Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Contex-Based Features for Semantic Mapping of Words

G. Pilato, G. Vassallo, G. Di Fatta, S. Gaglio

RT-ICAR-PA-03-09 **dicembre 2003**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: www.icar.cnr.it
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: www.na.icar.cnr.it
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: www.pa.icar.cnr.it

Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Contex-Based Features for Semantic Mapping of Words

G. Pilato[2], G. Vassallo[2], G. DiFatta[1], S. Gaglio[1 2]

Rapporto Tecnico N.09: **RT-ICAR-PA-03-09**

Data: **dicembre 2003**

[1] Istituto di Calcolo e Reti ad Alte Prestazioni, ICAR-CNR, Sezione di Palermo Viale delle Scienze edificio 11 90128 Palermo

[2] Università degli Studi di Palermo Dipartimento di Ingegneria Informatica Viale delle Scienze 90128 Palermo

Contex-Based Features for Semantic Mapping of Words

Giovanni Pilato[1], Giorgio Vassallo[2],
Giuseppe Di Fatta[1], and Salvatore Gaglio[1,2]

[1] Istituto di CAlcolo e Reti ad alte prestazioni
Italian National Research Council
Viale delle Scienze - 90128 Palermo - Italy
{pilato,difatta}@pa.icar.cnr.it
[2] Dipartimento di ingegneria INFOrmatica
University of Palermo
Viale delle Scienze - 90128 Palermo - Italy
vassallo@csai.unipa.it, gaglio@unipa.it

Abstract. It is introduced a semantic coding of words based the SVD technique for generating a semantic space in which words are then mapped. The proposed technique is based on the introduction of a *link energy* formula related to digrams frequency. To test the effectiveness of the porposed approach, the english translation of the set of Grimm Fairy Tales has been analyzed and a bi-dimensional visual representation has been obtained using the Sammon projection algorithm.

1 Introduction

Information retrieval systems are built from large document collections in order to organize and access information. Large text corpora are used for many purposes in computational linguistics and natural language processing.
Computational lingustics has recently pointed out that there is a close relation between underlying lexical-semantic structures and their associated synctactic behaviors[]. The information extracted from the text is built into some mathematical model: a typical one is the vector-space model.
The simplest method to associate uncorrelated codes to words is to assign a unit vector for each token.[7] However this method is not manageable when a large number of words has to be considered.
An alternative approach, called *latent semantic analysis*, has been introduced[9]. This methodology is based on co-occurrence statistics to generate word vectors that can be used to calculate similarity between words, or documents.
A term-document matrix is typically used to generate vector space models. The generic element of this matrix is simply the number of times each word occurs in each document. The rows of this matrix are usually interpreted as vectors representing words, while documents are usually represented by vectors calculated as a weighted sum of the vectors associated to the words appearing in the document.
This method usually leads to high-dimensional and very sparse term-document matrices[18]. It is therefore necessary to project each word onto a smaller subspace which

gives the best least-squares approximation to the original data. This goal is usually reached using the singular-value decomposition methodology. The result is to represent each word using the n most significant *latent variables*: this leads to the denomination of *latent semantic analysis*[9].
An alternative to the traditional term-document matrix has been proposed by [3] for the purpose of measuring semantic similarity between words. This technique, called Hyperspace Analogue to language uses a words-by-words matrix: the rows of the co-occurrence matrix can be interpreted as context vectors for the words in the vocabulary. In [12] a different methodology to construct the co-occurrence matrix has been developed. The technique, called Random indexing, uses distributed representations to accumulate context vectors from the distributional statistics of words. This is accomplished by first assigning a unique high-dimensional sparse random index vector to each word type in the text data. Then, every time a word occurs in the text data, it is added the index vectors to the n surroundings words to the context vector for the word in question.
In [7] it has been introduced a technique for encoding words in such a manner that the *i*-th word in a sequence of words was represented by a 270-dimensional real vector with random number components. The SOM algorithm has been used for creating word category maps describing relations of words based on their contexts. The result of the processing showed that interrelated words which have similar contexts appear close to each other on the map.
In this paper we introduce a semantic coding of words based on the assumption that the context of the words contains all the needed information. The method differs from the one of Kohonen: in that paper, Kohonen starts by a set of quasi-orthogonal vectors; in the proposed approach it is used the SVD technique for generating a semantic space in which words are then mapped.
The proposed technique uses a set of raw documents without any prior synctactic or semantic categorization of the words. A square matrix M will then built: both in the rows and in the columns there will be all the words w_k present in the document corpus. This matrix will code the relationship between a word w_i and its successive w_j in the sentences of the document corpus. A link energy formula will be introduced to establish the prediction capability of a word on the other one.
The Singular Value Decomposition is then applied, leading to the construction of a semantic space based on the the data-driven extraction of latent semantic information.
To test the effectiveness of the porposed approach, the set of Grimm Tales has been analyzed, then the 150 most frequent words have been mapped in this space, and a bi-dimensional visual representation has been obtained using the Sammon projection algorithm[13].
The following of the paper is organized as follows: in section 2 theoretical background concerning the SVD technique will be illustrated, in section 3 will be outilined the proposed approach, in section 4 experimental results will be described, and in section 5 conclusion will be given.

2 Theoretical Background: the SVD Technique

Latent Semantic Analysis (LSA) [9] is a paradigm to extract and characterize the meaning of words by statistical computations applied to a large corpus of texts. LSA is based on the *vector space method*: a text corpus is represented as a matrix A where rows are related to words, while columns are associated to documents or other contexts.
A truncated singular value decomposition (SVD) [9] is used to estimate the strucutre in word usage across documents. The technique can be outlined as follows: Let m be the number of words contained in all the n documents relevant to some domain of interest and composing the training corpus. Given an $m \cdot n$ matrix M, where, without loss of generality, $m \geq n$ and $rank\,(M) = r$, the singular value decomposition of M, denoted by SVD(M), is defined as:

$$M = U\Sigma V^T \tag{1}$$

where $U^TU = V^TV = I_n$ and $\Sigma = diag\,(\sigma_1, \cdots, \sigma_n)$, $\sigma_i > 0$ for $1 \leq i \leq r$, $\sigma_j = 0$ for $j \geq r + 1$.
The first r columns of the orthogonal matrices U and V define the orthonormal eigenvectors associated with the r non zero eigenvalues of MM^T and M^TM respectively.
The LSA paradigm defines a mapping between the n words and the m documents and a continuous vector space S, where each word w_i is associated to a vector u_i in S, and each document d_j is associated a vector v_j in S[18]..

3 The proposed solution

It is well known that semantic roles are reflected by the contexts in which they occur. It has been shown that contextual relations or roles of words are also statistically reflected in unrestricted natural expressions[7].
The idea is to start from a collection of documents, exploiting the contextual roles of words, i.e. their usage in short contexts formed by adjacent words, to define a syntactic and semantic space in which each word is therefore coded as a vector of reals. The main characteristic of this synctatic-semantic space is that words will be mapped both according to their synctatic role (i.e. nouns, verbs, adjectives will be near points in this space) and *also* according to their meaning (i.e. motion verbs, related nouns will be close n-dimensional points in the generated space).
The proposed technique is based on the extraction of sentences in a set of generic documents and the analysis of the occurrence of digrams in the extracted phrases. A matrix M is then generated from the available text data by analyzing each phrase of the raw text, hence, the Singular Value Decomposition is applied, leading to the construction of a semantic space based on the the data-driven extraction of latent semantic information according to the LSA paradigm: words will be mapped in this space, and a bi-dimensional visual representation will be given.

3.1 Source data

The source data consisted of a set of raw documents without any prior synctactic or semantic categorization of the words. The language can be considered arbitrary chosen and not formal by any means.

3.2 Preprocessing

The preprocessing phase can be outlined in the following steps:

- the texts of all documents are concatenated into one file.
- A set of very common words, called stopwords [**?**], which do not carry information, has been removed from the text.
- The text has been transformed in a list of phrases, each one extracted by it using the punctuation and the carriage-return as a delimiter: therefore each phrase will constitute a sort of "micro-document" to analyze and form which digrams will be extracted.

3.3 Learning process

We want to realize a semantic coding of words using the left and right context of a word in documents corpora. The fundamental assumption is that the context of the words contains all the needed information. The method differs from the one of Kohonen: in that paper, Kohonen starts by a set of quasi-orthogonal vectors; in the proposed approach it is used the SVD technique for generating a semantic space in which words are then mapped. A square matrix M will be built: both in the rows and in the columns there will be all the words w_k present in the document corpus. This matrix will code the relationship between a word w_i and its successive w_j (and, as a consequence, the word and its precedent looking at the columns of the matrix M). The question is how to define the generic element of the matrix M. Let m_{ij} the (i, j) element of the M matrix: m_{ij} is related to the occurrence of the digram $w_i w_j$. Trivially choosing m_{ij} as the number of occurrences of the digram $w_i w_j$ in the text is not the best way to identify the "prediction capability" of the word w_i on the word w_j: it is necessary to find a function that can give a better information. The goal is to define some sort of "link energy" about the prediction capability of a word on another one. As an example, the following are some of the digrams present in the previous sentence; for clarity, also stop-words (like the) have also been considered:

- *The goal*
- *goal is*
- *is to*

If $|w_i|$ is the number of occurrences of the word w_i, $|w_j|$ is the number of occurrences of the word w_j, and $|w_i w_j|$ is the number of times in which the word w_j follows the word w_i, then:

$$\frac{|w_i w_j|}{|w_i|} \tag{2}$$

is the probability that the word w_j will follow the word w_i (i.e. $\frac{|thecat|}{|the|}$ is the probability that the word cat will follow the word the, which is small, given that the number of occurrencies of the word the is typically high. The quantity:

$$\frac{|w_i w_j|}{|w_j|} \tag{3}$$

is the probability that the word w_i will precede the word w_j (i.e. $\frac{|the\ cat|}{|cat|}$ is the probability that the word the will precede the word cat, which is high, given that the number of occurrencies of the word cat is typically low. The goal is to define a quality index which evaluates the "best" permutation among a set of words, given a known digram frequency, which will lead to the max extent in a digram which is commonly used in natural language. A natural choice is therefore to define the following formula:

$$m_{ij} = max\left(\frac{|w_i w_j|}{|w_i|}, \frac{|w_i w_j|}{|w_j|}\right) \tag{4}$$

The proposed approach is different by the one used in [7]: in that work quasi-orthogonal vectors have been used, while here it is used the SVD technique to generate a semantic space in which words are then mapped. The technique also differs from the classical LSA techniques, where a matrix of word-documents is used. Here the technique is to generate a word-word matrix. The matrix is non symmetric, in fact the occurrence of the digram *"the dog"* is different from the occurrence of the digram *"dog the"* in the common use of the words. The algorithm was able to create diagrams that seem to comply reasonably well with the traditional synctactical categorizations and human intuition about the semantics of the words.

4 Experimental Results

To test the effectiveness of the proposed approach, a set of English translations of Grimm's Fairy tales has been used. The corpus is composed by over 7000 words. The language is not formal and the strenght of its use is the fact that it is used a set of sentences effectivbely used in the natural language, rather than using a set of artificially built sentences. The results of the processing are shown in fig. 4. The general organization of the map reflects both synctactical and semantical categories. Formation of syntactic categories on the map can be explained by sentential context. The context of a word is dependent on the syntactical restrictions that govern the position of the words in the text. In particular, figure 4 shows some semantic associated words, like *heart, head, eves, hand*, which are all nouns indicating body parts, *sat, lay, fell*, which are past participles of closely related verbs; other significant groups are shadowed in figure.

5 Conclusions

A semantic coding of words based the SVD technique to generate a semantic space in which words are then mapped has been presented. The introduction of a *link energy* formula, related to digrams frequency in texts, leads to an effective sub-symbolic coding of words. As a case-study, the english translation of the set of Grimm Fairy Tales has been analyzed and a bi-dimensional visual representation has been obtained using the Sammon projection algorithm.
Experimental results show that the general organization of the obtained map reflects both synctactical and semantical categories.

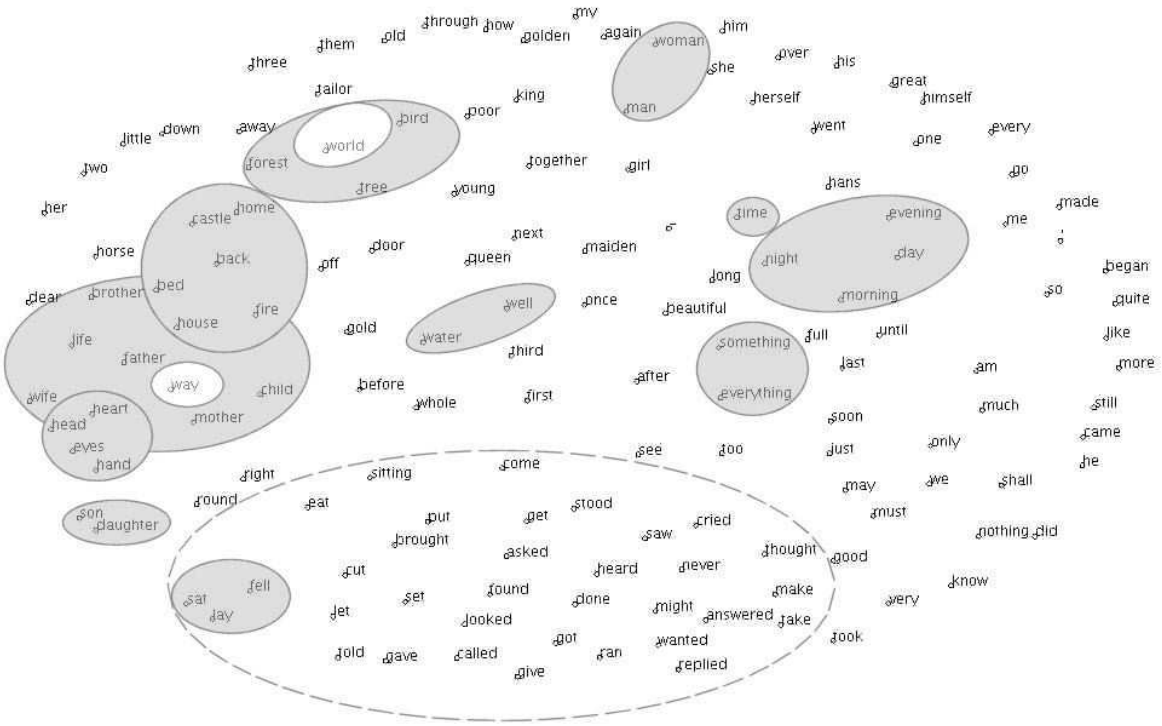

Fig. 1. Procedure of encoding the *k*-th synset of the *i*-th lexical set

Acknowledgments

Authors would thank Engineering Ingegneria Informatica SpA for its partial contribution to this research within the Teschet project.

References

1. J.R. Bellegarda. Aug 2000. *Exploiting latent semantic information in statistical language modelling*, volume 88(8), pages 1279-1296. Proceedings of the IEEE.
2. M. Berry, T. Do, G. O'Brien, V. Krishna, S. Varadhan. 1993. *SVDPACKC (version 1.0) User's Guide*. http://www.netlib.org/svdpack/index.html.
3. C. Burgess, K. Lund. 2000. *The Dynamics of Meaning in Memory*. Cognitive dynamics: Conceptual and Representational Change in Humans and Machines. E. Dietrich and A. Markman, Hillsdale, N.J, Lawrence Erlbaum Associates.
4. J. Didion. 2002. *JWNL (Java WordNet Library*. http://www.sourceforge.net.
5. D. Gildea and T. Hofmann. 1999. *Topic-based language modeling using EM*. Proc. 6th Eur. Conf. Speech Commun. Technol., vol. 5, Budapest, Hungary, pp. 2167-2170.
6. T. Hofmann. 2000. *Learning the Similarity of Documents: An Information-Geometric Approach to Document Retrieval and Categorization*. 914-920. Advances in Neural Information Processing Systems, S.A. Solla, T.K. Leen and K.R. Muller (eds). pp.914-920, MIT press.
7. T. Honkela, V. Pulkki, T. Kohonen. 1995. *Contextual Relations of Words in Grimm Tales, Analyzed by Self-Organizing Map*. Proceedings of International Conference on Artificial Neural Networks, ICANN-95. F. Fogelman - Soulie and P. Gallinari (eds.). EC2 et Cie, (Paris, 1995) 3-7.
8. T.G. Kolda, D.P. O'Leary. 2000. *Computation and Uses of the Semidiscrete Matrix Decomposition*. Trans. Math. Software.
9. T.K. Landauer, P.W. Foltz, D. Laham. 1998. *Introduction to Latent Semantic Analysis*. Discourse Processes, vol 25, pp.259-284.
10. G.A. Miller, R. Beckwidth, C. Fellbaum, D. Gross, K.J. Miller. 1990. *Introduction to WordNet: An On-line Lexical Database*. International Journal of Lexicography, vol. 3(4), pp.235-244.
11. D. Mladenic, J. Institute. 1999. *Text learning and related intelligent agents: A survey*. IEEE Intelligent Systems, pages 44–54, 1999.
12. M. Sahlgren, J. Karlgren, R. Cster, T. Jrvinen. 2002. *SICS at CLEF 2002: Automatic Query Expansion Using Random Indexing*. The CLEF 2002 Workshop, September 19-20, 2002 , Rome, Italy.
13. J.W. Sammon Jr. 1969. *A Nonlinear Mapping for Data Structure Analysis*. IEEE Transactions on Computers, Vol. C-18, no. 5. (May 1969) 401-409
14. F. Sebastiani. 2002. *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1), pages 1-47.
15. V. Siivola. 2000. *Language modeling based on neural clustering of words*. IDIAP-Com 02, Martigny, Switzerland.
16. G. Siolas, F. d'Alche-Buc. *Support Vector Machines based on a semantic kernel for text categorization*. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks, IJCNN 2000. Vol.5, pp.205 -209.
17. K.R. Sloan Jr., and S.L. Tanimoto. 1979. *Progressive Refinement of Raster Images*. IEEE Transactions on Computers. Vol.28(11), pp.871-874.
18. D. Widdows, S. Cederberg, B. Dorow. 2002. *Visualisation Techniques for Analysing Meaning*. Fifth International Conference on Text, Speech and Dialogue. Brno, Czech Republic, September 2002. pp. 107-115.

19. H. Yang, C. Lee. 2000. *Automatic category generation for text documents by self-organizing maps*. Proc. of IEEE-INNS-ENNS International Joint Conference on Neural Networks. Vol.3, pp.581 -586.