Noun Phrase Detection by Repeated Chunking

Erik F. Tjong Kim Sang
Center for Dutch Language and Speech
Universitaire Instelling Antwerpen
Belgium
erikt@uia.ua.ac.be

overhead sheets: http://lcg-www.uia.ac.be/~erikt/talks/

Goal

Developing a method for detecting arbitrary noun phrases (NPs) with reasonable precision and recall.

Possible approaches

1. Determine the position of NP brackets and derive balanced structures from these positions.
2. Find as many NP brackets as possible and derive balanced structures from these brackets.
3. Perform repeated NP chunking: build NP chunks which may contain other NP chunks.

Basis

A good method for detecting non-recursive baseNPs.

Algorithm

We have used the memory-based learning algorithm IB1-IG. This algorithm uses the context of a word (neighboring words with their part-of-speech tags) for determining whether the word is at the beginning or the end of an NP.

After NP brackets have been found, they are combined and converted to balanced structures with other software.

Evaluation measures

The performance of the NP detection method will be determined by four rates:

1. Precision: the percentage of NPs that were found that were correct.
2. Recall: the percentage of correct NPs that were found.
3. F: a combination of precision and recall:
   F = (2*precision*recall) / (precision+recall).
4. Crossing rate: the average number of found NPs per sentence that cross correct NPs.

Data

The training data consists of four sections from the Wall Street Journal corpus (15-18). The test data contains one section of this corpus (20).

The best approach and the best parameters for this approach have been determined by using only the sections 15-18 (15-17 for training and 18 for testing).

Data analysis

The test data contains 16180 noun phrases of which 11815 are non-recursive NPs (73%). A baseline score for this data set can be determined by applying a NP chunker to it. Our NP chunker obtains the following rates: P: 94.93; R: 65.97; F: 77.84.

Best results for the three approaches

	P	R	F
Guess positions	84.41%	76.66%	80.35
Guess brackets	87.99%	76.56%	81.88
Repeated chunking	90.97%	76.20%	82.94

Observed optimization rule: throw away every bracket that might be wrong.

Chunking performance progression

	P	R	F
BaseNPs	94.61%	66.60%	78.17
Level 1	92.35%	74.25%	82.32
Level 2	90.97%	76.20%	82.94
Level 3	89.89%	76.73%	82.79

Chunks have been represented with the (guessed) head word with the tag NPCHUNK. A head word is the final word of the first noun cluster in the chunk or the final word of the chunk. Non-baseNP levels were trained with data from all available chunk levels.

Performance on all data

	C	P	R	F
BaseNPs	0.03	94.93%	65.97%	77.84
Level 1	0.10	92.78%	73.80%	82.21
Level 2	0.14	91.28%	76.06%	82.98
Level 3	0.16	90.15%	76.76%	82.92

Concluding remark

Our method for detecting arbitrary noun phrases has improved the F rate of an NP chunker from 77.84 to 82.98. This error reduction of 23% is less than we had expected.

Address

address: Center for Dutch Language and Speech
Universitaire Instelling Antwerpen
Universiteitsplein 1
B-2610 Wilrijk
Belgium

phone: +31 3 8202765

e-mail: erikt@uia.ua.ac.be

www: http://lcg-www.uia.ac.be/ erikt/

Overhead sheets

http://lcg-www.uia.ac.be/˜erikt/talks/

Software (TiMBL)

http://ilk.kub.nl/

Data

http://lcg-www.uia.ac.be/conll99/npb/