Selforganizing classification on the Reuters news corpus

Stefan Wermter
The Informatics Centre
School of CET
University of Sunderland
St. Peter's Way, Sunderland SR6 0DD
United Kingdom
Stefan.wermter@sunderland.ac.uk

Chihli Hung[1]
The Informatics Centre
School of CET
University of Sunderland
St. Peter's Way, Sunderland SR6 0DD
United Kingdom
Chihli.hung@sunderland.ac.uk

Abstract

In this paper we propose an integration of a selforganizing map and semantic networks from WordNet for a text classification task using the new Reuters news corpus. This neural model is based on significance vectors and benefits from the presentation of document clusters. The Hypernym relation in WordNet supplements the neural model in classification. We also analyse the relationships of news headlines and their contents of the new Reuters corpus by a series of experiments. This hybrid approach of neural selforganization and symbolic hypernym relationships is successful to achieve good classification rates on 100,000 full-text news articles. These results demonstrate that this approach can scale up to a large real-world task and show a lot of potential for text classification.

Introduction

Text classification is the categorization of documents with respect to a set of predefined categories. Traditional neural techniques for classification problems cannot present their results easily without adding extra modules but selforganizing memory networks (SOM) are capable of combining topological presentation with neural learning. We extract suitable relations from WordNet to present a semantic map of news articles and show that these relations can complement neural techniques in text categorization. This integration of SOM and WordNet is proposed to deal with the text classification of news articles.

The remainder of this paper is organised as follows. In Section 1, we give a brief review of SOM. Section 2 is dedicated to a description of methods of dimensionality reduction. In section 3 of our hybrid neural approach, the new version of the Reuters corpus and the results of our experiments are presented.

1 Selforganising Memory for Learning Classification

According to the theory and the organisation of biological systems, neurons with similar functions are placed together. Based on this idea, Kohonen proposed SOM (Kohonen 1982). SOM, based on an unsupervised learning principle, can map a multi-dimensional dataset into a low-dimensional space, usually 2-dimensional. SOM learns to place similar data on topologically close areas on the map. Therefore, people can choose the relevant clusters of documents on the map to get relevant documents. However, it is impossible for one map to encompass the continuously growing data source.

In such cases, the categories are often arranged in a hierarchy or an adaptive structure, e.g. Incremental Grid Growing model (Blackmore and Miikkulainen 1993), Growing Cell Structures (Fritzke 1993), Hierarchical SOM (Wan and Fraser 1994), and Adaptive Coordinates (Rauber 1996; Merkl and Rauber 1997). Presentation and explanation are a possibly weakness for most ANN models to text classification. The robustness of the SOM algorithm and its appealing visualization effects

[1] Hung is a lecturer of De Lin Institute of Technology as well.

make it a prime candidate in text classification (Lin *et al*. 1991; Ritter and Kohonen 1989; Honkela 1997).

2 Dimensionality Reduction

VSM (Vector Space Model) is a basic technique to transform text documents to numeric vectors. Often neural networks including the SOM model for text classification apply VSM on their pre-processing stage. SOM does not reduce the length of vectors but only presents the high dimensionality of input vectors by prearranged units on a low dimensional space. Dealing with a huge text collections means dealing with huge dimensionality that needs to be reduced for neural approaches such as SOM (Berry *et al*. 1999).

In the field of linear algebra, PCA (Principal Component Analysis), SVD (Singular Value Decomposition) and Random projection are effective for dimensionality reduction but suffer from two main side effects. The first one is that the results are difficult to interpret and the second one is a reduction of the accuracy.

Rather than introducing hierarchies from SOM we want to exploit existing semantic knowledge, especially here from WordNet. WordNet (Miller, 1985) is a network of semantic relationships between English words. Semantic relations among words construct a network. The sets of synonyms compose synsets, which are the very basic relations in WordNet. Words in the same synset have the same or similar concept and vice versa. In addition to synonymy, there are several different types of semantic relations such as antonymy, hyponymy, meronymy, troponomy, and entailment in each different syntactic category, i.e. nouns, verbs, adjectives and adverbs. This semantic dictionary is useful in extracting the real concept of a word, a query or a document in the field of text mining (Richardson 1994; Richardson and Smeaton 1995; Voorhees 1993; Voorhees 1998; Scott and Matwin 1998; Gonzalo *et al*. 1998; Moldovan and Mihalcea 1998; Moldovan and Mihalcea 2000). Using these semantic relations in WordNet, one index word may present its many synonyms, siblings or other relevant words. Therefore, by mapping words to more general concepts, WordNet can be used to reduce the dimensionality.

Instead of using these approaches to reduce multi-dimensional vectors, we apply significance vectors to present the importance of words in each semantic category and use pre-assigned topics as axes of multi-dimensional space. Thus a news article can be represented by a *n*-dimension vector, where *n* is the number of pre-assigned topics. This method offers a way to divert from the huge dimensionality curse. A more detail description is shown in section 3.2.

3 Selforganizing classification on the new Reuters corpus using WordNet

3.1 The New Version of Reuters Corpus

We work with the new version of Reuters corpus (Reuters 2000). This corpus is made up of 984 Mbytes of newspaper articles in compressed format from issues of Reuters between the 20th Aug., 1996 and 19th Aug., 1997. The number of total news articles is 806,791, which contain 9,822,391 paragraphs, 11,522,874 sentences and about 2 hundred million word occurrences.

Each document is saved in a standard XML format and is pre-classified by 3 different codes of categories, which are industry code, region code and topic code. We are currently interested in the topic code only. 126 topics are defined in this new corpus but 23 of them contain no articles. All articles except 10,186 of them are classified in at least one topic.

In our first experiments we concentrate on 8 major topics (Table 1). In order to get a comparison of the performance with and without the use of WordNet and the relation of headlines and full-text news articles, a series of experiments have been performed. First, we use the first 100,000 news headlines for training and another 100,000 news headlines for test. The second experiment is exactly the same as the first one but we use full-text instead of headlines. In the third experiment, we use 100,000 full-text news articles for training and use their headlines for test. The fourth experiment is opposite to the third one. An integration of SOM and WordNet will be presented in last two experiments.

Table 1. The description of chosen topics and their distribution over whole corpus

no	Topic	Description	Distribution
1	C15	performance	149,358
2	C151	accounts/earnings	81,200
3	CCAT	corporate/industrial	372,097
4	E21	government finance	42,573
5	ECAT	economics	116,205
6	GCAT	government/social	232,031
7	GCRIM	crime, law enforcement	32,036
8	GDIP	international relations	37,630

3.2 Presenting Text Documents by Significance Vectors

We use pre-assigned topics as axes of a multi-dimensional space and apply significance vectors to present the importance of words in each semantic category based on (Wermter 2000). Significance vectors are defined by the frequency of a word in different topics. A significance vector is presented with topic elements ($t_1 t_2 \ldots t_j$), where t_j presents the frequency of a word in j semantic category. Thus a document x is presented with:

$$x(w, t_j) = \sum_{i=1}^{n} \frac{\textit{Frequency for word } w_i \textit{ in topic } t_j}{\sum_{j=1}^{m} \textit{Frequency for word } w_i \textit{ in topic } t_j} \tag{1}$$

where n is the number of words and m is the number of topics. This Method1 vector is the summation of significance vectors.

Method 1 can be susceptible to the number of news documents observed in each topic. An alternative *method 2* of vector presentation can alleviate skewed distributions. Thus a document x is modified as:

$$x(w, t_j) = \sum_{i=1}^{n} \left(\frac{\textit{Frequency for word } w_i \textit{ in topic } t_j}{\sum_{j=1}^{m} \textit{Frequency for word } w_i \textit{ in topic } t_j} \times \ln \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \textit{Frequency for word } w_i \textit{ in topic } t_j}{\sum_{i=1}^{n} \textit{Frequency for word } w_i \textit{ in topic } t_j} \right) \tag{2}$$

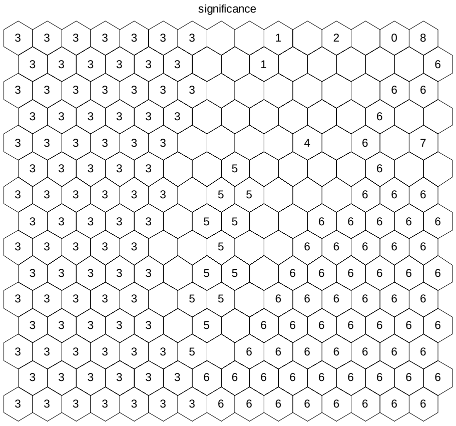
Because only nouns and verbs have the hypernym relation in WordNet and because nouns and verbs convey enough information of document concepts, we remove all words except nouns and verbs found in WordNet in our experiments. We also benefit by a function of WordNet, *morphword*, as a simple stemming tool. After above pre-processing, our 100,000 news article training set represents the total number of 8,920,287 (381,871) word occurrences and the total number of 22,848 (10,185) distinct words in full-text and headline experiments respectively. An example of these vector representation methods is shown in (Table 2). Note that the representation of "to" is the 0-vector since is not shown in nouns and verbs collections of WordNet.

Table 2. Examples of rounded significance vectors on news headline experiment. Topic codes are presented on number 1 to 8 (Table 1).

Word	1	2	3	4	5	6	7	8
Recovery	.13	.05	.33	.02	.29	.13	.04	.01
Excitement	.00	.00	.00	.00	1.0	.00	.00	.00
Brings	.01	.00	.19	.03	.14	.49	.05	.08
Mexican	.03	.01	.19	.02	.16	.42	.14	.01
Markets	.11	.04	.55	.04	.16	.09	.01	.00
To	.00	.00	.00	.00	.00	.00	.00	.00
Life	.16	.09	.39	.01	.04	.23	.07	.02
Method1	.44	.20	1.66	.12	1.79	1.35	.31	.13
Method2	1.01	.57	1.79	.38	3.76	1.85	1.01	.39

3.3 Classification and Presentation using SOM

Our work is based on the SOM algorithm (Vesanto *et al.* 1999). We give each news article a topic label. This label is determined by the most significant weights of topics in an input vector based on one of the above methods. Then input vectors are normalised. After the training process, a label of a map unit is assigned according to the highest number of assigned labels. For example, if 3 news articles of ECAT and 10 news articles of CCAT are mapped to unit 1, then the label of unit 1 will be associated with CCAT. Therefore, all units present their favourite news article labels. We adopt a semi-supervised SOM concept to add an extra semantic vector, x^s, with a small number 0.2 as its highest value to represent the desired class. In our case x^s has 8 elements, as has x. That is, the document vector *d* is represented as *d=[x^s x]*, e.g. [0 0 0 0 0.2 0 0 0 0.44 0.20 1.7 0.12 1.79 1.35 0.31 0.13]. This approach can make the border of SOM units more prominent and also can be used to verify the performance of text classification. A SOM map with 225 output units is shown in (Fig 1) based on classifying these 16 element document vectors. Other architectures (e.g. 25 x25) have been tested and show similar clear results.



SOM 01-Feb-2002

Fig. 1. SOM with 15*15 units. Reuters topic codes are presented on numbers (Table 1)

3.4 Composing Semantic Clusters from WordNet

WordNet physically builds the database according to syntactic categories and semantic relations among synsets. In our work, we use the hypernym-hyponymy relation. A hypernym of a term is a more general term where hyponymy is more specific. For example, an *apple* is a kind of *edible fruit*, so *edible fruit* is a hypernym of *apple* and an *apple* is a hyponymy of *edible fruit*. We use the hypernym relation because the concept of this relation is similar to the definition of news classification.

The concept of a category of news is more general than each distinct news article. News articles with a similar concept will be grouped in a same class, and each group member, i.e. each distinct news article, still has its own specific meaning. We use a 2-level hypernym to replace each word in a news article with its hypernym term in order to get a more general concept of its original word. Only nouns and verbs in WordNet consist of this hypernym relation. Polysemous and synonymous terms can be represented in several synsets and each synset may lie in a different hypernym hierarchy. It is difficult to decide the concept of a document that contains several ambiguous terms. Salton and Lesk give an example that offers a useful approach (Salton and Lesk 1971). The set of nouns *base*, *bat*, *glove*, and *hit* have each their own different senses, but putting them together means the game of baseball clearly. We use this idea and take advantage of synsets' glosses, which are an explanation of the meaning of each concept. Then the correct concept of a term is decided by comparing the similarity of each gloss with the semantic term-topic database of Reuters. For example, the first news article is pre-assigned to topic ECAT. The first term of the headline of this article is *recovery* that consists of 3 senses as Noun and 0 senses as Verb. Thus, there are 3 glosses for this word. We count the number of the co-occurrence of terms shown in each gloss and the pre-assigned term-topic database. Then we average the significance of terms by dividing by the total number of terms in each gloss. Thus, the most significance of the gloss means the most possibility of the sense. Finally every term is replaced by its 2-level hypernym. This approach is successful to reduce the total number of distinct words in the training set by 83.15% and 72.84% in full-text and headline experiments respectively (Table 3). Furthermore, this approach can also offer an easy way to extract a reasonable right word sense for an ambiguous word. We will represent our results in the experiment section.

Table 3. The total number of distinct words in training set with and without the help of WordNet

News source	without	With	reduction
Headline	10,185	2,766	72.84%
Full-text	22,848	3,851	83.15%

3.5 Evaluation Method

The label shown on a trained SOM is a preference and it is possible that several different labels are assigned to the same SOM unit. We consider that every input vector which is mapped to this unit will be reassigned the unit label to replace its original label. In our above example, those 3 news articles lose their label of ECAT and get the unit label of CCAT. Kohonen *et al.* (2000) define the classification error as "*all documents that represented a minority newsgroup at any grid point were counted as classification errors*." Our classification accuracy is very similar to Kohonen's, but we use the corpus itself to verify the performance. If the replaced input vector label matches ONE of the original labels assigned by Reuters, it is a correct mapping. The accuracy is calculated from the proportion of the number of relevant mappings to the number of input news articles. Some news articles have the label 0 because after pre-processing these articles are zero vectors.

3.6 Results of Experiments

3.6.1 Selforganization classification based on News Headline and Full-text

The first 100,000 news articles are used for training and the following 100,000 news articles are used for testing the generality. SOM represents the original distribution of source data so it is important to describe the distribution of data sets (Table 4). Because a news article can be classified in several topics, the distribution over chosen topics is inevitably not even.

Table 4. The distribution of articles from new Reuters corpus over the semantic categories

no	Training Set		Test Set	
	Number	Distribution	Number	Distribution
1	20,448	12.39%	25,810	14.84%
2	10,427	6.32%	13,876	7.98%
3	57,641	34.94%	61,120	35.15%
4	7,034	4.26%	7,061	4.06%
5	18,871	11.44%	19,312	11.11%
6	38,792	23.51%	35,983	20.70%
7	5,317	3.22%	4,588	2.64%
8	6,447	3.91%	6,120	3.52%

We have four experiments in this subsection. In the first experiment, the first 100,000 news titles are used for training and 100,000 successive news titles are used for test. The second experiment is same as the first one but full-text news articles are used instead of headlines only. We then try to use the trained SOM based on full-text news to test the coherence of news title sentences. The fourth experiment is inversely to the third one. The results are shown in Table 5-8 respectively. We find that our significance vector representation methods can achieve high accuracy. Second, even though full-text news articles contain more information than headlines there is no big difference in accuracy for a text classification task. Third, a trained SOM based on news headlines or based on full-text news can be highly generalised. However, the former is more general than the latter. Although the new version of Reuters news corpus is used in this work, this result is similar to the conclusion of Rodríguez *et al.* (1997) who use the old version of Reuters and confirms that the topic headings in Reuters corpus tend to consist of frequent words in the news document itself and this helps the task of news classification.

Table 5. Accuracy on 100,000 news titles for training and test set

Method	Training set	Test set
1	88.85%	87.55%
2	91.07%	89.03%

Table 6. Accuracy on 100,000 full-text news articles for training and test set

Method	Training set	Test set
1	85.70%	85.96%
2	92.77%	92.01%

Table 7. Accuracy on 100,000 full-text news for training and their headlines for test

Method	Full-text for training	Headline for test
1	85.70%	80.81%
2	92.77%	80.18%

Table 8. Accuracy on 100,000 news headlines for training and their full-text news for test

Method	Headline for training	Full-text for test
1	88.85%	84.11%
2	91.07%	89.95%

3.6.2 Selforganization classification with and without the help of WordNet

Our results using 2-level hypernym relation are significant for several reasons. First, we successfully reduce the total number of distinct words from 10,185 to 2,766 (22,848 to 3,851) in our training tests based on news headline and full-text news respectively (Table 3). Second, with the use of WordNet, this hybrid neural technique successfully improves the accuracy of news classification without any loss of categorisation ability (Table 9-10).

Table 9. Accuracy without and with the help of WordNet 2-level hypernym on 100,000 full-text for training set

Method	SOM	SOM with WordNet
1	85.70%	94.21%
2	92.77%	98.95%

Table 10. Accuracy without and with the help of WordNet 2-level hypernym on 100,000 news titles for training set

Method	SOM	SOM with WordNet
1	88.85%	89.94%
2	91.07%	90.65%

Discussion and Conclusion

In the past there had been no consistent conclusions about the value of WordNet for information retrieval tasks (Mihalcea and Moldovan 2000). Experiments performed using different methodologies led to various, sometime contradicting results (Voorhees 1998). This is probably because extracting the concept of a word is seriously dependent on other unambiguous words. Text classification is mapping documents with similar concepts to a cluster with a more general concept.

If a vector label matches ONE of the original labels assigned by Reuters, it is considered a correct mapping. Another test could be to consider a multi-topic a NEW topic. This adds many more classes and topics. In this case, we found 54.29% and 80.51% on 100,000 full-text news articles without and with the help of WordNet respectively, demonstrating the merit of using WordNet even more.

We have demonstrated that it is suitable to use the hypernym relation from WordNet for text classification. We successfully used this relation and improved the text classification performance substantially. By merging statistical neural methods and semantic symbolic relations, our hybrid neural learning technique is robust to classify real-word text documents and allows us to learn to classify above 98% of 100,000 documents to a correct topic.

References

Berry, M.W., Drmac, Z. and Jessup, E.R. (1999). Matrices, Vector Spaces, and Information Retrieval. *SIAM Review*, Vol. 41, No. 2, pp. 334-362.

Blackmore, J. and Miikkulainen, R. (1993). Incremental Grid Growing: Encoding High-Dimensional Structure into a Two-Dimensional Feature Map. In *Proceedings of the IEEE International Conference on Neural Networks* (*ICNN'93*), San Francisco, CA, USA.

Fritzke, B. (1993). Kohonen Feature Maps and Growing Cell Structures – a Performance Comparison. Advances in Neural Information Processing Systems 5, C.L. Gibs, S.J. Hanson, J.D. Cowan (eds.), Morgan Kaufmann, San Mateo, CA, USA.

Gonzalo, J., Verdejo, F., Chugur, I. and Cigarran, J. (1998). Indexing with WordNet Synsets Can Improve Text Retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal

Honkela, T. (1997). Self-Organizing Maps in Natural Language Processing. PhD thesis. Helsinki University of Technology.

Kohonen, T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43, pp.59-69.

Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela A. (2000). Self organization of a massive document collection. In *IEEE Transactions on Neural Networks*, Vol. 11, No. 3, pp. 574-585.

Lin, X., Soergel, D. and Marchionini, G. (1991). A Self-Organizing Semantic Map for Information Retrieval. In *Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp. 262-269, Chicago.

Merkl, D. and Rauber, A. (1997). Alternative Ways for Cluster Visualization in Self-Organizing Maps. In *Proc. Of the Workshop on Self-Organizing Maps (WSOM97)*, Helsinki, Finland.

Mihalcea, R. and Moldovan, D. (2000), Semantic Indexing Using WordNet Senses, In *Proceedings of ACL Workshop on IR & NLP*, Hong Kong, 2000.

Miller, G. A. (1985). WordNet: A Dictionary Browser. In *Proceedings of the First International Conference on Information in Data*, University of Waterloo, Waterloo.

Moldovan, D. and Mihalcea, R. (1998). A WordNet-Based Interface to Internet Search Engines. *In Proceedings of FLAIRS-98*, May 1998, Sanibel Island, FL.

Moldovan, D. and Mihalcea, R. (2000). Improving the Search on the Internet by Using WordNet and Lexical Operators. In *IEEE Internet Computing*, vol. 4 no. 1, pp.34-43.

Rauber, A. (1996). Cluster Visualization in Unsupervised Neural Networks. Diplomarbeit, Technische Universitat Wien, Austria.

Richardson, R. (1994). A Semantic-based Approach to Information Processing. PhD Dissertation, Dublin City University.

Richardson, R. and Smeaton, A.F. (1995). Using WordNet in a Knowledge-Based Approach to Information Retrieval. Working Paper CA-0395, School of Computer Applications, Dublin City University, Dublin.

Ritter, H. and Kohonen, T. (1989). Self-Organizing Semantic Maps. *Biological Cybernetics*, 61. pp. 241-254.

Rodríguez, Manuel de Buenaga, José María Gómez-Hidalgo and Belén Díaz-Agudo (1997). Using WordNet to Complement Training Information in Text Categorization. In Proc. RANLP-97, Standford, March 25-27.

Reuters Corpus (2000). Volume 1, English language, 1996-08-20 to 1997-08-19, release date 2000-11-03, Format version 1. http://about.reuters.com/researchandstandards/corpus/

Salton, G. and Lesk, M. E. (1971). Information Analysis and Dictionary Construction. In Salton, G. Eds. (1971). The SMART Retrieval System: Experiments in Automatic Document Processing, chapter 6, pp. 115-142. Prentice-Hall, Inc. Englewood Cliffs, New Jersey.

Scott, S. and Matwin, S. (1998). Text Classification Using WordNet Hypernyms. In Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Montreal.

Vesanto, J., Himberg, J., Alhoniemi, E. and Parhankangas, J. (1999). Self-Organizing Map in matlab: the Som Toolbox. In Proceedings of the Matlab DSP Conference 1999, pp. 35-40, Espoo, Finland.

Voorhees, E. M. (1993). Using WordNet to Disambiguate Word Senses for Text Retrieval. In *Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 171 – 180.

Voorhees, E. M. (1998). Using WordNet for Text Retrieval. In Fellbaum C. Eds. (1998). WordNet : an electronic lexical database. MIT Press, Cambridge, Mass. pp. 285-303.

Wan, W and Fraser, D. (1994). Multiple Kohonen Self-Organizing Maps: Supervised and Unsupervised Formation with Application to Remotely Sensed Image Analysis. In Proceedings of the 7th Australian Remote Sensing Conference, Melbourne, Australia.

Wermter, S. (2000). Neural Network Agents for Learning Semantic Text Classification. Information Retrieval, 3(2), pp.87-103.