Training Context-Sensitive Neural Networks With Few Relevant Examples for the TREC-9 Routing

Mathieu Stricker *,** Frantz Vichot * Gérard Dreyfus ** Francis Wolinski *

* Informatique-CDC – Groupe Caisse des Dépôts
Direction des Techniques Avancées
4 rue Berthollet
94114 Arcueil cedex - France
{forename.surname}@caissedesdepots.fr

** ESPCI
Laboratoire d'Electronique
10 rue Vauquelin
75005 Paris, France
{forename.surname}@espci.fr

1 Introduction

The present paper describes our second participation to the routing task; it features improvements over our previous approach [Stricker *et al.*, 2000]. Our former model used a "bag of words" for text representation with a feature selection, and a neural network without hidden neuron (i.e. a logistic regression), to estimate the probability of relevance of each document. This approach was close to the ones proposed by [Schütze *et al.*, 1995] or [Wiener *et al.*, 1995] but its original feature was the use of very few relevant features for text representation (25 features per topic on the average for the TREC-8 Routing).
In this paper, two main improvements are proposed:

- The feature selection defines target words for which vectors of local contexts are subsequently defined. These vectors help disambiguate the target words and are defined by an analysis of both the relevant and the irrelevant documents of the training set.
- This new representation requires large neural networks, which are therefore prone to overfitting. A regularization technique is applied during training to favor smoother network mappings, thereby avoiding overfitting. This was achieved by adding a weight decay term to the usual cost function.

This approach led to good results on the MeSH Sample topics (S2RNsamp) and on the OHSUMED topics (S2RNr1 and S2RNr2).

2 Problem and data description

The corpus for TREC-9 is the OHSUMED collection, which is a subset of the MEDLINE database.
This corpus features documents from medical journals of years 1987 to 1991, which usually have titles and abstracts; some of them, however, only have a title. The documents were manually indexed using subject categories (Medical Subject Headings or MeSH). All documents contain the assigned MeSH headings, which are manual annotations (called .M field in the documents).
There are three topic sets for the TREC-9 routing:

1. 63 topics from the original OHSUMED query set.
2. 4904 topics based on MeSH categories.
3. 500 topics chosen amongst the 4904 previous ones called *MeSH sample topics*.

The manual annotations could be used with the OHSUMED queries (as long as it was mentioned) but NOT with the MeSH queries (nor of course with the MeSH sample queries).

We submitted 2 runs for the OHSUMED queries (one without manual annotation and one with manual annotations) and one for the MeSH sample topics.

The 1987 OHSUMED collection is intended for training and contains 54,710 documents; the test set is the 1988-91 OHSUMED collection and contains 293,882 documents. The test set is just used for evaluation and is not used in any way for building the profiles.
For each topic, a set of relevant documents is available, and all other documents are considered irrelevant. Figure 1 presents statistics for the training set for each topic set.

	OHSUMED queries	MeSH Sample queries
Number of queries	63	500
Average number of relevant documents available for training	10.6	46.5
Median	8	25

Figure1: Figures for the training set

We may observe that the number of relevant documents available per topic is small, especially for the OHSUMED queries since the median is 8.

3 Feature Selection

Each document of the collection is first tokenized into single words, case being ignored. In the following, each word is considered as a single unit called feature. No stemming is performed.
The goal of feature selection is to define, for each topic, a vector of features that are neither too frequent nor too rare, and are typical of the relevant documents. The choice of these features must be done very carefully since the quality of the filter relies heavily on this choice, irrespective of the model. These features must be chosen so as to allow a classifier to discriminate between relevant and irrelevant documents. Their number results from a tradeoff between two requirements: the larger the number of features, the larger the number of examples required to have a significant estimate of the classifier parameters; however, discarding features leads to information loss.

For each topic, a ranked list of features is computed, in which the top features are specific to the relevant documents. The method has already been used for the TREC-8 routing [Stricker *et al.*, 2000] and is discussed in detail in [Stricker, 2000].
With this technique, rare and frequent words are discarded automatically; it is useful to discard rare words because it is not possible to compute reliable statistics from them, and to discard frequent words because they carry no information.
This method is fully automatic and relies only on the computation of corpus frequency for each feature. There is no need, for example, to define a list of stop words that will depend on the language.
Contrary to last year, a Gram-Schmidt orthogonalisation with a stopping criterion was not used, because the number of relevant documents per topic was too small.

The twenty-five first features were selected, and defined the target words whose specific local context will be considered in the next section.

Figure 2 shows two lists of the top 10 features obtained at the end of this step for the topics OHSU7 "young wf with lactase deficiency". The left column is the result when the manual annotations are ignored and the right one is the result when the manual annotations are taken into account.

OHSU7 (without manual annotations)	OHSU7 (with manual annotations)
lactose	lactose
lactase	intolerance
milk	lactase
galactosidase	galactosidase
malabsorption	milk
breath	galactosidases
hydrogen	breath
digestion	hydrogen
deficient	malabsorption
yogurt	digestion

Figure 2: Examples of 10 top ranked list.

We may observe in the right column the presence of the words *intolerance* and *galactosidases,* which arise from the manual annotations.

4 Determination of local contexts

Words are naturally prone to ambiguity, and can be used in many contexts. For example, the presence of the word *intolerance*, which has been selected for the topic OHSU7 (cf. Figure 2), does not imply that a text is about "young wf with lactase deficiency".
In the past, the use of dictionaries to help disambiguate words has not proved efficient for information retrieval [Voorhees, 1993]. But disambiguation with corpus-based methods has been used successfully in [Cohen and Singer, 1996] and more recently in [Jing and Tzoukermann, 1999]. The basic idea of these methods is to disambiguate words through their local context.
Therefore, in our approach, a local context of ten words (five words on either side of the word) is considered to define which words have the highest rate of co-occurrence near a target word in the training set. Actually, two context vectors are computed for each target word: one is computed from the set of relevant documents, and the other one from the irrelevant documents.
Of course, the words with the highest rates of co-occurrence near a given word are stop words, which are useless for disambiguation. Consequently, the same procedure as used to achieve feature selection is applied to give a weight to each potential context. This method has the additional benefit of discarding automatically frequent words and rare words from the context vectors.

To compute these vectors, all relevant documents available were used, and five thousands irrelevant documents were chosen randomly. The context vectors defined by the relevant documents are called *positive context vectors* and those defined by irrelevant documents are called *negative context vectors*.

Figure 3 shows examples of positive and negative context vectors for the topic OHSU7 with manual annotations. The left column shows context vectors computed from the relevant documents for five target words, and the right one shows context vectors for the same target words computed from the irrelevant documents. The contexts are taken into account only if they appear on more than two documents.
It is worth noting that the presence of the word *intolerance* with *lactose* in its local context must be in favor of relevancy for the topic OHSU7. But, if the local context of *intolerance* contains the word *glucose* instead of *lactose* the importance of the presence of *intolerance* must decrease.

Positive context vectors		Negative context vectors	
lactose		lactose	
	intolerance		
	lactose		
	milk		
	yogurt		
	digestion		
intolerance		intolerance	
	lactose		glucose
	ul		rats
	lactobacillus		diabetes
	hydrogen		fructose
	milk		cyclosporin
lactase		lactase	
	deficient		
	milk		
	lactase		
	osteoporosis		
	organisms		
galactosidase		galactosidase	
	beta		beta
	derived		detection
	microbial		
	yogurt		
	cattle		

Figure3: Examples of local context for topic OHSU7 with manual annotations.

4.1 Choosing the size of the context vector

For each target word, the local context is chosen according to several criteria: the five first positive contexts are selected if they appear in more than two documents, and the five first negative contexts are chosen if they appear on more than ten documents. The number of documents required to take into account a context is larger for the irrelevant documents than for the relevant documents, due to the fact that irrelevant documents are much more numerous.

5 Neural Networks

5.1 Definition of the architecture

The neural network architecture must reflect the representation defined above: the influence of a target word must decrease or increase according to its local context. Therefore, instead of having a single input per target word, the local context is included as indicated in the left side of Figure 4; the right side shows the entire neural network.
Each hidden neuron is a sigmoid function and the output of the network is a logistic function in order to keep the output in the range [0,1].
Thus, this architecture contains one hidden neuron for each target words i.e. twenty-five in our case as explained in section 3.

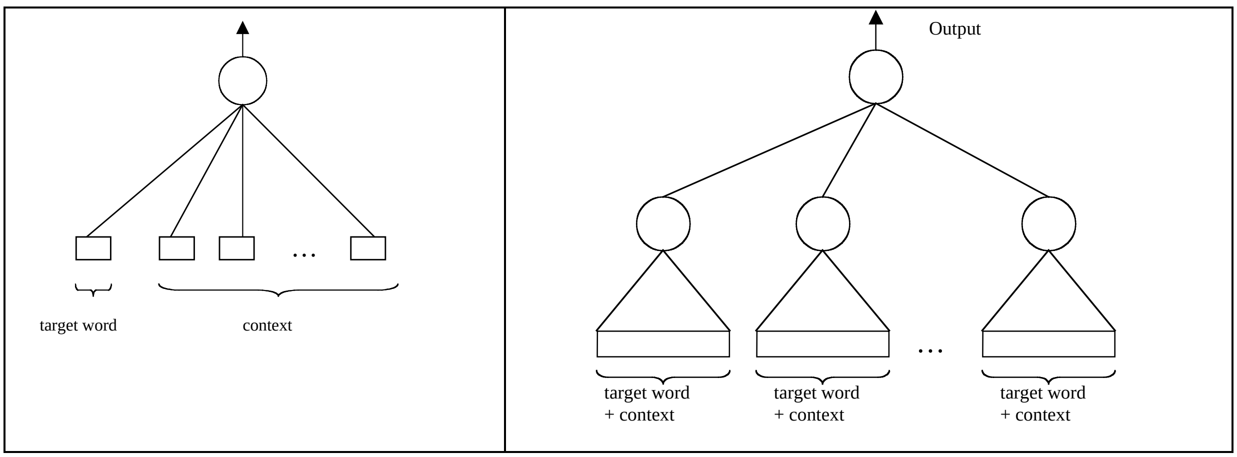

Figure 4: Neural network architecture.

5.2 Choice of irrelevant documents

Previous experiments [Stricker, 2000] have shown that it was desirable to exclude from the training set irrelevant documents for which all the target words are absent. Therefore, amongst the five thousand irrelevant documents chosen randomly, only those that share words with the relevant documents are kept.

The components of the vector are coded using the Lnu scheme defined by [Singhal, 1996].

5.3 Training with regularization

Our text representation increases the number of weights of the neural network; in addition, few relevant documents are available for training; therefore, the risk of overfitting increases, which makes the use of a regularization scheme mandatory. A penalty term is added to the usual cost function, which favors smoother functions. In our case, we use a weight decay regularization which has proven to be efficient [Krogh and Hertz, 1992][Gallinari and Cibas, 1999] and is very simple to implement.

To summarize, training is performed by minimizing a cost function *G* defined as:

$$G = J + \frac{\alpha}{2}\sum_{i=1}^{p} w_i^2$$

J is an of cross-entropy term appropriate for classification problems [Bishop, 1995]; the sum runs over all weights. α is a hyperparameter that defines the tradeoff between the two terms: if α is too small, the penalty term is negligible and overfitting tends to occur, whereas, if α is too large, weights will decay rapidly to zero and no training will occur.

The computation of the gradient of the new cost function is very simple since:

$$\nabla G = \nabla J + \alpha w$$

Where the quantity is computed by the well-known backpropagation algorithm.

In practice, all weights do not have the same dynamics, so that it is desirable not to use the same hyperparameter for all weights [MacKay, 1992b][Bishop, 1995]. Therefore, three hyperparameters are used according to the following relation:

$$G = J + \frac{\alpha_1}{2}\sum_{\omega \in W_0} w_i^2 + \frac{\alpha_2}{2}\sum_{\omega \in W_1} w_i^2 + \frac{\alpha_3}{2}\sum_{\omega \in W_2} w_i^2$$

W_0 denotes the bias of the first layer, W_1 denotes the weights of the first layer of weights except for the bias, and W_3 denotes the weights of the second layer plus the bias of the output.

5.4 Values of the hyperparameters

The values of (α_1, α_2, α_3) must be chosen appropriately in order to achieve a satisfactory training. A solution would be to test several values and to pick up the best ones by cross-validation. Unfortunately, this method is intractable since there are three different parameters.
A theoretical approach based on Bayesian inference has been proposed, in order to determine automatically the values of these hyperparameters during training [MacKay, 1992a]. The results of the theory rely on the estimation of integrals that cannot be computed easily. MacKay [MacKay, 1992b] has proposed several approximations in a theoretical framework known as the *evidence framework* to make the computation feasible. Unfortunately, these results did not provide good results on previous experiments.
Consequently, the values of the hyperparameters were chosen according to experience that we gathered previously on other corpuses (TREC-8 and Reuters21578):

$$\alpha_1 = 0.001 \qquad \alpha_2 = 0.1 \qquad \alpha_3 = 5.0$$

6 Results

We proposed three runs for the TREC-9 routing:

1. S2RNr1: 63 OHSUMED queries without using the manual annotations.
2. S2RNr2: 63 OHSUMED queries using the manual annotations.
3. S2RNsamp: 500 MeSH sample queries (without manual annotations).

The score is computed thanks to the *uninterpolated average precision* as described in [Hull and Robertson, 2000].
Figure 5 shows the comparisons of our runs with the other officials runs submitted for the TREC-9 Routing.

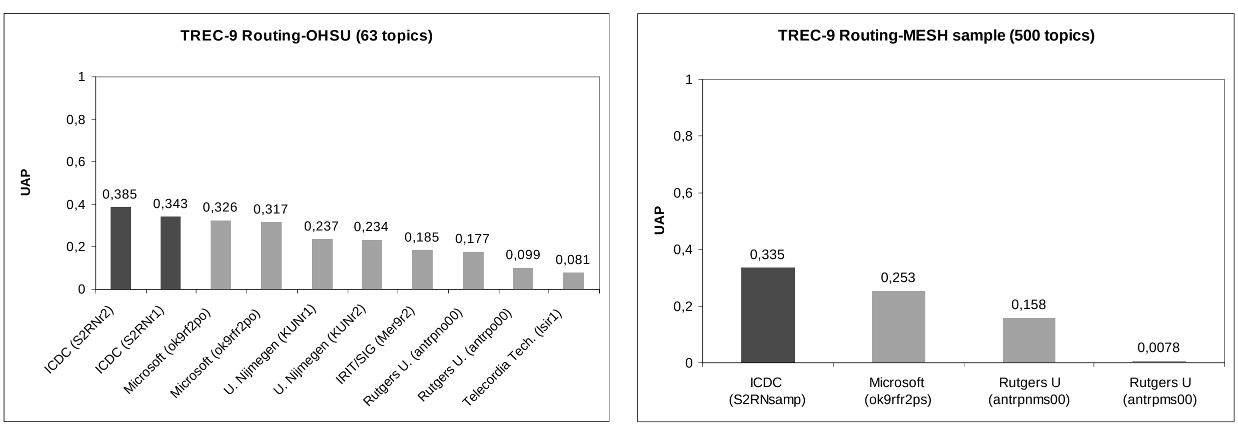

Figure 5: Comparisons of officials runs for the TREC-9 Routing.

For each subtopics (OHSU and MeSH sample), our method achieved the top scores. It is worth noting that the run S2RNr2 has better results than S2RNr1. It shows that our model can take advantage of the manual annotations without changing anything to our approach, since the difference relies only in the reading of the ".M field" which is considered as part of the text for S2RNr2.
In the case of the MeSH sample, the gap between our run and the second one is bigger than in the case of the OSHU topics ; it seems that our method has taken advantage of the greatest number of relevant documents available for training on the MeSH sample.

REFERENCES

[Bishop, 1995] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[Cohen and Singer, 1996] W. W. Cohen, Y. Singer. Context-sensitive methods for text categorization. *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR '96)*, 307-315, 1996.

[Gallinari and Cibas, 1999] P. Gallinari, T. Cibas. Practical complexity control in multilayer perceptrons. *Signal Processing*, 74, 29-46, 1999.

[Hull and Robertson, 2000] D. A. Hull, S. Robertson. The TREC-8 Filtering Track Final Report. *Proceedings of the Eighth Text REtrieval Conference (TREC-8)*. NIST Special Publication 500-246, 35-56, 2000.

[Jing and Tzoukermann, 1999] H. Jing, E. Tzoukermann. Information Retrieval Based on Context Distance and Morphology. *Proceedings of the 22nd Annual International Conference on Research and Development in Information Retrieval (SIGIR '99)*, 90-96, 1999.

[Krogh and Hertz, 1992] A. Krogh, J.A. Hertz. A Simple Weight Decay Can Improve Generalization. *Advances in Neural Information Processing Systems*, 4, J.E. Moody, S.J. Hanson and R.P. Lippmann, eds., Morgan Kauffmann Publishers, San Mateo CA, 950-957, 1992.

[MacKay, 1992a] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3), 415-447, 1992.

[MacKay, 1992b] D. J. C. MacKay. A Practical Bayesian Framework for Backpropagation Networks. *Neural Computation*, 4(3), 448-472, 1992.

[Schütze *et al.*, 1995] H. Schütze, D. A. Hull, J. O. Pedersen. A Comparison of Classifiers and Document Representations for the Routing Problem. *Proceedings of the 18th Annual International Conference on Research and Development in Information Retrieval (SIGIR'95)*, 229-238, 1995.

[Singhal, 1996] A. Singhal. Pivoted Length Normalization. *Proceedings of the 19th Annual International Conference on Research and Development in Information Retrieval (SIGIR'96)*, 21-29, 1996.

[Stricker, 2000] M. Stricker. *Réseaux de neurones pour le traitement automatique du langage : conception et réalisation de filtres d'informations*. Thèse de l'université Paris VI, 2000.

[Stricker *et al.*, 2000] M. Stricker, F. Vichot, G. Dreyfus, F. Wolinski,. Two Step Feature Selection for the TREC-8 Routing. *Proceedings of the Eighth Text REtrieval Conference (TREC-8).* NIST Special Publication 500-246, 425-430, 2000.

[Voorhees, 1993] E. M. Voorhees. Using WordNet to disambiguate words senses for text retrieval. *Proceedings of the 16th Annual International Conference on Research and Development in Information Retrieval (SIGIR'93)*, 171-180, 1993.

[Wiener *et al.*, 1995] E. D. Wiener, J. O. Perdersen, A. S. Weigend. A Neural Network Approach for Topic Spotting. *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 317-332, 1995.