Learning Natural Language Filtering under Noisy Conditions

Stefan Wermter
Department of Computer Science
University of Hamburg
22765 Hamburg, Germany

Abstract

This paper describes a novel AI technique, called plausibility networks, that allows for learning to filter natural language phrases according to predefined classes under noisy conditions. We describe the automatic knowledge acquisition for representing the words of natural language phrases using significance vectors and the learning of filtering of phrases according to ten different domain classes. We particularly focus on examining the filtering performance under noisy conditions, that is the degradation of these filtering techniques for incomplete phrases with unknown words. Furthermore, we show that this technique already scales up for a few thousand real-world phrases, that it compares favorably to some classification techniques from information retrieval, and that it can deal with unknown words as they might occur based on incomplete lexicons or speech recognizers.

1 Introduction: The importance of learning robust language filtering from an application context

So far there have been relatively few natural language systems that have been used in real practice compared to the number of developed natural language models and prototypes. Two of the most important general problems with current applications in natural language processing are the issues of scaling up and robustness. First, many clever natural language prototypes have demonstrated clear abilities in their restricted domains but do not tackle the problem of scaling up vertically or horizontally. By horizontal scaling we mean the transfer of knowledge from one domain area to a new one, by vertical scaling we mean the extension for dealing with hundreds, thousands, or even millions of words, phrases, or sentences. In both, vertical and horizontal scaling, manual knowledge acquisition has been a major factor restricting the prototypes from scaling up quickly. One promising approach to tackle this problem is to incorporate learning capabilities and automatic knowledge acquisition directly into the natural language systems. Without such techniques it seems unlikely that natural language systems can be scaled up easily.

Second, besides the pressing need for scaling up and learning, there is a second evenly important problem of processing incomplete natural language. Even if the domain is restricted the potential number of text or speech utterances which do not follow the assumed natural language constructions is usually large. The violation of syntactic, semantic, or even pragmatic regularities in natural language is the norm and therefore should receive a primary place in processing natural language. Furthermore, dealing with unknown words is an important general requirement for natural language systems since either knowledge sources may not be complete (e.g., lexicons) or may not be able to analyze a word (e.g., speech recognizers). In this paper we concentrate on learning the semantic filtering of natural language phrases under noisy incomplete conditions and we particularly focus on learning a semantic classification of complete and incomplete phrases. We will describe the training and testing of a substantial number of phrases taken from a real-world library corpus and we examine to what extent the network can deal with gradually increasing noise in the form of unknown words. Finally, we will relate our approach to previous natural language techniques for dealing with ill-formed incomplete input and to potential benefits for building speech language systems.

2 Learning natural phrases in plausibility networks

In this paper we address the issues of scaling up and robust processing in natural language systems by incorporating automatic knowledge acquisition and fault-tolerant training directly into plausibility networks. Plausibility networks are part of the overall framework SCAN, a **S**ymbolic **C**onnectionist **A**pproach to learning structural, semantic, and contextual interpretations of **N**atural language [9] [10]. Here we concentrate on learning semantic class assignments as an example for integrating techniques of automatic knowledge acquisition and fault-tolerant training into natural language systems.

2.1 Automatic knowledge acquisition as the basis for scaling up

In order to deal with the problem of scaling up we first had to choose an area which contained a whole variety of real-world natural language constructions from different domains. Furthermore, this area should enable us to clearly evaluate the performance. We chose the task of filtering natural language titles from a university library classification since this classification contains many thousand German and English titles as natural language phrases and since the predefined association of each title with a certain semantic class allows for a clear performance evaluation. The semantic classes (domains) were theology/religion TR, history/politics HP, law LA, mathematics MA, chemistry CH, computer science CS, electrical engineering EE, materials/geology MG, art/architecture AA, and music MU.

Each word in a phrase was represented with an automatically acquired 10-element significance vector which is part of the incremental input to plausibility networks (see figure 1). Each element in the significance vector of a word represented the normalized frequency of this particular word across the ten semantic classes of the library corpus. This word representation may be less detailed than a manually encoded semantic feature vector specifically designed for each word, but significance vectors have the advantages of automatic acquisition and primary encoding of class significances for words. Furthermore, our word representations reflect more directly the significance of certain words for certain classes: domain-independent words like "in" or "and" show relatively low significance values across all classes while domain-dependent words like "theorem" are mainly significant for mathematics and partially for computer science but not for arts and religion.

After the word representation had been acquired automatically, the main task of filtering phrases according to their semantic classes had to be learned. Simple recurrent networks have been shown to be powerful methods for processing 3-word sequences from a small artificial corpus [2]. Based on simple recurrent networks we derived the general plausibility networks which use supervised learning [5] for generalizing regularities from single instances and which contain simple feedforward models and simple recurrent networks as special cases [10]. Plausibility networks can consist of an arbitrary number of hidden layers and an arbitrary number of recurrent connections (context layers CL) which therefore allow to process multiple preceding contexts. Plausibility networks are used for plausible classification tasks and the general version of plausibility networks is shown in figure 1.

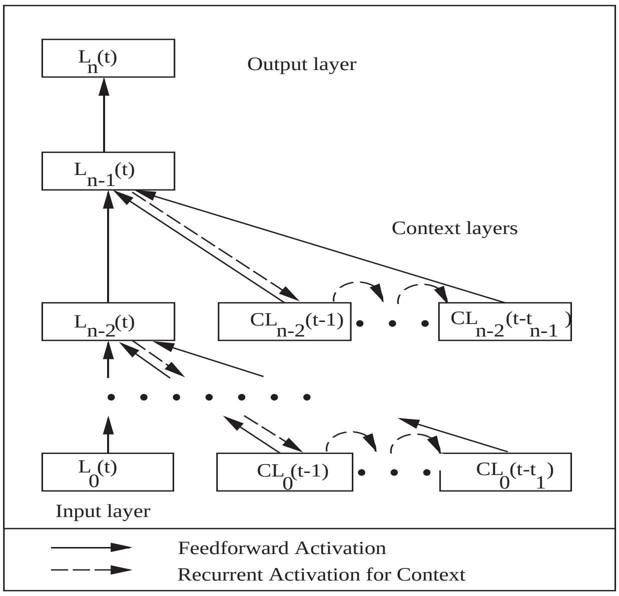

Figure 1: General structure of a recurrent plausibility network

Each phrase was incrementally presented to a plausibility network: at the input layer each word representation of a phrase was associated with its desired semantic class at the output layer. The input layer contained 10 units for the word representation, the output layer 10 units for the desired semantic classes. In each training step the current activation of the units of the hidden layer was saved in a context layer and recurrently fed into the network for the subsequent word/class association. These recurrent connections enabled the network to integrate the preceding context of a word during the learning of the class assignment to a phrase. Furthermore, this training regime allowed the network to learn to make a class assignment as early as possible.

In our experiments we used 2000 English and German phrases from the library corpus. 1000 phrases were used for training and 1000 were reserved for testing new phrases. The phrases of the test set (and their representations) were not part of the training set. Various architectures and parameter settings were tested and a network with 10 units in the hidden layer and a single context layer performed best. After training the plausibility network for 400 cycles using the BP-training procedure [5] and a learning rate of 0.000001 for 200 epochs and 0.00001 for 200 epochs, the network could assign the correct semantic context class at the end of most phrases. Only 2.4% of the phrases of the training set and 5.5% of the test set had an incorrect final class assignment. These results were taken as a benchmark for our question to what extent this network architecture could deal with noisy incomplete and incorrect phrases.

2.2 Fault-tolerant incomplete training for robust processing

We will now focus on examining fault-tolerant training in order to examine robust processing. We introduced noise to the network in the form of unknown "empty" words. This represents an important test for the robustness behavior since unknown words can occur because of incomplete lexicons, incomplete prior input analysis as in speech recognizers, etc. Furthermore, unknown words modify the sequential order of semantic preferences in phrases. Unknown words were represented by a significance vector whose units had the value 0. We introduced several degrees of noise in the form of unknown words into our original corpus of 2000 titles. Randomly we replaced 5%, 10%, and 20% of the words of the title phrases before training and testing. The same network architecture with the same learning rate was used in order to allow a clear comparison of the degradation under noisy conditions. A summary of the results on the training phrases and the test phrases is shown in table 1.

As we described above the bottom line for complete phrases is 2.4% error rate on the training phrases and 5.5% on the test phrases. The network can deal with noisy unknown phrases depending on the added noise. Table 1 shows that 5% more noise just lead to 1.5% (3.9% - 2.4%) performance loss on the training phrases and 0.9% performance loss on the test phrases. Similarly 10% noise lead to only 4.2% performance loss on the training phrases and only 2% on the test phrases. Finally 20% noise provide just 4.7% and 6.7% less performance. That is, of course unknown words reduce the performance of the network, however, it is important to point out that the degradation of the network is graceful since the performance of the network degrades much less than the percentage of added noise. The reason for this graceful degradation and robust behavior is the preceding context of phrases learned in the context layer in the plausibility network. For unknown words a correct hypothesis about the current semantic context class can only be made based on the context since the vector of an unknown word does not provide any activation for the network. In the following examples we will analyze various noisy examples in more detail. These examples were all taken from the test set using 10% noise of unknown words.

Percentage of added noise	Error rate training	Performance loss training
0	2.4	-
5	3.9	1.5
10	6.6	4.2
20	7.1	4.7

Percentage of added noise	Error rate testing	Performance loss testing
0	5.5	-
5	6.4	0.9
10	7.5	2.0
20	12.2	6.7

Table 1: Performance of the recurrent plausibility network for unrestricted phrases

2.2.1 Complete examples without noise

First, we show two phrases without unknown words. Since the start of the two phrases ("the" and "introduction to") does not contain words with a significance for a certain class a class is not yet assigned (marked by "-*"). Only after significant words have been found the network can correctly assign the music class MU and the mathematics class MA respectively.

```
1.
   The                 -*
   music               MU
   of                  MU
   africa              MU
```

```
2.
   Introduction        -*
   to                  -*
   numerical           MA
   linear              MA
   algebra             MA
   and                 MA
   optimisation        MA
```

2.2.2 Examples with single unknown words

While the first two phrases do not contain any noise all the following phrases demonstrate the ability of the network to deal with unknown words. Examples 3 and 4 show two phrases which start with "introduction to" for which the class electrical engineering EE and mathematics MA can only be assigned after significant words for these classes have been seen ("robot" and "probability"). However, there is also one unknown word within each example 3 through 7. The unknown words are illustrated as "- - -"; the original word is shown in brackets behind the unknown words. In these examples the network assigns the correct class even for noisy phrases with one unknown word. This process is independent from the semantic classes: electrical engineering EE, mathematics MA, chemistry CH, and arts and architecture AA in these examples.

```
3.
   Introduction        -*
   to                  -*
   robot               EE
   programming         EE
   in                  EE
   --- (Basic)         EE

4.
   Introduction        -*
   to                  -*
   probability         MA
   --- (models)        MA

5.
   The                 -*
   chemistry           CH
   of                  CH
   the                 CH
   catalyzed           CH
   hydrogenation       CH
   of                  CH
   carbon              CH
   --- (monoxide)      CH

6.
   Improvement         MA
   --- (of)            MA
   the                 MA
   average             MA
   linkage             MA
   method              MA

7.
   A                   -*
   history             -*
   of                  -*
   --- (western)       -*
   architecture        AA
```

2.2.3 Examples with double unknown words

The examples so far have illustrated that the network can still assign the correct class to phrases if a single word is unknown. The recurrent knowledge of the context layer of the network allows for representing the preceding context and for keeping the current class. Without the recurrent architecture of the network this behavior would not be possible. In the next set of phrases we examine the behavior for more unknown words.

```
8.
   Photometric          CH
   methods              CH
   in                   CH
   inorganic            CH
   --- (trace)          CH
   --- (analysis)       CH

9.
   Hybrid               MA
   and                  MA
   --- (mixed)          MA
   --- (finite)         MA
   element              MA
   methods              MA

10.
   Communicating        EE* CS
   --- (with)           EE* CS
   databases            CS
   in                   CS
   --- (natural)        CS
   language
```

```
11.
   --- (the)             -*
   --- (politics)        -*
   of                    -*
   public                -*
   expenditure           HP
```

Examples 8 to 11 show that the recurrent hidden layer of the plausibility network can also bridge two unknown words, even if they occur in a row as in examples 8 and 9. Example 10 shows that the network can also assign multiple classes if the words are significant for different classes. In this case "communicating - - -" could in fact start a phrase for electrical engineering EE and computer science CS. Only when more specific knowledge is available ("databases ...") the network votes for a computer science class alone. Example 11 illustrates that initial unknown words can be dealt with ("-*") as long as some class-specific knowledge about the class history/politics HP is available later.

2.2.4 Examples with triple unknown words and examples for mistakes

Examples 12 and 13 show that the plausibility network can even deal with three unknown words in a row and assign the desired classes computer science CS and music MU. Examples 14 and 15 show a similar behavior for two German phrases. Finally, we also show examples 15 and 16 that illustrate two of the remaining mistakes. In example 15, the subphrase "Die Polizei" (the police) is assigned to the law LA class although the given library classification assigned it to the history/politics HP class. Since this incomplete phrase does not contain further significant knowledge the network stays with this class assignment. However, we should note, that - although this example is counted as a mistake with respect to the library classification - this title might in fact be part of a law class. Finally, the last example shows another final mistake based on the underspecified contents of the word "engineering". Using only this initial word followed by three unknown words the network can not assign a particular class, since engineering occurs across many different classes (e.g. electrical engineering EE, mathematics MA, computer science CS, materials/geology MG...). Since this is the only specific knowledge for the network, it is not possible to assign a certain class, although the complete title "engineering composite materials" could be assigned to the MG class by the plausibility network.

```
12.
   Diagonalization       CS
   --- (over)            CS
   --- (polynomial)      CS
   --- (time-computable)CS
   sets                  CS
```

```
13.

   --- (a)               -*
   --- (generative)      -*
   --- (theory)          -*
   of                    -*
   tonal                 MU
   music                 MU
```

```
14.
   Historische           HP
   Leitlinien            HP
   fuer                  HP
   --- (das)             HP
   --- (Militaer)        HP
   der                   HP
   --- (neunziger)       HP
   Jahre                 HP

   English: Historical guidelines for
            the armed forces of the 90ies
```

```
15.
   Die                   -*
   Polizei               LA* (should be HP)
   --- (in)              LA*
   --- (der)             LA*
   --- (Bundesrepublik) LA*

   English: The police in Germany
```

```
16.
   Engineering           -*
   --- (composite)       -*
   --- (materials)       -*
```

3 Discussion: comparison with previous efforts in robust natural language filtering

There are several approaches for natural language processing which address the issues of scaling up and robustness. First, approaches from information retrieval compute superficial representations for matching queries and documents, that is, text filtering or classification according to the desired response for a query (e.g., [6]). Information retrieval approaches like boolean keyword techniques or statistical weighting techniques have the potential of being fast, easy to compute, and robust but rely crucially on larger documents rather than phrases. In many classification techniques from information retrieval, single terms are combined, for instance with boolean operators or statistical measures. However, usually the sequential order in phrases is not taken into account and single phrases like "computer century" and "century computer" are interpreted equally for retrieval. One main reason why this can be done in information retrieval is the larger size of documents for classification. However, for classifying phrases there is much less context than in documents so that the sequential order should be exploited for finding the class of a phrase. Techniques from information retrieval like average vector weighting and average vector weighting after the elimination of stop word lists do not consider sequentiality and the results for these techniques are independent of the order of the words. Average vector weighting and average vector weighting after stop word elimination were tested on our test set and showed just 49.4% and 72% of correct class assignments compared to the 94.5% of the plausibility networks. These examinations suggest that information retrieval techniques - while fast, robust, and scalable for document filtering - cannot directly be used for phrase filtering.

Second, approaches of symbolic semantic analyzing emphasized the predictive semantics of a language segment and therefore relaxed constraints on the required order of grammatical constituents. These approaches tackled the classification of complete stories based on sketchy scripts [1], pattern rules [3], and conceptual analyzing [4]. For instance, a system CONSTRUE is described in [3] which focused on the classification of stories into specified classes based on handcoded patterns. A pattern contained weighted words and phrases that could occur in the stories. A pattern match was suggested by probable and possible patterns of a story. Furthermore detailed patterns could then decide about the more specific class. In general, this approach worked well, but relied a lot on manual encoding of patterns. For instance, the knowledge engineering effort for a redevelopment of a story classification system CONSTRUE [3] has been estimated as 8 person months. Our described filtering techniques would need much less knowledge engineering due to the automatically generated word representations and the learning plausibility networks. In summary, techniques from information retrieval and symbolic semantic analyzing as outlined above have a potential for scaling up and robustness due to their broad and weak techniques. However, techniques from information retrieval crucially rely on larger text documents for classification and techniques from symbolic semantic analyzing often use manually-encoded representations rather than automatically acquired or learned representations.

Furthermore, our techniques and experiments provide a test for integrating robustness and learning capabilities directly in bigger architectures for speech and language (e.g., [11] [7] [8]). Current speaker-independent continuous speech recognizers are still unreliable due to the complexity of the mapping from the signal to a sequence of words. For instance, different speeds, dialects, accents, moods, word use, grammatical competence, use of prosody etc. influence this mapping. Therefore, in many cases speech recognizers produce word hypotheses together with their confidence values but many confidence values are under a certain threshold of reliability. In these cases of uncertain incomplete hypotheses of word sequences a further semantic or contextual analysis has to deal with unknown words. Since connectionist plausibility networks have been shown to be able to process phrases with unknown words in an incremental fault-tolerant manner, they have the potential for dealing with uncertain or unknown speech input.

4 Conclusions

We have used plausibility networks as a novel AI technique for addressing the crucial issues of scaling up and robustness in practical natural language systems. Scaling up and robustness were identified as two major problems that have restricted the development of real-world natural language systems. In order to address these issues we have used automatic knowledge acquisition for scaling up and fault-tolerant training for providing robustness. These techniques showed good performance using several versions of a corpus of 2000 real-world title phrases. We believe that this kind of bottom-up learning in plausibility networks (1) has demonstrated real-world capabilities based on a substantial amount of library titles (2) and more generally has a lot of potential for related real-world tasks like processing language based on incomplete knowledge sources (e.g., lexicons) and incomplete analysis (e.g., speech recognizers).

Acknowledgements

This research was supported in part by the Federal Secretary for Research and Technology under contract #01IV101A/O.

References

[1] G. F. DeJong. Skimming stories in real time: an experiment in integrated understanding. Technical Report Research Report 158, Yale University, New Haven, CT, 1979.

[2] J. L. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[3] P. J. Hayes, L. E. Knecht, and M. J. Cellio. A news story categorization system. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 9–17, Austin, TX, 1988.

[4] W. G. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff, and S. Soderland. University of massachusetts: Description of the circus system as used for muc-4. In *Proceedings of the Fourth Message Understanding Conference*, 1992.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume Vol. 1, pages 318–362. MIT Press, Cambridge, MA, 1986.

[6] G. Salton. *Automatic Text Processing.* Addison Wesley, New York, 1989.

[7] W. von Hahn and C. Pyka. System architectures for speech understanding and language processing. In G. Heyer and H. Haugeneder, editors, *Applied Linguistics.* Wiesbaden, 1992.

[8] A. Waibel, A. N. Jain, A. McNair, J. Tebelskis, L. Osterholtz, H. Saito, O. Schmidbauer, T. Sloboda, and M. Woszczyna. Janus: speech-to-speech translation using connectionist and non-connectionist techniques. In J. E. Moody, S. J. Hanson, and R. P. Lippmann, editors, *Advances in Neural Information Processing Systems 4*, pages 183–190. Morgan Kaufmann, San Mateo, CA, 1992.

[9] S. Wermter. A hybrid and connectionist architecture for a scanning understanding. In *Proceedings of the 10th European Conference on Artificial Intelligence*, pages 188–192, Vienna, Austria, 1992.

[10] S. Wermter. A hybrid connectionist approach for a scanning understanding of natural language phrases. Technical Report Doctoral thesis, University of Hamburg, Hamburg, FRG, 1993.

[11] S. R. Young, A. G. Hauptmann, W. H. Ward, E. Smith, and P. Werner. High level knowledge sources in usable speech recognition systems. *Communications of the ACM*, 32:183–194, 1989.