Towards Explicit Semantic Features using Thresholded Independent Component Analysis

Jaakko J. Väyrynen and **Timo Honkela** and **Lasse Lindqvist**
Adaptive Informatics Research Centre
Helsinki University of Technology, Finland
{jjvayryn,tho,llindqvi}@cis.hut.fi

Abstract

Latent semantic analysis (LSA) can be used to create an implicit semantic vectorial representation for words. Independent component analysis (ICA) can be derived as an extension to LSA that rotates the latent semantic space so that it becomes explicit, that is, the features correspond more with those resulting from human cognitive activity. This enables nonlinear filtering of the features, such as thresholding that forces sparse ICA components for words. We will demonstrate this with multiple choice semantic vocabulary tests generated from a multilingual thesaurus. The experiments are conducted in English, Finnish and Swedish.

1 Introduction

Latent semantic analysis (LSA) (Landauer and Dumais, 1997) is a very popular method for extracting information from text corpora. The mathematical method behind LSA is singular value decomposition (SVD) (Deerwester et al., 1990), that removes second order correlations from data and can be used to reduce dimension. LSA has been shown to produce reasonably low-dimensional latent semantic spaces that can handle various tasks, such as vocabulary tests and essay grading, at human level (Landauer and Dumais, 1997). The found latent components, however, are implicit and cannot be understood by humans. In fact, as typical distance measures are rotation-invariant, any rotation of the latent space would not be seen.

Independent component analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001) is a method for removing higher order correlations from data and it can be seen as whitening followed by a rotation, where whitening can be produced with SVD. The rotation should find components that are statistically independent of each other and that we think are meaningful. In case the components are not truly independent, ICA should find "interesting" components. ICA has been demonstrated to produce unsupervised structures that well-align with that resulting from human cognitive activity in text, images, social networks and musical features (Hansen et al., 2005). We will show that the components found by the ICA method can be further processed by simple nonlinear methods, such as thresholding, that give rise to a sparse feature representation of words. An analogical approach can be found from the analysis of natural images, where a soft thresholding of sparse coding is as a denoising operator (Oja et al., 1999).

The ICA can be, e.g., used to detect topics in document collections (Isbell and Viola, 1999; Bingham et al., 2001). Earlier we have shown that the ICA analysis results into meaningful word features (Honkela and Hyvärinen, 2004; Honkela et al., 2004) and that these features correspond to a reasonable extent with categorizations created through human linguistic analysis (Väyrynen et al., 2004).

In this paper, we present experimental results that show how the ICA method produces explicit semantic features instead of the implicit features created by the LSA method. We show through practical experiments that this approach exceeds the capacity of the LSA method.

2 Data

We use collection of texts as our source of natural language for English, Finnish and Swedish. Our unsupervised learning methods are singular value decomposition and independent components analysis. The semantic representations learned with the methods are applied to multiple choice vocabulary tasks that measure how well the word representations capture semantics.

2.1 Europarl Corpus

The Europarl corpus (Koehn, 2005) contains texts from the Proceedings of the European Parliament in 11 languages. We concentrated in English, Finnish and Swedish in our experiments. XML tags and special characters were removed from the texts and uppercase characters were replaced with respective lowercase ones. The English text had 26 million tokens (word forms in running text) and 83 thousand types (unique word forms). The Finnish text had 19 million tokens and 480 thousand types. The Swedish text had 24 million tokens and 240 thousand types.

2.2 Gutenberg Corpus

A more general example of a natural text is a collection of 4966 free English e-books that were extracted from the Project Gutenberg website[1]. The texts were pruned to exclude poems and the e-book headers and footers were removed. The texts were then concatenated into a single file and preprocessed by removing special characters and replacing numbers by a special symbol and uppercase characters with respective lowercase ones. The final corpus had 319 million tokens and 1.41 million types. For computational reasons, a subset of the types was selected as the vocabulary to be analyzed.

2.3 Vocabulary Test Sets

Semantic word representations can be evaluated with multiple choice vocabulary tests that measure some semantic concept, such as synonymity. In a multiple choice test, the task is to select the correct word from a list of alternatives when given a stem word or a cue word.

For the English language, there exists free electronic resources that can be used to conduct such tests. For many other languages of interest, however, such resources may not be directly available. We briefly introduce one famous but small and two large semantic resources for English, as well as one for many European languages.

Performance of the compared methods is measured with precision, the ratio of correct answers to the number of questions in the test set. The higher the precision is, the better the method has captured the part semantics the questions cover. The vocabulary and the test questions were chosen so that recall was 100 percent. Especially this means that only single word terms occurring in the analyzed vocabulary were considered for test questions.

2.3.1 TOEFL Synonyms

A famous test case for English is the synonym part of the TOEFL data set[2]. It was provided for us by the Institute of Cognitive Science, University of Colorado, Boulder. The task is to select the synonym for each stem word from four alternatives. For the TOEFL data set, LSA has been shown to get 64.4% correct which is statistically at the same level as for a large sample of applicants to US colleges from non-English speaking countries (Landauer and Dumais, 1997). Even precision level of 97.5% has been reached by combining several methods, including LSA and an online thesaurus (Turney et al., 2003).

However, the TOEFL test set has only 80 questions and comparison of methods with only this test set is not sufficient. Also, the baseline precision with guessing from four alternatives is 25% and chance might play a big role in the precision. An example question with the correct answer emphasized is shown below.

figure: list, *solve*, divide, express

2.3.2 Moby Synonyms and Related Words

The Moby Thesaurus II[3] of English words and phrases has more than 30 000 entries with 2.5 million synonyms and related terms. We generated multiple choice questions by selecting a stem from the Moby thesaurus, and mixing one of the listed synonyms with a number of random words from our vocabulary as alternatives. This method allows us

[1] http://www.gutenberg.org

[2] http://www.ets.org

[3] http://www.dcs.shef.ac.uk/research/ilash/Moby/

to have more questions and alternatives than the TOEFL data set, which makes the test more robust in terms of confidence intervals for precision. On the other hand, the generated questions are very likely to lack the finesse of the hand-crafted TOEFL questions and no human level performance is known. An example entry in the thesaurus is shown below.

approve: OK, accede to, accept, accord to, accredit, admire, adopt, affiliate, affirm, . . .

Our vocabulary overlapped with 16 638 stems in the Moby thesaurus and one multiple choice question with 16 alternatives was generated for each entry. The baseline precision is 6.25% with guessing from 16 alternatives. An example of a generated question is shown below.

constitute: *validate*, washington, wands, paper-based, convention, aérospatiale, vanhecke, indifference, kaklamanis, possess, criminalization, grouping, shari, reorganisations, diluents

2.3.3 Idiosyncratic Associations

The free association norms data set[4] from the University of South Florida contains idiosyncratic responses in English, that is, responses given only by one human subject, to more than five thousand cue words. On average, there are approximately 22.15 idiosyncratic responses per cue word with high variation and more idiosyncratic responses are produced than responses given by two or more participants. An example entry is shown below.

early: before, classes, frost, on time, prompt, sleepy, sun, tired, years

Similarly to the generated Moby questions, the idiosyncratic association data set was used to generate 4 582 multiple choice questions with 16 alternatives. An example of a generated question is shown below.

corrupt: *crook*, plaice, wfp, a5-0058, administrated, vega, 1871, a5-0325, h-0513, toolbox, compelling, 1947,crashing, vac, illating, indemnity

2.3.4 Eurovoc Thesaurus

The multilingual Eurovoc thesaurus[5] covers fields that are of importance for the activities of the European institutions. It is available in many European languages and contains different semantics relationships between the terms in the thesaurus. Each field is divided into several microthesauri, e.g., the field “trade” contains seven microthesauri, including “tariff policy” and “consumption”. An excerpt of a microthesaurus is shown below.

- political system

  RT political science (3611)

  NT1 authoritarian regime
  NT1 change of political system
    RT political reform (0431)
    RT transition economy (1621)
  NT1 constitutional monarchy
    RT parliament (0421)

We decided the task to be identification of terms in the same microthesaurus, but not including the related terms (RT) in other microthesauri. For each pair of terms in a microthesaurus, one term was selected as a cue word and the other was mixed with a number of random words from the analyzed vocabulary as alternatives. Only fields “finance”, “law”, “politics” and “trade” were included in these experiments. This procedure gave 2 312 questions for English, 1 848 for Finnish, and 7 564 for Swedish. An example of a generated question in English is shown below.

republic: *oligarchy*, alps, spits, seventy, greeks, progressivity, deflationary, endorsing, renowned, understate, cogently, miscalculations, 0306, range, heralding, lèse-majesté

3 Methods

It has been known already for some time that statistical analysis of the contexts in which a word appears in text can provide reasonable amount of information on the syntactic and semantic roles of the word (Ritter and Kohonen, 1989; Church and Hanks, 1990). A typical approach is to calculate a

[4] `http://w3.usf.edu/FreeAssociation/`

[5] `http://europa.eu/eurovoc/`

document-term matrix in which the rows correspond to the documents and the columns correspond to the terms. A column is filled with the number of occurrences of the particular term in each document. The similarity of use of any two terms is reflected by the relative similarity of the corresponding two columns in the document-term matrix. Instead of considering the whole documents as contexts, one can also choose the neighboring words, a sentence, a paragraph or some other contextual window. An alternative approach, that is taken here, is to calculate the number of co-occurrences of the particular term with number of other terms in a contextual window around the analyzed term. This produces a context-term matrix, where each context is defined using terms instead of documents.

3.1 Contextual Information

Contextual information is a standard way of filtering more dense data from running text. Frequencies of term occurrences, or co-occurrences, in different chunks of texts are typically calculated. The idea behind this is that relations of words manifest themselves by having related words occur in similar contexts, but not necessary together. Raw contextual data is too sparse for practical use and it has been shown that finding a more compact representation from the raw data can increase the information content by generalizing the data (Landauer and Dumais, 1997).

A context-term matrix $\mathbf{X}$ was calculated using the Gutenberg corpus or one languages in the Europarl corpus. The rows in the matrix correspond to contexts and the columns represent the terms in the analyzed vocabulary. The context contained frequencies of the 1 000 most common word forms in the selected corpus in a 21 word window centered around each occurrence of the analyzed terms. The terms included the 50 000 most common word forms. For the Gutenberg corpus, an additional 29 words were included in the analyzed terms so that all of the questions in the TOEFL set could be utilized.

The contextual information was encoded with a bag-of-words model and the matrix $\mathbf{X}$ was of size $1\,000 \times 50\,029$ for the English Gutenberg corpus and of size $1\,000 \times 50\,000$ for each language in the Europarl corpus.

The raw frequency information of the terms is typically modified using stop-word lists and term weighting, such as the tf·idf method that is suitable for document contexts. We did not use stop-word lists and frequency rank information was preserved by taking the logarithm of the frequencies increased by one.

3.2 Singular Value Decomposition

Singular value decomposition learns a latent structure for representing data. Input to singular value decomposition is a $m \times n$ matrix $\mathbf{X}$. The SVD method finds the decomposition $\mathbf{X} = \mathbf{UDV}^T$, where $\mathbf{U}$ is an $m \times r$ matrix of left singular vectors from the standard eigenvectors of square symmetric matrix $\mathbf{XX}^T$, $\mathbf{V}$ is an $n \times r$ matrix of right singular vectors from the eigenvectors of $\mathbf{X}^T\mathbf{X}$, $\mathbf{D}$ is a diagonal $r \times r$ matrix whose non-zero values are the square roots of the eigenvalues of $\mathbf{XX}^T$ or (equivalently) $\mathbf{X}^T\mathbf{X}$, and $r = \min(n, m)$ is the rank of $\mathbf{X}$. A lossy dimension reduction to $l \leq r$ components can be achieved by discarding small eigenvalues.

In SVD-based latent semantic analysis, the input matrix $\mathbf{X}$ is a context-term matrix representing the weighted frequencies of terms in text passages or other contexts. The method can handle tens of thousands of terms and contexts. Dimension is typically lowered to a few hundred components, that reduces noise and generalizes the data by finding a latent semantic representation for words. Words and texts can be compared by their respective vectorial representations in the latent space.

3.3 Independent Component Analysis

Independent component analysis uses higher-order statistics compared to singular value decomposition that only removes second-order correlations. ICA finds a decomposition $\mathbf{Z} = \mathbf{BS}$ for a data matrix $\mathbf{Z}$, where $\mathbf{B}$ is a mixing matrix of weights for the independent components in the rows of matrix $\mathbf{S}$. The task is usually to find a separating matrix $\mathbf{W} = \mathbf{B}^{-1}$ that produces independent components $\mathbf{S} = \mathbf{WZ}$.

If data $\mathbf{Z}$ is white, i.e., covariance matrix is the identity matrix, it suffices to find a rotation that produces maximally independent components (Hyvärinen et al., 2001). The right singular values $\mathbf{V}$ produced by SVD are uncorrelated and thus SVD can be seen as a direct preprocessing step to ICA, if the data $\mathbf{X}$ has zero mean. This mathematical relation is showed in Figure 1. The ICA rotation should find components that are more interesting and structure the semantic space in a meaningful manner, as illustrated in Figure 2.

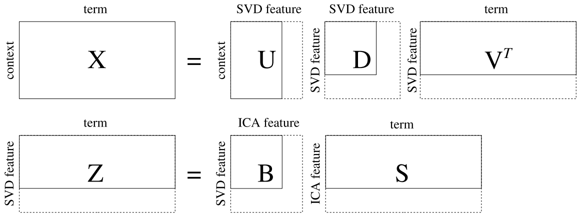


Figure 1: Mathematically, for zero-mean data $\mathbf{X}$, ICA can be represented as an extension of SVD, where the white SVD components $\mathbf{Z} = \sqrt{n}\mathbf{V}^T$ for the n terms are transformed with a rotation matrix $\mathbf{B}$ to find the ICA components $\mathbf{S}$. SVD is approximated for a reduced dimension from the original dimension of the data matrix $\mathbf{X}$, marked here with the solid and dashed lines, respectively.

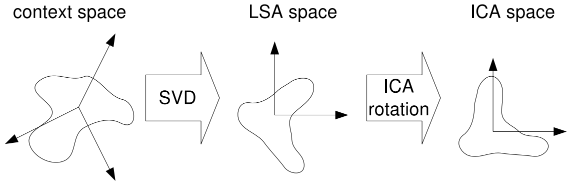
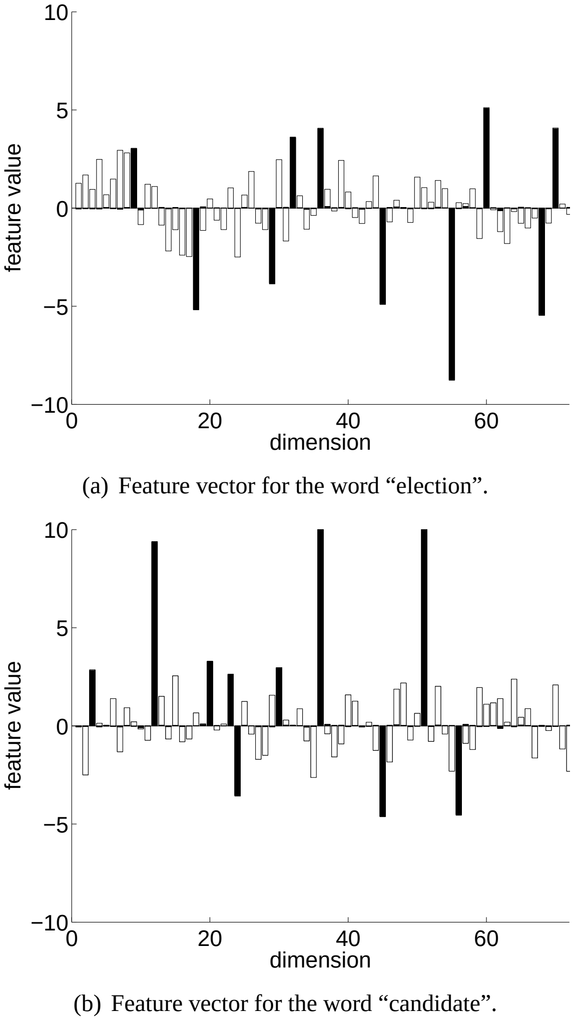


Figure 2: The distribution of terms in contexts can be approximated by a low-dimensional LSA space. ICA can be seen as an additional rotation of the latent space that finds interesting components.

3.4 Thresholding

Thresholding is an example of a nonlinear filtering method. It forces a word representation to be more sparse by retaining only a subset of the features. For a successful usage of such thresholded feature representation in a semantic task, it is necessary that those features that contain most of the semantic information are kept while less informative features are discarded. It is also important that the underlying representation models each word with as less features as possible, which can be said to be a definition of sparseness.

Our features have zero mean and have the same variance. For each term in our vocabulary, the features with the lowest absolute values were considered inactive and were thresholded. Thus the remaining active features depend on the particular term. For comparison purposes, the same number of active features were selected for each word. An example of thresholded word features is shown in Figure 3. We compare thresholded ICA and thresholded SVD with different number of dimension. Results are also reported for standard SVD, that is also used for selecting the dimensionality for the thresholded versions.



(a) Feature vector for the word "election".

(b) Feature vector for the word "candidate".

Figure 3: ICA feature vectors for the word "election" (a) and "candidate" (b). The outlined bars show the original feature values and the filled bars show the thresholded values with ten active dimensions. Any comparison based on the dot product of the thresholded feature vectors depends wholly on the common active dimensions 36 and 45.

4 Results

Here we will compare SVD and ICA as feature extraction methods by evaluating the emerging semantic word representations using multiple choice vocabulary tests in three languages and different semantic vocabulary tests. In order to show how ICA finds an explicit feature representation, we threshold the word features and show that ICA produces better results than SVD. In our experiments, the similarity of words was measured as the cosine of the angle between the respective words vectors.

We have previously reported results for the English Gutenberg corpus and the Moby and idiosyncratic test sets that are reproduced here (Väyrynen et al., 2007). We present here additional results for representations learned from the English, Finnish and Swedish parts of the Europarl corpus. Suitable tests sets for the Europarl were generated from the multilingual Eurovoc thesaurus. The dimension for the thresholded versions of ICA and SVD was selected as approximately the dimension that produced the highest precision with the basic SVD method without thresholding. Additionally, results with other dimensions are shown. In this section, the number of active components for each word, i.e., the level of thresholding, is varied and precision of the thresholded representation is measured in a multiple choice vocabulary test. The ICA and SVD methods converge when no thresholding is done. The fewer active dimensions there are, the sparser the word representations are.

The representation learned from the Gutenberg corpus were evaluated with the Moby and the idiosyncratic test sets. The results indicate that thresholding with ICA outperforms standard SVD and that thresholding with SVD does not improve the results. The reproduced results are shown in Figure 4 and Figure 5.

Results for the TOEFL data set with the Gutenberg corpus in (Väyrynen et al., 2007), are similar to the Eurovoc test with the Finnish part of Europarl in Figure 6. In both cases the thresholded ICA and SVD have very similar performance. The handmade questions in the TOEFL would make the semantics of the alternatives closer to each other, that would make the thresholding process more accurate as the word vector would have more similar features. It is still unclear why this happens also with the Finnish Eurovoc test.

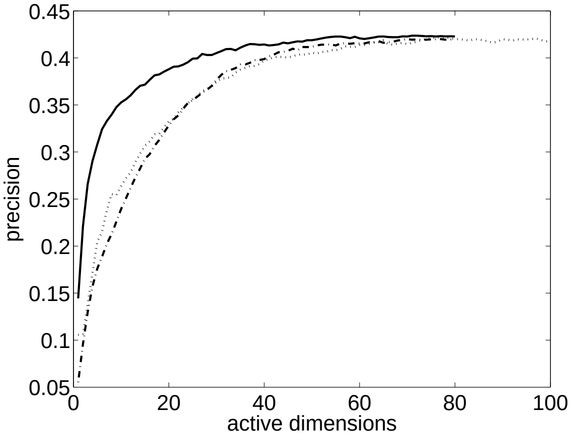


Figure 4: Precisions of the SVD (dotted), SVD with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) with the Moby data set w.r.t. the number of active components. The representations were learned from the Gutenberg corpus.

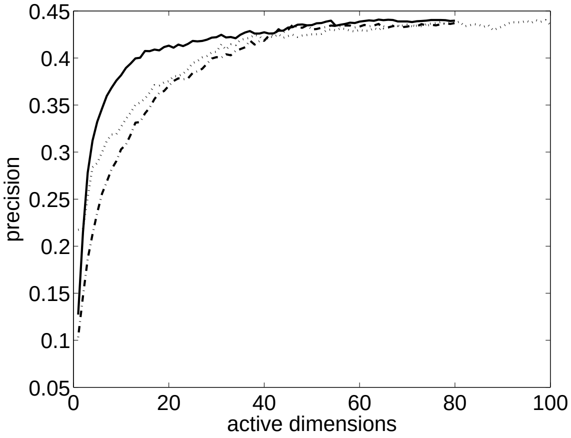


Figure 5: Precisions of the SVD (dotted), SVD with thresholding with 80 components (dashed) and ICA with thresholding with 80 components (solid) with the idiosyncratic association data set w.r.t. the number of active components. The representations were learned from the Gutenberg corpus.

The English and Swedish word representations learned from the Europarl corpus behave more like the Gutenberg results. The Swedish test, shown in Figure 8, is a good example how the thresholded ICA can maintain high precision even when more

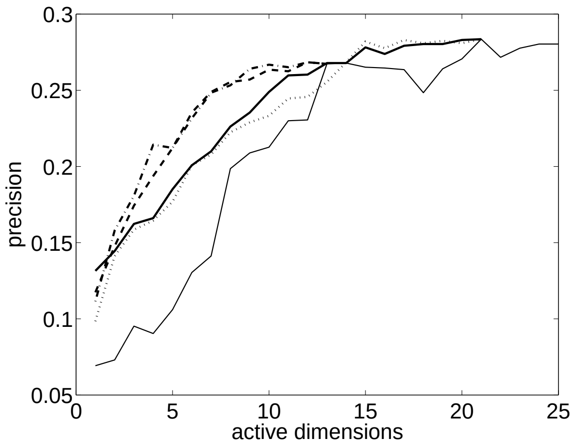


Figure 6: Precisions of the SVD (thin solid), SVD with thresholding with 21 components (dotted) and 13 components (dash dotted) and ICA with thresholding with 13 components (thick solid) and 13 components (dashed) with the Finnish Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

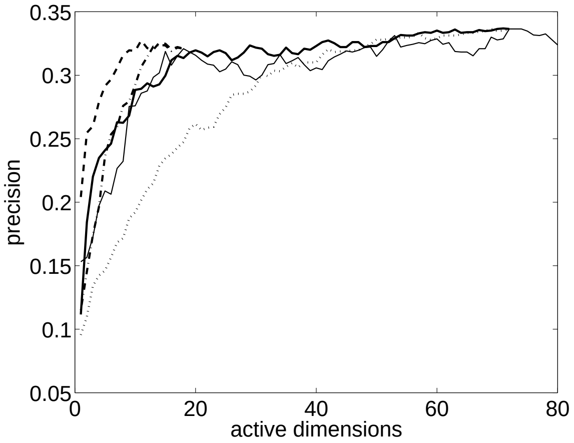


Figure 7: Precisions of the SVD (thin solid), SVD with thresholding with 72 components (dotted) and 18 components (dash dotted) and ICA with thresholding with 72 components (thick solid) and 18 components (dashed) with the English Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

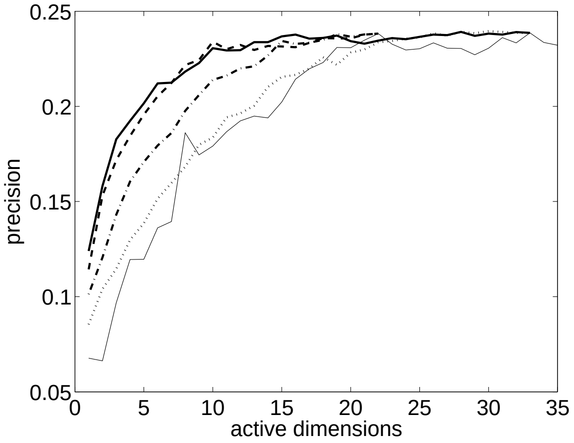


Figure 8: Precisions of the SVD (thin solid), SVD with thresholding with 33 components (dotted) and 22 components (dash dotted) and ICA with thresholding with 33 components (thick solid) and 22 components (dashed) with the Swedish Eurovoc test set w.r.t. the number of active components. The representations were learned from the Europarl corpus.

than half of the features in each word are ignored. The English test with Europarl did not give equally clear results, but even here the thresholded ICA method does not worse than the standard SVD and outperforms the thresholded SVD method.

5 Conclusions

In this paper, we showed how the explicit semantic features for words produced by independent component analysis align more to cognitive components resulting from human activity. We applied a nonlinear filtering, thresholding, to the word vectors produced by ICA and SVD and studied these thresholded semantic representations in multiple choice vocabulary tests.

The results shown in this article indicate that it is possible to create automatically a sparse representation for words. Moreover, the emergent features in this representation seem to correspond with some linguistically relevant features. When the context is suitably selected for the ICA analysis, the emergent features mostly correspond to some semantic selection criteria. Traditionally, linguistic features have been determined manually. For instance, case grammar is a classical theory of grammatical analysis (Fillmore, 1968) that proposes to analyze sentences as constituted by the combination of a verb plus a set of deep cases, i.e., semantic roles. Numerous different theories and grammar formalisms exist that provide a variety of semantic or syntactic categories into which words need to be manually classified.

Statistical methods such as SVD and ICA are able to analyze context-term matrices to produce automatically useful representations. ICA has the additional advantage, especially when combined with some additional processing steps reported in this article, over SVD (and thus LSA) that the resulting representation is explicit and sparse: each active component of the representation is meaningful as such. As the LSA method is already very popular, we assume that the additional advantages brought by this method will further strengthen the movement from a manual analysis to an automated analysis.

References

Ella Bingham, Ata Kabán, and Mark Girolami. 2001. Finding topics in dynamical text: application to chat line discussions. In *Poster Proc. 10th Intern. World Wide Web Conf. (WWW'10)*, pages 198–199.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Pierre Comon. 1994. Independent Component Analysis, a new concept? *Signal Processing*, 36(3):287–314.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *J. American Sociecty of Information Science*, 41(6):391–407.

Charles J. Fillmore, 1968. *Universals in Linguistic Theory*, chapter The Case for Case, pages 1–88. Holt, Rinehart, and Winston, New York, USA.

Lars Kai Hansen, Peter Ahrendt, and Jan Larsen. 2005. Towards cognitive component analysis. In Timo Honkela, Ville Könönen, Matti Pöllä, and Olli Simula, editors, *Proc. International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning AKRR'05*, pages 148–153.

Timo Honkela and Aapo Hyvärinen. 2004. Linguistic feature extraction using independent component analysis. In *Proc. Intern. Joint Conf. on Neural Networks (IJCNN)*, pages 279–284.

Timo Honkela, Aapo Hyvärinen, and Jaakko Väyrynen. 2004. Emergence of linguistic features: Independent component analysis of contexts. In *Proc. 9th Neural Computation and Psychology Workshop (NCPW9): Modeling Language Cognition and Action*, pages 129–138.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. 2001. *Independent Component Analysis*. John Wiley & Sons.

Charles Lee Isbell, Jr. and Paul Viola. 1999. Restructuring sparse high dimensional data for effective retrieval. In *Proc. Conf. on Advances in Neural Information Processing Systems (NIPS 1998)*, pages 480–486.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, pages 79–86.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

Erkki Oja, Aapo Hyvärinen, and Patrik Hoyer. 1999. Image feature extraction and denoising by sparse coding. *Pattern Analysis & Applications*, 2(2):104–110.

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

Peter D. Turney, Michael L. Littman, J. Bigham, and V. Schnayder. 2003. Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proc. Intern. Conf. on Recent Advances in Natural Language Processing (RANLP-03)*, pages 482–489.

Jaakko J. Väyrynen, Timo Honkela, and Aapo Hyvärinen. 2004. Independent component analysis of word contexts and comparison with traditional categories. In *Proc. 6th Nordic Signal Processing Symposium (NORSIG 2004)*, pages 300–303.

Jaakko J. Väyrynen, Lasse Lindqvist, and Timo Honkela. 2007. Sparse distributed representations for words with thresholded independent component analysis. In *International Joint Conference on Neural Networks (IJCNN 2007)*. To appear.