Introduction to the CoNLL-2000 Shared Task: Chunking

Erik Tjong Kim Sang, University of Antwerp
Sabine Buchholz, Tilburg University

Motivation

A shared task is an interesting method for comparing the performance of machine learning method in one specific domain.

Possible topics

- Full parsing: hard and requires a lot of training data.
- Part-of-speech tagging: has been done by many.
- Chunking: a reasonable amount of work has been done on noun phrase chunking but relatively few have worked on recognizing arbitrary text chunks.

Chunks

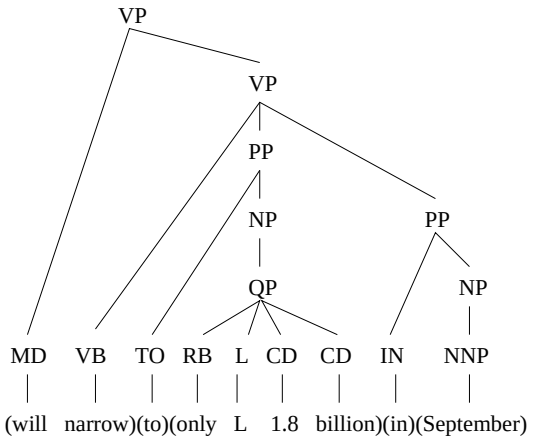
Our goal is to divide texts into sequences of syntactically related words (chunks). A word can only be member of one chunk.

[$_{\mathrm{NP}}$ He] [$_{\mathrm{VP}}$ reckons]
[$_{\mathrm{NP}}$ the current account deficit]
[$_{\mathrm{VP}}$ will narrow] [$_{\mathrm{PP}}$ to]
[$_{\mathrm{NP}}$ only £ 1.8 billion]
[$_{\mathrm{PP}}$ in] [$_{\mathrm{NP}}$ September] .

This sentence contains eight base chunks.

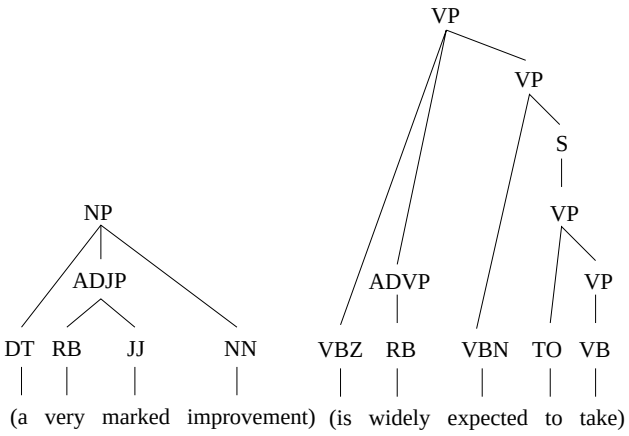
The chunks will be derived from a parsed corpus (Penn Treebank).

Deriving chunks (1)



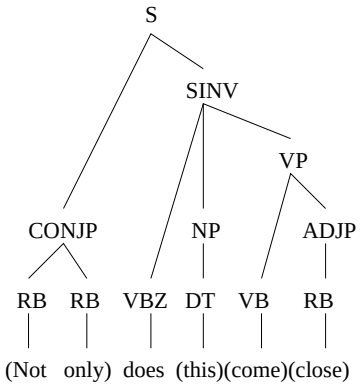
- words are member of at most one phrase
- related verbs are combined

Deriving chunks (2)



- internal adjective phrases are included in noun phrases
- internal adverbial phrases are included in verb phrases

Deriving chunks (3)



- some words are not assigned to a chunk
- outside-chunk entities are mainly punctuation signs

Chunk types

count	%	type
55081	51%	NP (noun phrase)
21467	20%	VP (verb phrase)
21281	20%	PP (prepositional phrase)
4227	4%	ADVP (adverbial phrase)
2207	2%	SBAR (subordinated clause)
2060	2%	ADJP (adjective phrase)
556	1%	PRT (particles)
56	0%	CONJP (conjunction phrase)
31	0%	INTJ (interjection)
10	0%	LST (list marker)
2	0%	UCP (unlike coordinated phrase)

The training data (WSJ sections 15-18) contains 211727 tokens and 106978 chunks of eleven types (see table). The test data (WSJ section 20) contains 47377 tokens.

Data representation

The training and test data have been represented with so-called chunk tags:

He/B-NP reckons/B-VP
the/B-NP current/I-NP account/I-NP deficit/I-NP
will/B-VP narrow/I-VP to/B-PP
only/B-NP £/I-NP 1.8/I-NP billion/B-NP
in/B-PP September/B-NP ./O

Tag B-X is used for the first word of a chunk of type X, tag I-X for non-initial words in an X chunk and tag O for words outside of any chunk.

Goal

Use the training data for creating a model that is able to find text chunks in unseen data.

Test this model with the test data.

Evaluation method

$F_{\beta=1}$: a combination of precision and recall scores for typed chunks.

All participants use the same evaluation software.

Participants

Eleven systems have been applied to the CoNLL-2000 shared task. The systems used a wide variety of techniques.

- Marc Vilain and David Day
- Christer Johansson (*)
- Ferran Pla, Antonio Molina and Natividad Prieto
- Jorn Veenstra and Antal van den Bosch
- Miles Osborne (*)
- Rob Koeling
- Hervé Déjean
- GuoDong Zhou, Jian Su and TongGuan Tey
- Erik F. Tjong Kim Sang
- Hans van Halteren
- Taku Kudoh and Yuji Matsumoto

The authors will present their systems themselves.

Results

test data	precision	recall	$F_{\beta=1}$
Kudoh and Matsumoto	93.45%	93.51%	93.48
Van Halteren	93.13%	93.51%	93.32
Tjong Kim Sang	94.04%	91.00%	92.50
Zhou, Tey and Su	91.99%	92.25%	92.12
Déjean	91.87%	91.31%	92.09
Koeling	92.08%	91.86%	91.97
Osborne	91.65%	92.23%	91.94
Veenstra and Van den Bosch	91.05%	92.03%	91.54
Pla, Molina and Prieto	90.63%	89.65%	90.14
Johansson	86.24%	88.25%	87.23
Vilain and Day	88.82%	82.91%	85.76
baseline	72.58%	82.14%	77.07

Seven of the eleven systems have obtained an $F_{\beta=1}$ score between 91.5 and 92.5. Two systems performed a lot better. Both used some kind of system combination.

The $F_{\beta=1}$ rates for NPs of the top-two systems are the same as the best reported result for NP chunking (93.8).

Future work

1. Evaluate combinations of the results presented
2. Extend this approach to full parsing

Concluding remark

In the CoNLL-2000 shared task, recognizing arbitrary phrases in text, two systems have outperformed the other nine: Support Vector Machines used by Kudoh and Matsumoto and Weighted Probability Distribution Voting used by Van Halteren.