

# Memory-Based Shallow Parsing for Text Mining

*Walter Daelemans*

walter.daelemans@ua.ac.be

<http://cnts.uia.ac.be>

CNTS, University of Antwerp, Belgium  
ILK, Tilburg University, Netherlands

*PASCAL Text Mining Workshop Tutorial  
Grenoble January 27, 11:20-12:50*

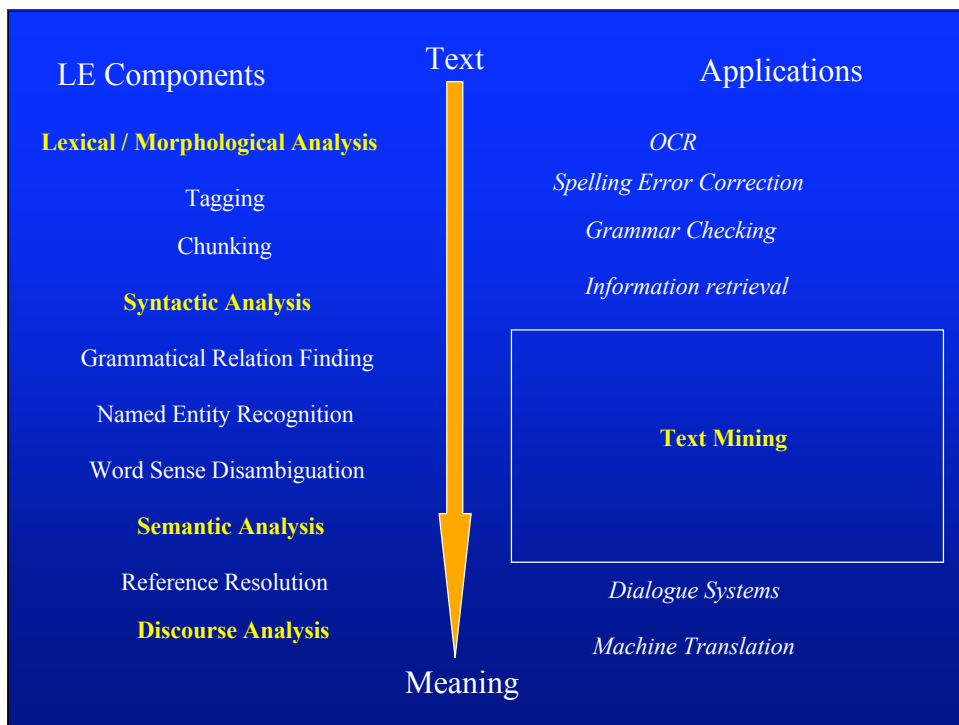
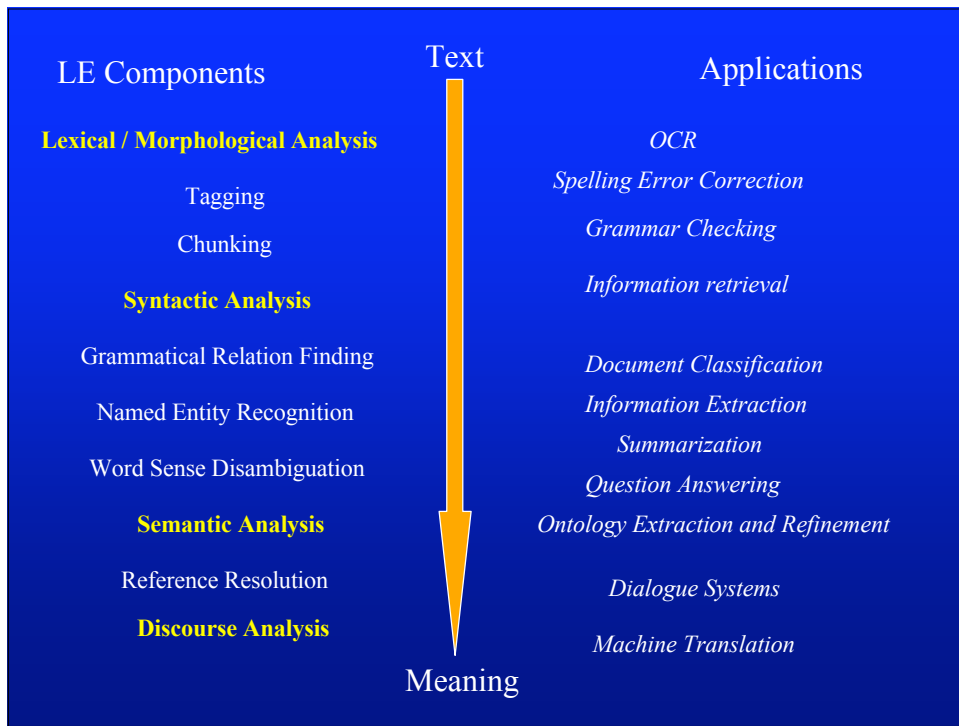
## Outline

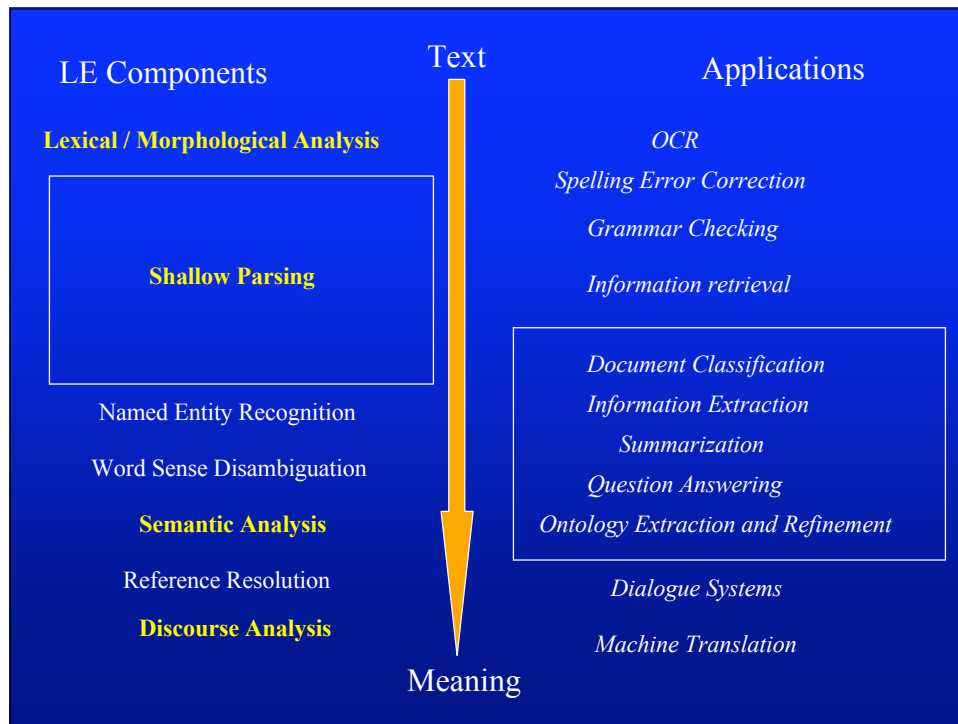
- Shallow Parsing for Text Mining Applications
  - Ontology Extraction
  - Question Answering
  - Automatic Summarization
- Shallow Parsing as memory-based classification
  - Eager and lazy learning
  - Tagging
  - Chunking
  - Relation-finding
  - Information Extraction and question answering as classification
- An interesting optimization problem

## (1) Shallow Parsing for Text Mining

### Text Mining

- Automatic extraction of reusable information (knowledge) from text, based on linguistic analysis of the text
- Goals:
  - Data mining (KDD) from unstructured and semi-structured data
  - Knowledge Management and retrieval
  - “Intelligence”
- Examples:
  - Email routing and filtering
  - Finding protein interactions in biomedical text
  - Matching on-line resumes and vacancies





## 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://www.ai.mit.edu/projects/jmlr/papers/special/shallowparsing.html>
- Example:
- The woman will give Mary a book

## POS Tagging

The/**Det** woman/**NN** will/**MD** give/**VB**  
Mary/**NNP** a/**Det** book/**NN**

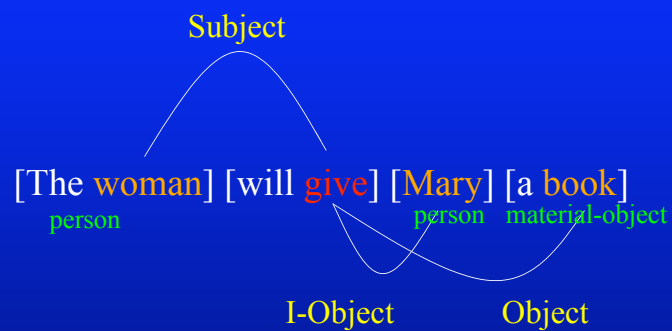
## Chunking

[The/**Det** woman/**NN**]<sub>NP</sub> [will/**MD** give/**VB**]<sub>VP</sub>  
[Mary/**NNP**]<sub>NP</sub> [a/**Det** book/**NN**]<sub>NP</sub>

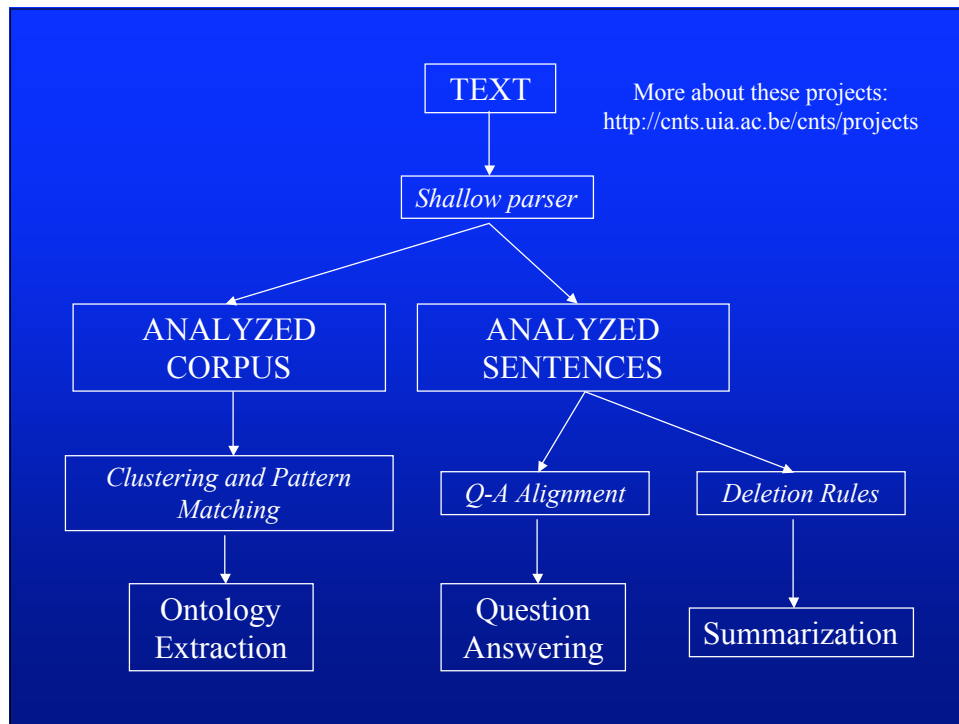
## Named Entity Recognition / Sense Tagging

[The/Det woman/NN]<sub>NP-PERSON</sub> [will/MD  
give/VB]<sub>VP</sub> [Mary/NNP]<sub>NP-PERSON</sub> [a/Det  
book/NN]<sub>NP-MATERIAL-OBJECT</sub>

## Relation Finding



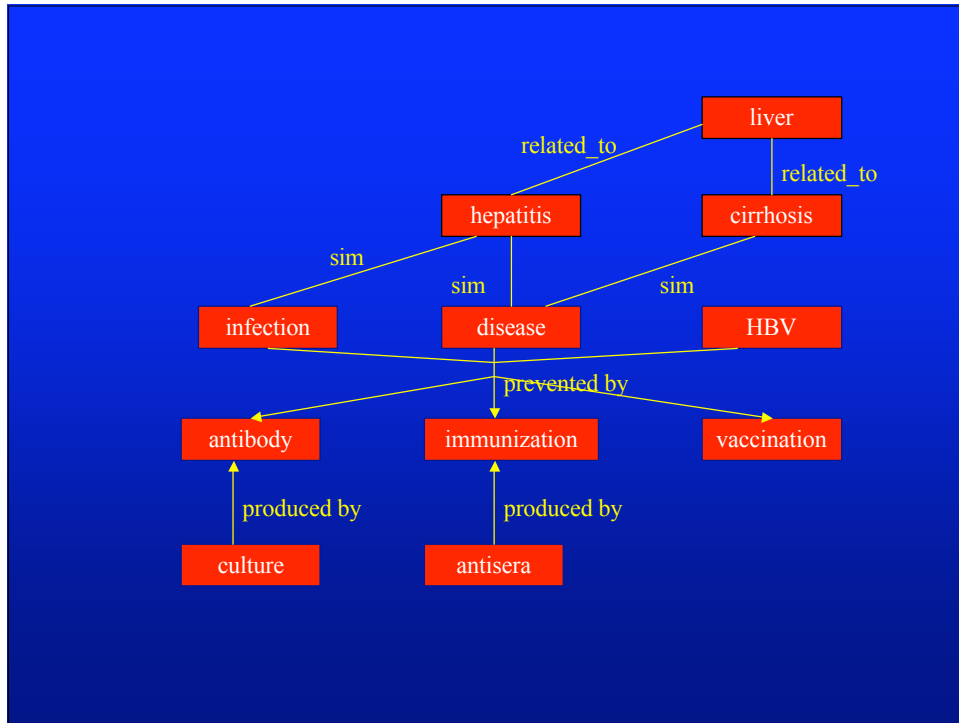
*WHO, WHAT, WHERE, WHEN ... ?*



## Application: Ontology Extraction

- Clustering of head nouns of Subject-Verb and Verb-Object relations
- Combine with pattern matching and heuristics
- Case study: Medline 4 million words hepatitis
- Results:
  - Better clusters with shallow parsing
  - Useful in knowledge management, thesaurus development, ...

Ontobasis (IWT): Marie-Laure Reinberger



## Application: Question Answering

- Give answer to question  
(document retrieval: find documents relevant to query)
- Who invented the telephone?  
– Alexander Graham Bell
- When was the telephone invented?  
– 1876

PhD project (NWO): Sabine Buchholz

## QA System: Shapaqa

- Parse question
  - When was the telephone invented?*
  - Which slots are given?
    - Verb           invented
    - Object         telephone
  - Which slots are asked?
    - Temporal phrase linked to verb
- Document retrieval on internet with given slot keywords
- Parsing of sentences with all given slots
- Count most frequent entry found in asked slot (temporal phrase)

## Shapaqa: example

- *When was the telephone invented?*
- Google: **invented** AND “**the telephone**”
  - produces 835 pages
  - 53 parsed sentences with both slots and with a temporal phrase

is through his interest in Deafness and fascination with acoustics that **the telephone** was **invented in 1876** , with the intent of helping Deaf and hard of hearing

**The telephone** was **invented** by Alexander Graham Bell **in 1876**

When Alexander Graham Bell **invented the telephone in 1876** , he hoped that these same electrical signals could

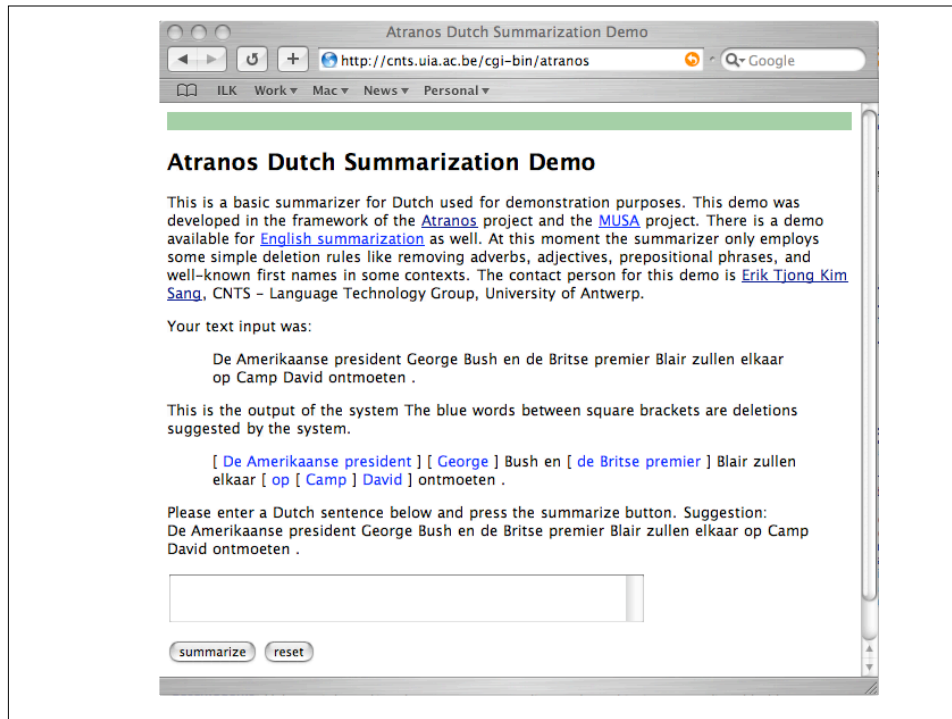
## Shapaqa: example (2)

- So when was the phone invented?
- Internet answer is noisy, but robust
  - 17: 1876
  - 3: 1874
  - 2: ago
  - 2: later
  - 1: Bell
  - ...
- System was developed quickly
- Precision 76% (Google 31%)
- International competition (TREC): MRR 0.45

## Application: Automatic Summarization

- Approach
  - Shallow parsing of sentences
  - Paraphrase table
  - Word information values
  - Deletion Rules
- Application: automatic subtitling from speech input

MUSA project (EU): Anja Höthker  
ATRANOS project (IWT): Erik Tjong Kim Sang



## (2) Shallow Parsing as memory-based classification

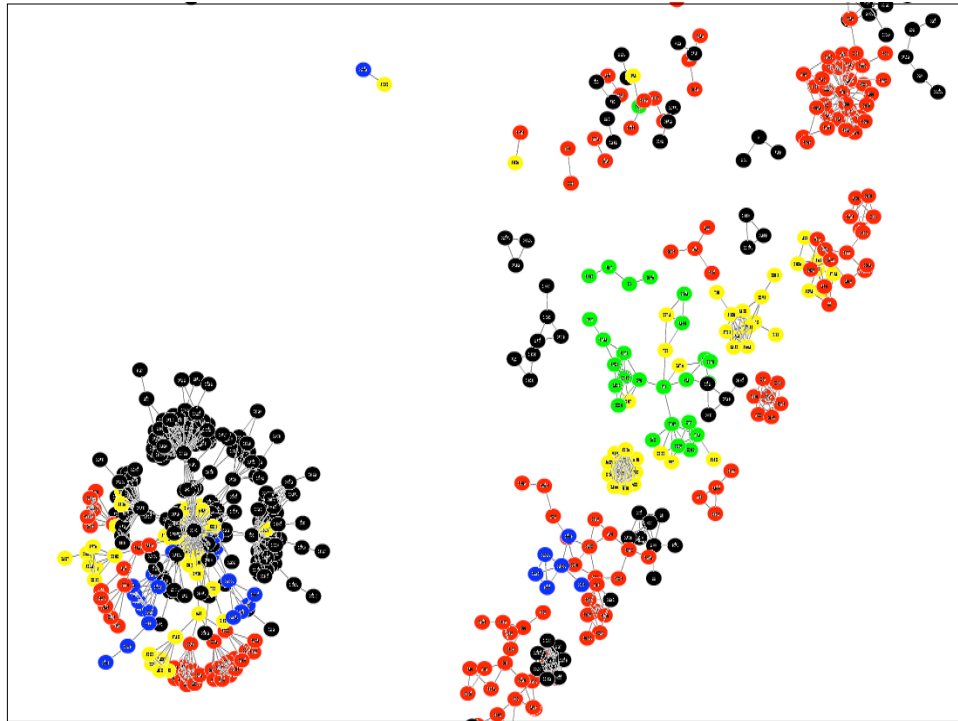
## Eager vs Lazy Learning

- Eager: Decision tree learning, Rule induction, Hyperplane discriminators, Probabilistic classifiers
  - Use MDL principle to guide search for representation
  - Learning is compression
- Lazy: example-based (MBL)
  - Retains every piece of information available at training time



## Generalization = Abstraction?

- In language data, what is core? What is periphery?
- (Sub-)regularities, pockets of exceptions
- Disjunctiveness, poly-morphic concepts
- Zipf-distributions
- Hard to distinguish noise from exceptions on the basis of
  - Frequency
  - Typicality
- Forgetting Exceptions is harmful in Language Learning (Daelemans, van den Bosch, Zavrel, 1999, *Machine Learning*)



MBL: Use memory traces of experiences as a basis for analogical reasoning, rather than using rules or other abstractions extracted from experience and replacing the experiences.

*This “rule of nearest neighbor” has considerable elementary intuitive appeal and probably corresponds to practice in many situations. For example, it is possible that much medical diagnosis is influenced by the doctor's recollection of the subsequent history of an earlier patient whose symptoms resemble in some way those of the current patient. (Fix and Hodges, 1952, p.43)*

## Operationalization

- Basis:  $k$  nearest neighbor algorithm:
  - store all examples in memory
  - to classify a new instance  $X$ , look up the  $k$  examples in memory with the smallest distance  $D(X, Y)$  to  $X$
  - let each nearest neighbor vote with its class
  - classify instance  $X$  with the class that has the most votes in the nearest neighbor set
- Choices:
  - similarity metric
  - number of nearest neighbors ( $k$ )
  - voting weights

### ib1

$$D(X, Y) = \sum_{i=1}^n w(f) D(x_i, y_i)$$

$$D(x_i, y_i) = \frac{|x_i - y_i|}{\max_i + \min_i}$$

$$D(x_i, y_i) = 0 \text{ if } x_i = y_i \text{ else } 1$$

### ib1-mvdm

$$D(v1, v2) = \sum_c |p(C | v1) - p(C | v2)|$$

### Metrics

### ib1-ig

$$w(f) = \sum_c P(C) \log_2 P(C)$$

$$\sum_v \sum_c P(V_f) \sum_c P(C | V_f) \log_2 P(C | V_f)$$

## Distance weighting

- Relation between larger k and smoothing
- Make more distant neighbors contribute less in the class vote
  - Linear inverse of distance (w.r.t. max)
  - Inverse of distance
  - Exponential decay

## The properties of NLP tasks ...

- NLP tasks are mappings between linguistic representation levels that are
  - context-sensitive (but mostly local!)
  - complex (sub/ir/regularity), pockets of exceptions
- Similar representations at one linguistic level correspond to similar representations at the other level
- Several information sources interact in (often) unpredictable ways at the same level
- Data is sparse

## ... fit the bias of MBL

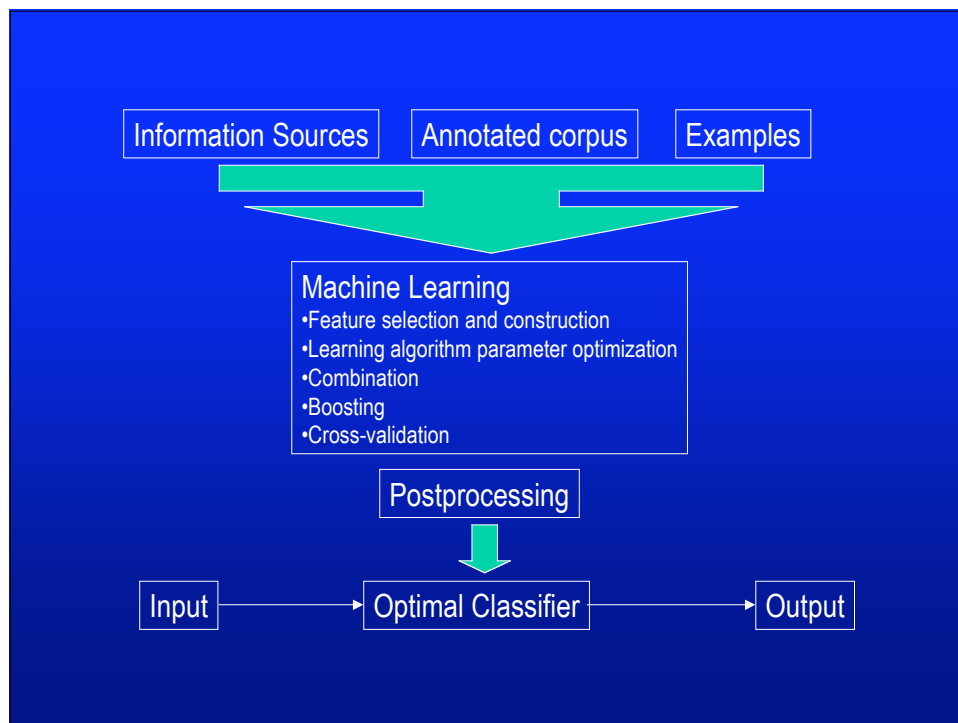
- The mappings can be represented as (cascades of) *classification* tasks (*disambiguation* or *segmentation*)
- Locality is implemented through windowing over representations
- Inference is based on Similarity-Based / Analogical Reasoning
- Adaptive data fusion / relevance assignment is available through feature weighting
- It is a non-parametric approach
- Similarity-based smoothing is implicit
- Regularities and subregularities / exceptions can be modeled uniformly

## TiMBL 5.0

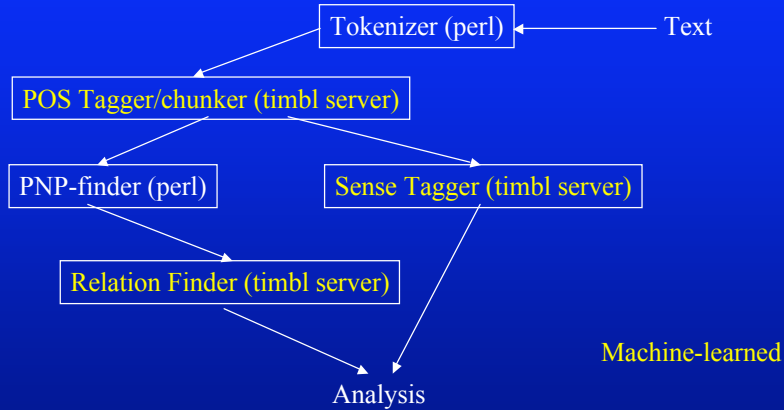
- <http://ilk.uvt.nl> (includes detailed reference guide and guide to related literature)
- Different algorithms (ib1, ib2, igtree, tribl, tribl2)
- Different feature and value weighting methods (information gain, chi-square, mvdM, Jeffrey Divergence...)
- API, server version
- C++
- Free for research (with source code)

# Machine Learning of Shallow Parsing

- Requirements
  - Robust
  - Adaptable to new domains
  - Fast and efficient
- Approach: MBSP (Memory-Based Shallow Parser)
  - Cascade of TiMBL servers
  - Domain adaptation at level of
    - POS tagger (lexicon)
    - Named-Entity Recognition / sense tagger (domain ontology)
  - ~ 1000 words per second



# MBSP



## Memory-Based POS Tagger

Assigning morpho-syntactic categories (parts-of-speech) to words in context:

The	green	train	runs	down	that	track	.
Det	Adj/NN	NNS/VBZ	NN/VB	Prep/Adv/Adj	SC/Pron	NN/VB	.
Det	Adj	NNS	VB	Prep	Pron	NN	.

Disambiguation: resolution of a combination of lexical and local contextual constraints.

- Lexical representations: Frequency-sensitive ambiguity class lexicon.
- Convert sentences to MBL cases by 'windowing'. Local constraints are modeled by features of neighboring words.

## Memory-Based POS Tagger

- Case base for known words. Features:

$\text{tag}_{-2}, \text{tag}_{-1}, \text{lex}_{\text{focus}}, \text{word}^{(\text{top}100)}_{\text{focus}}, \text{lex}_{+1}, \text{lex}_{+2} \square \text{POS tag}$

- Case base for unknown words. Features:

$\text{tag}_{-2}, \text{tag}_{-1}, \text{pref}, \text{cap}, \text{hyp}, \text{num}, \text{suf1}, \text{suf2}, \text{suf3}, \text{lex}_{+1}, \text{lex}_{+2}$   
 $\square \text{POS tag}$

## Memory-Based POS Tagger

- Experimental results:

language	tagset size	train	test	accuracy
English WSJ	44	2000	200	96.4
English LOB	170	931	115	97.0
Dutch	13	611	100	95.7
Czech	42	495	100	93.6
Spanish	484	711	89	97.8
Swedish	23	1156	11	95.6

## Memory-Based XP Chunker

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

[ <sub>NP</sub> The	woman <sub>NP</sub> ]	[ <sub>VP</sub> will	give <sub>VP</sub> ]	[ <sub>NP</sub> Mary <sub>NP</sub> ]	[ <sub>NP</sub> a	book <sub>NP</sub> ]	.
Det	NN	MD	VB	NNP	Det	NN	.
I-NP	I-NP	I-VP	I-VP	I-NP	B-NP	I-NP	

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

Features:

POS<sub>-2</sub>, IOBtag<sub>-2</sub>, word<sub>-2</sub>,  
 POS<sub>-1</sub>, IOBtag<sub>-1</sub>, word<sub>-1</sub>,  
 POS<sub>focus</sub>, word<sub>focus</sub>,  
 POS<sub>+1</sub>,  
 word<sub>+1</sub>, POS<sub>+2</sub>, word<sub>+2</sub>,  IOB tag

## Memory-Based XP Chunker

- Results (WSJ corpus)

type	prec	recall	F1
NP	92.5	92.2	92.3
VP	91.9	91.7	91.8
ADJP	68.4	65.0	66.7
ADVP	78.0	77.9	77.9
Prep	95.5	96.7	96.1
PP	91.9	92.2	92.0
ADVFunc	78.0	69.5	73.5

- Useful for: Information Retrieval, Information Extraction, Terminology Discovery, etc.

## Memory-Based GR labeling

Assigning labeled Grammatical Relation links between words in a sentence:

The	woman	will	give	Mary	a	book	.
Det	NN	MD	VB	NNP	Det	NN	.
I-NP	I-NP	I-VP	I-VP	I-NP	B-NP	B-NP	
	SUBJ-1	VP-1	VP-1	OBJ-1			

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

Focus: prep, adv-func, word<sub>+1</sub>, word<sub>0</sub>, word<sub>-1</sub>, word<sub>-2</sub>, POS<sub>+1</sub>, POS<sub>0</sub>, POS<sub>-1</sub>,  
POS<sub>-2</sub>, Chunk<sub>+1</sub>, Chunk<sub>0</sub>, Chunk<sub>-1</sub>, Chunk<sub>-2</sub>.  
Verb: POS, word,  
Distance: words, VPs, comma's  
□ GRtype

## Memory-Based GR labeling

- Results (WSJ corpus)

features	prec	recall	F1
words+POS only	60.7	41.3	49.1
+NPs	65.9	55.7	60.4
+VPs	72.1	62.9	67.2
+ADJPs +ADVPs	72.1	63.0	67.3
+Preps	72.5	64.3	68.2
+PPs	73.6	65.6	69.3
+ADVFunc	74.8	67.9	71.2

- 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]

## Conclusion (1)

- Text Mining tasks benefit from linguistic analysis (shallow parsing)
- Problems ranging from shallow parsing to application-oriented tasks (Information Extraction) can be formulated as classification-based learning tasks
- These classifiers can be trained on annotated corpora
- Memory-Based Learning seems to have the right bias for this type of task (can cope with rich feature sets and exceptions)

## An interesting optimization problem

## Parameter and Feature Optimization: A Problem

- Importance of comparative Machine Learning experiments
- Methodological problem: effect of feature and algorithm parameter interaction
- Walter Daelemans, Véronique Hoste, Fien De Meulder and Bart Naudts, Combined Optimization of Feature Selection and Algorithm Parameter Interaction in Machine Learning of Language. In: Proceedings of the 14th European Conference on Machine Learning (ECML-2003), Cavtat-Dubrovnik, Croatia, 2003.
- Walter Daelemans and Véronique Hoste, Evaluation of Machine Learning Methods for Natural Language Processing Tasks . In: Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), pp. 755-760, Las Palmas, Gran Canaria, 2002.

## Why Comparative ML experiments in NLP ?

- Evaluate bias of ML method for some (class of) NLP tasks
- Evaluate the role of different information sources in solving a ML of NL task
- Examples:
  - EMNLP, CoNLL, ACL, ...
  - Competitions:
    - SENSEVAL
    - CoNLL shared tasks
    - TREC / MUC / DUC / ...

## What influences the outcome of a (comparative) ML experiment?

- Information sources
  - feature selection
  - feature representation (data transforms)
- Algorithm parameters
- Training data
  - sample selection
  - sample size
- Combination methods
  - bagging, boosting
  - output coding
- Interactions
  - Algorithm parameters and sample selection
  - Algorithm parameters and feature representation
  - Feature representation and sample selection
  - Sample size and feature selection
  - *Feature selection and algorithm parameters*
  - ...

### Current Practice Comparative ML Experiments

- Methodology: k-fold cross-validation, McNemar, paired t-test, learning curves, etc.
- Use default algorithm parameters
- Sometimes: algorithm parameter optimization
- Sometimes: feature selection
- Never: combined feature selection and parameter optimization  
= combinatorial optimization problem

## Hypothesis

The variability in accuracy resulting from interactions of algorithm parameter settings and feature selection is higher than the accuracy difference between two algorithms given constant input features and default algorithm parameter settings.

Therefore: many published comparative machine learning experiment results (and their interpretation) are not reliable.

## Case Study: Word Sense Disambiguation

- Decide on the contextually appropriate word sense given local information (collocations, keywords, pos tags, syntactic structure, ...)
- Supervised ML methods outperform knowledge-based and unsupervised learning approaches
- Senseval-1, Senseval-2 lexical sample and all-word tasks, different languages
- Which information sources?
- Which machine learning method?

## Experiment 1

- Investigate the effect of
  - algorithm parameter optimization
  - feature selection (forward selection)
  - interleaved feature selection and parameter optimization
- ... on the comparison of two inductive algorithms (lazy and eager)
- ... for a selection of NLP task datasets
  - Word Sense Disambiguation, tagging known words and unknown words, diminutive morphology

## Algorithms compared

- Ripper
  - *Cohen, 95*
  - Rule Induction
  - Algorithm parameters: different class ordering principles; negative conditions or not; loss ratio values; cover parameter values
- TiMBL
  - *Daelemans/Zavrel/van der Sloot/van den Bosch, 98*
  - Memory-Based Learning
  - Algorithm parameters: ib1, igtrees; overlap, mvdm; 5 feature weighting methods; 4 distance weighting methods; 10 values of k

## WSD (line)

Similar: little, make, then, time, ...

	Ripper	TiMBL
Default	21.8	20.2
Optimized parameters	22.6	27.3
Optimized features	20.2	34.4
Optimized parameters + FS	33.9	38.6

## Experiment 2

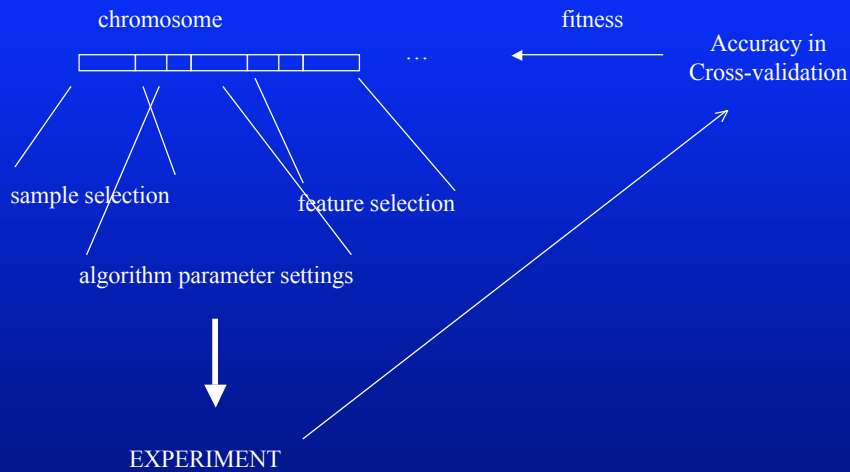
- Investigate the effect of
  - algorithm parameter optimization
- ... on the comparison of different knowledge sources for one inductive algorithm (TiMBL)
- ... for WSD
  - Local context
  - Local context and keywords

## TiMBL-WSD (do)

Similar: experience, material, say, then, ...

	Local Context	+ keywords
Default	49.0	47.9
Optimized parameters LC	60.8	59.5
Optimized parameters	60.8	61.0

## Solution: Genetic Algorithms?



## Conclusion (2)

- Optimizing algorithm parameter setting and feature selection interaction has a huge effect on generalization accuracy and on the comparison of ML algorithms and information sources
- Current published results are methodologically correct but nevertheless unreliable
- For many problems and algorithms, this optimization is computationally not feasible
- Current research: optimization using GAs
- Is the ML of NL field in need of new goals?