Artificial Neural Networks
T. Kohonen, K. Mäkisara, O. Simula and J. Kangas (Editors)

Learning to Classify Natural Language Titles in a Recurrent Connectionist Model

Stefan Wermter
Department of Computer Science
University of Dortmund
4600 Dortmund 50
Federal Republic of Germany

Abstract

This paper describes a recurrent connectionist model which learns to classify book titles from a library. This task poses several difficult constraints to the recurrent network: learning sequences of words, detecting the context of preceding words, assigning a class, and dealing with variable length, syntax, and semantics of available titles. We describe our underlying word representation, the connectionist model, and experiments with titles from an online library classification. The model *learned* to classify almost perfectly in comparison with the existing library classification. This research shows that a recurrent connectionist model can learn the necessary knowledge for scaling up to "real-world" title classifications in natural language processing.

1 Introduction

Connectionist networks have been demonstrated to be powerful approaches for certain low-level tasks like signal processing and pattern recognition in vision and speech. On the other hand, high-level cognitive tasks like natural language processing, still show many difficult challenges for connectionist models (e.g., [2] [4] [5] [8] [9]). However, processing natural language has a number of properties which can be supported particularly well in a connectionist framework, among them graceful degradation of graded concepts, associative retrieval of words, integration of ambiguous constraints, learning, and generalization.

In this paper we will focus on the task of classifying natural language concepts of book titles. For classifying book titles a model should be able to process arbitrarily long phrases. Furthermore, it should be able to identify significant content words for certain classes depending on the preceding context in the title. Three simple examples of book titles from three classes are "Optical properties of glass" from a materials/geology class (MG), "History and tradition in afro-american culture" from a history/politics class (HP), and "Learning to use the spss batch system" from a computer science class (CS). In this paper we will describe a connectionist model for learning to classify arbitrary book titles based on a recurrent connectionist network.

2 Learning Library Titles in a Recurrent Connectionist Network

There were several reasons why we chose the task of classifying titles as a testbed for classifying natural language concepts. First, this is a theoretically important problem which contains many natural language constraints. Second, we can use online available data for practical tests. Third, we can compare the learned classification of the model with the existing library classification that has been used for many years. For our experiments we used titles from the online catalog of the library of Dortmund University. We collected titles from the three classes CS, HP, and MG. The titles were in English and German, did not exceed 140 characters and did not contain punctuation marks,

brackets, and other special symbols. Double titles were not included in our collection. There were 880 titles (3898 words) in class CS, 1230 titles (5865 words) in class HP, and 383 titles (2729 words) in class MG. Totally there were 4583 different words in the three classes. Each word was represented as a significance vector of three real numbers and each number represented the relative frequency of the word in each class. The following examples show some words with their significance vector. Significant domain-dependent content words usually have a high value for one particular class (e.g., "world": HP: 0.83, CS: 0.17, MG: 0.00; "processing": HP: 0.00, CS: 0.94, MG: 0.06; "fibres": HP: 0.00, CS: 0.00, MG: 1.00). General domain-independent words have low values for several classes (e.g., "interpretation": HP: 0.20, CS: 0.40, MG: 0.40 "from": HP: 0.33, CS: 0.33, MG: 0.33). Each title is represented as a sequence of significance vectors for individual words.

For representing sequences of arbitrary length in a connectionist network we designed a recurrent network based on SRN-networks [1]. The network shown in figure 1 has three layers. The input layer consists of a word bank and a context bank. While the word bank of three units represents the current word, the context bank contains the context of the preceding words in the title. The hidden layer represents a reduced description of the preceding words and provides the initialization for the context bank of the following word. Therefore, the number of hidden units is equal to the number of units in the context bank. The output layer consists of three real-valued units for the three classes.

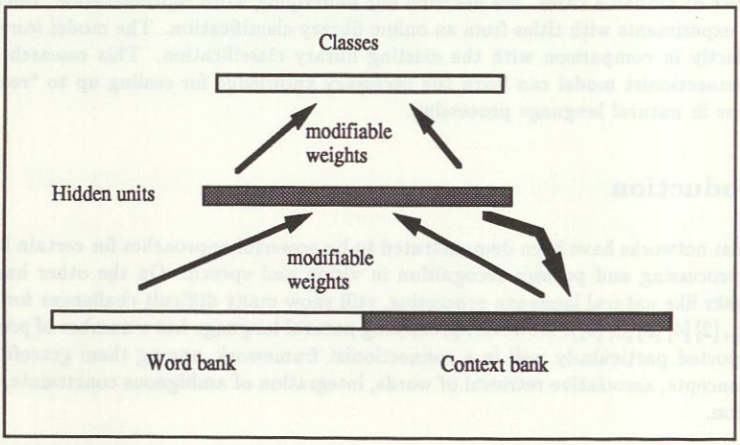

Figure 1: The architecture of the recurrent network

The training regime was to present one word of a title and its preceding context in each training step and to assign the class of the complete title. We arbitrarily collected 266 titles (1319 words) for a training set and 286 titles (1441 words) for a test set. The recurrent network was trained with the training set for 100 epochs using the backpropagation learning rule [6]. After testing several configurations of the number of hidden units and the learning rate we found that an architecture with three hidden units and a learning rate 0.0001 performed best. If the number of hidden units was higher than three units then the context bank received too much importance and reduced the classification performance. Higher values of the learning rate led to high weights in the network, lower values slowed down the learning process too much. The percentage of incorrectly assigned classes after each word from the training and test set was 2.8% and 4.2%. The percentage of incorrectly assigned classes after each complete training and test title was 0.4%. That is, the network learned the classification task quite well, the generalization performance for new titles in the test set almost reached the performance for the training set, and preceding context reduced the error rates at the end of a title.

3 Analysis of the Learned Internal Representation

The change of activation in the hidden units after each word of a complete title is shown in figure 2 as well as the currently assigned class. Black hidden units stand for a value 1, light grey hidden units for a value 0. For titles starting with content words the network can assign the final class from the beginning of the title. For instance, the words "optical", "history", and "learning" from the first three examples are significant for assigning the classes MG, HP, and CS respectively and subsequent words do not change the originally assigned class.

Examples	Hidden Units / Words	Assigned Class
1)	Optical	MG
	properties	MG
	of	MG
	glass	MG
2)	History	HP
	and	HP
	tradition	HP
	in	HP
	afro-american	HP
	culture	HP
3)	Learning	CS
	to	CS
	use	CS
	the	CS
	spss	CS
	batch	CS
	system	CS
4)	Die (The)	HP
	raeumliche (local)	MG
	und (and)	MG
	zeitliche (temporal)	MG
	verteilung (distribution)	MG
	der (of)	MG
	diffusen (diffuse)	MG
	und (and)	MG
	direkten (direct)	MG
	sonnenstrahlung (sunshine)	MG
	in (in)	MG
	der (the)	MG
	bundesrepublik (federal republic)	MG
	deutschland (germany)	MG
5)	Probleme (Problems)	HP
	des (of)	HP
	konfessionalismus (confessionalism)	HP
	in (in)	HP
	deutschland (germany)	HP
	seit (since)	HP
	1800 (1800)	HP
6)	Commonsense	HP
	and	HP
	the	HP
	theory	CS
	of	CS
	international	HP
	politics	HP
7)	The	CS
	new	CS
	cartography	MG
8)	C (C)	CS
	in (in)	CS
	der (the)	HP
	praxis (practice)	HP*

Figure 2: Sequential analysis of internal representations

The fourth to eighth example illustrate that class assignments can be changed. At the beginning of the fourth title the network assigns the class HP since the determiner "die" (femininum of "the") provided more support for class HP. However, after the word "raeumliche (local)" has been seen, the network changes the currently assigned class to MG. If we compare the fourth and fifth example we

see the same word "deutschland" which assigns the class MG in the fourth example, but the class HP in the fifth example. This behavior is caused by the preceding context of the words. While in phrase 4 the context before "deutschland" belongs to MG, in phrase 5 the context belongs to HP. In the sixth title the network assigns the class HP after the word "commonsense". Then, after "theory" the network weakly assigns the CS class but drops back to the HP class after more evidence for this class is provided by the words "international politics". The last title 8 shows the only title from the training set for which an incorrect class was assigned at the end of the title. The title "C in der praxis" (literally: "C in the practice") is assigned to the HP class although it belongs to the CS class. This was due to two *weak* preferences for CS at the beginning ("C in") and a subsequent weak preference for HP ("der Praxis"). All the other 265 titles of the training set have been classified correctly at the end of a title.

4 Conclusion

We have presented a new model for learning to classify titles. We used a recurrent network and grounded the meaning of individual words in the corpus frequency from a medium-size library corpus (12492 words total, 4583 different words). Our model differs from purely symbolic artificial intelligence approaches to natural language classification (e.g., [3]) since we make less assumptions on syntax and semantics. Our model also differs from information retrieval approaches (e.g., [7]) since we make use of the sequential context in a title and since our model *learns* to assign the class. In conclusion, the model demonstrated that a simple recurrent network and a representation grounded in corpus frequency can be used effectively for learning a classification of natural language phrases.

Acknowledgements

Part of this research has been done while the author was in the Natural Language Processing Laboratory at the University of Massachusetts, Amherst, MA, USA. I would like to thank Wendy G. Lehnert for her support. Furthermore, I would like to thank Ruth Hannuschka for her work on the library classification and Volker Weber for commenting on an earlier draft of this paper.

References

1. Elman J.L. 1989. Structured representations and connectionist models. *Proceedings of the Conference of the Cognitive Science Society.*
2. Feldman J.A., Lakoff G., Stolcke A., Hollbach Weber S. 1990. Miniature Language Acquisition: A Touchstone for Cognitive Science. *Proceedings of the Conference of the Cognitive Science Society.*
3. Hayes P.J., Knecht L.E., Cellio M.J. 1988. A News Story Categorization System. *Proceedings of the Conference on Applied Natural Language Processing.*
4. Jain A.N., Waibel A.H. 1990. Incremental Parsing by modular recurrent connectionist networks. In: Touretzky D.S. (Ed.) *Advances in neural information systems 2.* Morgan Kaufmann, San Mateo, CA.
5. Pollack J. 1988. Recursive Auto-Associative Memory: Devising Compositional Distributed Representations. Technical Report MCCS-88- 124. New Mexico State University.
6. Rumelhart D.E., Hinton G.E., Williams R.J. 1986. Learning Internal Representations by Error Propagation. In: Rumelhart D.E., McClelland J.L. (Eds.) *Parallel distributed Processing Vol. 1.* MIT Press, Cambridge, MA.
7. Salton G. 1989. *Automatic Text Processing.* Addison Wesley Publishing Company. Reading, MA.
8. Wermter, S. 1989. Integration of Semantic and Syntactic Constraints for Structural Noun Phrase Disambiguation. *Proceedings of the International Joint Conference on Artificial Intelligence.*
9. Wermter S., Lehnert W.G. 1989. A Hybrid Symbolic/Connectionist Model for Noun Phrase Understanding. *Connection Science 1 (3).*