Chunk/Shallow Parsing

Miriam Butt
October 2002

PP-Attachment

Recall the PP-Attachment Problem (demonstrated with XLE):

The girl saw the monkey with the telescope. — 2 readings

The ambiguity increases exponentially with each PP.

The girl saw the monkey with the telescope in the garden. — 4 readings

The girl saw the monkey with the telescope in the garden under the tree. — 9 readings

PP-Attachment

It would be nice to not have to solve the *Attachment* relations.

Or, at least, to leave them to a later component so that one can already do some *shallow parsing*.

The girl saw the monkey with the telescope.

[The girl] saw [the monkey] [with the telescope].
NP NP PP

Chunk Parsing

Steve Abney pioneered the notion of *chunk parsing*.

[I begin] [with an intuituion]: [when I read] [a sentence],
[I read it] [a chunk] [at a time].

- *Chunks* appear to correspond roughly to prosodic phrases (though nobody has really researched this from the phonology-syntax interface perspective).
- Strong stress will be once a chunk.
- Pauses are most likely to fall between chunks.

Chunk Parsing

Chunks are difficult to define precisely.

- "A typical chunk consists of a content word surrounded by a constellation of function words, matching a fixed template." (Abney 1991).

  [I begin] [with an intuituion]: [when I read] [a sentence], [I read it] [a chunk] [at a time].

  [The girl] saw [the monkey] [with the telescope].

- "By contrast, the relationship *between* chunks is mediated more by lexical selection than by rigid templates." (Abney 1991).

Chunk Parsing:A Three-Step Process

Abney envisioned chunk parsing as a *cascade* of (finite-state) processes:

- Word Identification (in practice mostly done via POS-Tagging)
- Chunk Identification (via rules)
- Attachment of Chunks (Parsing, also via rules)

Chunk Parsing:Word Identification

The effort to establish such a conclusion must have two foci, ...

[the/Det] [effort/N] [to/Inf-to/P] [establish/V] [such/Predet/Det/Pron] [a/Det] [conclusion/N] [of course/Adv] [must/N/V] [have/V] [two/Num] [foci/N] [,/Comma]

Problems: Ambiguity, Multiwords (need lexicons and stochastic methods)

Chunk Parsing:Chunk Identification

The effort to establish such a conclusion must have two foci, ...

[the effort/DP] [to establish/CP-inf] [such a conclusion/DP] [of course must have/CP] [two foci/DP] [,/Comma]

Problems: Where does a chunk end? What do I want to call my chunks?

Chunk Parsing:Attachment

The effort to establish such a conclusion must have two foci, ...

IP: DP: [the effort]

CP-inf: [[to establish] [such a conclusion]]

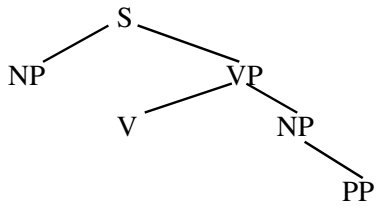
VP: [[of course must have] [two foci]]

[,/Comma]

Chunk Parsing:Attachment

Attachment Preferences:

- Prefer argument attachment, prefer verb attachment.
- Prefer low attachment.



The girl saw the monkey with the telescope.

Chunk Parsing:Finding an End

Unlike the ends of sentences, the ends of chunks are not overtly marked.

What to do?

Solution:

• Mark when you are the beginning (<B>) of a chunk or in the middle (<I>).

•"Hallucinate" end-of-chunk markers everywhere and then take those that fit the given template/pattern (this turns out not be costly).

• Prefer longer chunks to shorter chunks.

Chunk Parsing:Demo

1. TreeTagger (IMS) --- see handout
2. POS-Tagger and Chunker from Infogistics (on the web: www.infogistics.com)

Why is Shallow Parsing Good?

Shallow Parsing or *Partial Parsing* can be used for:

- Bootstrapping a more complete parser
- Constructing a tree bank (annotated text) which other applications can use.
- Extraction of specialized terminology or multi-words
- Information Retrieval (delimits the search space)

Noun Chunking Applications

Many applications seem to concentrate on *Noun Chunking* rather than trying to analyze a whole sentence.

That is: only the noun chunks are identified and labeled.

Church (1988): uses input of an HMM POS-tagger and brackets those that are noun chunks.

[		]		[	]	
DT	NN	VBD	IN	NN		CS
the	prosecutor	said	in	closing		that

Noun Chunking Applications

Example: A German noun chunker (Schmid and Schulte im Walde 2000)

Correct Identification: 92%

With Syntactic Category and Case Information: 83%

What is NP Chunking Good For?

Names tend to be a problem for parsers: new ones are constantly being made up and there is a lot of specialized terminology in medical texts or manuals.

NP Chunkers can help by identifying these special terms so that they can be listed in a specialized lexicon.

What is NP Chunking Good For?

Date/Time NPs and Name NPs have a very different syntax from normal NPs.

I can meet you on [Thursday the fifth of November at 12:30 pm].

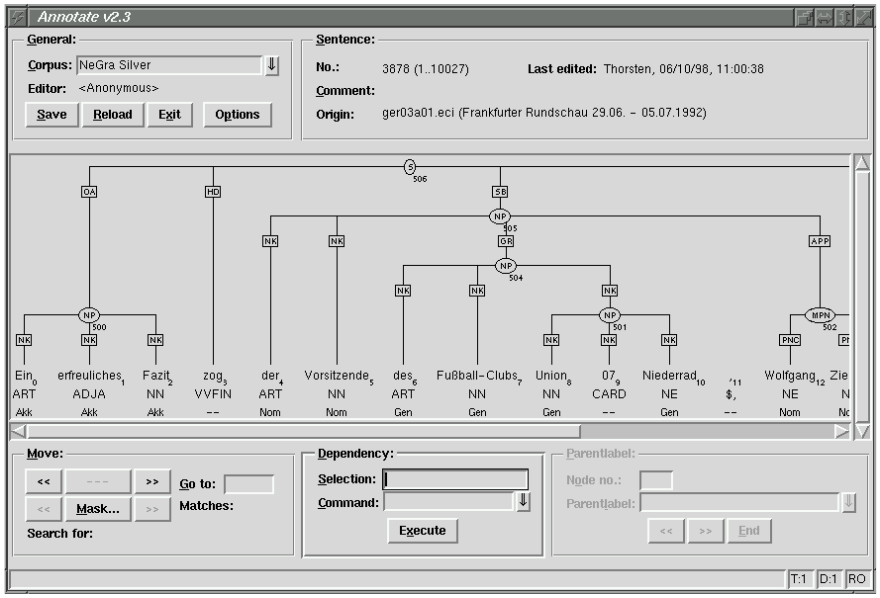
[Professor Dr. John A. Smith] arrived at the university.

NP Chunkers can again help to identifying these special NPs without needing to use a specialized, painfully handwritten grammar.

Interactive Annotation: Annotate

- Developed by Thorsten Brants and Oliver Plaehn at the DFKI (German Center for Artificial Intelligence).
- Includes a Stochastic POS-Tagger
- Allows for further manual annotation/parsing (identification of NPs/PPs, etc.)
- "Watches" the user with a statistical component and learns the grammar the user is working with.
- Provides the user with guesses as to an analysis which is correct about 70% of the time.

Interactive Annotation: Annotate



Types of Shallow Parsers

- Sophisticated versions of POS-Taggers
- Chunk Parsers (also rely on POS-Tags)
- Finite-State Parsers
- Stochastic Processes which "learn" a grammar
- Combinations of all of the above.