In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 136–138, Lisbon, Portugal, 2000.

A Context Sensitive Maximum Likelihood Approach to Chunking

Christer Johansson
Electrotechnical Laboratories Machine Understanding Division
1-1-4 Umezono, Tsukuba. 305 Ibaraki, JAPAN

1 Introduction

In Brill's (1994) groundbreaking work on parts-of-speech tagging, the starting point was to assign each word its most common tag. An extension to this first step is to utilize the lexical context (i.e., words and punctuation) surrounding the word. This approach could obviously be used for ordering tags into higher order units (referred to as chunks) using chunk labels.

This paper will investigate the performance of simply picking the most likely tag for a given context, under the condition that a larger context is allowed to override the most likely label of a smaller context. The results could be extended by secondary error correction as in Brill's tagger, but this exercise is left to the reader to allow us to concentrate on the performance based on storing and retrieving the most likely examples only.

More sophisticated methods may use more than one stored context to determine the label that best fits the current context (Van den Bosch and Daelemans, 1998; Zavrel and Daelemans, 1997; Skousen, 1989, inter al.). The method of this paper uses only one context to determine the best label, but may decrease the size of the context until a full match is found.

2 Outline of the procedure

2.1 "Training"

The training of this mechanism is to determine which patterns in the training set are the most likely. Only tag information is used. A filter to convert a tag with a context into a chunk-label is constructed as follows:

0) Construct symmetric n-contexts from the training corpus. A 1-context is simply the most common chunk-label for each tag. A 3-context is the tag followed by the tag before and after it, i.e., *[*t_0 t_{-1} t_{+1}*]:label.* Similarly, a 5-context, (i.e., [t_{-2} [t_{-1} [t_0] t_{+1}] t_{+2}]: label (of t_0)), is represented *[*t_0 t_{-1} t_{+1} t_{-2} t_{+2}*]:label.* Finally, a 7-context is represented as *[*t_0 t_{-1} t_{+1} t_{-2} t_{+2} t_{-3} t_{+3}*]:label.* It was verified that results do not significantly improve using larger contexts than 5-contexts.

1) For each set of n-contexts, determine the most frequent label for each occurring n-context. For example, the tag CC most frequently has the label *B-NP* if the context is *PRP CC RP*. The most frequent label for *CC* without extra context is *"O"*.

2) To save some storage space, the most frequent label in an n-context is only added if it is different from its nearest lower order context. For example, the label *B-NP* can be added for a 3-context since *PRP CC RP* gives a different result from *CC* alone.

2.2 Testing

Testing is done by constructing the maximum context for each tag, and look it up in the database of the most likely patterns. If the largest context cannot be found the context is diminished step-by-step.

3) In the test phase we need to form the longest contexts used in training (e.g., 7-contexts). The first word to get a chunk label is 'Rockwell' (*Rockwell International Corp. 's*) and its corresponding 7-context (without its label) is *NNP = NNP = NNP = POS*, where '=' is a tag for a blank line (i.e., no text tag) since this is the very first few words.

4) The only rule for chunk-labeling is to look up the closest surviving n-context and output its label. Simply look up [t_0 t_{-1} t_{+1} t_{-2} t_{+2} t_{-3} t_{+3}] ... [t_0] in that order until the context is found. The [t_0] context alone produces a $F_{\beta=1}$ of 77.

3 Results

The evaluation program shows that this simple procedure reaches its best result for 5-contexts (table 1) with 92.46% label accuracy and phrase correctness measured by $F_{\beta=1} = 87.23$. However, the improvement from 3-contexts to 5-contexts is insignificant, as 3-contexts reached 92.41% accuracy and $F_{\beta=1}$=87.09. The results for 7-contexts is almost identical to 5-contexts (92.44% and $F_{\beta=1}$=87.21). This is taken as the limit performance due to the size of the training corpus.

In a larger training corpus, the most common longer contexts are likely to be useful but in a small set the longer contexts may occur with very low frequencies making it hard to determine if the label of such contexts is the best guess for unseen samples.

These results are the best that could be expected without generalization. In order to do better, the method has to generalize to unseen contexts, e.g., by using some notion of close matching contexts (instances), to be able to use longer context even when some of that context has not been previously recorded. In addition, the tag-structure could be productively utilized. The presented method has treated all labels as arbitrary, atomic and independent symbols.

3.1 Computational complexity

Using rule 2 from section 2.1, 45 patterns 'survived' for 1-contexts, and 3225, 71022, 38541 for 3-,5- and 7-contexts respectively, i.e., a total of 45, 3270, 74292, 109563 using all contexts up to and including 1-, 3-, 5- and 7-contexts. Each unique context can be retrieved in one logical step (i.e., a hash-table lookup). There are obviously many patterns in the database - but the complexity of the task is limited to the number of look-ups necessary.

There is a maximum of four hash-table lookups for each tag (i.e., when the 7-, 5-, and 3-contexts does not exist in the database the most likely label of the current tag will be used). Good performance can be obtained within a maximum of 2 look-ups for each label (i.e., using only 1- and 3-contexts) and the best results were obtained with a maximum of 3 look-ups per label.

4 Discussion

The memory-based approach seemingly postulates innate tags in the processing machinery. The author has found very little discussion on how the tags are thought to correspond to reality, a fact that was also pointed out, not so long ago, by Palmeri (1998). However, a few papers aiming towards automatic 'label', 'feature' or 'tag' creation are available (Miikkulainen and Dyer, 1991; Johansson, 1999).

It is undeniable that, from a practical perspective, it is possible to reach very high performance on tasks, such as tagging, that demand a choice from a known set of alternatives by estimating statistical properties (e.g., the most likely label) from a large enough training set. This makes the method extremely useful for quick development of tools, which can be used in practical applications such as text retrieval and machine translation; but also in linguistic research; e.g., finding examples of specific grammatical constructions in large collections of data.

A challenge for future research is how tags could be constructed automatically, and what kind of information would be necessary to detect the relevant tag dimensions for some linguistically motivated task.

5 Conclusion

It was shown that using context made it possible to improve performance of maximum likelihood prediction. It was suggested that the limit of performance for this method is implicitly given by the size of the training set, as this determines the significance of larger contexts, and increases the chance of finding a matching longer context. In smaller collections, large patterns are a) likely to occur at a low frequency with few competing labels and b) likely to not exist in the test set. A larger collection will increase the number of different contexts, as well as the significance of picking the best, most frequent, prediction from a set of (identical) competitors with different labels.

The presented method does not generalize beyond what is recorded in the training set as the most likely alternative. However, it is expected to • improve with the size of the training set, as this makes it feasible to use longer contexts, and • have a low computational complexity, as the process is always limited to use a low number of hash table look-ups (determined by the largest size of context). Training is limited to detecting the most likely outcome of each context (i.e., a sorting operation).

References

Antal van den Bosch and Walter Daelemans. 1998. Do not forget: Full memory in memory-based learning of word pronunciation. In D.M.W. Powers, editor, *proceedings of NeMLap3/CoNLL98*, pages 195–204, Sydney, Australia.

Eric Brill. 1994. Some advances in rule-based part of speech tagging. In *proceedings of AAAI.*

Christer Johansson. 1999. Noise resistance in processing center-embedded clauses: A question of representation? In *proceedings of ICCS'99*, pages 253–258, Tokyo. Waseda University.

Risto Miikkulainen and Michael G. Dyer. 1991. Natural language processing with modular PDP networks and distributed lexicon. *Cognitive Science*, 15(3):343–399.

Thomas J. Palmeri. 1998. Formal models and feature creation. *Behavioral and Brain Sciences*, 21:33–34.

Royal Skousen. 1989. *Analogical Modeling of Language.* Kluwer Academic, Dordrecht, the Netherlands.

Jakub Zavrel and Walter Daelemans. 1997. Memory based learning: using similarity for smoothing. In *proceedings of the 35th annual meeting of the Association of Computational Linguistics (ACL) and the 8th conference of the European Chapter of the ACL*, pages 436–443, Madrid, Spain.

test data	precision	recall	$F_{\beta=1}$
ADJP	58.33%	52.74%	55.40
ADVP	67.98%	71.59%	69.74
CONJP	0.00%	0.00%	0.00
INTJ	33.33%	50.00%	40.00
LST	0.00%	0.00%	0.00
NP	88.09%	90.53%	89.30
PP	88.18%	93.39%	90.71
PRT	36.14%	28.30%	31.75
SBAR	54.97%	33.08%	41.31
VP	88.27%	91.28%	89.75
all	86.24%	88.25%	87.23

Table 1: Results using at most 5-contexts