Seminar: Efficient Natural Language Processing

Overview of Topics

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Overview

- The following is a **short!** overview of each topic to give you a rough idea
- Exact definition of the problem: your task
- Some topics leave room for suggestions from your side
- In general: we are interested in solutions that provide **reasonable quality AND efficiency** (speed)

1. POS Tagging

- POS = **p**art-**o**f-**s**peech ("*word category*")
- POS tagging: assign each word in a sentence its word-category
- Example:

While	in	Prague	he	met	Albert	Einstein	for	the	first	time
IN	IN	NP	PP	VBD	NP	NP	IN	DT	JJ	NN

Tag	Category
IN	Preposition
NP	Proper noun
VBD	Verb in past-tense
…	…

2. Text Chunking

- Task of chunking a text into phrases that contain *“syntactically related words”*
- *syntactically related words:* noun-phrases, verb-phrases, …
- Example:
  - *“($_{\text{SBAR}}$While) ($_{\text{PP}}$ in) ($_{\text{NP}}$Prague) ($_{\text{NP}}$he) ($_{\text{VP}}$met) ($_{\text{NP}}$Albert Einstein) ($_{\text{PP}}$for) ($_{\text{NP}}$the first time) ($_{\text{O}}$.)“*

Tag	Phrase
SBAR	subordinate clause
PP	prepositional phrase
NP	noun phrase
VP	verb phrase
…	…

3. Clause Identification

- Task of “*dividing text into clauses*”
- Clauses usually contain subject and predicate
- Clauses form a hierarchical structure (a clause can contain another clause)
- Example:
  - *“(Coach them in (handling complaints) (so that (they can resolve problems immediately)).)”*

4. Entity Recognition

- Identify entites (persons, locations, organizations …) in a text, based on a list of known entities
- Example:
  - “[$_{\text{ENT1}}$ Steve Jobs] was the CEO of [$_{\text{ENT2}}$ Apple], situated in [$_{\text{ENT3}}$ California].”

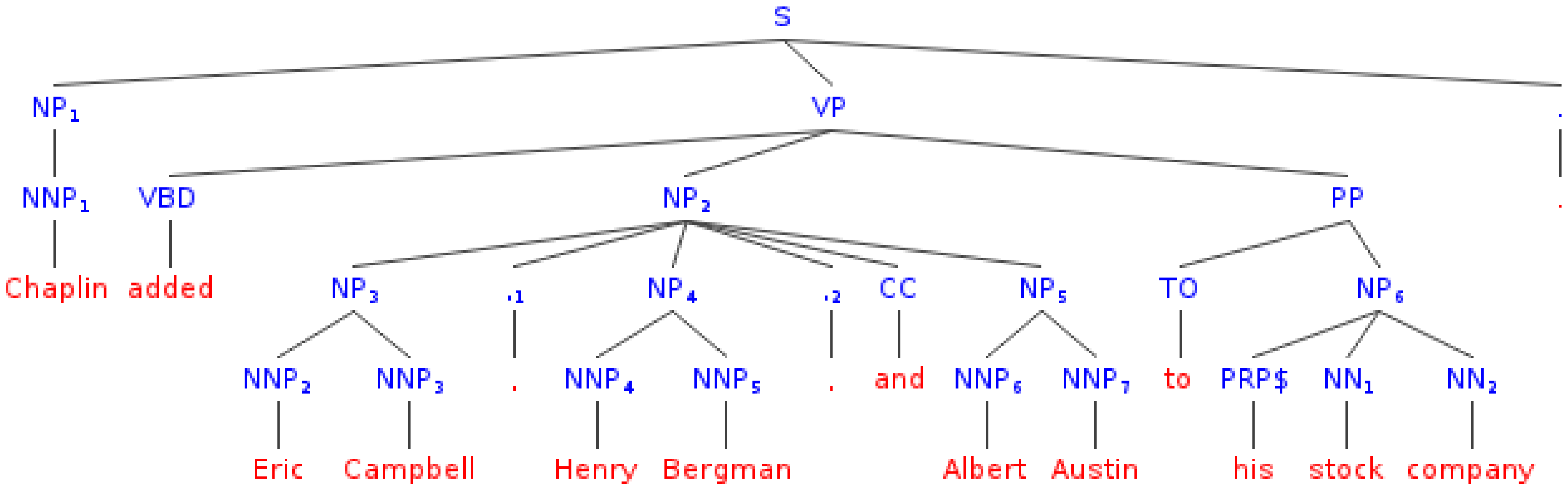
Tag	Exact entity
ENT1	http://en.wikipedia.org/wiki/Steve_Jobs
ENT2	http://en.wikipedia.org/wiki/Apple_Inc.
ENT3	http://en.wikipedia.org/wiki/California
…	…

5. Semantic Role Labeling

- *"recognize arguments of verbs in a sentence, and label them with their semantic role"*
- Example:
  - *"[$_{A0}$Chaplin]* ***added*** *[$_{A1}$Eric Campbell, Henry Bergman, and Albert Austin] [$_{A2}$to his stock company]."*
  - Arguments of "***added***":
    - Subject (A0): "*Chaplin*"
    - Object (A1): "*Eric Campbell, Henry Bergman, and Albert Austin*"
    - Indirect Object (A2): "*to his stock company*"

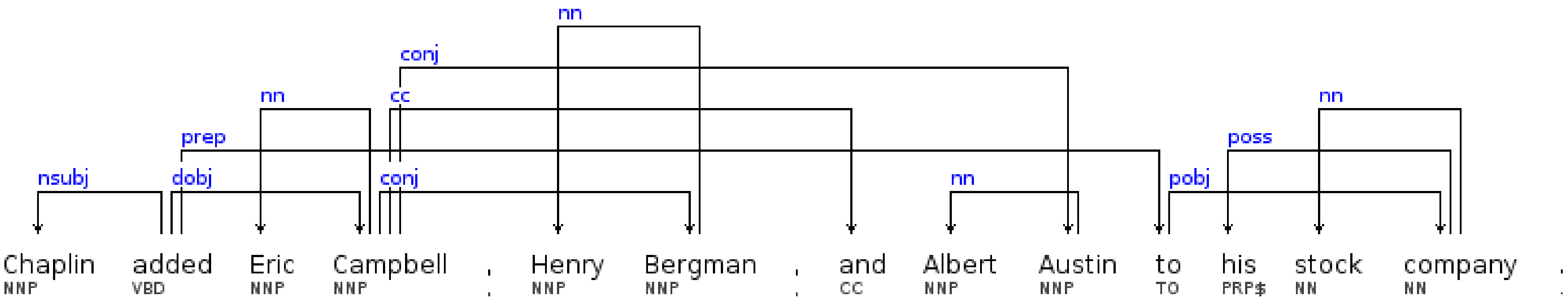
6. Constituent Parsing

- Recursively identify all grammatical parts/constituents of a sentence
- Output the syntax tree (constituent tree)
- Example:



7. Dependency Parsing

- For each word of a sentence identify its „head“ - the word in the sentence it depends on
- Output is a (di)graph
- Example:



8. Link Grammar Parsing

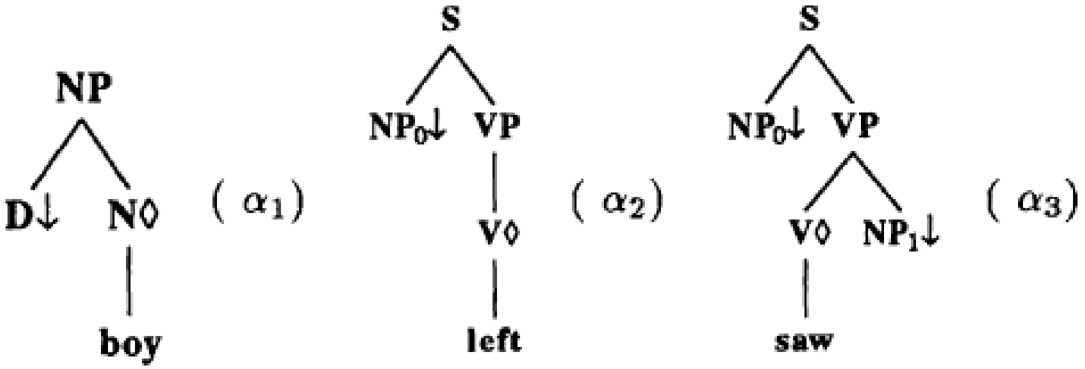
- Link Grammar: a special kind of grammar expressing relationships between the words of a sentence
- Example:

```
 +--------------------------------------Xp-----------------------------------------+
    |                +-----------------MVp-----------------+--------Jp-------+     |
    |                +------Os-----+                       |  +------D*u-----+     |
    +---Wd---+---Ss--+      +---G--+                       |  |     +---AN---+     |
    |        |       |      |      |                       |  |     |        |     |
LEFT-WALL Chaplin added.v Eric Campbell and Albert Austin to his stock.n company.n .
```

- “MV connects verbs (and adjectives) to modifying phrases like adverbs” -> “*added*” modified by “*to his stock company*”

9. LTAG Parsing

- LTAG = **L**exicalized **T**ree-**A**djoining **G**rammar
- Tree-adjoining grammar: a grammar that consists of trees



- Parsing: tree operations
- Result similar to a constituent parse

10. Text Simplification

- Simplify complex sentence by applying lexical and syntactical operations
- Main motivation: make text readable for humans with reading problems (aphasics)
- Other motivation: shorter sentences are easier/faster to parse

“Mr. Anthony, who runs an employment agency, decries program trading, but he isn’t sure it should be strictly regulated.”

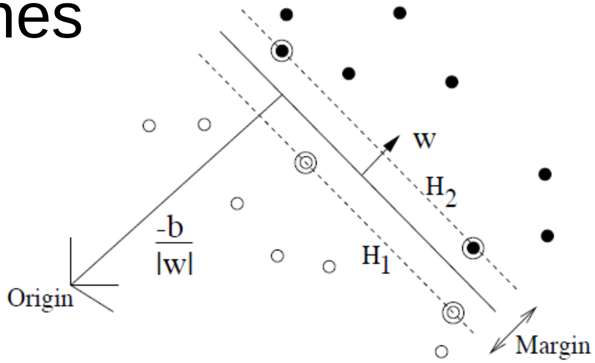
⇩

“Mr Anthony runs an employment agency. Mr Anthony decries program trading. But he isn’t sure it should be strictly regulated.”

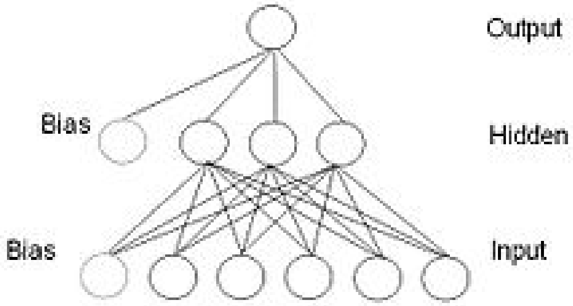
11. Machine Learning

- Important machine learning techniques for NLP

- especially Support Vector Machines
  - Map instances into vector space
  - Find hyperplane separating classes



- and perceptrons
  - Train a neural network to classify examples



12. Question Answering

- Answer a question given in natural language
- What is the capital of Mongolia?

DEMO

13. Entity Retrieval

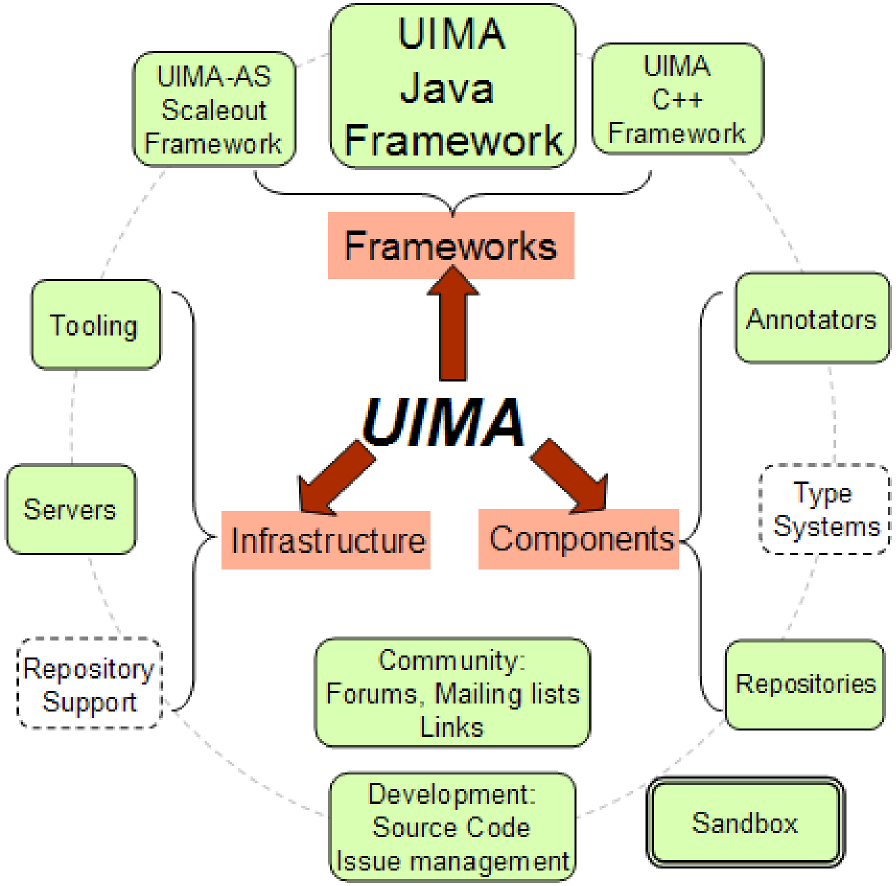
- Answer a factoid question with a list of entities

- What cities in Germany have more than 100.000 inhabitants?

DEMO

14. Apache UIMA

- UIMA = **U**nstructured **I**nformation **M**anagement **A**pplications
- A framework for NLP applications providing well-defined interfaces
- Manages the data flow between components
- Used by “*Watson*“, the IBM computer playing the Jeapordy! Competition



15. NLP Applications

- Any application relying heavily on NLP, like *Contentus*
- Contentus: a research project sponsored by the german government
- Goal:
  - Digitalise and analyze information (text, pictures, audio, video … archived in libraries)
  - Providing an efficient way of thematic research in discovered information

List of Topics

1. POS tagging
2. Text chunking
3. Clause identification
4. Entity recognition
5. Semantic role labeling
6. Constituent parsinging
7. Dependency parsing
8. Link grammar parsing
9. LTAG parsing
10. Text simplification
11. Machine learning
12. Question answering
13. Entity retrieval
14. Apache UIMA
15. NLP application

References

- Barbella, D., Benzaid, S., Christensen, J. M., Jackson, B., Qin, X. V., and Musicant, D. R. 2009. Understanding support vector machine classifications via a recommender system-like approach. In *Int. Conf. on Data Mining*. 305–311.
- Burges, C. 1999. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*.
- Carreras, X. 2005. Learning and inference in phrase recognition: A filteringranking architecture using perceptron. Ph.D. thesis, Universitat Politècnica de Catalunya.
- Carreras, X. and Màrques, L. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of CoNLL-2004*. Boston, MA, USA, 89–97.
- Chandrasekar, R., Doran, C., and Srinivas, B. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th conference on Computational linguistics - Volume 2*. COLING '96. Association for Computational Linguistics, Stroudsburg, PA, USA, 1041–1044.
- Sang, E. F. T. K. and Déjean, H. 2001. Introduction to the conll-2001 shared task: Clause identification. *Computing Research Repository cs.CL/0107*.
- Siddharthan, A. 2003. Syntactic simplification and text cohesion. Ph.D. thesis.
- Tjong Kim Sang, E. F. and Buchholz, S. 2000. Introduction to the conll-2000 shared task: chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*. ConLL '00. Association for Computational Linguistics, Stroudsburg, PA, USA, 127–132.