Text Mining

2004-2005
Master TKI

Antal van den Bosch en Walter Daelemans
http://ilk.uvt.nl/~antalb/textmining/

Dinsdag, 10.45 - 12.30, SZ33

Timeline

- [22 februari 2005]
  - Concept chunking (Sander Canisius)
- [1 maart 2005]
  - Syntactic pipeline 2: chunking, relation finding (WD)
- [8 maart 2005]
  - Named-entity recognition (Toine Borgers)

Outline

- Shallow Parsing
  - (Tokenization)
  - (POS Tagging)
  - Chunking
  - Relation-finding
- Applications
  - Information Extraction [15/3]
  - Ontology Extraction [26/4]
  - Question Answering
  - Factoid Extraction [3/5]

Shallow Parsing

- Steve Abney 1991 (FST)
  - http://www.vinartus.net/spa/
- Ramshaw & Marcus 1995 (TBL)
- CoNLL Shared tasks 1999, 2000, 2001
  - http://cnts.uia.ac.be/signll/shared.html
- JMLR special issue 2002
  - http://jmlr.csail.mit.edu/papers/special/shallow_parsing02.html

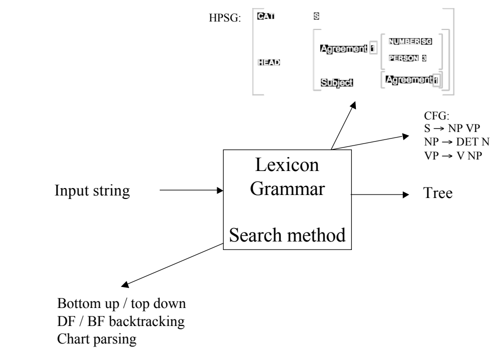
Formalisms for Computational Linguistics

Orthography	finite-state	spelling rules
Phonology	finite-state	text to speech
Morphology	finite-state	synthesis / analysis
	context-free	compounds
Syntax	context-free	parsing
	+ extensions	
Semantics	FOPC / CD	interpretation
Pragmatics		

- Classes of grammars are differentiated by means of a number of restrictions on the type of production rule
  - **Type-0-grammar** (unrestricted rewrite system). Rules have the form $\alpha \rightarrow \beta$
  - **Type-1-grammar** (context-sensitive). Rules are of the type $\alpha \rightarrow \beta$, where $|\alpha| \leq |\beta|$
  - **Type-2-grammar** (context-free). Rules are of the form $A \rightarrow \beta$, where $\beta \neq e$
  - **Type-3-grammar** (regular, finite). Rules are of the form $A \rightarrow a$ or $A \rightarrow aB$
- A grammar *generates* strings of *L(G)*, an automaton *accepts* strings of *L(M)*. Structure may be assigned as a side-effect.

The problem with full parsing

- Vicious trade-off coverage - ambiguity
  - The larger the grammar (more coverage), the more spurious ambiguity
- Why parsing ?
  - Structure of sentence determines its meaning



HPSG:

$$\begin{bmatrix} \text{CAT} & \text{S} \\ \text{HEAD} & \begin{bmatrix} \text{Agreement } \boxed{1} & \begin{bmatrix} \text{NUMBER SG} \\ \text{PERSON 3} \end{bmatrix} \\ \text{Subject} & \begin{bmatrix} \text{Agreement } \boxed{1} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

CFG:
S → NP VP
NP → DET N
VP → V NP

Input string

Lexicon
Grammar

Search method

Tree

Bottom up / top down
DF / BF backtracking
Chart parsing

Shallow parsing

- Approximate expressive power of CFG and feature-extended CFG by means of a *cascade* of simple transformations
- Advantages
  - deterministic (no recursion)
  - efficient (1600 words per second vs. 1 word per second for a typical comparison)
  - accurate
  - robust (unrestricted text, partial solutions)
  - can be learned

Cascade

- POS tagging
- NP chunking
- XP chunking
- Grammatical relation assignment
- Function assignment
- Parsing

Chunk Parsing

Pierre Vinken, 61 years old, will join the board of directors as a non-executive director November 29.

Pierre/NNP Vinken/NNP ,/, 61/CD years/NNS old/JJ ,/, will/MD join/VB the/DT board/NN of/IN directors/NNS as/IN a/DT non-executive/JJ director/NN November/NNP 29/CD ./.

[$_{NP}$ Pierre Vinken $_{NP}$] , [$_{NP}$ 61 years $_{NP}$] old , [$_{VP}$ will join $_{VP}$] [$_{NP}$ the board $_{NP}$] of [$_{NP}$ directors $_{NP}$] as [$_{NP}$ a non-executive director $_{NP}$] [$_{NP}$ Nov 29 $_{NP}$]

Approaches

<u>Deductive</u>

CASS-parser (Abney, 1991)
Finite-State

Fidditch (Hindle, 1994)
Rule-based

<u>Inductive</u>

Ramshaw & Marcus, 1995
Transformation Rules

Daelemans/Buchholz/Veenstra, 1999; Tjong Kim Sang, 2000
Memory-based

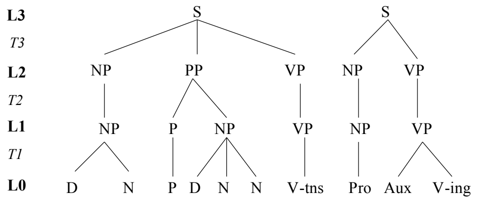
Abney (1991): CASS-parser

- Chunk = maximal, continuous, non-recursive syntactic segment around a head
- Comparable to morphologically complex word in synthetic languages
- Motivation
  - Linguistic (incorporate syntactic restrictions)
  - Psycholinguistic
  - Prosodic (phonological phrases)

Levels and Transformations

Levels

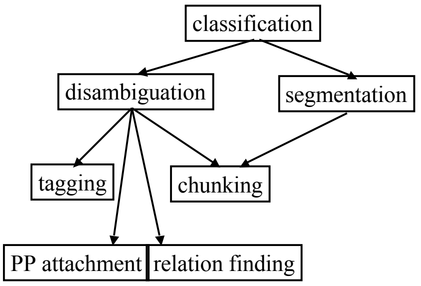
- words and their part of speech tags
- chunks (kernel NP, VP, AP, AdvP)
  - NP → D? N* N
  - VP → V-tns | Aux V-ing
- simple phrases (transforming embedding to iteration)
  - PP → P NP
- complex phrases
  - S → PP* NP PP* VP PP*



The woman in the lab coat thought you were sleeping

- Pattern = category + regular expression
- Regular expression is translated into FSA
- For each T_i we take the union of the FSAs to construct a recognizer for level $\mathbf{L_i}$
- In case of more than one end state for the same input, choose the longest
- In case of blocking, advance one word
- “Easy-first parsing” (islands of certainty)
- Extensions: add features by incorporating actions into FSAs

MBLP Cascade: shallow parsing



classification

disambiguation

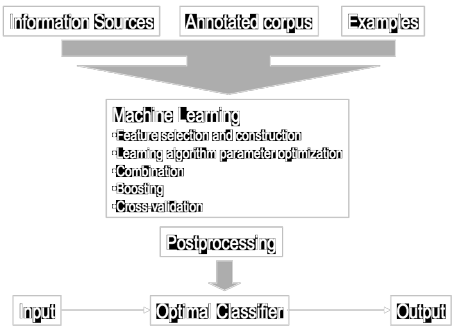
segmentation

tagging

chunking

PP attachment

relation finding



Information Sources

Annotated corpus

Examples

Machine Learning

- Feature selection and construction
- Learning algorithm parameter optimization
- Combination
- Boosting
- Cross-validation

Postprocessing

Input

Optimal Classifier

Output

NP Chunking as tagging

[$_{NP}$ Pierre Vinken $_{NP}$] , [$_{NP}$ 61 years $_{NP}$] old , [$_{VP}$ will join $_{VP}$] [$_{NP}$ the board $_{NP}$] of [$_{NP}$ directors $_{NP}$] as [$_{NP}$ a non-executive director $_{NP}$] [$_{NP}$ Nov 29 $_{NP}$]

Pierre/I Vinken/I ,/O 61/I years/I old/O ,/O will/O join/O the/I board/I of/O directors/I as/O a/I non-executive/I director/I Nov/B 29/I ./O

I	Inside chunk
O	Outside chunk
B	Between chunks

Memory-Based XP Chunker

Assigning non-recursive phrase brackets (Base XPs) to phrases in context:

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	

Convert NP, VP, ADJP, ADVP, PrepP, and PP brackets to classification decisions (I/O/B tags) (Ramshaw & Marcus, 1995).

Features:

POS $_{-2}$, IOBtag$_{-2}$, word $_{-2}$,
POS $_{-1}$, IOBtag$_{-1}$, word $_{-1}$,
POS $_{focus}$, word$_{focus}$,
POS $_{+1}$,
word $_{+1}$, POS $_{+2}$, word $_{+2}$, → IOB tag

Memory-Based XP Chunker

- Results (WSJ corpus)

[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

- One-pass segmentation and chunking for all XP
- Useful for: Information Retrieval, Information Extraction, Terminology Discovery, etc.

Finding subjects and objects

- Problems
  - One sentence can have more than one subject/object in case of more than one VP
  - One VP can have more than one subject/object in case of conjunctions
  - One NP can be linked to more than one VP
  - subject/verb or verb/object can be discontinuous

Task Representation

- From tagged and chunked sentences, extract
  - Distance from verb to head in chunks
  - Number of VPs between verb and head
  - Number of commas between verb and head
  - Verb and its POS
  - Two words/chunks context to left, word + POS
  - One word/chunk context to right
  - Head itself

Memory-Based GR labeling

Assigning labeled Grammatical Relation links between words in a sentence:

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]	[illegible]	[illegible]			

GR's of Focus with relation to Verbs (subject, object, location, …, none)

Features:

Focus: prep, adv-func, word$_{+1}$, word$_0$, word$_{-1}$, word$_{-2}$, POS$_{+1}$, POS$_0$, POS$_{-1}$, POS$_{-2}$, Chunk$_{+1}$, Chunk$_0$, Chunk$_{-1}$, Chunk$_{-2}$.
Verb: POS, word,
Distance: words, VPs, comma's
→ GRtype

Memory-Based GR labeling

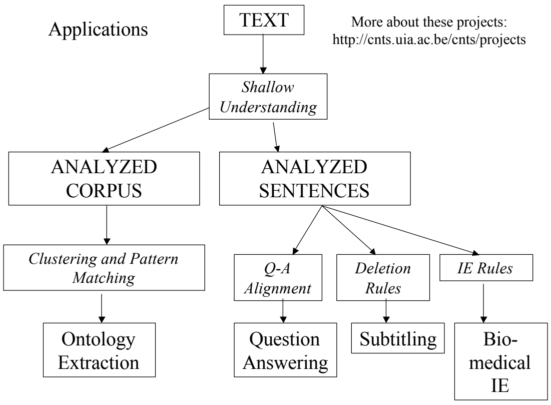
- Results (WSJ corpus)

[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]

- Subjects: 83%, Objects: 87%, Locations:47%, Time:63%
- Completes shallow parser. Useful for e.g. Question Answering, IE etc.

From POS tagging to IE Classification-Based Approach

- POS tagging
  The/Det woman/NN will/MD give/VB Mary/NNP a/Det book/NN
- NP chunking
  The/I-NP woman/I-NP will/I-VP give/I-VP Mary/I-NP a/B-NP book/I-NP
- Relation Finding
  [NP-SUBJ-1 the woman] [VP-1 will give] [NP-I-OBJ-1 Mary] [NP-OBJ-1 a book]]
- Semantic Tagging = Information Extraction
  [Giver the woman][will give][Givee Mary][Given a book]
- Semantic Tagging = Question Answering
  Who will give Mary a book?
  [Giver ?][will give][Givee Mary][Given a book]



Applications

TEXT

More about these projects:
http://cnts.uia.ac.be/cnts/projects

Shallow Understanding

ANALYZED CORPUS

ANALYZED SENTENCES

Clustering and Pattern Matching

Q-A Alignment

Deletion Rules

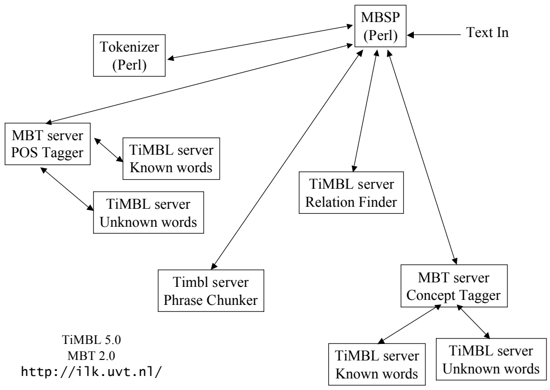
IE Rules

Ontology Extraction

Question Answering

Subtitling

Bio-medical IE



MBSP (Perl)

Text In

Tokenizer (Perl)

MBT server POS Tagger

TiMBL server Known words

TiMBL server Unknown words

TiMBL server Relation Finder

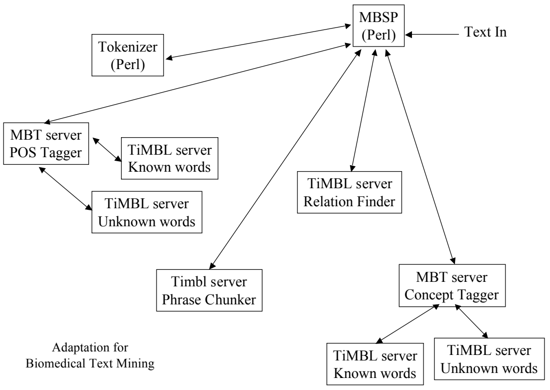
Timbl server Phrase Chunker

MBT server Concept Tagger

TiMBL server Known words

TiMBL server Unknown words

TiMBL 5.0
MBT 2.0
`http://ilk.uvt.nl/`



MBSP (Perl)

Text In

Tokenizer (Perl)

MBT server POS Tagger

TiMBL server Known words

TiMBL server Unknown words

TiMBL server Relation Finder

Timbl server Phrase Chunker

MBT server Concept Tagger

TiMBL server Known words

TiMBL server Unknown words

Adaptation for
Biomedical Text Mining

What should be in?

- Shallow parsing (tagging, chunking, grammatical relations)
- Semantic roles
- Domain semantics (NER / concept tagging)
- Negation, modality, quantification can be solved as classification tasks?

Conclusions

- Text Mining tasks benefit from linguistic analysis (shallow understanding).
- Understanding can be formulated as a flexible heterarchy of classifiers.
- These classifiers can be trained on annotated corpora.

Assignment 1

- http://ilk.uvt.nl/~antalb/textmining/assignment1.html