Memory-Based Text Chunking

Jorn Veenstra
ILK, Tilburg University
veenstra@kub.nl

Introduction

The abundance of data often hinders computational efficiency in information retrieval, information extraction and parsing. Several data reduction techniques have been proposed to overcome this problem. A recently blooming approach to data reduction in computational linguistics is chunking.

In (Daelemans, Buchholz, and Veenstra, 1999), we describe experiments on NP and VP (noun and verb phrase) chunking using Memory-Based Learning (MBL) carried out in the context of subject/object identification. In this paper we will show that NP, VP and PP (prepositional phrase) chunking is possible with a precision and recall around 94-95%.

Methods and Data

For our experiments we used the WSJ corpus, an annotated (with syntactic structure) and POS (Part-of-Speech) tagged corpus with over one million words taken from the Wall Street Journal. We used TiMBL[1], an implementation of MBL, as learning tool (Daelemans et al., 1998).

Memory-Based Learning MBL is a form of *supervised, inductive learning from examples*. Examples or cases are represented as vectors of feature values with associated class labels. During training, a set of cases (the training set) is presented in an incremental fashion to MBL and then added to memory (the case base). During testing, a set of previously unseen feature-value patterns (the test set) is presented to MBL. For each test pattern its distance to all cases in memory is computed and the category of the nearest case(s) is used to predict the class of the test pattern. The approach assumes that reasoning is based on direct reuse of stored experiences rather than on the application of knowledge (such as rules or decision trees) abstracted from experience.

In AI, the concept of MBL has appeared in several disciplines (from computer vision to robotics) using names such as: example-based, case-based, similarity-based, exemplar-based, instance-based, lazy, analogical or nearest-neighbour learning/reasoning.

The Data Following (Ramshaw and Marcus, 1995) we defined chunking as a tagging task, each word in a sentence is assigned a tag which indicates whether this word is inside or outside a chunk. We used as tagset:

I_NP inside a baseNP.

O outside a baseNP,

B_NP inside a baseNP, but the preceding word is in another baseNP.

[1]The TiMBL package is freely available for academic and educational purposes from http://ilk.kub.nl/.

I_VP, B_VP, I_PP and B_PP are used in a similar fashion.

Since baseNPs, baseVPs and basePPs are non-overlapping and non-recursive these five tags suffice to unambiguously chunk a sentence. For example, the sentence:

> $[_{NP}$ Pierre Vinken $_{NP}]$, $[_{NP}$ 61 years $_{NP}]$ old , $[_{VP}$ will join $_{VP}]$ $[_{NP}$ the board $_{NP}]$ $[_{PP}$ as $_{PP}]$ $[_{NP}$ a nonexecutive director $_{NP}]$ $[_{NP}$ Nov. 29 $_{NP}]$.

will be IOB-tagged as:

> Pierre_{I_NP} Vinken_{I_NP} $,_{O}$ 61_{I_NP} years_{I_NP} old_{O} $,_{O}$ will_{I_VP} join_{I_VP} the_{I_NP} board_{I_NP} as_{I_PP} a_{I_NP} $\text{nonexecutive}_{I_NP}$ director_{I_NP} Nov._{B_NP} 29_{I_NP} $._{O}$

We trained a separate classifier to decide whether a sequence of a PP and an NP chunk can be combined to a PNP (full PP) chunk (e.g. *$[_{PP}$ as $_{PP}]$ $[_{NP}$ a nonexecutive director $_{NP}]$* will be re-chunked as: *$[_{PNP}$ as a nonexecutive director $_{PNP}]$*).

Results

For all these experiments we used the POS tags as provided by the WSJ corpus. The features we used are the word form and the POS tag of the two words to the left, the focus word, and one word to the right. The results given in Table 1 are obtained with 25 fold cross validation over the entire WSJ corpus, where each fold is a separate WSJ section; precision and recall values are around 94-95%.

	recall	precision	accuracy
NPs	93.7	94.0	98.0
VPs	94.0	95.5	99.2
PNPs	95.0	95.1	–

Table 1: Overview of the score of 25-cv on the WSJ corpus, using overlap with feature weighting and a context of two words and POS right and one left. The accuracy is measured at the level of IOB tags; since for PNP chunking the IOB tags are used only indirectly, no accuracy is given.

References

Daelemans, W., S. Buchholz, and J. Veenstra. 1999. Memory-based shallow parsing. In *proceedings of CoNLL*, Bergen, N. submitted.

Daelemans, W., J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 1998. TiMBL: Tilburg Memory Based Learner, version 1.0, reference manual. Technical Report ILK-9803, ILK, Tilburg University.

Ramshaw, L.A. and M.P. Marcus. 1995. Text chunking using transformation-based learning. In *Proc. of third workshop on very large corpora*, pages 82–94, June.