Improving Dependency Parsing Using Sentence Clause Charts

Linguistic Mondays, 10.10.2016
MFF UK

Vincent Kríž
Motivation

- large collections of documents
- efficient browsing & querying
- typical approaches
  - full-text search
  - meta-data search

no semantics
• Intelligent Library (INTLIB)
  – founded by
  – 2012–2015
  – partners
• New search approach
  - semantic interpretation of documents
  - suitable DB & query language
  - user-friendly browsing & querying
INTLITB

• **New search approach**
  - semantic interpretation of documents
  - suitable DB & query language
  - user-friendly browsing & querying

• **Knowledge base**
  - set of entities and relations between them
INTL LIB

• New search approach
  – semantic interpretation of documents
  – suitable DB & query language
  – user-friendly browsing & querying

• Knowledge base
  – set of entities and relations between them

• RExtractor
  – information extraction system
RExtractor

- entity and relation extraction from plain-texts
- server architecture
  - process client's requests
  - REST API
  - web interface (~ demo)

http://quest.ms.mff.cuni.cz:14280
RExtractor

• extract entities and relations
  – queries over dependency trees
  – domain and language independent

• real use-case defined by INTLIB
  – definitions, rights and obligations in Czech laws
  – Czech extraction strategy
RExtractor

• extract entities and relations
  - queries over dependency trees
  - domain and language independent

• real use-case defined by INTLIB
  - definitions, rights and obligations in Czech laws
  - Czech extraction strategy
  - English extraction strategy
Evaluation

Czech Legal Text Treebank 1.0 (CLTT)

- Accounting Act (563/1991 Coll.)
- Decree on Double-entry Accounting for undertakers (500/2002 Coll.)

- automatically parsed, then manually checked
  - 1,133 manually annotated dependency trees
  - 35,085 tokens
Evaluation

Czech Legal Text Treebank 1.0 (CLTT)

Evaluation

Error analysis

<table>
<thead>
<tr>
<th>Error</th>
<th># of errors</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parser</td>
<td>145</td>
<td>59.7%</td>
</tr>
<tr>
<td>Query</td>
<td>93</td>
<td>38.3%</td>
</tr>
<tr>
<td>Entity</td>
<td>5</td>
<td>2.1%</td>
</tr>
</tbody>
</table>
Baseline

Automatic parser for Czech

- MST parser
  - trained on newspaper texts
  - long sentences still problematic
Baseline

Long sentences dependency parsing

- as the sentence length increases, the unlabeled attachment score (UAS) decreases
Related Work

- **segmentation of complex sentences**
  - Kuboň (2001), Kuboň et al. (2007)
  - segments – easily detectable and linguistically motivated units
  - may be combined into clauses
  - provide a structure of a complex sentence with regard to the mutual relationship of individual clauses
Related Work

- **segmentation of complex sentences**
  - Lopatková and Holan (2009)
  - a new module between morphological and syntactic analysis
  - determine the overall sentence structure
  - **segmentation chart**
    - relationship among segments
    - especially relations of coordination, apposition and subordination
S tím byly trochu problémy, protože starosta v řeči rád zdůrazňoval své vzdělání.

Credits: Lopatková and Holan (2009)
Related Work

S tím byly trochu problémy, protože starosta v řeči rád zdůrazňoval své vzdělání.

- **split sentence into segments**
  - rule-based boundaries identification
    - punctuation marks, coordinating conjunctions, brackets, ...
Related Work

- determine mutual relations
  - manually designed rules
    - finite verb
    - subordinating expression
    - opening bracket

Credits: Lopatková and Holan (2009)
Related Work

S tím byly trochu problémy

protože starosta ... vzdělání

Credits: Lopatková and Holan (2009)
Related Work

- segmentation chart
  - captures the layer of embedding for individual segments

Credits: Lopatková and Holan (2009)
## Related Work

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>S tím byly trochu problémy</td>
<td>protože starosta … vzdělání</td>
</tr>
</tbody>
</table>

• **segmentation chart principles**
  
  - main segments belong to layer 0
  - segments that depend on segment on layer $k$ belong to $k+1$
  - coordinated segments have the same layer
  - segments in parenthesis/brackets belong to $k+1$ layer

Credits: Lopatková and Holan (2009)
Related Work

- **sentence clause structure**
  - Lopatková et al. (2012)
  - manual clause structure annotation based on the concept of segments
  - 2,699 annotated sentences
Related Work

• sentence clause structure
  – Krůza and Kuboň (2014)
  – automatic procedure for recognizing clauses and their mutual relationship \textit{from plain-texts}
  – Bejček et al. (2013)
  – automatic procedure for recognizing clauses and their mutual relationship \textit{from dependency trees}
  – used for clause annotation in PDT 3.0
Related Work

• clause annotation in PDT 3.0
Clause Charts

• analogous to a segmentation chart
  – Lopatková and Holan (2009)
  – two differences
    • subordinating conjunctions at the beginning of each clause are considered as boundaries
    • clauses split into two parts (by an embedded clause) are considered as two different clauses
While failure is usually an orphan, the success tends to have many fathers, claiming eagerly that particularly they were present at its conception.

Credits: Kuboň et al. (2007)
Clause Charts

- generating clause charts
  - from dependency trees with the clause annotation
  - a layer of embedding $\rightarrow$ number of different clauses on the path from the clause to the root in the dependency tree
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Ačkoliv jsou si jisti, že se jim podařilo rozdělit molekuly dekaboranu na části, nepodařilo se zmíněným vědcům zatím určit, jaké produkty přitom vznikly.
Clause Charts

- exploring clause charts
Clause Charts

- exploring clause charts
Clause Charts

- exploring clause charts
  - sentence with 36 clauses
  - sentence with 7 layers of embedding
    - 0B1B2B3B4B5B6
Clause Chart Parsing

- **new method for dependency parsing**
  - exploit an existing dependency parser
    - trained on complete sentences
  - exploit gold-standard clause charts

- **two specific strategies**
  - parsing coordinated clauses
  - parsing subordinated clauses
Clause Chart Parsing

• parsing coordinated clauses
  - let's explore the most simple sentences with coordinated clauses – 0B0
  - how good is the full-scale parser on individual clauses from 0B0?
• parsing coordinated clauses
  - let's explore the most simple sentences with coordinated clauses – 0B0
  - how good is the full-scale parser on individual clauses from 0B0?
Clause Chart Parsing

- parsing coordinated clauses
  - $C_1, C_2, \ldots, C_n$
    - neighboring coordinated clauses
    - on the same layer
  - parse $C_i$ individually
    - obtain dependency tree $T_i$ with root node $r_i$
  - create a sequence of tokens
    - $S = r_1 B_{1,2} r_2 B_{2,3} \ldots r_n$
  - parse $S$, obtain $T_s$
  - build a final dependency tree using $T_i$ and $T_s$
Clause Chart Parsing

• parsing coordinated clauses
  – *John loves Mary and Linda hates Peter.*
  – $C_1 = \{\text{John loves Mary}\}$, $C_2 = \{\text{Linda hates Peter}\}$
  – parse individual clauses
    • $C_1 \rightarrow T_1$, $r_1 = \text{loves}$
    • $C_2 \rightarrow T_2$, $r_2 = \text{hates}$
  – create a sequence of tokens
    $S = \{\text{loves and hates}\}$
  – parse $S \rightarrow T_s$

  – build a final dependency tree
Clause Chart Parsing

- parsing coordinated clauses
Clause Chart Parsing

- parsing coordinated clauses

![Chart showing improvement in UAS]

+1.4% of UAS in average
Clause Chart Parsing

- **parsing subordinated clauses**
  - exploring 0B1 sentences
  - almost no improvement when parse individual clauses
Clause Chart Parsing

- parsing subordinated clauses
  - exploring 0B1 sentences
  - almost no improvement when parse individual clauses
  - UAS is significantly higher then overall UAS
Clause Chart Parsing

- parsing subordinated clauses
  - $C_1, C_2, \ldots, C_n$
    - the longest sequence of neighboring subordinated clauses
      - $\text{layer}(C_{i+1}) = \text{layer}(C_i) + 1$
    - create a sequence of tokens
      - $S = C_1 B_{1,2} C_2 B_{2,3} \ldots C_n$
    - parse $S$, obtain $T_s$
Clause Chart Parsing

- parsing subordinated clauses
  - evaluation on 0B1B0 sentences
    - parse 0B1
    - parse 0B0
Clause Chart Parsing

• parsing subordinated clauses
  – evaluation on 0B1B0 sentences
    • parse 0B1
    • parse 0B0

+1.6% of UAS in average
Clause Chart Parsing

- **CCP as full-scale parsing**
  - work in cycles
  - check the deepest layer
    - if there are coordinated clauses → apply 0B0 strategy
    - otherwise identify the longest sequence of subordinated clauses → apply 0B1 strategy
  - use standard full-scale parsing as a fall-back
Clause Chart Parsing

- final evaluation (excl. 0 sentences)
Clause Chart Parsing

- final evaluation (excl. 0 sentences)

[Bar chart showing comparison between Full-scale and CCP with labels for PDT dtest, PDT etest, and CAC 2.0.]

+1.0% of UAS in average
Clause Chart Parsing

- final evaluation
Clause Chart Parsing

- final evaluation

+0.7% of UAS in average
Future Work

• evaluation
  - Czech Legal Text Treebank 1.0
  - relation extraction in RExtractor

• clause charts
  - extraction from plain-text

• special parsers
  - train on individual clauses
Conclusion

- sentence clause structure helps with dependency parsing
- 1% increase of UAS on complex sentences
Conclusion

- sentence clause structure helps with dependency parsing
- 1% increase of UAS on complex sentences

In the real parsing task, automatically detected clause structures must be used, not gold-standard
Conclusion

- sentence clause structure helps with dependency parsing
- 1% increase of UAS on complex sentences

- in the real parsing task, automatically detected clause structures must be used, not gold-standard
- we can train specialized clause-parsers – for main clauses, subordinated clauses, merge clauses, ...
Conclusion

- sentence clause structure helps with dependency parsing
- 1% increase of UAS on complex sentences

- in the real parsing task, automatically detected clause structures must be used, not gold-standard
- we can train specialized clause-parsers – for main clauses, subordinated clauses, merge clauses, ...
- we can find out better strategies for parsing sequences of subordinated clauses