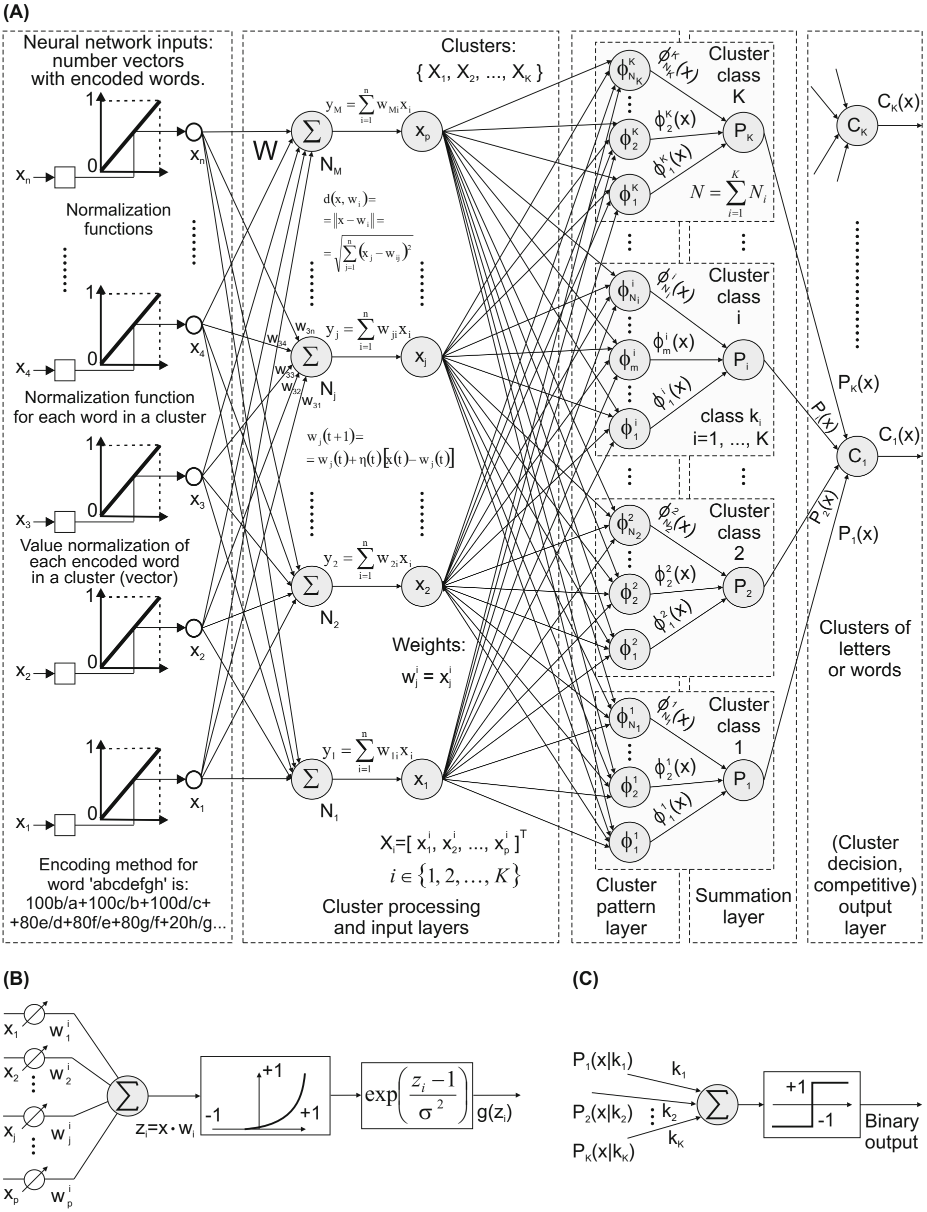

Fig. 2. (A) The hybrid probabilistic neural networks for recognition of clusters of letters or words, (B) Neuron of the pattern layer, (C) Neuron of the output layer

to beyond the given cluster training samples. As the number of cluster training samples and their Gaussians increases the estimated probability density function approaches the true function of the cluster training set.

The classification decision for a cluster of letters or words is taken according to the inequality (5):

$$\begin{aligned}&\sum_{j=1}^{N_i} \exp\left(-\tfrac{1}{2\sigma^2}\left(x - x_j^i\right)^T\left(x - x_j^i\right)\right)\\ &\quad > \sum_{j=1}^{N_k} \exp\left(-\tfrac{1}{2\sigma^2}\left(x - x_j^k\right)^T\left(x - x_j^k\right)\right) \quad for\ \ all\ \ i\ \ and\ \ k\end{aligned} \tag{5}$$

Before classification, the sums in Eq. (5) are multiplied by their respective prior probabilities (P_i and P_k) calculated as the relative frequency of the cluster samples in each cluster class [18]. The decision layer classifies the cluster pattern x in accordance with the Bayes's decision rule based on the output of all the summation layer neurons using (6):

$$\begin{aligned}&\hat{C}(x) = \arg\ \max\left\{\tfrac{1}{(2\pi)^{s/2}\sigma^s}\tfrac{1}{N_i}\sum_{j=1}^{N_i}\exp\left(\tfrac{-\left(x-x_j^i\right)^T\left(x-x_j^i\right)}{2\sigma^2}\right)\right\}\\ &i = 1,\ 2,\ ...,\ K\end{aligned} \tag{6}$$

where $\hat{C}(x)$ denotes the estimated class of the cluster pattern x and K is the total number of classes in the cluster training samples [18].

The smoothing factor σ is the only factor that needs to be selected for training. A σ too small causes a very spiky approximation, which will not generalize clusters of letters or words well, whereas a σ too large will smooth out details of cluster structures. An appropriate σ is chosen empirically.

4 Experimental Results

Our corpus comprising 1,376 articles contains clusters of the types from 3 to 8 word clusters. The experimental results show the numbers of clusters in the corpus, which are presented in (Fig. 3A).

The proposed system allowed for recognition of any combination of meaningful word clusters with similar meanings but different lexico-grammatical patterns. The tests measured the performance of the cluster meaning recognition. The effectiveness of the system was achieved to a satisfactory level. As shown in Fig. 3B, the ability of the hybrid probabilistic neural network to recognize a cluster depends on the number of words in that cluster. For best performance, the neural network requires a minimum number of words of each cluster to be recognized as its input.

Important factors are both the neural network design (i.e., selection of the smoothing factor (σ)) and development of representative training patterns of word clusters by the proposed system.

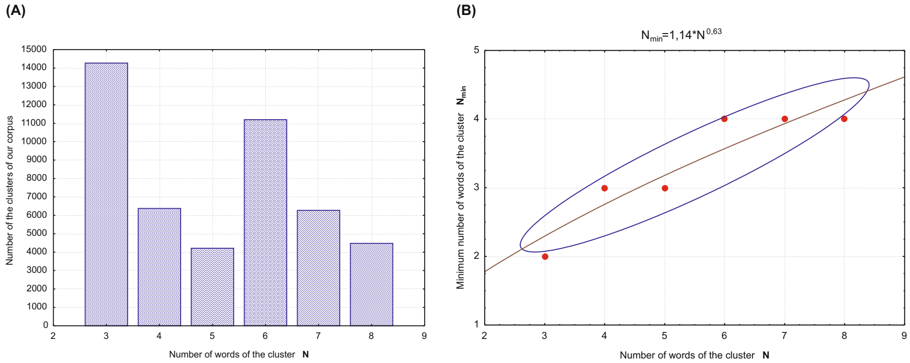

Fig. 3. (A) number of clusters of our corpus vs. number of words of the cluster, (B) sensitivity of word cluster meaning recognition: minimum number of words of the cluster being recognized vs. number of cluster component words

5 Conclusions and Perspectives

It is assumed that language processing is closely tied to a user's experience, and that distributional frequencies of words and structures play an important role in learning. Therefore the interest in the statistical profile of language usage plays an important role in research. This paper has developed a method which allows for extraction of possible word cluster components in a corpus for training hybrid probabilistic neural networks. The networks are capable of recognizing word clusters with similar meaning but different lexico-grammatical patterns. It has long been an ambition of corpus linguistics to investigate fully relationships between form and meaning, sense and syntax [17]. The patterns of language have been revealed by corpus linguistics through concordance lines, word clusters, collocation and colligation but there is no automated way of generating these word clusters. It might be useful for corpus linguistics to learn from neural networks how to generate word clusters automatically based on the training of the aforementioned networks with corpus examples and thereby bridge the gap between data-driven Hallidayan approaches to language and the more formalized Chomskyan predictive approach.

References

1. Biber, D.: Variation across speech and writing. Cambridge University Press, Cambridge (1988)
2. Biber, D.: Using register-diversified corpora for general language studies. Computational Linguistics 19(2), 219–241 (1993)
3. Biber, D., Conrad, S., Reppen, R.: Corpus linguistics: Investigating language structure and use. Cambridge University Press, Cambridge (1998)
4. Cacoullous, R.: Estimation of a probability density. Annals of the Institute of Statistical Mathematics (Tokyo) 18(2), 179–189 (1966)

5. Carter, R., Hughes, R., McCarthy, M.: Exploring grammar in context. Cambridge University Press, Cambridge (2000)
6. Jurafsky, D.: A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science 20(2), 137–194 (1996)
7. Jurafsky, D., Martin, J.H.: Speech and language processing: An introduction to natural language processing. In: Speech Recognition, and Computational Linguistics. Prentice-Hall, New Jersey (2000)
8. Kacalak, W., Stuart, K., Majewski, M.: Intelligent natural language processing. In: Jiao, L., Wang, L., Gao, X.-b., Liu, J., Wu, F. (eds.) ICNC 2006. LNCS, vol. 4221, pp. 584–587. Springer, Heidelberg (2006)
9. Kacalak, W., Stuart, K., Majewski, M.: Selected problems of intelligent handwriting recognition. Advances in Soft Computing 41, 298–305 (2007)
10. Kennedy, G.: An introduction to corpus linguistics. Longman, London (1998)
11. Lapata, M., Keller, F., Schulte, S.: Verb frame frequency as a predictor of verb bias. Journal of Psycholinguistic Research 30(4), 419–435 (2001)
12. Manning, C., Schtze, H.: Foundations of statistical natural language processing. MIT Press, Cambridge (1999)
13. Murthy, V.K.: Estimation of a probability density. The Annals of Mathematical Statistics 36(3), 1027–1031 (1965)
14. Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33(3), 1065–1076 (1962)
15. Roland, D., Elman, J.L., Ferreira, V.S.: Why is that? Structural prediction and ambiguity resolution in a very large corpus of English sentences. Cognition 98(3), 245–272 (2006)
16. Sampson, G.: English for the computer. Oxford University Press, Oxford (1995)
17. Sinclair, J.: Corpus, Concordance, Collocation. Oxford University Press, Oxford (1991)
18. Specht, D.F.: Probabilistic neural networks. Neural Networks 3(1), 109–118 (1990)
19. Specht, D.F.: A general regression neural network. IEEE Transactions on Neural Networks 2(6), 568–576 (1991)
20. Specht, D.F.: Enhancements to probabilistic neural networks. In: Proceedings of the IEEE International Joint Conference on Neural Networks, Baltimore Maryland USA, vol. 1, pp. 761–768 (1992)
21. Specht, D.F., Romsdahl, H.: Experience with adaptive probabilistic neural networks and adaptivegeneral regression neural networks. In: IEEE World Congress on Computational Intelligence, IEEE International Conference on Neural Networks, Orlando Florida USA, vol. 2, pp. 1203–1208 (1994)
22. Stuart, K., Majewski, M.: Selected problems of knowledge discovery using artificial neural networks. In: Liu, D., Fei, S., Hou, Z., Zhang, H., Sun, C. (eds.) ISNN 2007. LNCS, vol. 4493, pp. 1049–1057. Springer, Heidelberg (2007)
23. Stuart, K., Majewski, M.: A new method for intelligent knowledge discovery. Advances in Soft Computing 42, 721–729 (2007)
24. Stuart, K., Majewski, M.: Artificial creativity in linguistics using evolvable fuzzy neural networks. In: Hornby, G.S., Sekanina, L., Haddow, P.C. (eds.) ICES 2008. LNCS, vol. 5216, pp. 437–442. Springer, Heidelberg (2008)
25. Stuart, K., Majewski, M.: Evolvable neuro-fuzzy system for artificial creativity in linguistics. In: Huang, D.-S., Wunsch II, D.C., Levine, D.S., Jo, K.-H. (eds.) ICIC 2008. LNCS (LNAI), vol. 5227, pp. 46–53. Springer, Heidelberg (2008)
26. Stuart, K.D., Majewski, M., Trelis, A.B.: Selected problems of intelligent corpus analysis through probabilistic neural networks. In: Zhang, L., Lu, B.-L., Kwok, J. (eds.) ISNN 2010. LNCS, vol. 6064, pp. 268–275. Springer, Heidelberg (2010)