Semantic analysis in word vector spaces with ICA and feature selection

Tiina Lindh-Knuutila and **Jaakko J. Väyrynen** and **Timo Honkela**
Aalto University School of Science
Department of Information and Computer Science
PO Box 15400, FI-00076 AALTO, Finland
firstname.lastname@aalto.fi

Abstract

In this article, we test a word vector space model using direct evaluation methods. We show that independent component analysis is able to automatically produce meaningful components that correspond to semantic category labels. We also study the amount of features needed to represent a category using feature selection with syntactic and semantic category test sets.

1 Introduction

The concept of semantic similarity or the broader concept of semantic relatedness is central in many NLP-related applications. The representations that are used range from thesauri to vector space models (Budanitsky and Hirst, 2006). Semantic similarity can be measured as a distance between representations of words; as vector similarity in a vector space model, or as a path length in a structured representation e.g, an ontology. For humans, the notion of semantic similarity often means perceiving two words sharing similar traits: synonyms like *car:automobile*, hypernyms *vehicle:car* or antonyms *short:long* (Turney and Pantel, 2010).

Earlier research has shown that it is possible to learn automatically linguistic representations that reflect syntactic and semantic characteristics of words. In particular, independent component analysis (ICA) can be used to learn sparse representations in which components have a meaningful linguistic interpretation. However, earlier results cannot be considered conclusive especially when semantic relations and semantic similarity are considered. We present results of a systematic analysis that focuses on semantic similarity using manually built resources as a basis for evaluation of automatically generated vector space model (VSM) representation. We concentrate on word vector spaces based on co-occurrence data. These can be evaluated directly by comparing distances between word pairs or groups of words judged similar by human evaluators. In this article, we describe several test sets used for semantic evaluation of vector space models, and validate our model with them. We then explore how many features are required to distinguish the categories with a feature selection algorithm. Further, we measure how well ICA is able to automatically find components that match semantic categories of words.

2 Methods

2.1 Vector space models

VSMs contain distributional information about words derived from large text corpora. They are based on the idea that words that appear in similar contexts in the text are semantically related, and that relatedness translates into proximity of the vector representations (Schütze, 1993). In information retrieval and many other tasks, topic representations using term-document matrices are often employed. In the same fashion, word vector spaces are built using the more immediate context of a word. Similarity measures for vector spaces are numerous. For VSMs, they have been extensively tested for example by Bullinaria and Levy (2007). In this work, we use the cosine similarity (Landauer and Dumais, 1997), which is

most commonly used (Turney and Pantel, 2010).

The simple way of obtaining a raw word co-occurrence count representation for N target words is to consider C context words that occur inside a window of length l positioned around each occurrence of the target words. An accumulation of the co-occurrences creates a word-occurrence matrix $X_{C\times N}$. Different context sizes yield representations with different information. Sahlgren (2006) notes that small contexts (of a few words around the target word), give rise to paradigmatic relationships between words, whereas longer contexts find words with syntagmatic relationship between them. For a review on the current state of the art for vector space models using word-document, word-context or pair-pattern matrices using singular value decomposition-based approaches in dimensionality reduction, see Turney and Pantel (2010).

2.2 Word spaces with SVD, ICA and SENNA

The standard co-occurrence vectors for words can be very high-dimensional even if the intrinsic dimensionality of word context information is actually low (Karlgren et al., 2008; Kivimäki et al., 2010), which calls for an informed way to reduce the data dimensionality, while retaining enough information. In our experiments, we apply two computational methods, singular value decomposition (SVD) and ICA, to reduce the dimensionality of the data vectors and to restructure the word space.

Both the SVD and ICA methods extract components that are linear mixtures of the original dimensions. SVD is a general dimension reduction method, applied for example in latent semantic analysis (LSA) (Landauer and Dumais, 1997) in the linguistic domain. The LSA method represents word vectors in an orthogonal basis. ICA finds statistically independent components which is a stronger requirement and the emerging features are easier to interpret than the SVD features (Honkela et al., 2010).

Truncated SVD approximates the matrix $X_{C\times N}$ as a product UDV^T in which $D_{d\times d}$ is a diagonal matrix with square roots of the d largest eigenvalues of X^TX (or XX^T), $U_{C\times d}$ has the d corresponding eigenvectors of XX^T, and $V_{N\times d}$ has the d corresponding eigenvectors of X^TX. The rows of $V_{N\times d}$ give a d-dimensional representation for the target words.

ICA (Comon, 1994; Hyvärinen et al., 2001) represents the matrix $X_{C\times N}$ as a product AS, where $A_{C\times d}$ is a mixing matrix, and $S_{d\times N}$ contains the independent components. The columns for the matrix $S_{d\times N}$ give a d-dimensional representation for the target words. The FastICA algorithm for ICA estimates the model in two stages: 1) dimensionality reduction and whitening (decorrelation and variance normalization), and 2) rotation to maximize the statistical independence of the components (Hyvärinen and Oja, 1997). The dimensionality reduction and decorrelation step can be computed, for instance, with principal component analysis or SVD.

We compare the results obtained with dimension reduction to a set of 50 feature vectors from a system called SENNA (Collobert et al., 2011)[1]. SENNA is a labeling system suitable for several tagging tasks: part of speech tagging, named entity recognition, chunking and semantic role labeling. The feature vectors for a vocabulary of 130 000 words are obtained by using large amounts of unlabeled data from Wikipedia. In training, unlabeled data is used in a supervised setting. The system is presented a target word in its context of 5+5 (preceding+following) words with a 'correct' class label. An 'incorrect' class sample is constructed by substituting the target word with a random one and keeping the context otherwise intact. The results in the tagging tasks are at the level of the state of the art, which is why we want to compare these representations with the direct evaluation tests.

2.3 Direct evaluation

In addition to indirect evaluation of vector space models in applications, several tests for direct evaluation of word vector spaces have been proposed see e.g., Sahlgren (2006) and Bullinaria and Levy (2007). First we describe the semantic and syntactic category tests. Here, a *category* means a group of words with a given class label. The precision P in the category task is calculated according to (Levy et al., 1998). A centroid for each category is calculated as an arithmetic mean of the word vectors belonging to that category.

[1] http://ronan.collobert.com/senna/

The distances from each word vector to all category centroids are then calculated, recalculating the category centroid for the query to exclude the query vector. The precision is then the percentage of the words for which the closest centroid matches the category the word is labeled with.

The semantic category test (Semcat) set[2] is used for example in (Bullinaria and Levy, 2007). This set contains 53 categories with 10 words in each category, based on the 56 categories collected by Battig and Montague (1969). Some word forms appear in more than one category, for example *orange* in FRUIT and in COLOR, and *bicycle* in TOY, and in VEHICLE. We made some slight changes to the test set by changing the British English spelling of a limited number of words back into American English (e.g., *millimetre-millimeter*) to better conform to the English used in the Wikipedia corpus.

We also consider two different syntactic category test alternatives. Bullinaria *et al.* (1997; 2007) use ten narrow syntactic categories, separating noun and verb forms in own categories, whereas Sahlgren (2006) uses eight broad categories. In this article, we compare both of these approaches. As neither of these test sets were publicly available, we constructed our own applying the most common part-of-speech (POS) tags from the Penn Treebank tag set (Marcus et al., 1993) to 3 000 most frequent words in our vocabulary. We call the built test sets Syncat1 and Syncat2. In Syncat1, the 50 most frequent words in 10 narrow POS categories: Singular or mass nouns (NN), plural nouns (NNS), singular proper nouns (NNP), adjectives in base form (JJ), adverbials in base form (RB), verbs in base form (VB), verbs in past participle form (VBN), verbs in ing-form (VBG), cardinal numbers (CD), and prepositions or subordinating conjunctions (IN). In (Levy et al., 1998) the last category contains only prepositions, but the Penn Treebank tagset does not separate between subordinating conjunctions and prepositions. Syncat2 contains 20 most frequent words in seven broader POS categories: nouns, verbs, adjectives, adverbs, prepositions, determiners, and conjunctions. In the open categories; nouns, verbs, adjectives and adverbs, the words can be in any of the tagged forms. The original experiments (Sahlgren, 2006) contain a category named 'conjunctions', which we created by combining the aforementioned IN-category which contains also prepositions with coordinating conjunctions (CC). The interjection category (UH) from the original work was left out due to the infrequency of such words in the Wikipedia corpus.

In addition to category tests, synonymy can also be directly measured, for example with a multiple choice test, where synonyms should be closer to each other than the alternatives. The most commonly used test for VSMs is the Test of English as a Foreign Language (TOEFL) (Landauer and Dumais, 1997), although other similar tests, such as English as a Second Language (ESL) and SAT college entrance exam (Turney, 2001; Turney, 2005), are also available. In addition, one can easily construct similar multiple-choice tests based on, for example, thesauri and random alternatives. Bullinaria et al. (1997; 2007; 2012) use a test which consists of related word pairs, e.g., *thunder-lightning, black-white, brother-sister*. The distance from a cue word to the related word is then compared to randomly picked words with a similar frequency in the corpus[3]. In addition to semantic similarity, Deese antonym pairs (Deese, 1954), have been used for VSM evaluation (Grefenstette, 1992). We employ a similar procedure described above for the related words. The precision is calculated by checking how often the cue and the correct answer are closest – with a comparison to eight randomly picked words for each cue word.

2.4 Forward feature selection with entropy

The forward feature selection method is a simple greedy algorithm. At each step, the algorithm selects the single feature that best improves the result measured by an evaluation criteria, without ever removing already selected features. The algorithm stops when all features are selected, or a stopping criterion is triggered. Compared to an exhaustive feature search, which has to evaluate all possible feature combinations, only $d(d+1)/2$ different input sets have to be evaluated, where d is the desired number of features (Sorjamaa, 2010). Reaching the global optimum is not guar-

[2] http://www.cs.bham.ac.uk/ jxb/corpus.html

[3] http://www.cs.bham.ac.uk/~jxb/corpus.html

anteed because the forward feature selection algorithm can get stuck in a local minimum. However, an exhaustive search is not computationally feasible given the large number of features in our case.

In our experiments, Shannon's entropy, based on category distributions in cluster evaluation, is our selected criterion for feature selection (Celeux and Govaert, 1991). The number of clusters is set equal to the number of categories, with the cluster centroid as the centroid for words in the category. Each word is assigned to the cluster with the closest cluster centroid (Tan et al., 2005, Ch. 8, p. 549). The entropy is highest if each cluster contains only one word from each category, and lowest when each cluster contains only words from one category. In our feature selection task, we measure a single category against all other categories and have only two clusters. We then compare the performance of provided categories (class labels) and groups of words selected randomly.

3 Data and preprocessing

Our data consists of all the documents in the English Wikipedia[4] that are over 2k in size after removal of the wikimedia tags. In preprocessing, all words are lowercased and punctuation is removed, except for hyphens and apostrophes. The vocabulary consists of 200 000 most frequent types. The context vocabulary used to build the word vectors consists of 5 000 most frequent types. The total number of times the 200 000 types occur together in the corpus is slightly over 326 million, and the least frequent word occurs 26 times in the corpus. The least frequent context word occurs 6 803 times in the corpus.

In previous work, small windows have corresponded the best to paradigmatic relationship between words, which is what most of the direct evaluation tests described above measure. This why we use a small window of three ($l = 3$) in the experiments, which corresponds to capturing the previous and the following word around a target word. The word frequency in the whole corpus biases the raw co-occurrence frequency counts. Several weighting schemes that dampen the effect of the most frequent words have been proposed. We use positive point-wise mutual information (PPMI) weighting (Niwa and Nitta, 1994), which gave best results in the evaluation tests by Bullinaria and Levy (2007).

[4]The October 2008 edition, which is no longer available at the Wikipedia dump download site http://dumps.wikimedia.org/enwiki/

4 Experiments and results

4.1 Direct VSM evaluation

We first validate our model with the syntactic and semantic tests (Semcat, Syncat1, Syncat2, TOEFL, related word pairs (Distance), and Deese antonyms (Deese)). The results for the semantic and syntactic tests are summarized in Table 1. The results for the our VSM for the semantic categorization, distance, and TOEFL tests are in line with results for the same number of features in (Bullinaria and Levy, 2012). The SENNA results for the syntactic tests beat our simple system and even though the training does not contain semantic information, the results in the semantic test are only slightly worse than our results, which may also be due to the fact that the system has a larger context window. We also reduced the dimensionality of the vector space to 50 with SVD and ICA, and performed the testing. Using only 50 dimensions is not enough to capture all the meaningful information from the 5 000 original dimensions, compared to the 50 dimensions of SENNA, but the SVD and ICA results can be produced in about 10 minutes, whereas SENNA training takes weeks (Collobert et al., 2011). To obtain results equivalent to those with the 5000 features, we tested a growing number of features: using approximately 500 ICA or SVD components would be enough. We repeated the experiments for the Semcat test and only took those word vectors that represented the 530 words of the test set. In this case, ICA and SVD are able to better represent the dimensions of the interesting subset instead of the whole vocabulary of 200 000 words, which makes the error decrease considerably.

ICA and SVD perform equally well in dimensionality reduction. This is not surprising, as ICA can be thought as (1) whitening and dimensionality reduction followed by (2) an ICA rotation, where (1) is usually computed with PCA. As (2) does not change distances or angles between vectors, any method that utilizes distances or an-

gles between vectors finds very little difference between PCA and ICA. As PCA and SVD are closely related, the same reasoning can be applied there. (Vicente et al., 2007)

4.2 Feature selection

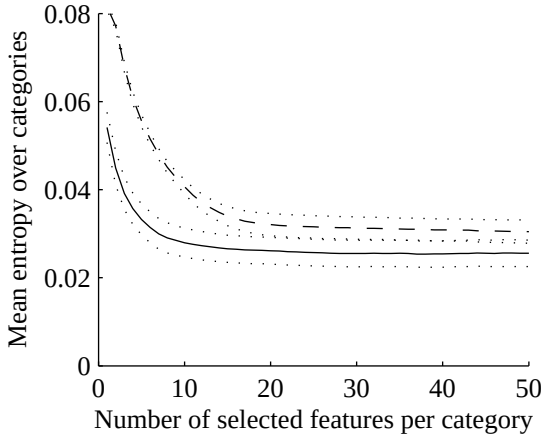

Figure 1: The average means with 95% confidence intervals for entropy for the 53 semantic categories (lower curve) and random categories (upper curve).

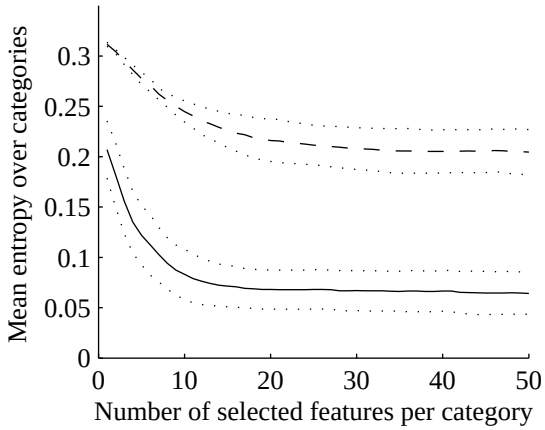

Figure 2: The average means with 95% confidence intervals for entropy for the 10 syntactic categories of Syncat1 (lower curve) and random categories (upper curve).

Figures 1, 2 and 3 show the results for the semantic and syntactic feature selection experiments. To help the visualization, only the confidence intervals around the mean of the semantic/syntactic categories and random categories are shown. The results indicate that each category can be easily separated from the rest by a few features

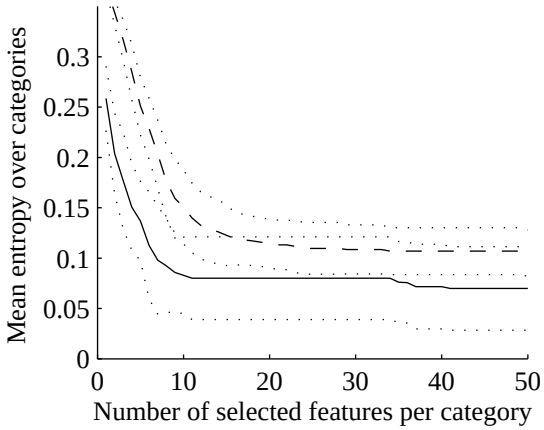

Figure 3: The average means with 95% confidence intervals for entropy for the 7 syntactic categories of Syncat2 (lower curve) and random categories (upper curve).

only. The randomly constructed categories cannot be as easily separated, but the difference to semantic categories diminishes as more features are added.

Syncat1, where each category corresponds to a single part of speech, gives very good separation results, whereas the separation in the case of Syncat2 is less complete. The separation of adjective, adverb and determiner categories cannot be distinguished from the results of random categories. Looking more closely at the most common words labeled as adjectives, we see that most of them can be also used as adverbs, nouns or even verbs, e.g., *first*, *other*, *more*, *such*, *much*, *best*, and unambiguous adjectives e.g., *british*, *large*, *early* are a minority. Similar reasoning can be applied to the adverbs. Determiners, on the other hand, may exist in so many different kind of contexts, and finding few context words (i.e. features) to describe them might be difficult.

We also looked at the first context word which was selected for each semantic category, and in 40 cases out of 53 the first selected feature was somehow related to the word of the category. We found out that it was either a semantically related noun or verb: A PRECIOUS STONE:*ring*, A RELATIVE:*adopted*; the name or part of a name for the category: A CRIME:*crime*, AN ELECTIVE OFFICE:*elected*; or a word that belongs to that category: A KIND OF CLOTH:*cotton*, AN ARTICLE OF FURNITURE:*table*.

	5 000 feat	ICA50	SVD50	SENNA
Semcat	0.22	0.31/0.19	0.32/0.19	0.25 (R=0.98)
Syncat1	0.17	0.25	0.25	0.10
Syncat2	0.26	0.38	0.37	0.21
TOEFL	0.22 (R=0.95)	0.38 (R=0.95)	0.38 (R=0.95)	0.34 (R=0.91)
Distance	0.11	0.19	0.19	0.11
Deese	0.07	0.12	0.13	0.04

Table 1: The error, $Err = 1 - P$ for different test sets and data sets. Recall is 1 unless otherwise stated. For the Distance test the values reported are mean values over 50 runs. For ICA and SVD, first values for Semcat report the error when the dimensionality was reduced from the full $200\,000 \times 5\,000$ whereas the second values are calculated for the Semcat subset $530 \times 5\,000$ only.

4.3 Feature selection with ICA

As previous results suggest (Honkela et al., 2010), ICA can produce components that are interpretable. An analysis using 50 independent components obtained by ICA was carried out for the vector representations of the 530 words in the Semcat data set, and forward feature selection was then applied for each of the categories. The results are shown in Fig. 4. The semantic categories separate better than the random categories with only a few features in this experiment as well, but as more features are added, the difference decreases. In Fig. 4 we show the reader also some examples of the semantic categories for which the entropy is smallest and largest.

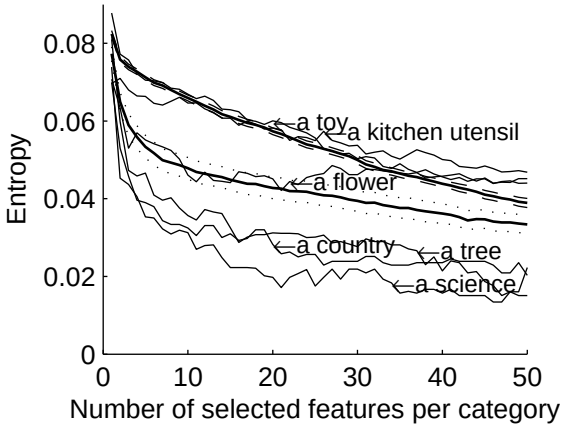

Figure 4: Entropy for sample semantic categories using ICA components in feature selection, mean entropy for random categories with 95% confidence interval shown on dashed lines and mean entropy over semantic categories with 95% confidence interval shown with dotted lines over random categories.

In this experiment, we used 10-fold stratified cross validation, choosing 90 % of the 530 words as the training set and the rest was used as the test set. I.e., when separating one category from all the rest, 9 words were used as the representative of the tested category, 468 words as the representative of the other categories from which we separate, and the remaining 10% as the test data set in the same relation. Reported results are the averaged entropy over the different folds of the cross validation. Between different folds, a few same features were always chosen first, after which there was considerable variation. This seems to indicate that the first 2-3 features selected are the most important when a category is separated from the rest. The results in case of the SENNA data and SVD, left out due to space constraints are similar.

4.4 Semantic information from independent components

Earlier research has shown that ICA is able to produce cognitively meaningful components (Hansen et al., 2005) that correspond to, for instance, noun concepts (Chagnaa et al., 2007), phonological categories (Calderone, 2009), personal traits (Chung and Pennebaker, 2008) and syntactic categories (Honkela et al., 2010). In this work, we studied the interpretability of the ICA components in a semantic task, using the Semcat set, and compare the results to those obtained by SVD and SENNA. For this, we study the activations of the components in the S-matrix. For each component, we take a group of ten words, for which the component has the highest activation

and compare that group to the given categories.

The ICA components are usually skewed in one direction. To see which words have most activation, it is often enough to check the skewness of the distribution. We then selected the 10 words for which the component activation was largest in that direction. We applied two criteria, *strict* and *lax* (similar to Sahlgren (2006)) to see how well the components corresponded to the semcat category labels. To fill the *strict* criterion a minimum of $9/10$ words should belong to the same category, and the *lax* criterion is defined as majority, i.e., a minimum of $6/10$ words must belong to the same category.

The selection of the subset of word vectors can be done either before or after dimensionality reduction. It is obvious that if ICA or SVD is applied after the subset selection (subset+ICA), the components are a better representation of only those words, than if the subset selection is carried out after a dimension reduction for the complete matrix of 200 000 word vectors (ICA+subset). As FastICA produces results which may vary from a run to another due to random initialization, we verified that the results stayed consistent on subsequent runs.

The results are shown in Table 2. In subset+dimensionality reduction case, ICA is able to find 17 categories out of 53 with the strict criterion, and 37 categories with the lax criterion. Those categories that filled the strict criterion had also the smallest entropy in the feature selection described earlier. We repeated the above-described analysis evaluating separately the 10 smallest and 10 largest values of each SVD component. SVD found found only two categories which passed the strict test. For the relaxed condition, 19 categories passed. In the dimensionality reduction+subset case, ICA is slightly better with the lax criterion, whereas SVD finds three categories with the strict criterion. These results can also be compared to SENNA results, in which no categories passed the strict test, and 4 categories passed the lax test.

As a separate experiment for subset + dimension reduction, we checked whether the features that best represented these categories were also those that were first selected by the feature selection algorithm. We found out that for 15 categories, the most prominent feature was also selected first and for the two remaining categories, the feature was selected second. For the relaxed condition, for the 37 categories, the best feature was selected first in 27 cases and second in 6 categories, and third in 2 categories.

	Strict	Lax
subset+ICA	17/53	37/53
ICA+subset	1/53	12/53
subset+SVD	2/53	19/53
SVD+subset	3/53	8/53
SENNA	0/52	4/52

Table 2: Fraction of categories which filled the strict and lax condition for ICA, SVD and SENNA

A further analysis of the categories that failed the relaxed test suggests several reasons for this. A closer analysis shows that words from certain categories tend to occur together, and these categories contain a common superordinate category: For example NONALCOHOLIC and ALCOHOLIC BEVERAGES are all beverages. Among the top 10 activations of a component, there are five words from each of these categories, and among 20 highest activations for this component, 18 of them come from one of these two categories. Similarly, a component shows high activations for words that form the female half of the RELATIVE: *aunt, sister, mother*, together with *daisy, tulip, rose, lily* from FLOWER, which are also used as female names. There are more overlapping categories in which words may also belong to another category than for which they are assigned, for example TYPE OF DANCE and TYPE OF MUSIC; SUBSTANCE FOR FLAVORING FOOD and VEGETABLE; the TYPE OF SHIP and VEHICLE; and FOOTGEAR and ARTICLE OF CLOTHING and the TYPE OF CLOTH. There results are in line with the earlier result with the feature selection, where unambiguous categories separated better than the more ambiguous ones. There are four categories which cannot be described by a single component in any way: KITCHEN UTENSIL, ARTICLE OF FURNITURE, CARPENTER'S TOOL and TOY. The words in these categories have activations in different ICA components, which suggests that the most common usage is not the one invoked by the given category. For example, in TOY category,

words *bicycle* and *tricycle* go with the words from VEHICLE and *block* has an active feature which also describes PARTS OF A BUILDING or FURNITURE.

5 Discussion

The semantic category test set is based on studies of human similarity judgment, which for even with a large group of responses, is quite subjective. The ICA analysis shows that meaning of the surface forms for some words based on the corpus data are different than the one it is labeled with. For example, *bass* from the FISH category had a strong activation for a feature which represented MUSICAL INSTRUMENT. This is obviously a downside of our method which relies on the bag-of-words representation without taking into account the sense of the word.

We saw that the dimension reduction applied as drastically as we did worsens the evaluation results considerably. The 50-dimensional feature vectors of SENNA produce better results in many of the tasks excluding the TOEFL and the semantic categorization, but definite conclusions on the performance cannot be made, as the SENNA is not trained with the exactly same data. Another downside of SENNA is the very long training. In this paper, we opted to have the simplest word space without taking into account word senses, elaborate windowing schemes or such. The current paper does not address the feature selection as a means for reducing the dimensionality as such, but it is an interesting direction for future work. Karlgren et al. (2008) suggest studying the local dimensionality around each word, as most vectors in a high-dimensional vector space are in an orthogonal angle to each other. We found out that the first features are most important in representing a semantic category of 10 words, and an unreported experiment with 300 ICA features showed that the features included last had a negative impact to the separation of the categories. Cross-validation results showed that the selected ICA features were also useful with a held-out set.

6 Conclusions

This paper describes direct evaluation tests for word vector space models. In these tests, ICA and SVD perform equally well as dimensionality reduction methods. Further, the work shows that only a small number of features was needed to distinguish a group of words forming a semantic category from others. Our experiments with the random categories show that there is a clear difference between the separability between most of the semantic categories and the random categories. We found the gap surprisingly large.

Some of the semantic categories separated very badly, which were analyzed to stem from differences in frequency for the different senses of the word collocations. Our premise is that a good latent word space should be able to separate different cognitively constructed categories with only a few active components, which is related to the sparse coding generated by ICA. Further, we have shown that we could find interpretable components that matched semantic categories reasonably well using independent component analysis. Compared to SVD, ICA finds a fixed rotation where the components are also maximally independent, and not only uncorrelated. This facilitates the analysis of the found structure explicitly, without relying on implicit evaluation methods. The interpretability of the ICA components is an advantage over SVD, demonstrated by the quantitative *strict/lax* evaluation.

The main motivation of this work is to support the development towards automatic processes for generating linguistic resources. In this paper, we focus on independent component analysis to generate the sparse linguistic representations, but similar conclusions can be made with closely related methods, such as non-negative matrix factorization (NMF).

Acknowledgments

We gratefully acknowledge the support from the Department of Information and Computer Science at Aalto University School of Science. In addition, Tiina Lindh-Knuutila was supported by the Finnish Cultural Foundation and Jaakko J. Väyrynen and Timo Honkela were supported by the META-NET Network of Excellence.

References

W.F. Battig and W.E. Montague. 1969. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology Monograph*, 80(3, part 2.):1–45.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

J.A. Bullinaria and J.P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

J.A. Bullinaria and J.P. Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44.

B. Calderone. 2009. Learning phonological categories by independent component analysis. *Journal of Quantitative Linguistics*, 17(2):132–156.

G. Celeux and G. Govaert. 1991. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8:157–176.

A. Chagnaa, C.-Y. Ock, C.-B. Lee, and P. Jaimai. 2007. Feature extraction of concepts by independent component analysis. *International Journal of Information Processing Systems*, 3(1):33–37.

C. K. Chung and J. W. Pennebaker. 2008. Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Research in Personality*, 42(1):96–132.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 12:2493–2537.

P. Comon. 1994. Independent component analysis—a new concept? *Signal Processing*, 36:287–314.

J.E. Deese. 1954. The associative structure of some common English adjectives. *Journal of Verbal Learning and Verbal Behavior*, 3(5):347–357.

G. Grefenstette. 1992. Finding the semantic similarity in raw text: The Deese antonyms. Technical Report FS-92-04, AAAI.

Lars Kai Hansen, Peter Ahrendt, and Jan Larsen. 2005. Towards cognitive component analysis. In *Proceedings of AKRR'05, International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, pages 148–153, Espoo, Finland, June.

T. Honkela, A. Hyvärinen, and J.J. Väyrynen. 2010. WordICA — emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16:277–308.

A. Hyvärinen and E. Oja. 1997. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492.

A. Hyvärinen, J. Karhunen, and E. Oja. 2001. *Independent Component Analysis*. John Wiley & Sons.

J. Karlgren, A. Holst, and M. Sahlgren. 2008. Filaments of meaning in word space. In *Proceedings of the European Conference on Information Retrieval*, pages 531–538.

I. Kivimäki, K. Lagus, I.T. Nieminen, J.J. Väyrynen, and T. Honkela. 2010. Using correlation dimension for analysing text data. In *Proceedings of ICANN 2010*, pages 368–373. Springer.

T.K. Landauer and S.T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

J.P. Levy, J.A. Bullinaria, and M. Patel. 1998. Explorations in the derivation of semantic representations from word co-occurrence statistics. *South Pacific Journal of Psychology*, 10:99–111.

M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Y. Niwa and Y. Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 304–309.

M. Patel, J.A. Bullinaria, and J.P. Levy. 1997. Extracting semantic representations from large text corpora. In *Proceedings of Fourth Neural Computation and Psychology Workshop: Connectionist Representations*, pages 199–212. Springer.

M. Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis, Stocholm University, Department of Linguistics.

H. Schütze. 1993. Word space. In *Advances in Neural Information Processing Systems 5*, pages 895–902. Morgan Kaufmann.

A. Sorjamaa. 2010. *Methodologies for Time Series Prediction and Missing Value Imputation.* Ph.D. thesis, Aalto University School of Science.

P.-N. Tan, M. Steinbach, and V. Kumar. 2005. *Introduction to Data Mining*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

P.D. Turney and P. Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

P.D. Turney. 2001. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings*

of the Twelfth European Conference on Machine Learning (EMCL2001), pages 491–502, Freiburg, Germany. Springer-Verlag.

P.D. Turney. 2005. Measuring semantic similarity by latent relational analysis. In *Proceedings of the 19th International Joint conference on Artificial intelligence (IJCAI-05)*, pages 1136–1141.

M. A. Vicente, P. O. Hoyer, and A. Hyvärinen. 2007. Equivalence of some common linear feature extraction techniques for appearance-based object recognition tasks. *IEEE Transactions Pattern Analusis Machine Intelligence*, 29(5):896–900.