Seminar: Efficient NLP
Session 2, NLP behind Broccoli

November 2nd, 2011

Elmar Haußmann
Chair for Algorithms and Data Structures
Department of Computer Science
University of Freiburg

UNI FREIBURG

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Conclusion / Current and Future Work

Motivation and Problem Definition

- The idea of semantic full-text search
  - Search in full-text
  - But combined with “structured information“
- Broccoli performs the following NLP-tasks:
  - Entity recognition
    - Based on the links inside Wikipedia articles and heuristics
  - Anaphora resolution
    - Based on simple, yet efficient heuristics
  - Contextual Sentence Decomposition
    - This talk

Motivation and Problem Definition

- The motivation for Contextual Sentence Decomposition – the “heavy” NLP-task behind Broccoli

Example Query

plant edible leaves

Result Sentence

The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

Motivation and Problem Definition

- Many false-positives caused by words, appearing in same sentence, but part of a different *context*

➡ Apply natural language processing to decompose sentence based on context and search resulting „sentences“ independently

Result Sentence

The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

Motivation and Problem Definition

UNI FREIBURG

Original Sentence

The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

Decomposed Sentence

- *The usable parts of rhubarb are the medicinally used roots*
- *The usable parts of rhubarb are the edible stalks*
- *its leaves are toxic*

UNI FREIBURG

Problem Definition

Contextual Sentence Decomposition

Contextual Sentence Decomposition

is the process of performing

1. Sentence Constituent Identification

followed by

2. Sentence Constituent Recombination

Motivation and Problem Definition

Sentence Constituent Identification

- Identify specific parts of sentence
- Differentiate 4 types of constituents

- Relative clauses	*Albert Einstein, who was born in Ulm, ...*
- Appositions	*Albert Einstein, a well-known scientist, ...*
- List items	*Albert Einstein published papers on Brownian motion, the photelectric effect and special relativity.*
- Separators	*Albert Einstein was recognized as a leading scientist and in 1921 he received the Nobel Prize in Physics.*

Motivation and Problem Definition

UNI FREIBURG

Original Sentence with Identified Constituents

The usable parts of rhubarb are
the medicinally used roots
and
the edible stalks,
however
its leaves are toxic.

list item separator

Motivation and Problem Definition

Sentence Constituent Recombination

- Recombine identified constituents into *sub-sentences*
  - Split sentences at separators
  - Attach relative clauses and appositions to noun (-phrase) they describe
  - Apply „distributive law“ to list items

Motivation and Problem Definition

Original Sentence

The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

Decomposed Sentence

- *its leaves are toxic*
- *The usable parts of rhubarb are the medicinally used roots*
- *The usable parts of rhubarb are the edible stalks*

Motivation and Problem Definition

Remarks

- Given identified constituents, recombination comparably simple - identification challenging part
- Constituents possibly nested, e.g. relative clause can contain enumeration etc.
- Resulting sub-sentences often grammatically correct but not required to be
- Approach must be feasible in terms of efficiency (English Wikipedia ~ 30GB raw text)

Motivation and Problem Definition

And…Natural Language is Tricky

- Ambiguous, even for humans:
  - "Time flies like an arrow; fruit flies like a banana."
  - "Flying planes can be dangerous."
  - "I once shot an elephant in my pajamas.

    How he got into my pajamas, I'll never know."

➢ Focus: large part of less complicated sentences

Motivation and Problem Definition

…Natural Language is Tricky

- Even if meaning is clear to a human: arbitrarily deep nesting and syntactic ambiguity

Difficult Sentence

Panofsky was known to be friends with Wolfgang Pauli, one of the main contributors to quantum physics and atomic theory, as well as Albert Einstein, born in Ulm and famous for his discovery of the law of the photoelectric effect and theories of relativity.

- Apposition similar to an element of enumeration
- Relative clause contains enumeration and starts in reduced form

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Conclusion / Current and Future Work

Rule based Approach

Idea

- Devise hand-crafted rules by closely inspecting sentence structure

Sentence containing Relative Clause

Koffi Annan, who is the current U.N. Secretary General, has spent much of his tenure working to promote peace in the Third World.

- Example: relative clause is set off by comma, starts with word „*who*“ and extends to the next comma

Rule based Approach

Basic Approach

- Identify „stop-words“

Original Sentence with marked Stop-words

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

- For each marked word decide if and which constituent it starts
- Determine corresponding constituent ends

Rule based Approach

Determine Constituent Starts

Original Sentence with Identified Stop-words

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Rule based Approach

Determine Constituent Starts

- If a verb follows but a noun preceeds it:
  separator

Original Sentence with Identified Separator

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Rule based Approach

Determine Constituent Starts

- If a verb follows but a noun preceeds it:
  separator
- If it is no relative clause or apposition:
  next word list item start

Original Sentence with Identified List Item Start

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Rule based Approach

Determine Constituent Starts

- If a verb follows but a noun preceeds it:
  separator
- If it is no relative clause or apposition:
  next word list item start
- First list item starts at noun-phrase preceeding already discovered list item start

Original Sentence with all Identified List Item Starts

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Rule based Approach

Determine Constituent Ends

- For each start assign a matching end

Original Sentence with all Identified List Item Starts

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Rule based Approach

Determine Constituent Ends

- For each start assign a matching end
- A list item extends to the next constituent start or the sentence end

Original Sentence with Identified Constituents

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

UNI FREIBURG

Determine Constituent Ends

- For each start assign a matching end
- A list item extends to the next constituent start or the sentence end

Original Sentence with Identified Constituents

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Conclusion / Current and Future Work

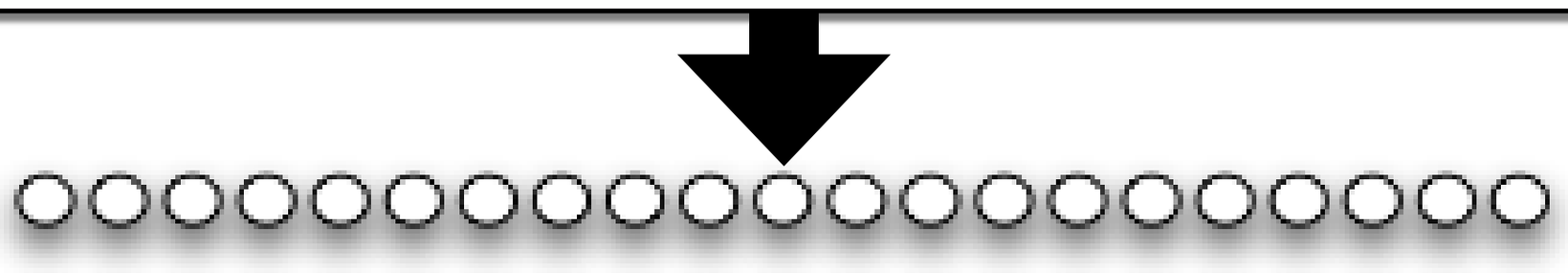
Machine Learning based Approach

Idea

- Use supervised learning to train classifiers that identify the start and end of constituents
- Train Support Vector Machines for each constituent start and end

Original Sentence

The usable parts of rhubarb are the medicinally used roots and the edible stalks , however its leaves are toxic.

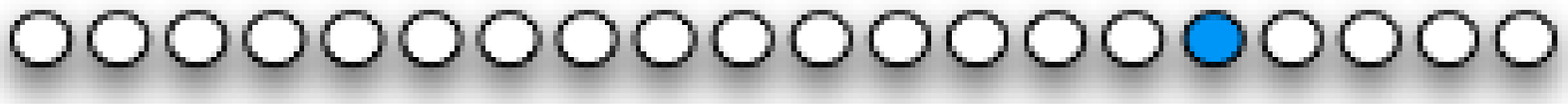


Machine Learning based Approach

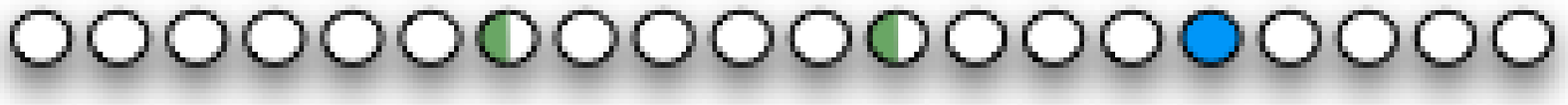
Basic Approach

- Apply classifiers in turn to each word
- Ideally this would already give a correct solution

1. Apply separator classifier



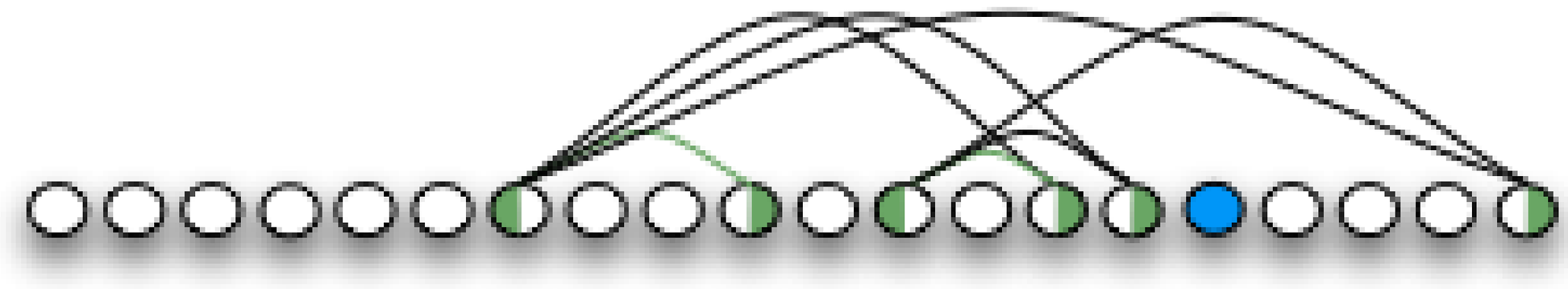
2. Apply list item start classifier



3. Apply list item end classifier

Machine Learning based Approach

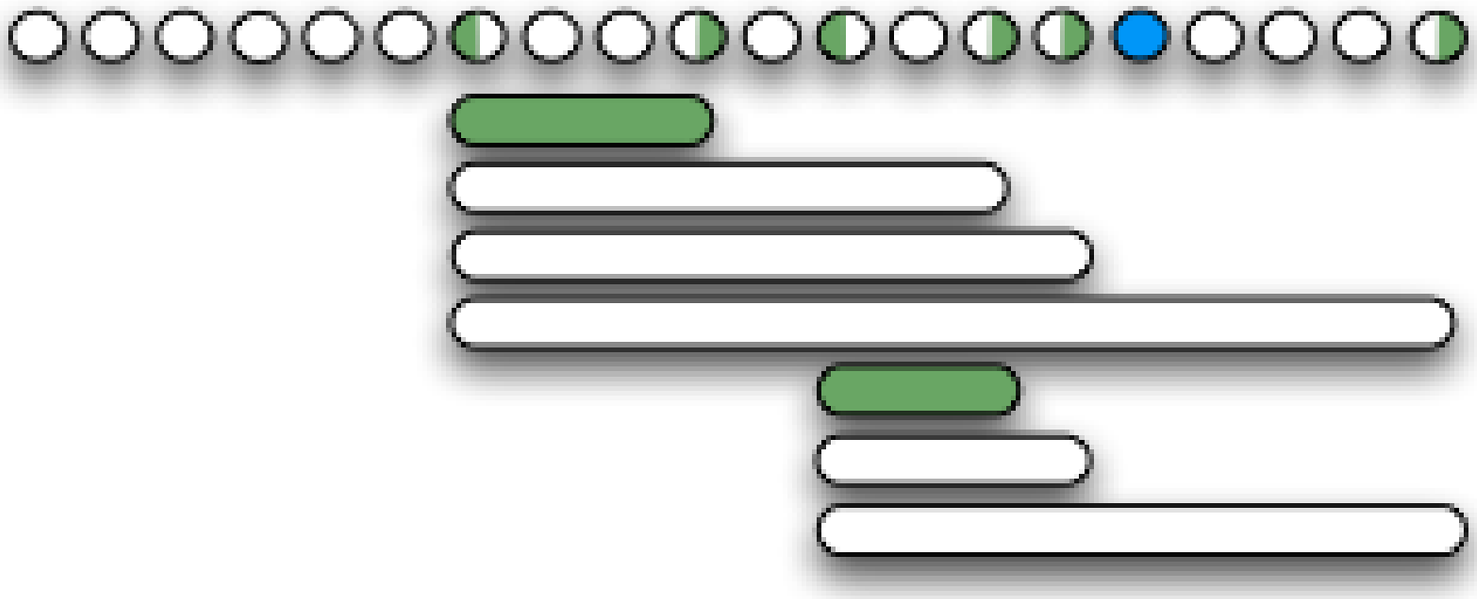
- However classifiers are not perfect
- Some additional ends and beginnings might be identified
- Decisions are local and do not consider admissible constituent structure



Machine Learning based Approach

- Train classifiers that identify whether a span of the sentence denotes a valid constituent

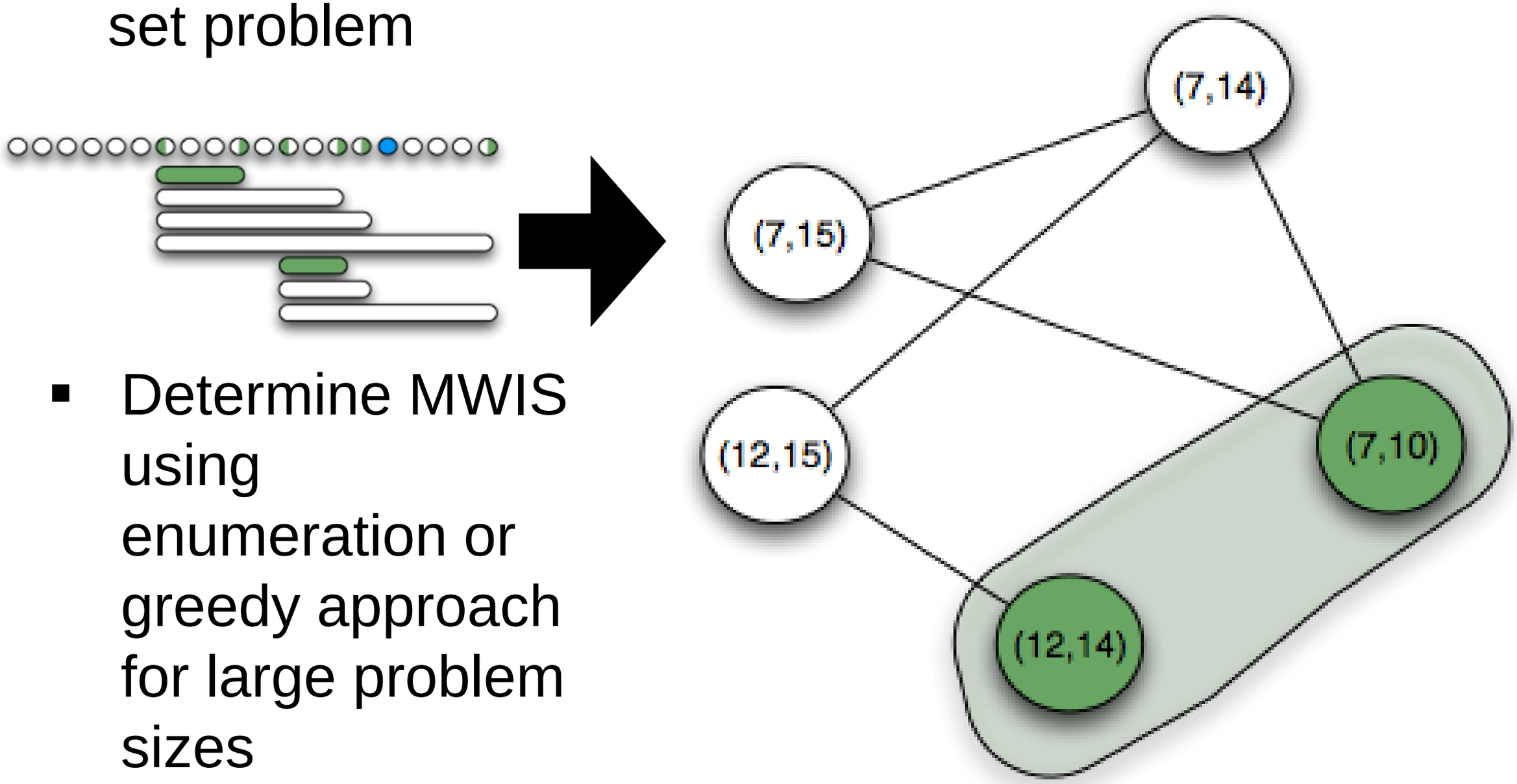
Apply list item classifier



- Still, identified constituents might overlap
- Structural constraints must be satisfied

Machine Learning based Approach

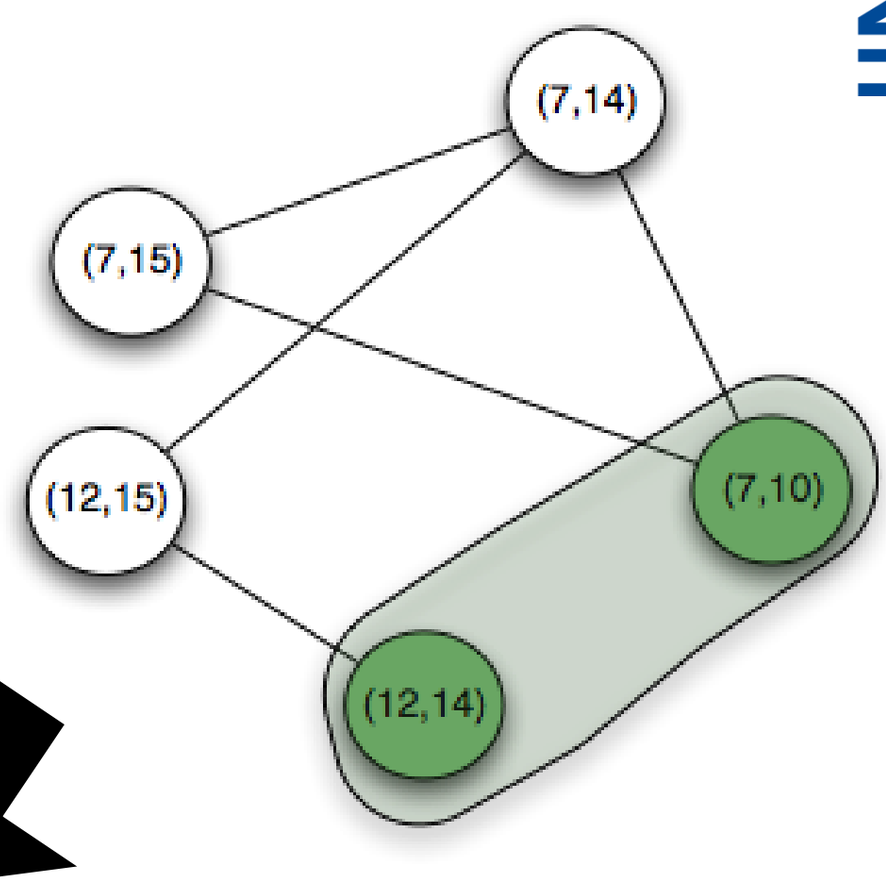
➡ Reduce to the maximum weight independent set problem



- Determine MWIS using enumeration or greedy approach for large problem sizes

Machine Learning based Approach

- Final result adheres to structural constraints
- More resistant to wrong „local“ classifications



Original Sentence with Identified Constituents

The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Conclusion / Current and Future Work

UNI FREIBURG

Evaluation on three levels

1. Compare identification using a ground truth
2. Compare resulting decomposition using a ground truth
3. Evaluate influence on search quality against ground truth

Evaluation / Conclusion

Results

- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective
- Search quality increases depend on exact query, but go up to doubling precision, with hardly loss in recall
- Contextual Sentence Decomposition integral part of Semantic Full-Text Search

Evaluation / Conclusion

Current Work

- Increasing quality of decomposition by:
  - efficient additional NLP (deep-parsers?…)
  - improvements of rules
  - better understanding what extent of decomposition is reasonable and necessary

Thank you

Thank you
for your attention!