Contextual Sentence Decomposition with Applications to Semantic Full-Text Search

Elmar Haussmann
August 9th, 2011
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

\[ \text{plant edible leaves} \]

Result Sentence

The usable parts of \textit{rhubarb} are the medicinally used roots and the \textit{edible} stalks, however its \textit{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.*
The usable parts of *rhubarb* are the medicinally used roots and the *edible* stalks, however *its leaves* are toxic.
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

- *The usable parts of rhubarb are the medicinally used roots*
- *The usable parts of rhubarb are the edible stalks*
- *its leaves are toxic*
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)
Motivation and Problem Definition

Sentence Constituent Identification
Motivation and Problem Definition

Sentence Constituent Identification

- Identify specific parts of sentence
Sentence Constituent Identification

- Identify specific parts of sentence
- Differentiate 4 types of constituents
Sentence Constituent Identification

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

- Identify specific parts of sentence
- Differentiate 4 types of constituents
  - Relative clauses  \( \text{Albert Einstein, who was born in Ulm, ...} \)
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
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.
Sentence Constituent Identification

- Identify specific parts of sentence
- Differentiate 4 types of constituents
  - Relative clauses  
    \[\text{Albert Einstein, who was born in Ulm, ...}\]
  - Appositions  
    \[\text{Albert Einstein, a well-known scientist, ...}\]
  - List items  
    \[\text{Albert Einstein published papers on Brownian motion, the photelectric effect and special relativity.}\]
  - Separators
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.
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.
Motivation and Problem Definition

Sentence Constituent Recombination
Sentence Constituent Recombination

- Recombine identified constituents into *sub-sentences*
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
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
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.

- 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

Remarks

• Given identified constituents, recombination comparably simple - identification challenging part
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
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
Devise hand-crafted rules by closely inspecting sentence structure.

August 9, 2011

Contextual Sentence Decomposition with Applications to Semantic Full-Text Search
Rule based Approach

Idea

• Devise hand-crafted rules by closely inspecting sentence structure

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

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

Basic Approach

- Identify „stop-words“
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.*
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
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
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.
The usable parts of rhubarb are the medicinally used roots and the edible stalks, however its leaves are toxic.
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

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

Natural Language is Tricky
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:
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
Natural Language is Tricky

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
Rule based Approach

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
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
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
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

**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
Basic Approach

- Apply classifiers in turn to each word
Basic Approach

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

Ideally this would already give a correct solution

1. Apply separator classifier
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
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
However classifiers are not perfect

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

• Some additional ends and beginnings might be identified

• Decisions are local and do not consider admissible constituent structure
• However classifiers are not perfect
• Some additional ends and beginnings might be identified
• Decisions are local and do not consider admissible constituent structure
• 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
• Train classifiers that identify whether a span of the sentence denotes a valid constituent

Apply list item classifier
• 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
• Final result adheres to structural constraints

• More resistant to wrong „local“ classifications
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
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

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>63.6%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

- 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
I. Identification of Constituents

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>63.6%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

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)
I. Identification of Constituents

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>63.6%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

- 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
2. Resulting Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>64.5%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>
Evaluation

2. Resulting Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>64.5%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

3. Search Quality
Evaluation

2. Resulting Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>64.5%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

3. Search Quality

- Rule based approach
2. Resulting Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>64.5%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

3. Search Quality

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

2. Resulting Decomposition

<table>
<thead>
<tr>
<th></th>
<th>Rule based</th>
<th>ML based</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-measure</td>
<td>64.5%</td>
<td>41.8%</td>
</tr>
</tbody>
</table>

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%**
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
• 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 for your attention!
Identification Results

<table>
<thead>
<tr>
<th>Type</th>
<th>SCI</th>
<th>True</th>
<th>False-Neg</th>
<th>False-Pos</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>REL</td>
<td>RULE-SCI</td>
<td>16</td>
<td>7</td>
<td>2</td>
<td>88.9%</td>
<td>69.6%</td>
<td>78%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>13</td>
<td>10</td>
<td>4</td>
<td>76.5%</td>
<td>56.5%</td>
<td>65%</td>
</tr>
<tr>
<td>RELA</td>
<td>RULE-SCI</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>22.2%</td>
<td>40%</td>
<td>28.6%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>3</td>
<td>2</td>
<td>13</td>
<td>18.8%</td>
<td>60%</td>
<td>28.6%</td>
</tr>
<tr>
<td>LIT</td>
<td>RULE-SCI</td>
<td>41</td>
<td>36</td>
<td>24</td>
<td>63.1%</td>
<td>53.2%</td>
<td>57.7%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>24</td>
<td>53</td>
<td>24</td>
<td>50%</td>
<td>31.2%</td>
<td>38.4%</td>
</tr>
<tr>
<td>SEP</td>
<td>RULE-SCI</td>
<td>23</td>
<td>2</td>
<td>14</td>
<td>62.2%</td>
<td>92.5%</td>
<td>74.2%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>71.4%</td>
<td>60%</td>
<td>65.2%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>RULE-SCI</td>
<td>82</td>
<td>48</td>
<td>47</td>
<td>63.6%</td>
<td>63.1%</td>
<td>63.3%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>55</td>
<td>75</td>
<td>47</td>
<td>53.9%</td>
<td>42.3%</td>
<td>47.4%</td>
</tr>
</tbody>
</table>

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.
### Identification Begin/End Results

<table>
<thead>
<tr>
<th>Type</th>
<th>SCI</th>
<th>True</th>
<th>False-Neg</th>
<th>False-Pos</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>REL(</td>
<td>RULE-SCI</td>
<td>18</td>
<td>5</td>
<td>0</td>
<td>100%</td>
<td>78.3%</td>
<td>87.8%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>16</td>
<td>7</td>
<td>1</td>
<td>94.1%</td>
<td>69.6%</td>
<td>80%</td>
</tr>
<tr>
<td>REL)</td>
<td>RULE-SCI</td>
<td>16</td>
<td>7</td>
<td>2</td>
<td>88.9%</td>
<td>69.6%</td>
<td>78.0%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>13</td>
<td>10</td>
<td>4</td>
<td>76.5%</td>
<td>56.2%</td>
<td>64.8%</td>
</tr>
<tr>
<td>RELA(</td>
<td>RULE-SCI</td>
<td>3</td>
<td>2</td>
<td>6</td>
<td>33.3%</td>
<td>60%</td>
<td>42.9%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>5</td>
<td>0</td>
<td>11</td>
<td>31.3%</td>
<td>100%</td>
<td>47.7%</td>
</tr>
<tr>
<td>RELA)</td>
<td>RULE-SCI</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>22.2%</td>
<td>40%</td>
<td>28.6%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>3</td>
<td>2</td>
<td>13</td>
<td>18.8%</td>
<td>60%</td>
<td>28.6%</td>
</tr>
<tr>
<td>LIT(</td>
<td>RULE-SCI</td>
<td>48</td>
<td>29</td>
<td>17</td>
<td>73.8%</td>
<td>62.3%</td>
<td>67.6%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>32</td>
<td>45</td>
<td>16</td>
<td>66.7%</td>
<td>41.6%</td>
<td>51.2%</td>
</tr>
<tr>
<td>LIT)</td>
<td>RULE-SCI</td>
<td>50</td>
<td>27</td>
<td>15</td>
<td>76.9%</td>
<td>64.9%</td>
<td>70.4%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>31</td>
<td>46</td>
<td>17</td>
<td>64.6%</td>
<td>40.3%</td>
<td>49.6%</td>
</tr>
<tr>
<td>SEP</td>
<td>RULE-SCI</td>
<td>23</td>
<td>2</td>
<td>14</td>
<td>62.2%</td>
<td>92.5%</td>
<td>74.2%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>71.4%</td>
<td>60%</td>
<td>65.2%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>RULE-SCI</td>
<td>160</td>
<td>75</td>
<td>61</td>
<td>72.4%</td>
<td>68.1%</td>
<td>70.2%</td>
</tr>
<tr>
<td></td>
<td>ML-SCI</td>
<td>115</td>
<td>120</td>
<td>45</td>
<td>71.9%</td>
<td>48.9%</td>
<td>58.2%</td>
</tr>
</tbody>
</table>

**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.
Begin/End Classifier Performance

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

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.
Machine Learning based Approach

Constituent Classifier Performance

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

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.
## Evaluation

### Decomposition Results

<table>
<thead>
<tr>
<th></th>
<th>True</th>
<th>False-Pos</th>
<th>False-Neg</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML-CD</td>
<td>56</td>
<td>63</td>
<td>93</td>
<td>47%</td>
<td>37.6%</td>
<td>41.8%</td>
</tr>
<tr>
<td>RULE-CD</td>
<td>98</td>
<td>57</td>
<td>51</td>
<td>63.4%</td>
<td>65.8%</td>
<td>64.5%</td>
</tr>
</tbody>
</table>

**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

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

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.
### Evaluation

#### Excerpt of Query Results

<table>
<thead>
<tr>
<th>Query</th>
<th>Index</th>
<th>True</th>
<th>False-Neg</th>
<th>False-Pos</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>A3</td>
<td>BASE</td>
<td>48</td>
<td>11</td>
<td>1409</td>
<td>3.29%</td>
<td>81.36%</td>
<td>4.37%</td>
</tr>
<tr>
<td></td>
<td>ML-CDS</td>
<td>48</td>
<td>11</td>
<td>1262</td>
<td>3.66%</td>
<td>81.36%</td>
<td>7.01%</td>
</tr>
<tr>
<td></td>
<td>RULE-CD</td>
<td>47</td>
<td>12</td>
<td>1103</td>
<td>4.09%</td>
<td>79.66%</td>
<td>7.78%</td>
</tr>
<tr>
<td>M4</td>
<td>BASE</td>
<td>21</td>
<td>0</td>
<td>7</td>
<td>75%</td>
<td>100%</td>
<td>85.71%</td>
</tr>
<tr>
<td></td>
<td>ML-CDS</td>
<td>20</td>
<td>1</td>
<td>4</td>
<td>83.33%</td>
<td>95.23%</td>
<td>88.89%</td>
</tr>
<tr>
<td></td>
<td>RULE-CD</td>
<td>19</td>
<td>2</td>
<td>1</td>
<td>95%</td>
<td>90.48%</td>
<td>92.68%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>A1-A5</td>
<td>BASE</td>
<td>519</td>
<td>201</td>
<td>18731</td>
<td>2.7%</td>
<td>72.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML-CDS</td>
<td>498</td>
<td>222</td>
<td>16546</td>
<td>2.9%</td>
<td>69.2%</td>
</tr>
<tr>
<td></td>
<td>RULE-CD</td>
<td>484</td>
<td>236</td>
<td>14654</td>
<td>3.2%</td>
<td>64.7%</td>
<td>6.1%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>M1-M5</td>
<td>BASE</td>
<td>160</td>
<td>0</td>
<td>173</td>
<td>48%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ML-CDS</td>
<td>136</td>
<td>24</td>
<td>131</td>
<td>50.9%</td>
<td>85%</td>
</tr>
<tr>
<td></td>
<td>RULE-CD</td>
<td>116</td>
<td>44</td>
<td>49</td>
<td>70.3%</td>
<td>72.5%</td>
<td>71.4%</td>
</tr>
</tbody>
</table>
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
Mapping to the maximum weight independent set problem

- Build a graph
  - Insert node for each span
Mapping to the maximum weight independent set problem

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

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
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 Fates, and 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

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