Contextual Sentence Decomposition with Applications to Semantic Full-Text Search

Elmar Haussmann
August 9th, 2011

Albert-Ludwigs-Universität Freiburg

UNI FREIBURG

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Evaluation
- Conclusion

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Evaluation
- Conclusion

Motivation and Problem Definition

- To understand the motivation for Contextual Sentence Decomposition (CSD) we must understand the idea of Semantic Full-Text Search

Motivation and Problem Definition

- To understand the motivation for Contextual Sentence Decomposition (CSD) we must understand the idea of Semantic Full-Text Search

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*

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

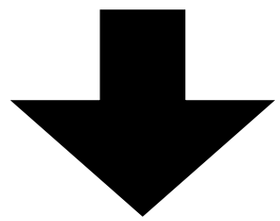
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

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.

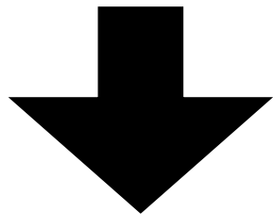


Motivation and Problem Definition

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 (CSD)
is the process of performing

1. Sentence Constituent Identification (SCI)
   followed by
2. Sentence Constituent Recombination (SCR)

Sentence Constituent Identification

Motivation and Problem Definition

Sentence Constituent Identification

- Identify specific parts of sentence

Motivation and Problem Definition

Sentence Constituent Identification

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

Motivation and Problem Definition

Sentence Constituent Identification

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

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, ...*

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

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, ...*

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

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.*

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	

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

Sentence Constituent Recombination

Motivation and Problem Definition

Sentence Constituent Recombination

- Recombine identified constituents into *sub-sentences*

Motivation and Problem Definition

Sentence Constituent Recombination

- Recombine identified constituents into *sub-sentences*
  - Split sentences at separators

Sentence Constituent Recombination

- Recombine identified constituents into *sub-sentences*
  - Split sentences at separators
  - Attach relative clauses and appositions to noun (-phrase) they describe

UNI FREIBURG

Sentence Constituent Recombination

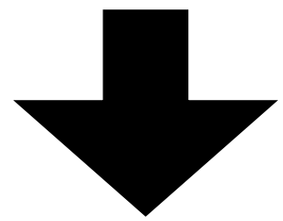
- Recombine identified constituents into *sub-sentences*
  - Split sentences at separators
  - Attach relative clauses and appositions to noun (-phrase) they describe
  - Apply „distributive law“ to list items

Motivation and Problem Definition

UNI FREIBURG

Original Sentence

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

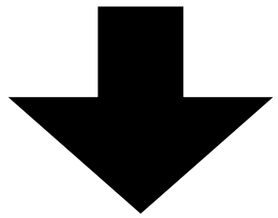


Motivation and Problem Definition

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

- *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

Motivation and Problem Definition

UNI FREIBURG

Remarks

- Given identified constituents, recombination comparably simple - identification challenging part

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.

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 require to be

Motivation and Problem Definition

UNI FREIBURG

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 require to be
- References within a sentence have to be resolved beforehand

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Evaluation
- Conclusion

Rule based Approach

Idea

- Devise hand-crafted rules by closely inspecting sentence structure

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.

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“

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.

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

UNI FREIBURG

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

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.

UNI FREIBURG

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.

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.

Natural Language is Tricky

Natural Language is Tricky

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.

Natural Language is Tricky

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.

- Problems:

UNI FREIBURG

Natural Language is Tricky

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.

- Problems:
  - Apposition similar to an element of enumeration

UNI FREIBURG

Natural Language is Tricky

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.

- Problems:
  - Apposition similar to an element of enumeration

UNI FREIBURG

Natural Language is Tricky

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.

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

UNI FREIBURG

Natural Language is Tricky

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.

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

Natural Language is Tricky

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.

- Problems:
  - 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
- Evaluation
- Conclusion

Machine Learning based Approach

UNI FREIBURG

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

Machine Learning based Approach

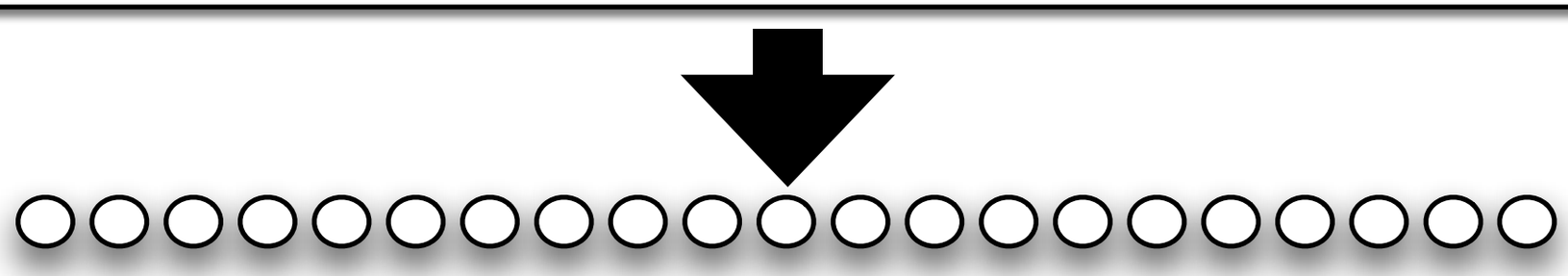
UNI FREIBURG

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.



Basic Approach

Machine Learning based Approach

Basic Approach

- Apply classifiers in turn to each word

Machine Learning based Approach

UNI FREIBURG

Basic Approach

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

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

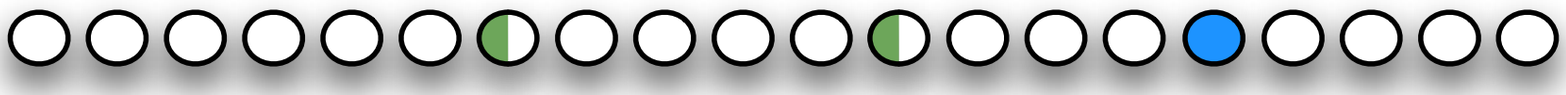
Machine Learning based Approach

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



Machine Learning based Approach

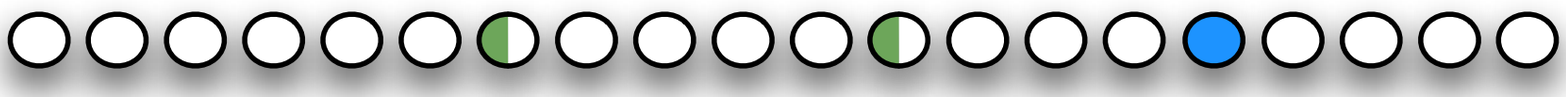
UNI FREIBURG

Basic Approach

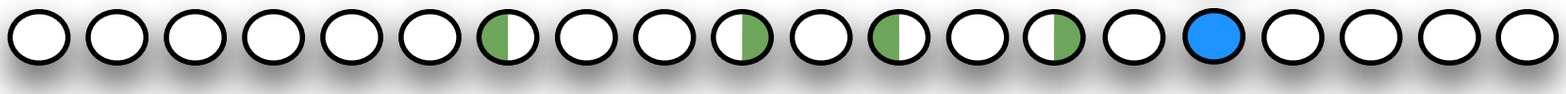
- Apply classifiers in turn to each word
- Ideally this would already give a correct solution

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

Machine Learning based Approach

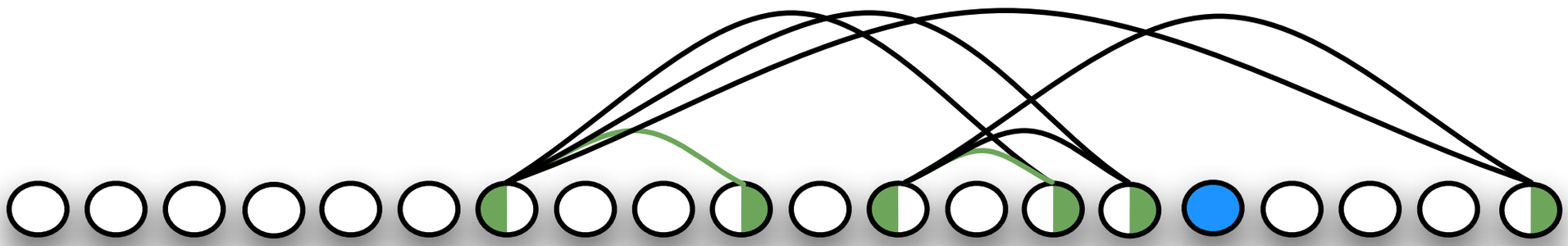
- However classifiers are not perfect
- Some additional ends and beginnings might be identified

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

- 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

Machine Learning based Approach

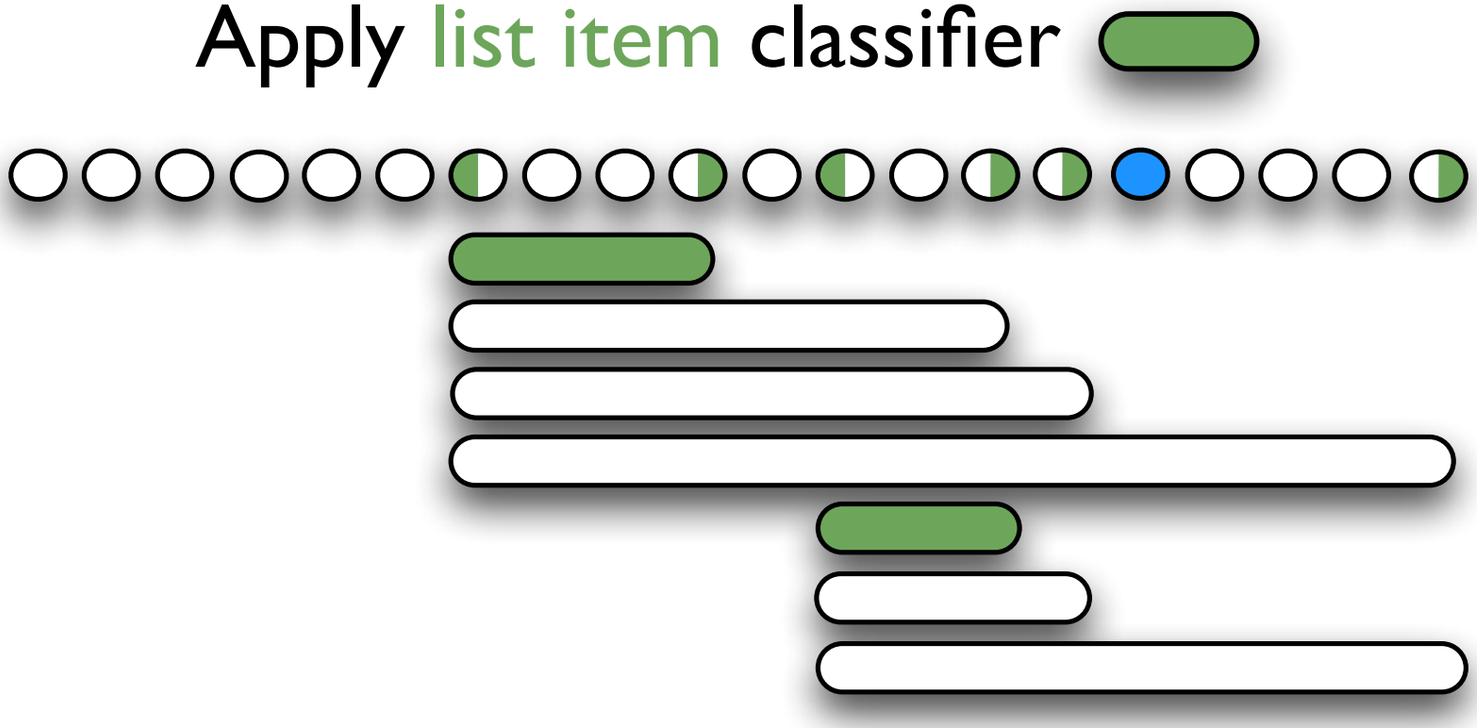
- Train classifiers that identify whether a span of the sentence denotes a valid constituent

  Apply list item classifier 

Machine Learning based Approach

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

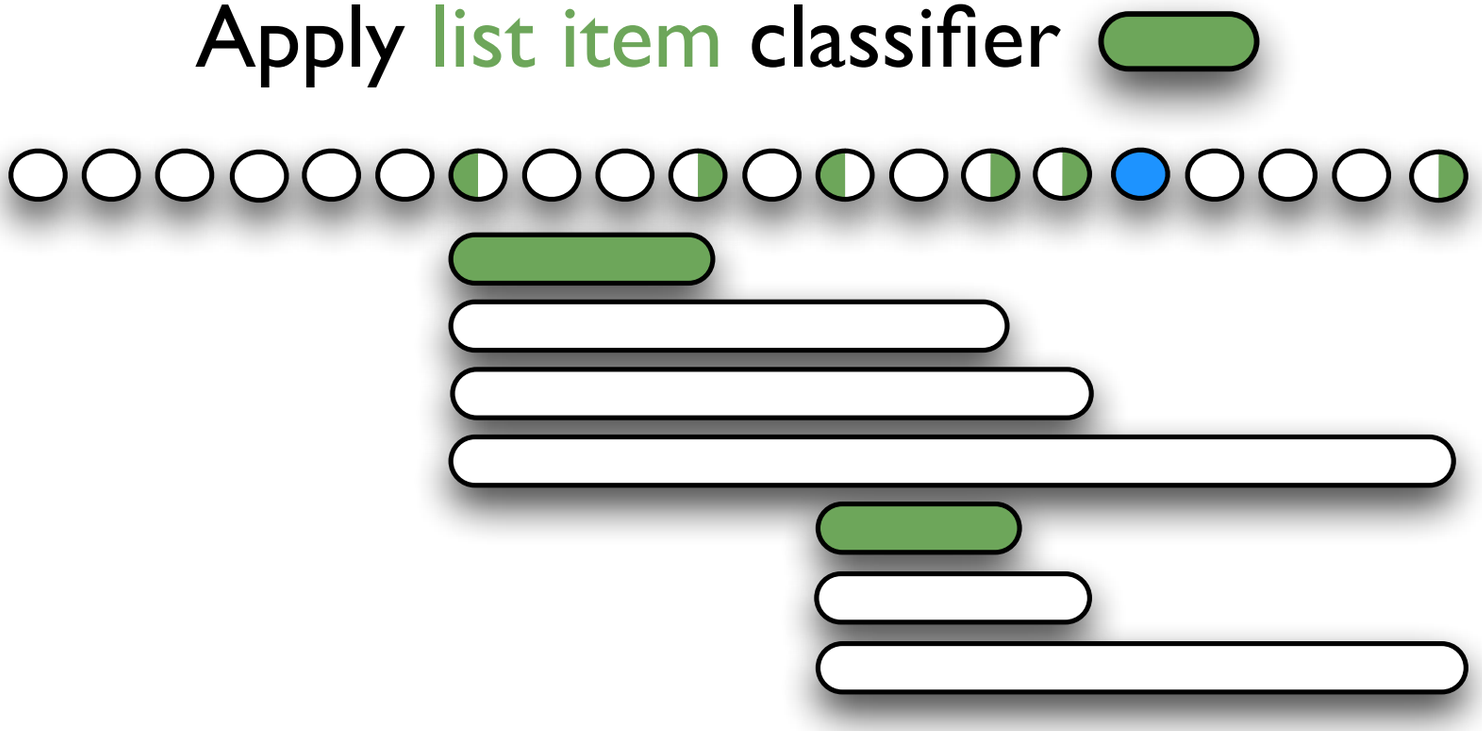
Apply list item classifier



Machine Learning based Approach

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

Apply list item classifier

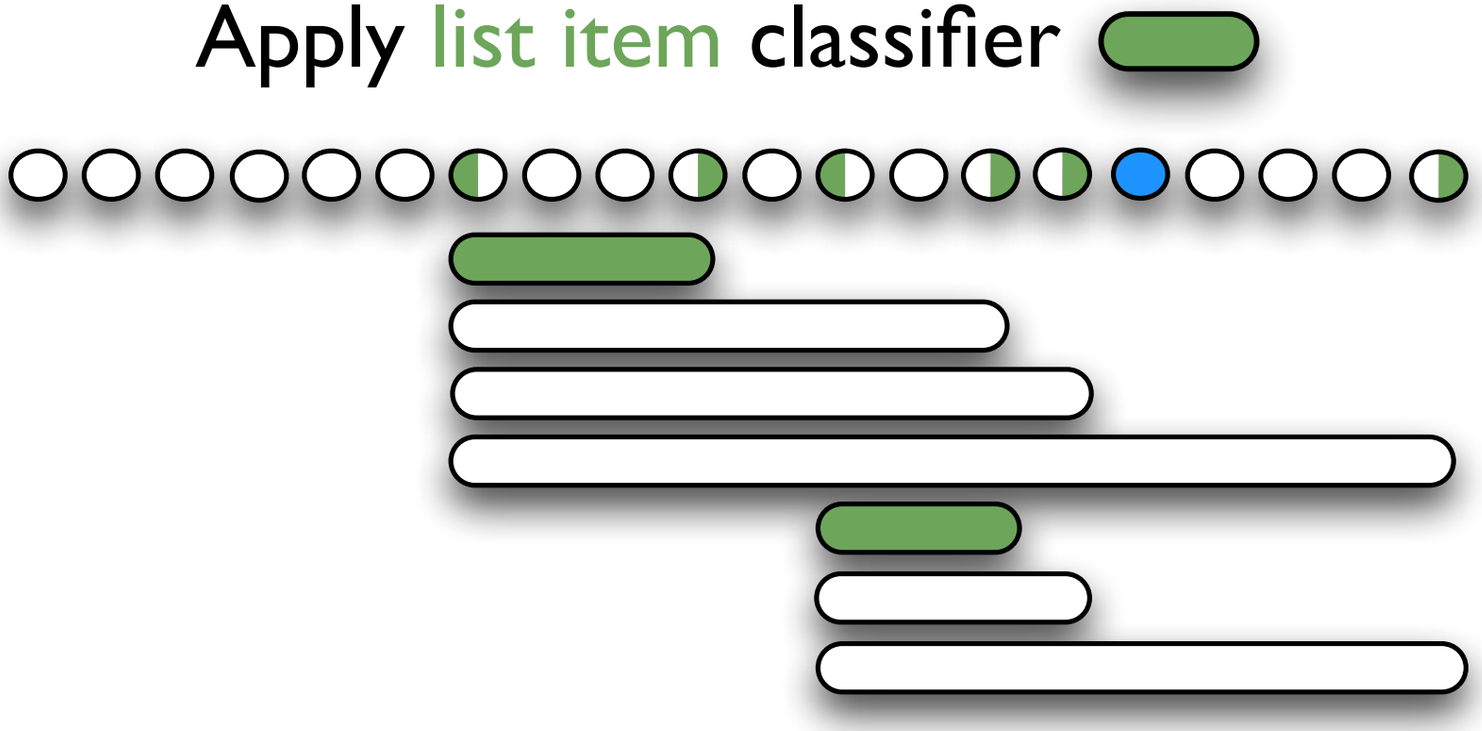


- Still, identified constituents might overlap

Machine Learning based Approach

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

Apply list item classifier



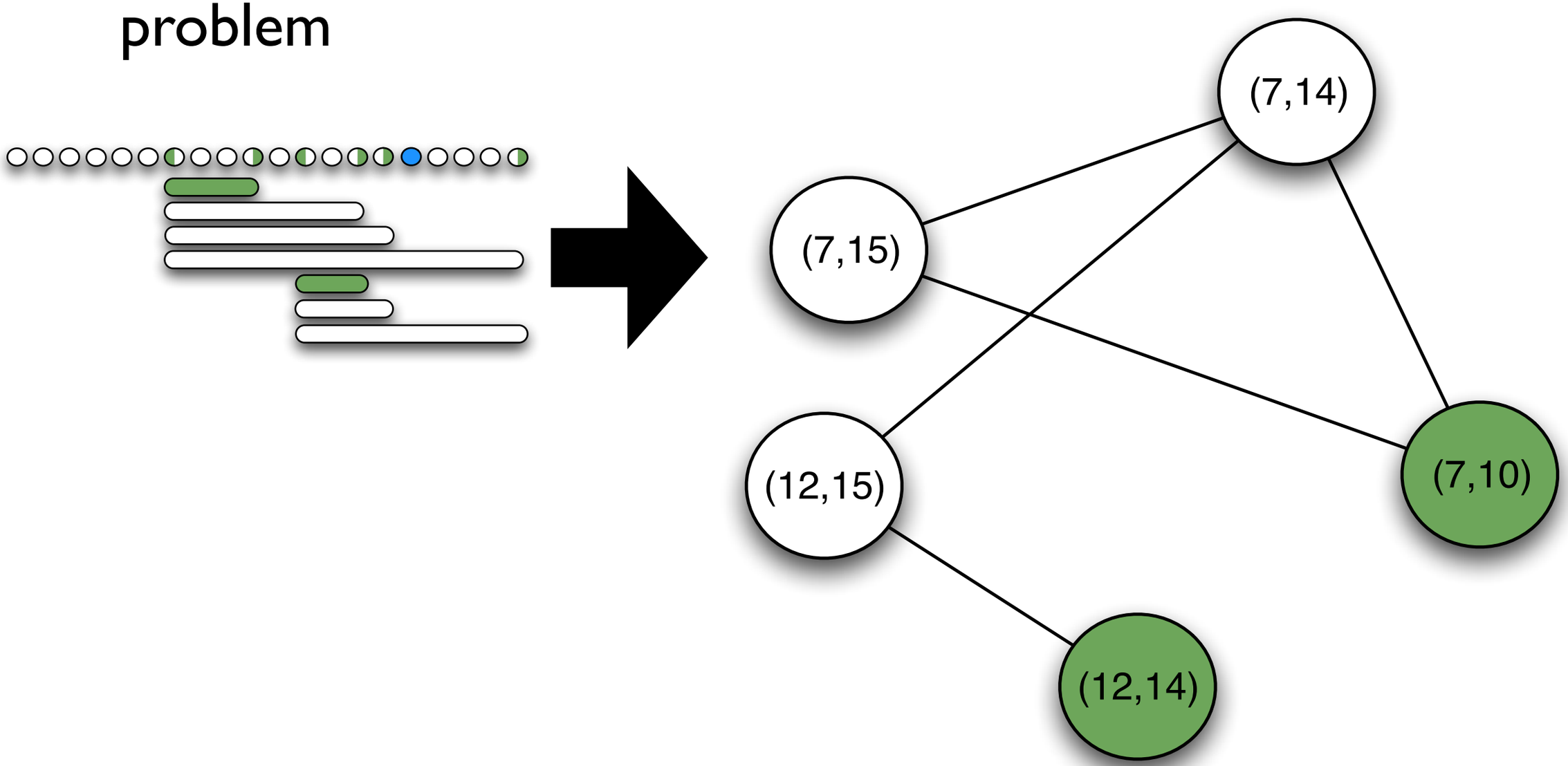
- Still, identified constituents might overlap
- Structural constraints must be satisfied

Machine Learning based Approach

➡ Reduce to the maximum weight independent set problem

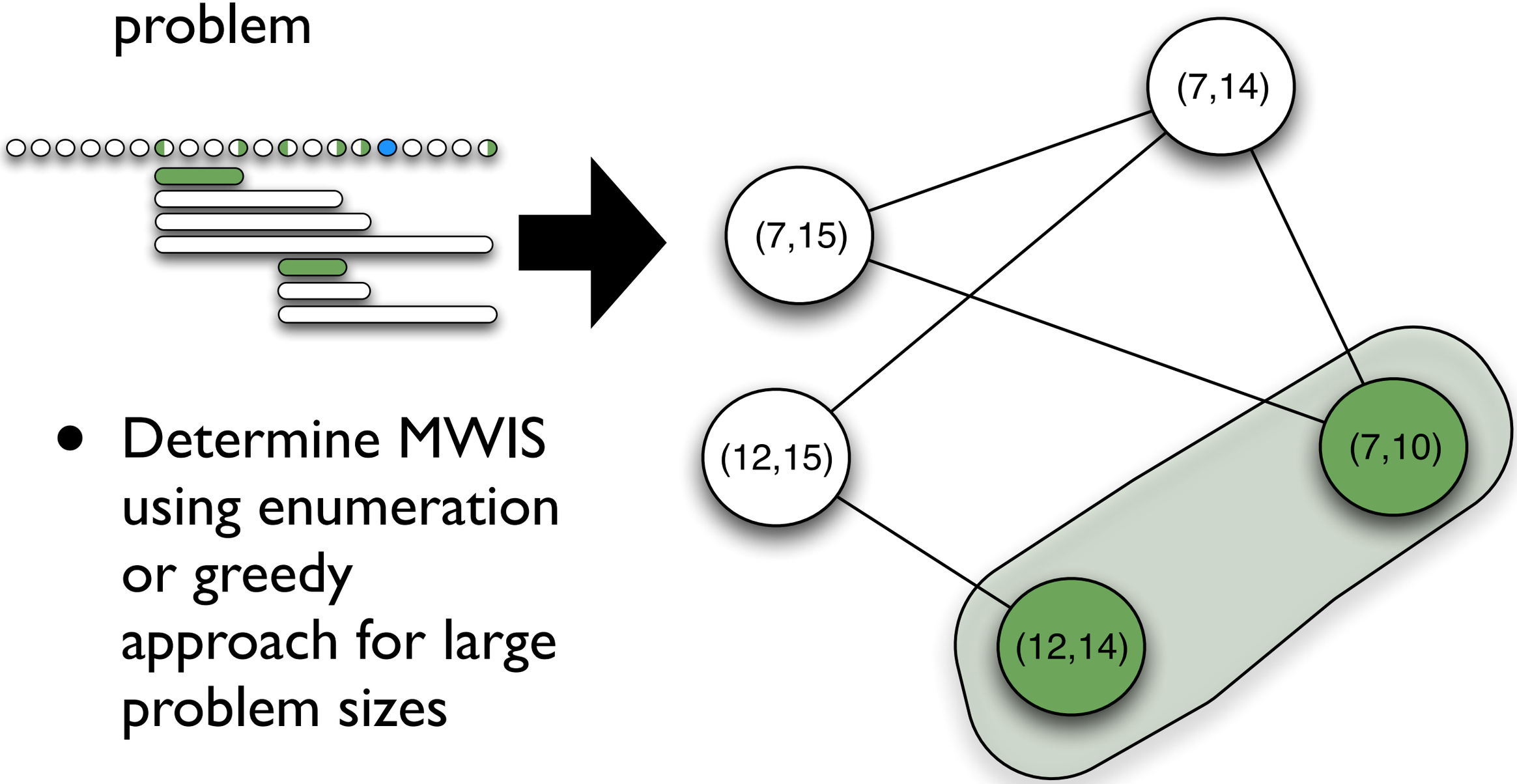
Machine Learning based Approach

➡ Reduce to the maximum weight independent set problem



Machine Learning based Approach

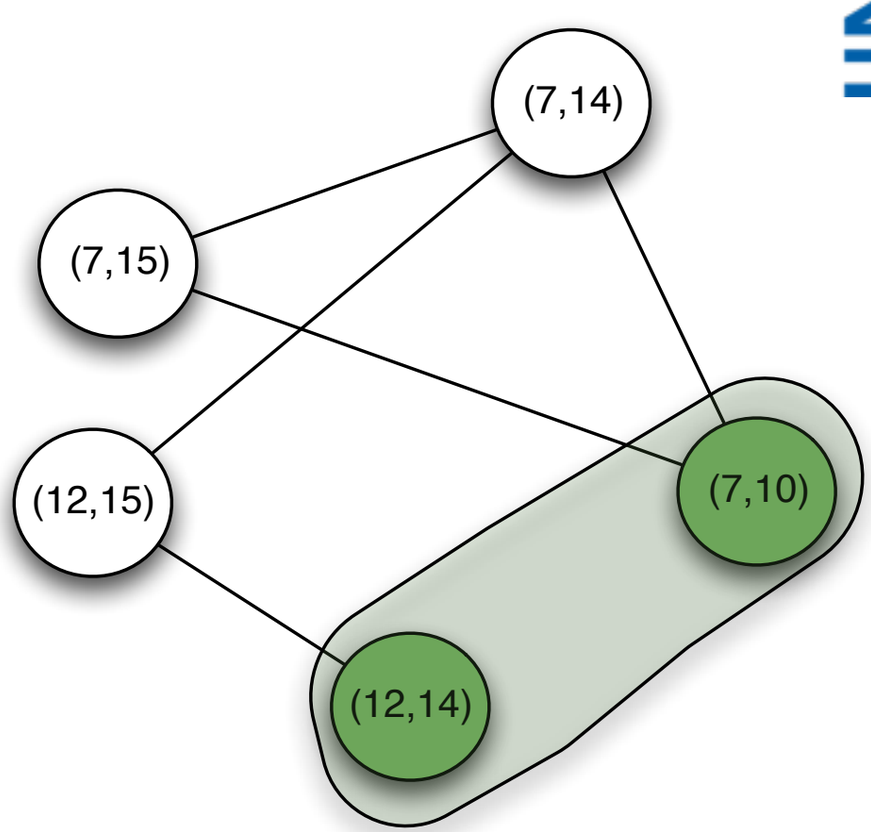
➡ Reduce to the maximum weight independent set problem



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

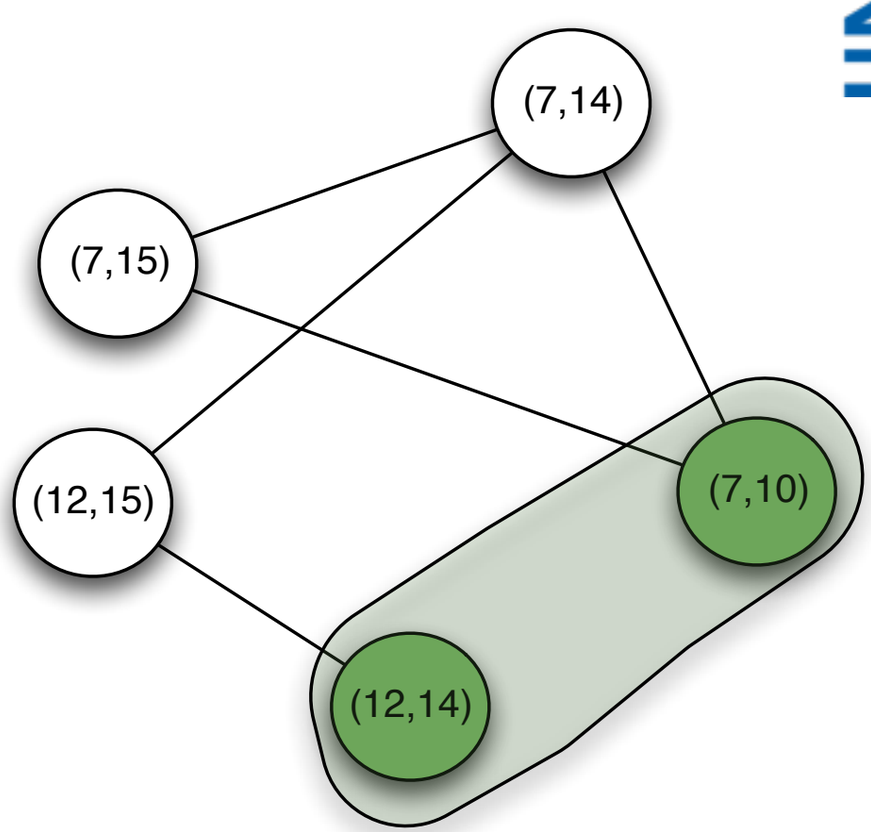
Machine Learning based Approach

- Final result adheres to structural constraints



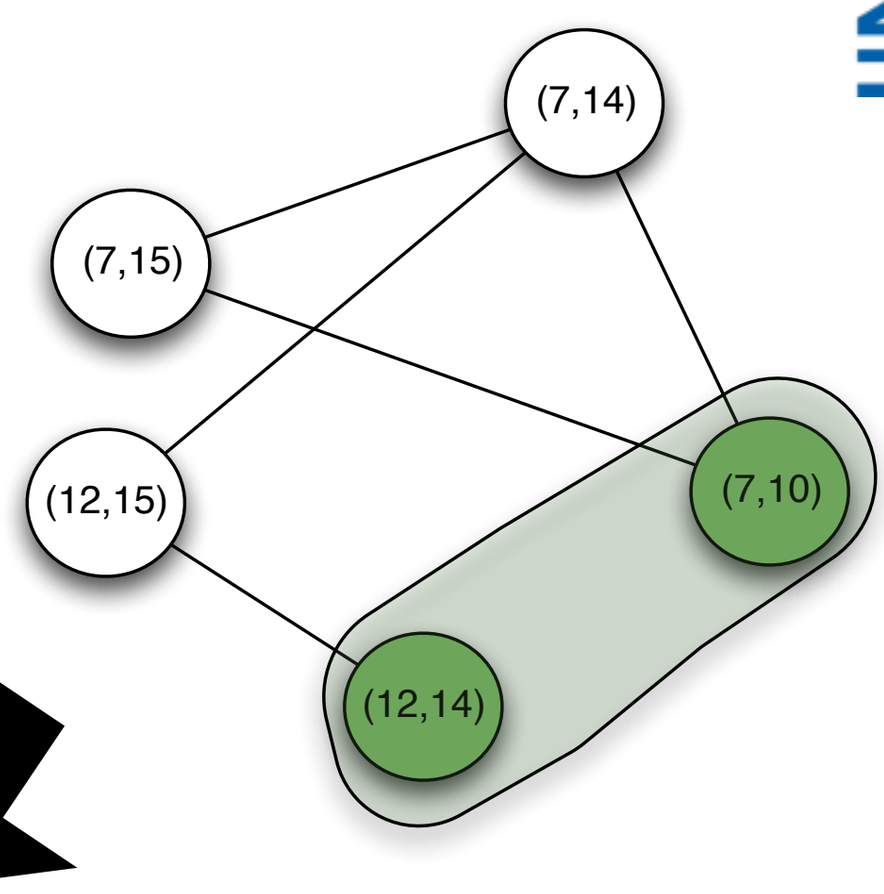
Machine Learning based Approach

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



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
- Evaluation
- Conclusion

Evaluation

- Evaluation of the different approaches on three levels

Evaluation

- Evaluation of the different approaches on three levels

1. Compare identification using a ground truth

Evaluation

- Evaluation of the different approaches on three levels

1. Compare identification using a ground truth
2. Compare resulting decomposition using a ground truth

Evaluation

- Evaluation of the different approaches on three levels

1. Compare identification using a ground truth
2. Compare resulting decomposition using a ground truth
3. Evaluate influence on search quality by integrating with a search engine

I. Identification of Constituents

- Main problem for ML approach:
  classifier performance, too many false-positive/ negative classifications (e.g. F-measure for relative clause start 87% vs. 57% for list item start)
- Main problem for rule based approach:
  recognition of embedded structures, and complex long sentences

Evaluation

1. Identification of Constituents

	Rule based	ML based
F-measure	63.6%	47.4%

- Main problem for ML approach:
  classifier performance, too many false-positive/ negative classifications (e.g. F-measure for relative clause start 87% vs. 57% for list item start)
- Main problem for rule based approach:
  recognition of embedded structures, and complex long sentences

Evaluation

I. Identification of Constituents

	Rule based	ML based
F-measure	63.6%	47.4%

- Main problem for ML approach:
  classifier performance, too many false-positive/ negative classifications (e.g. F-measure for relative clause start 87% vs. 57% for list item start)

Evaluation

1. Identification of Constituents

	Rule based	ML based
F-measure	63.6%	47.4%

- Main problem for ML approach:
  classifier performance, too many false-positive/ negative classifications (e.g. F-measure for relative clause start 87% vs. 57% for list item start)
- Main problem for rule based approach:
  recognition of embedded structures, and complex long sentences

2. Resulting Decomposition

Evaluation

2. Resulting Decomposition

	Rule based	ML based
F-measure	64.5%	41.8%

Evaluation

UNI FREIBURG

2. Resulting Decomposition

	Rule based	ML based
F-measure	64.5%	41.8%

3. Search Quality

Evaluation

2. Resulting Decomposition

	Rule based	ML based
F-measure	64.5%	41.8%

3. Search Quality

- Rule based approach

Evaluation

2. Resulting Decomposition

	Rule based	ML based
F-measure	64.5%	41.8%

3. Search Quality

- Rule based approach
  - Average relative increase in precision 35.15%

Evaluation

2. Resulting Decomposition

	Rule based	ML based
F-measure	64.5%	41.8%

3. Search Quality

- Rule based approach
  - Average relative increase in precision 35.15%
  - Absolute F-measure increase between 0.8% and 14%

3. Search Quality

3. Search Quality

- Machine Learning approach

Evaluation

3. Search Quality

- Machine Learning approach
  - Average relative increase in precision 8.6%, absolute F-measure increase between 0.3% and 5%

Evaluation

3. Search Quality

- Machine Learning approach
  - Average relative increase in precision 8.6%, absolute F-measure increase between 0.3% and 5%
  - Decrease in precision and F-measure for 3 of 10 queries

Agenda

- Motivation and Problem Definition
- Rule based Approach
- Machine Learning based Approach
- Evaluation
- Conclusion

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective
- A set of promising future work includes:

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective

- A set of promising future work includes:
  - Larger training set for ML approach, better feature selection and parameter tuning

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective

- A set of promising future work includes:
  - Larger training set for ML approach, better feature selection and parameter tuning
  - Further improvements of rules

Conclusion

- Contextual Sentence Decomposition integral part of Semantic Full-Text Search
- Rule based approach viable, clear improvement
- Machine Learning based approach viable, currently less effective

- A set of promising future work includes:
  - Larger training set for ML approach, better feature selection and parameter tuning
  - Further improvements of rules
  - Hybrid approach combining effective rules with ML classifiers

Thank you

Thank you
for your attention!

Backup

Machine Learning based Approach

UNI FREIBURG

Identification Results

Type	SCI	True	False-Neg	False-Pos	Precision	Recall	F-measure
REL	RULE-SCI	16	7	2	88.9%	69.6%	78%
	ML-SCI	13	10	4	76.5%	56.5%	65%
RELA	RULE-SCI	2	3	7	22.2%	40%	28.6%
	ML-SCI	3	2	13	18.8%	60%	28.6%
LIT	RULE-SCI	41	36	24	63.1%	53.2%	57.7%
	ML-SCI	24	53	24	50%	31.2%	38.4%
SEP	RULE-SCI	23	2	14	62.2%	92.5%	74.2%
	ML-SCI	15	10	6	71.4%	60%	65.2%
TOTAL	RULE-SCI	82	48	47	63,6%	63,1%	63,3%
	ML-SCI	55	75	47	53,9%	42,3%	47,4%

Table 7.1: Evaluation of sentence constituent identification. Results for the rule based SCI (RULE-SCI) and machine learning based SCI (ML-SCI) are shown. Matched constituents must have same start and end.

Machine Learning based Approach

Identification Begin/End Results

Type	SCI	True	False-Neg	False-Pos	Precision	Recall	F-measure
REL(	RULE-SCI	18	5	0	100%	78.3%	87.8%
	ML-SCI	16	7	1	94.1%	69.6%	80%
REL)	RULE-SCI	16	7	2	88.9%	69.6%	78.0%
	ML-SCI	13	10	4	76.5%	56.2%	64.8%
RELA(	RULE-SCI	3	2	6	33.3%	60%	42.9%
	ML-SCI	5	0	11	31.3%	100%	47.7%
RELA)	RULE-SCI	2	3	7	22.2%	40%	28.6%
	ML-SCI	3	2	13	18.8%	60%	28.6%
LIT(	RULE-SCI	48	29	17	73.8%	62.3%	67.6%
	ML-SCI	32	45	16	66.7%	41.6%	51.2%
LIT)	RULE-SCI	50	27	15	76.9%	64.9%	70.4%
	ML-SCI	31	46	17	64.6%	40.3%	49.6%
SEP	RULE-SCI	23	2	14	62.2%	92.5%	74.2%
	ML-SCI	15	10	6	71.4%	60%	65.2%
TOTAL	RULE-SCI	160	75	61	72,4%	68,1%	70,2%
	ML-SCI	115	120	45	71,9%	48,9%	58,2%

Table 7.2: Results for the evaluations of identified starts and ends of constituents. The results for the rule based SCI (RULE-SCI) and machine learning based SCI (ML-SCI) are shown. For ML-SCI this represents the final result after inference and not an intermediate classification.

Machine Learning based Approach

UNI FREIBURG

Begin/End Classifier Performance

Type	True	False-Neg	False-Pos	Precision	Recall	F-measure	Accuracy
REL(	18	5	1	94.7%	78.3%	85.7%	99.5%
REL)	16	7	44	26.7%	69.6%	38.6%	95.5%
RELA(	5	0	16	23.8%	100%	38.4%	98.7%
RELA)	3	2	5	37.5%	60%	46.2%	99.4%
LIT(	39	36	22	63.9%	52%	57.3%	95.1%
LIT)	53	22	5	91.4%	70.7%	79.7%	97.8%

Table 7.3: Filtering phase classifier performance on the test set. For each constituent type we show the number of constituent starts and ends correctly identified, missed and erroneously identified. Accuracy in the last column is based on all 1189 instances of words classified.

Constituent Classifier Performance

Type	True	False-Pos	False-Neg	Precision	Recall	F-measure	Accuracy
REL	14	19	9	42.4%	60.9%	50%	87.4%
RELA	2	5	3	28.6%	40%	33.3%	96.4%
LIT	57	18	20	76%	74%	75%	82.9%

Table 7.4: Inference phase classifier performance on the test set. Accuracy in the last column is based on all 222 instances of constituents classified.

Decomposition Results

	True	False-Pos	False-Neg	Precision	Recall	F-measure
ML-CD	56	63	93	47%	37.6%	41.8%
RULE-CD	98	57	51	63.4%	65.8%	64.5%

Table 7.5: Results for the evaluation of contextual sentence decomposition using the Machine Learning (ML-CD) and Rule Based (RULE-CD) sentence constituent identification.

Evaluation Search Queries

ID	Query
A1	*drug=died/death=:e:entity:[...]:person:**
A2	*united=states=elected=:e:entity:[...]:president:**
A3	*english=:e:entity:[...]:sovereign:**
A4	*political=:e:entity:p[...]:writer:**
A5	*computer=:e:entity:[...]:scientist:**
M1	*edible=leaf/leaves=:e:entity:[...]:plant**
M2	*friend*=:ee:entity:alberteinstein:*=:e:entity:[...]:person:**
M3	*blood=sugar/glucose/:e:entity:[...]:monosaccharide:* =:e:entity:[...]:hormone:**
M4	*die*/death=:ee:entity:diabetes:*=:e:entity:[...]:politician:**
M5	*disqualif*=doping=:e:entity:[...]:athlete:**

Table 7.7: Queries for search quality evaluation. The prefix M indicates evaluation using a manually generated ground truth, and analogously the prefix A evaluation against an automatically generated ground truth. For brevity the queries have been shortened. The full queries can be found in the appendix.

Excerpt of Query Results

Query	Index	True	False-Neg	False-Pos	Precision	Recall	F-measure
A3	BASE	48	11	1409	3.29%	81.36%	4.37%
	ML-CD	48	11	1262	3.66%	81.36%	7.01%
	RULE-CD	47	12	1103	4.09%	79.66%	7.78%
M4	BASE	21	0	7	75%	100%	85.71%
	ML-CD	20	1	4	83.33%	95.23%	88.89%
	RULE-CD	19	2	1	95%	90.48%	92.68%
TOTAL A1-A5	BASE	519	201	18731	2,7%	72,1%	5,2%
	ML-CD	498	222	16546	2,9%	69,2%	5,6%
	RULE-CD	484	236	14654	3,2%	64,7%	6,1%
TOTAL M1-M5	BASE	160	0	173	48%	100%	64,9%
	ML-CD	136	24	131	50,9%	85%	63,7%
	RULE-CD	116	44	49	70,3%	72,5%	71,4%

Machine Learning based Approach

Mapping to the maximum weight independent set problem

Machine Learning based Approach

Mapping to the maximum weight independent set problem

- Build a graph

Machine Learning based Approach

Mapping to the maximum weight independent set problem

- Build a graph
  - Insert node for each span

Machine Learning based Approach

Mapping to the maximum weight independent set problem

- Build a graph
  - Insert node for each span
  - Insert edge betweens spans that overlap or start at same word

Machine Learning based Approach

Mapping to the maximum weight independent set problem

- Build a graph
  - Insert node for each span
  - Insert edge betweens spans that overlap or start at same word
  - Assign high weight if span was correctly classified

Rule based Approach

UNI FREIBURG

Difficult Sentence

Woodwards hypothesis is related to Dennis William Sciama‘s formulation of Mach's principle, a rather vague concept propounded by the philosopher Ernst Mach, which Albert Einstein viewed as something along the lines of "inertia originates in a kind of interaction between bodies”

- Problem: embedded relative clauses
- need to know what they attach to in order to embed correctly

Rule based Approach

UNI FREIBURG

Difficult Sentence

Harrison asserts the existence of female trinities, discusses the Horae as chronological symbols representing the phases of the Moon and goes on to equate the Horae with the Seasons, the Graces and the Fatesand the three seasons of the ancient Greek year, and notes that "The matriarchal goddess may well have reflected the three stages of a woman's life."

- Problem: embedded list items

Rule based Approach

UNI FREIBURG

Difficult Sentence

The shooting left Jack with a lot of resentment towards people who dealt with drugs and caused him to let an innocent man fall to his death

- Problem: list items start with verbs

Rule based Approach

Difficult Sentence

Four popular cocktails that require the use of a muddler are the Old-Fashioned made with whiskey, the mojito made with light rum, the caipirinha made with cachaca and the mint julep made with Bourbon Whiskey.

- Problem: list items contain verb phrases and look like the start of a new self-sufficient sentence