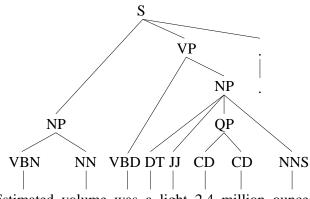
Parsing

From Chunking to Parsing

Erik Tjong Kim Sang CNTS - Language Technology Group University of Antwerp Belgium



Estimated volume was a light 2.4 million ounces

 Creating a syntactic analysis for a sentence can be divided in three tasks: POS tagging (level 1), chunking (level 2) and generating the rest of the tree.

CLIN 2000 1

Tjong Kim Sang 03/11/2000

Research questions

We have developed methods for training classifiers to divide sentences in groups of syntactically related words (chunking).

- Can these methods be used for parsing sentences?
- If so, how well do they perform?

Bottom-up Parsing is repeated chunking

03/11/2000

Tjong Kim Sang

We compute phrase positions and replace the phrases by their head and a tag denoting their type. After that we attempt to find more phrases.

Example

W	Estimated	volume	was	a	light	2.4	million	ounces	٠
р	VBN	NN	VBD	DT	JJ	CD	CD	NNS	
0	(NP	NP)				(QP	QP)		
0w		volume	was	а	light		million	ounces	
0р		NP	VBD	DT	JJ		QΡ	NNS	
1				(NP				NP)	
1w		volume	was					ounces	•
1р		NP	VBD					NP	÷
2			(VP					VP)	
2w		volume	was						
2p		NP	VP						
3		(S							S)

Finding phrases

Standard context-free grammars contain rules like

$$NP \rightarrow Det Adj N$$

Studies have shown that for text chunking a different rule format is more suitable (Muñoz et.al., 1999 vs. Cardie and Pierce, 1999):

$$N V Det Adj \rightarrow N V (NP Det Adj$$

Instead of looking for word/POS sequences that make up phrases, we look for word/POS sequences that make up phrase boundaries.

CLIN 2000 4

Tjong Kim Sang 03/11/2000

Finding brackets

The brackets are generated by a memory-based learner based on training examples.

It uses a sliding window of a word and its POS tag and a context of two left and two right word/POS tag pairs.

Identifying head words

The head words for a phrase are generated by a list of about 25 rules based on the POS tags that make up the phrase (Collins, 1997).

Balancing brackets

The opening and closing brackets of text chunks can be identified by independent processes. They will contain errors which means that the bracketing structures might be imbalanced.

We obtain balanced bracket structures by removing all brackets which cannot be matched with the closest possible candidate bracket.

Example:

(NP (NP VP) NP) (NP NP)

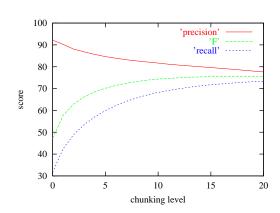
will become:

(NP NP)

CLIN 2000 5

Tjong Kim Sang 03/11/2000

Results



The performance is measured by comparing the labeled phrases found by the parser with the phrases in a parsed corpus (Penn Treebank).

Our current best result is F=77.7 (P=80.7, R=74.9) compared to a world's best of F=89.5.

Answers to research questions

- The techniques used for chunking can also be used for parsing.
- Their performance is not spectacular.

Future work

- Use more training data.
- Evaluate other bracket combination algorithms.
- Combine different learners for bracket identification.
- Use an NP chunker instead of general base chunker.
- Generate more than one evaluation per sentence (?)

CLIN 2000 8