

Information Extraction

Natural Language Processing Basics

Agenda

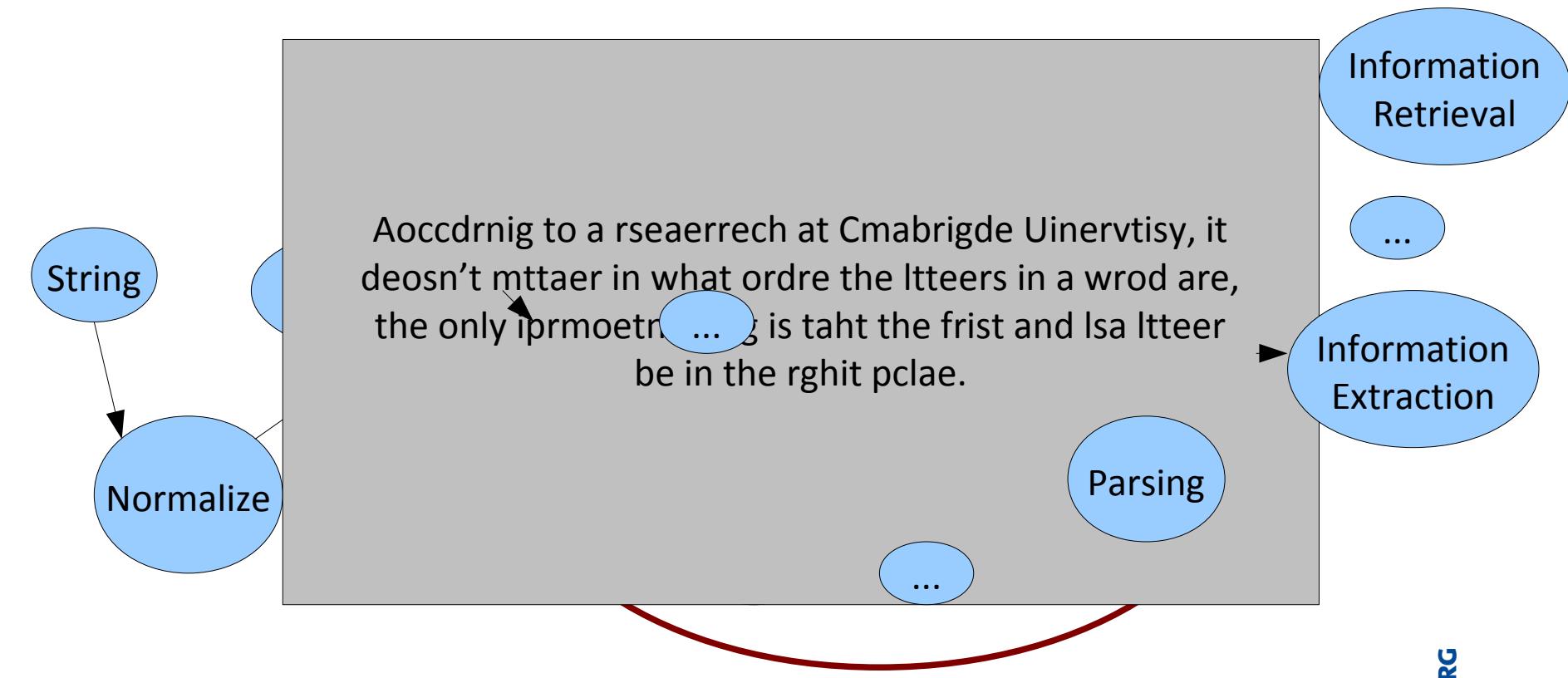
- Motivation
- Roadmap
- History
- Linguistics
- openNLP
- Detailed areas of NLP
 - Part-Of-Speech Tagging
 - Noun-Phrase Chunking
 - Named-Entity Recognition
 - Constituency Parsing
 - Dependency Parsing
- Benchmark
- Conclusion
- Literature

Motivation

- Semantic search engines (broccoli etc)
- Automated Translation

Roadmap

Mark left the window open to your left.



History

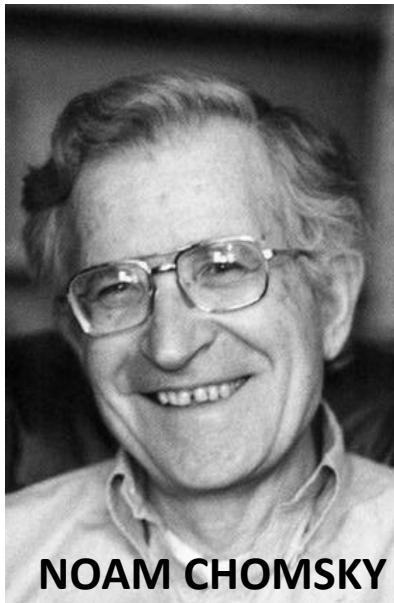
- 50's to 80's of the 19th Century
 - Alan Turing
 - Noam Chomsky
 - NSA Operation „Ivy Bells“ - Soviet Underwater Communication Lines (Déjà-vu ?)
 - Hand-Written Rules
- Late 80's
 - Machine Learning
- For a deeper look, consider wikipedia (as I do :-))

Linguistics (little freshup)

- Part Of Speech (Wortarten)
 - Noun
 - Pronoun
 - Adjective
 - Verb
 - Adverb
 - Preposition
 - Conjunction
 - Interjection
- But there are more subtypes

Linguistics (little freshup)

- Words or Phrases have a „Head“
- What??



It determines the syntactic type of the word/phrase

Examples:

Big red dog
birdsong

openNLP



- Implemented in java
- Apache License 2.0

Part-Of-Speech Tagging

- „Mark left the window open to your left.“
- How to tag „left“?
- Generally 2 types of approaches
 - Rule-based
 - There are many, e.g. [Brill1992]
 - Stochastic
 - e.g. k-breadth first search described in [Adwait1998] and used in openNLP

Part-Of-Speech Tagging

- Tag each word with its word type
- Possible at segmented writing systems
- Different Tagsets
 - Penn Treebank (English, also used in openNLP)
 - STTS (German, Stuttgart-Tübingen-Tagset)
 - And more ...

Part-Of-Speech Tagging

- „Mark_NNP left_VBD the_DT window_NN open_JJ to_TO your_PRP\$ left_VBN“
CC - Coordinating conjunction
CD - Cardinal number
DT - Determiner
EX - Existential there
FW - Foreign word
IN - Preposition or subordinating conjunction
JJ - Adjective
JJR - Adjective, comparative
JJS - Adjective, superlative
LS - List item marker
MD - Modal
NN - Noun, singular or mass
NNS - Noun, plural
NNP - Proper noun, singular
NNPS - Proper noun, plural
PDT - Predeterminer
POS - Possessive ending
PRP - Personal pronoun
PRP\$ - Possessive pronoun (prolog version PRP-S)
RB - Adverb
RBR - Adverb, comparative
RBS - Adverb, superlative
RP - Particle
SYM - Symbol
TO - to
UH - Interjection
VB - Verb, base form
VBD - Verb, past tense
VBG - Verb, gerund or present participle
VBN - Verb, past participle
VBP - Verb, non-3rd person singular present
VBZ - Verb, 3rd person singular present
WDT - wh-determiner
WP - wh-pronoun
WP\$ - Possessive wh-pronoun (prolog version WP-S)
WRB - wh-adverb|
- Rule based approach:
 - [...]
 - NN ate fhassuf 3 VB x
 - NNP ing fhassuf 3 VBG x
 - VBG is fgoodleft NN x
 - NN less fhassuf 4 JJ x
 - NN ary fhassuf 3 JJ x
 - Co. goodleft NNP x
 - NN ant fhassuf 3 JJ x
 - [...]
- Excerpt from [Brill.1994]

Part-Of-Speech Tagging

- „Mark_NNP left_VBD the_DT window_NN open_JJ to_TO your_PRP\$ left_VBN“
- Maximum Entropy Solution from openNLP
- Example: We are at „the“
- Get the feature vector

w=the
Suff=
Pre=
p=left
t=VBD
pp=Mark
t2=NNPVBD
n>window
nn=open

=>

Machine learn algorithm
which calculate a
probability
(Multinomial Logistic
Regression)

DEMO

Part-Of-Speech Tagging

■ State of the Art

System	Description	All Tokens	Unknown Words
SCCN	Semi-supervised condensed nearest neighbor	97.50%	Not available
Stanford Tagger 2.0	Maximum entropy cyclic dependency network	97.32%	90.79%
LAPOS	Perceptron based training with lookahead	97.22%	Not available

Performance measure: per token accuracy. (The convention is for this to be measured on all tokens, including punctuation tokens and other unambiguous tokens.)

Training data: sections 0-18 from Penn Treebank Wall Street Journal (WSJ) release 3

Testing data: sections 22-24 from Penn Treebank Wall Street Journal (WSJ) release 3

Excerpt from [AssociationforComputationalLinguistics.03.09.2013]

Noun-Phrase Chunking

- Noun-Phrase Chunking
 - Head of the phrase is a noun
 - „**Mark left the window open to your left.**“
 - [NP **Mark_NNP**]
[VP **left_VBD**]
[NP **the_DT window_NN**]
[ADJP **open_JJ**]
[PP **to_TO**]
[NP **your_PRP\$ left._NN**]

Noun-Phrase Chunking

- Chunks are non-overlapping partials
- Chunk Parsing (shallow Parsing)
 - No hierarchical relations
- WHY?
 - **In theory:** psycholinguistic shows that there is no full processing in speech recognition, just partial structures recognition
 - **In practice:** Full parsing has less performance and are not very accurate

[Carstensen.2010]

Noun-Phrase Chunking

- Problems:
 - The Begin and End of a Noun-Phrase
 - Participle
 - He enjoys baking potatoes
 - Conjunctions
 - ripe apples and bananas

Noun-Phrase Chunking

- State of the Art

System	Description	Report
SS05	specialized HMM + voting between different representations	95.23%
M05	Second order conditional random fields + multi-label classification	93.6%
S08	Second order latent-dynamic conditional random fields + an improved inference method based on A* search	94.34%

Performance measure: $F = 2 * \text{Precision} * \text{Recall} / (\text{Recall} + \text{Precision})$

Precision: percentage of NPs found by the algorithm that are correct

Training data: sections 15-18 of Wall Street Journal corpus (Ramshaw and Marcus)

Testing data: section 20 of Wall Street Journal corpus (Ramshaw and Marcus)

Excerpt from [AssociationforComputationalLinguistics.03.09.2013a]

Named-Entity Recognition

- Entities are:
 - People, Locations, Time, Quantities, Money, ...
 - „Mark left the window open to your left.“
 - <START:person> Mark <END> left the window open to your left.
- Gazetteer for cities
- Internal Evidence
 - Prefix, suffix e.g. **Rennweg**, **Georges-Köhler-Allee**
- External Evidence
 - Context e.g. [PERSON] thinks, [PERSON] visits [LOCATION]

Named-Entity Recognition

- Problem:
 - Ambiguities
 - Semantic e.g. „Look there, it's Willy Brandt.“
 - The former chancellor Willy Brandt
 - Or the under construction airport in Berlin
 - Structural e.g. „They're playing in the Volkswagen Halle.“
 - The Basketballstadion [Volkswagen Halle]
 - Or the brand [Volkswagen] and the Noun [Halle]

Named-Entity Recognition

- State of the Art

System	Description	System Type	Results
FIJZ	Best CONLL-2003 participant	supervised learning	88.76%
Baseline	Vocabulary transfer from training to testing	supervised learning	59.61%
Balie	Unsupervised approach: no prior training	unsupervised learning	55.98%

Performance measure: $F = 2 * \text{Precision} * \text{Recall} / (\text{Recall} + \text{Precision})$

Precision: percentage of named entities found by the algorithm that are correct

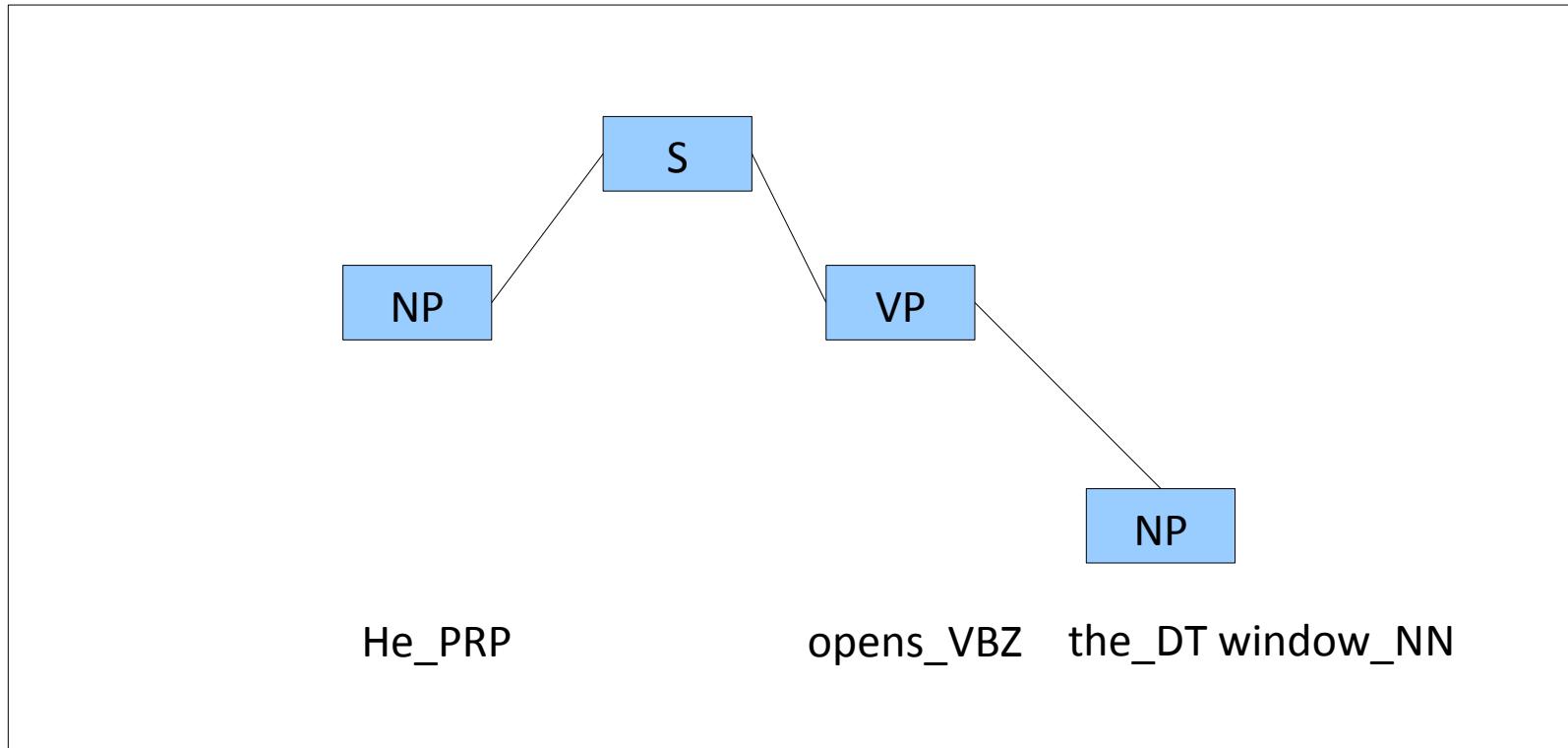
Training data: Train split of CONLL-2003 corpus

Testing data: Testb split of CONLL-2003 corpus

Excerpt from [AssociationforComputationalLinguistics.03.09.2013b]

Constituent Parsing

- The sentence is divided into smaller segments
 - „He open the window“

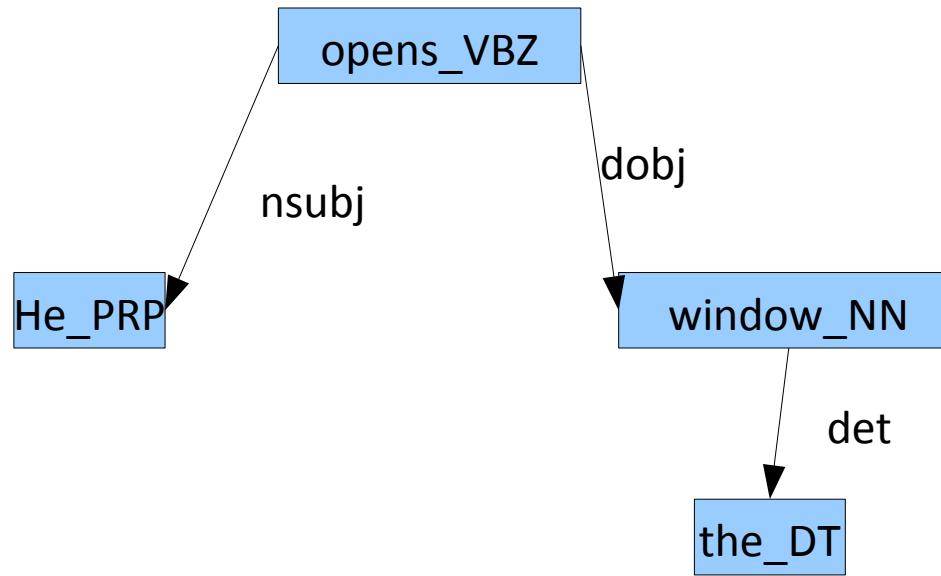


Constituency Parsing

- Probabilistic context free grammars
 - Sentence $x \rightarrow \text{tree } y$ with a probability
- A Constituent Tree is generated from a set of grammatical rules
- The probability of a constituent tree is the mathematical product of all probabilities of the sub rules
- The order of the words are not important
- The words/phrases are leafs

Dependency Parsing

- In the so called „dependency grammar“ all sentence parts gets related to the verbs
 - „He open the window.“



Dependency Parsing

- The words/phrases are nodes
- „syntactic functions“ are the edges
- <http://nlp.stanford.edu:8080/corenlp/>

DEMO

Parsing

■ State of the art

Type	Parser	Unlabeled attachment		Labeled attachment	
		P	R	P	R
Constituent	CJ	91,7 %	91,7 %	89,2 %	89,1 %
	Charniak	90,5 %	90,4 %	87,8 %	87,7 %
Dependency	Nivre Eager	85,4 %	84,2 %	81,7 %	80,5 %
	Feature Interact				
	MSTParser (Eisner)	83,0 %	82,2 %	79,2 %	78,4 %

Training data: Penn Treebank, section 2 to 21

Excerpt from [Cer et al. 2010]

Benchmark

NLP Disciplines in openNLP*

sentence	POS	Chunking	NER**	Constituency Parsing***
	s			
7567	3,094 s 2.445,7 sent/s	10,544 s 717,7 sent/s	16,531 s 457,7 sent/s	155,987 s ~ 2,5 min 39,2 sent/s (6.112 sent)
23.135	8,853 s 2.613,2 sent/s	30,904 s 748,6 sent/s	49,481 s 467,6 sent/s	479,371 s ~ 8 minutes 39,8 sent/s (19.063 sent)
58.252	21,625 s 2.693,7 sent/s	76,918 s ~ 1,2 min 757,3 sent/s	122,161 s ~ 2 min 476,8 sent/s	1199,058 s ~ 19 minutes 40,0 sent/s (47.924 sent)

* Using a AMD FX(tm)-8150 Eight-Core Processor, 16GB RAM

** Using date, location, money, organization, percentage, person and time models

*** The openNLP parser throws a lot of 'couldn't find parse for ..' errors, so not all sentences was parsed, just the count in the parentheses

Conclusion

- As seen, Part-of-Speech Tagging, and Chunking doing very well (> 95 %)
- Entity Recognition has still some problems (ambiguities etc)
- Full parsing is a lot slower than a shallow parsing

Literature

- **[Association for Computational Linguistics 03.09.2013b]** Association for Computational Linguistics: CONLL-2003 (State of the art) - ACLWiki.
[http://aclweb.org/aclwiki/index.php?title=CONLL-2003_\(State_of_the_art\)](http://aclweb.org/aclwiki/index.php?title=CONLL-2003_(State_of_the_art)), accessed 20 Nov 2013.
- **[Association for Computational Linguistics 03.09.2013a]** Association for Computational Linguistics: NP Chunking (State of the art) - ACLWiki.
[http://aclweb.org/aclwiki/index.php?title=NP_Chunking_\(State_of_the_art\)](http://aclweb.org/aclwiki/index.php?title=NP_Chunking_(State_of_the_art)), accessed 18 Nov 2013.
- **[Association for Computational Linguistics 03.09.2013]** Association for Computational Linguistics: POS Tagging (State of the art) - ACLWiki.
[http://aclweb.org/aclwiki/index.php?title=POS_Tagging_\(State_of_the_art\)](http://aclweb.org/aclwiki/index.php?title=POS_Tagging_(State_of_the_art)), accessed 18 Nov 2013.
- **[Bast.2013]** Bast, H.: Semantische Suche. In Informatik-Spektrum, 2013, 36; pp. 136–143.
- **[Brill.1992]** Brill, E.: Proceedings of the conference / Third Conference on Applied Natural Language Processing, 31 March - 3 April 1992, Trento, Italy. "A simple rule-based part of speech tagger.". ACL, Morristown, NJ, 1992.

Literature

- **[Brill.1994]** Brill, E.: Brills Rule Based Part-Of-Speech Tagger.
http://www.tech.plym.ac.uk/soc/staff/guidbugm/software/RULE_BASED_TAGGER_V.1.14.tar.Z.
- **[Carstensen.2010]** Carstensen, K.-U.: Computerlinguistik und Sprachtechnologie. Eine Einführung. In Computerlinguistik und Sprachtechnologie, 2010.
- **[Cer et al. 2010]** Cer, D. et al.: Parsing to Stanford Dependencies: Trade-offs between speed and accuracy, 2010.
- **[Computational Linguistics & Psycholinguistics Research Center 2011]** Computational Linguistics & Psycholinguistics Research Center: Chunking.
<http://www.clips.ua.ac.be/conll2000/chunking/>, accessed 10 Nov 2013.
- **[Ratnaparkhi.1996]** Ratnaparkhi, A.: "A maximum entropy model for part-of-speech tagging." Proceedings of the conference on empirical methods in natural language processing. Vol. 1. 1996. <http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf>, accessed 16 Nov 2013.
- **[Ratnaparkhi.1998]** Ratnaparkhi, A.: MAXIMUM ENTROPY MODELS FOR NATURAL LANGUAGE AMBIGUITY RESOLUTION.
http://repository.upenn.edu/cgi/viewcontent.cgi?article=1061&context=ircs_reports, accessed 12 Nov 2013.

Thank you for staying awake
and your attention!

Questions?

