Motivation

Forgetful & Self-Organizing Information Systems
(to support information management & knowledge work)

- continuous information value assessment
- continuous user activity tracking and evidence processing
- information extraction in (near) real-time

2013 – 2016 (EU project)

2016 – 2019 (DFG project)
Very Low System Response Time Needed


100 ms „limit for having the user feel that the system is reacting instantaneously“

1000 ms „limit for the user’s flow of thought to stay uninterrupted“

Problem of Inflections


DBpedia Spotlight:

## Related Work

<table>
<thead>
<tr>
<th>Inflection-Tolerant NER</th>
<th>Real-Time Capable NER</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Savary &amp; Piskorski (2010)</strong></td>
<td><strong>Dlugolinsky, Nguyen et al. (2013/2014)</strong></td>
</tr>
<tr>
<td>→ IE platform SProUT, Polish, explicitly listing all inflected forms</td>
<td>→ several gazetteer-based approaches</td>
</tr>
<tr>
<td><strong>Day &amp; Prukayastha (2013)</strong></td>
<td><strong>Al-Rfou &amp; Skiena (2012)</strong></td>
</tr>
<tr>
<td>→ NER for Indian languages, gazetteer-based &amp; ML &amp; hybrid</td>
<td>→ SpeedRead, 10x faster than CoreNLP, 153 tokens/sec.</td>
</tr>
</tbody>
</table>

**Al-Jumaily et al. (2013)**
→ NER for Arabic text mining, no details on performance given
Approach

NE recognizer
as a combination of several
multi-layer finite state transducers
having different tolerance levels

connection to knowledge graph(s)
[instance labels, types, …]

access to language information
[word types, flections, …]

arbitrary text

voter

named entities in text

Wiktionary
[ˈvikʃənəri], n
Das freie Wörterbuch

DBpedia

LT
Morphy
Multi-Layer FST with High Tolerance

Layer 1: Character Layer

Layer 2: Word Layer

Input: Deutsches Forschungszentrum für Künstliche Intelligenz

„deutsch“ (German) [w1]
„Forschungszentrum“ (research center) [w2]
„für“ (for) [w3]
„künstlich“ (artificial) [w4]
„Intelligenz“ (intelligence) [w5]

„Deutsches Forschungszentrum für Künstliche Intelligenz“ (German Research Center for Artificial Intelligence)
Multi-Layer FST with Low Tolerance

Input: Christian_Jilek

Layer 1: Character Layer

Layer 2: Word Layer

"Christian Jilek"
Word Node Processing

- **w6 processed**: $\emptyset \rightarrow w6 \rightarrow w7 \rightarrow w8 \rightarrow w9$
- **w7 processed**: $\emptyset \rightarrow w7 \rightarrow w8 \rightarrow X$
- **w8 processed**: $\emptyset \rightarrow w8 \rightarrow w9$
- **w9 processed**: $\emptyset \rightarrow X$

- **processor 1**
- **processor 2**
- **processor 3**
- **processor 4**
Additional Research Question

Our approach accepts 576 variations for the term „Deutsches Forschungszentrum für Künstliche Intelligenz“ (German Research Center for Artificial Intelligence)

only 4 of them are correct

→ increased false positive rate in real-world scenarios?
Evaluation Setting

• idea:
  • use the German Wikipedia as a large set of texts written by different people
  • use DBpedia types to decide whether to apply low or high inflection tolerance
  • use Wikipedia annotations as a “silver standard“
    • term used (often inflected form) manually annotated with its article name (often basic form)

      [[ Haus | Häuser ]]
      [[ Junktor | Junktoren ]]

• problems:
  • independent term-links-combinations
  • adjective-noun-combinations

→ use Levenshtein distance (LD) to identify samples (typically LD<=4)

• ambiguities (e.g. >1000 instances of „Jewish Cemetery“)
• terms not annotated in „their own“ article (e.g. „Berlin“ in article about „Berlin“)

• benefit: 3.9M articles having 50.4M annotations
Results: Recall

![Chart showing recall results for different LD levels with bars for HMT/CST, MLFST, and StemFST.](image)
Results: Measuring Precision

„A commercial personal information management tool is used in the project.“

- **PO**
  - only overlapping terms as false positives, ambiguities disregarded

- **PO**
  - only overlapping terms including ambiguities as false positives

- **PA**
  - all other terms as false positives, ambiguities disregarded

- **PA**
  - all other terms including ambiguities as false positives
Results: Precision

\[ P_O^* \]

- HMT/CST
- MLFST
- StemFST

<table>
<thead>
<tr>
<th></th>
<th>PO*</th>
<th>PO</th>
<th>PA*</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO*</td>
<td>HMT/CST</td>
<td>MLFST</td>
<td>StemFST</td>
<td></td>
</tr>
<tr>
<td>PO</td>
<td>HMT/CST</td>
<td>MLFST</td>
<td>StemFST</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>HMT/CST</td>
<td>MLFST</td>
<td>StemFST</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>HMT/CST</td>
<td>MLFST</td>
<td>StemFST</td>
<td></td>
</tr>
</tbody>
</table>

Christian Jilek – LDK 2019
Results: Processing Speed & Memory Consumption

<table>
<thead>
<tr>
<th>Method</th>
<th>Memory Consumption (GiB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMT</td>
<td>5048</td>
</tr>
<tr>
<td>CST</td>
<td>350</td>
</tr>
<tr>
<td>MLFST</td>
<td>6</td>
</tr>
<tr>
<td>OpenNLP (basic PL)</td>
<td>0</td>
</tr>
<tr>
<td>CoreNLP (basic PL)</td>
<td>0</td>
</tr>
</tbody>
</table>

Basic PL:
- tokenizer
- sentence splitter
- POS tagger

Total runtime for performing NER on German Wikipedia:
- HMT: 60 min
- CST: 45 min
- MLFST: 30 min
- StemFST: 15 min

[Graph showing processing speed and memory consumption for various methods]
Conclusion

• presented inflection-tolerant and real-time capable OB NER approach based on
  • Trie-based string matching
  • finite state cascades
  • exhaustive inflection listing
  • exploiting ontological background information

• comparably fast as available high speed methods

• outperforming them in recognizing terms that lexically vary slightly (e.g. inflection)

• narrowing the gap to more sophisticated but slower NLP pipelines without losing too much runtime performance
Outlook

• incorporate disambiguation mechanisms (exploiting user context)

• add more layers to scan for patterns (ToDos, appointments, Hearst patterns, …)

• improve language capabilities (rules, heuristics, multi-language support, …)

• incorporate StemFST into MLFST for multi-word terms (slightly better precision)
Selected References


Nielsen, J.: **Usability Engineering.** Morgan Kaufmann (1993)


Thanks for your attention! 😊

Parts of this work were funded by the DFG in the SPP on Intentional Forgetting in Organizations.

Christian Jilek – LDK 2019