Learning a Scanning Understanding for "Real-world" Library Categorization

Stefan Wermter*
Computer Science Department
University of Hamburg
2000 Hamburg 50
Federal Republic of Germany

Abstract

This paper describes, compares, and evaluates three different approaches for learning a semantic classification of library titles: 1) syntactically condensed titles, 2) complete titles, and 3) titles without insignificant words are used for learning the classification in connectionist recurrent plausibility networks. In particular, we demonstrate in this paper that automatically derived feature representations and recurrent plausibility networks can scale up to several thousand library titles and reach almost perfect classification accuracy (>98%) compared to a real-world library classification.

1 Introduction

Our goal is to examine hybrid symbolic/connectionist and connectionist approaches for classifying a substantial number of real-world title phrases. These approaches are embedded in the framework of SCAN (Wermter 92), a Symbolic Connectionist Approach for Natural language phrases, aimed towards a *scanning understanding* of natural language rather than focusing on an in-depth understanding. For our experiments we took an existing classification from the online catalog of the main library at Dortmund University and as a first subclassification we selected titles from three classes: "computer science" (CS), "history/politics" (HP), and "materials/geology" (MG).

2 Preprocessing of Title Phrases

2.1 Symbolic Syntactic Condensation

The first approach used syntactic condensation based on a chart parser and a headnoun extractor. The symbolic *chart parser* built a syntactic structure for a title using a context-free grammar and a syntactic lexicon. Then the *headnoun extractor* retrieved the sequence of headnouns for building a compound noun. For instance, the compound noun "software access guidelines" was generated from "guidelines on subject access to microcomputer software". This headnoun extractor was motivated by the close relationship between noun phrases and compound nouns and by the importance of nouns as content words (Finin 80).

Each noun in a compound noun was represented with 16 binary manually encoded semantic features, like measuring-event, changing-event, scientific-field, property, mechanism etc. The set of semantic features had been developed as a noun representation for a related scientific technical domain and had been used for structural disambiguation (Wermter 89). The first approach contained a relatively small set of 76 titles since for each noun 16 features had to be determined manually and for each word in the title the syntactic category had to be in the lexicon which contained 900 entries.

2.2 Unrestricted Complete Phrases

In our second approach, we used an automatically acquired *significance vector* for each word based on the occurrence of the words in certain classes. Each value $v(w, c_i)$ in a significance vector represented the frequency of occurrence of word w in class c_i divided by the total frequency of word w in all classes. These significance vectors were computed for the words of 2493 library titles from the three classes CS, HP, and MG.

2.3 Elimination of Insignificant Words

In the third approach we analyzed the most frequent words in the 2493 titles of the second approach. We eliminated words that occured more than five times in our corpus and that were prepositions, conjunctions, articles, and pronouns. Words were represented with the same significance vectors as in the second approach. This elimination of frequently occuring domain-independent words was expected to make classification easier since many domain-independent insignificant words were removed from the titles.

3 The Architecture of the Recurrent Plausibility Network

The semantic classification was learned by using a connectionist *recurrent plausibility network*. A recurrent plausibility network is similar to a simple recurrent network (Elman 89) but instead of learning to predict words, recurrent connections support the assignment of plausible classes (see figure 1). The recurrent plausibility network was trained in a supervised mode using

*This research was supported in part by the Federal Secretary for Research and Technology under contract #01IV101AO and by the Computer Science Department of Dortmund University.

the backpropagation learning algorithm (Rumelhart et al. 86). In each training step the feature representation of a word and its preceding context was presented to the network in the word bank and context bank together with the desired class. A unit in the output layer received the value 1 if the unit represented the particular class of the title, otherwise the unit received the value 0. The real-valued hidden layer represented the context of preceding words. At the beginning of a title the context bank was initialized with values of 0 since there was no preceding context. After the first word had been presented the context bank was initialized with the values of the hidden layer that encoded the reduced preceding context.

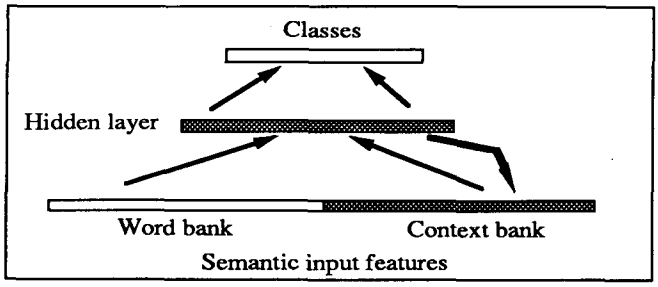

Figure 1: Recurrent Plausibility Network for Titles

4 Results and Conclusions

For the first approach the 76 titles were divided into 61 titles for training and 15 titles for testing. For the 2493 titles in the second and third approach we used 1249 titles for training and 1244 for testing. Using these training and test sets we examined different network architectures and training parameters. For the first approach a configuration with 6 hidden units and a learning rate of 0.001 showed the smallest number of errors on the training and test set. For the second and third approach 3 hidden units and the learning rate 0.0001 performed best.

Below we show example titles, the titles after preprocessing, and their sequential class assignment. The first two titles illustrate that two titles with the same final headnoun ("design") are assigned to different classes due to their different learned preceding context. The third title illustrates the second approach of classifying an unrestricted complete phrase. The network first assigns the CS class for the initial phrase "On the operating experience of the..." since such initial representations have occurred in the CS class. However, when more specific knowledge is available ("doppler sodar system...") the assigned class is changed to the MG class. In the fourth example the same title is shown for the third approach which eliminates insignificant domain-independent words. In general, the second and third approach have the potential to deal with unanticipated grammatical and even ungrammatical titles since they do not rely on a predefined grammar.

1. Title: Design of relational database schemes by deleting attributes in the canonical decomposition; Approach1: Compound noun: Decomposition (CS) attribute (CS) scheme (CS) design (CS)
2. Title: Design of bulkheads for controlling water in underground mines; Approach1: Compound noun: Mine (MG) water (MG) bulkhead (MG) design (MG)
3. Title: On the operating experience of the doppler sodar system at the Forschungszentrum Juelich; Approach2: Unrestricted complete title: On (CS) the (CS) operating (CS) experience (CS) of (CS) the (CS) doppler (MG) sodar (MG) system (MG) at (MG) the (MG) Forschungszentrum (MG) Juelich (MG)
4. Title: On the operating experience of the doppler sodar system at the Forschungszentrum Juelich; Approach3: Unrestricted reduced title: operating (CS) experience (CS) doppler (MG) sodar (MG) system (MG) Forschungszentrum (MG) Juelich (MG)

The overall performance of the three approaches as recorded in the best found configuration is summarized in table 1. The first approach performed worst for classifying new titles from the test set although the titles in the training set were learned completely. The second approach performed better on the test set for a much bigger training and test set of unrestricted phrases. The third approach demonstrated that the elimination of insignificant words from unrestricted phrases can improve performance for the big set of titles.

Performance	Approach1	Approach2	Approach3
Training	100%	98.4%	99.9%
Testing	93%	97.7%	99.4%

Table 1: Performance for Semantic Classification

In conclusion, we described and evaluated three different approaches for semantic classification which use hybrid symbolic/connectionist and connectionist representations. Our results show that recurrent plausibility networks and automatically acquired feature representations can provide an efficient basis for learning and generalizing a scanning understanding of real-world library classifications.

References

Elman J.L. 1989. Structured representations and connectionist models. *Proceedings of the Eleventh Annual Conference of the Cognitive Science Society*, Ann Arbor.

Finin T.W. 1980. *The semantic Interpretation of Compound Nominals.* PhD Thesis. University of Illinois at Urbana-Champaign.

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.

Wermter, S. 1989. Integration of Semantic and Syntactic Constraints for Structural Noun Phrase Disambiguation. *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit.

Wermter, S. 1992 (forthcoming). *Scanning Understanding: A Symbolic Connectionist Approach for Natural Language Phrases.* Technical Report. University of Hamburg.