Interactive Text Exploration

Günter Neumann, DFKI, Saarbrücken, Germany
Joined work with Sven Schmeier, DFKI, Berlin.
Overview of my talk

- Motivation and Background
- Interactive exploratory search
- Methods and technology
- Where we are, where we want to go
“The Big Idea”

- The extraction, classification, and talking about information from large-scale unstructured noisy multi-lingual text sources.

„Reading text and talking about it“
Motivation

- Today’s Web search is still dominated by one-shot-search:
  - Users basically have to know what they are looking for.
  - The documents serve as answers to user queries.
  - Each document in the ranked list is considered independently.

- Restricted assistance in content-oriented interaction
We consider a user query as a specification of a topic that the user wants to know and learn more about. Hence, the search result is basically a graphical structure of the topic and associated topics that are found.

The user can interactively explore this topic graph using a simple and intuitive (touchable) user interface in order to either learn more about the content of a topic or to interactively expand a topic with newly computed related topics.
Exploratory Search on Mobile Devices
Our Approach – On-demand Interactive Open Information Extraction

- Topic-driven Text Exploration
  - Search engines as API to text fragment extraction (snippets)

- Dynamic construction of topic graphs
  - Empirical distance-aware phrase collocation
  - Open relation extraction

- Interaction with topic graphs
  - Inspection of node content (snippets and documents)
  - Query expansion and eventually additional search
  - Guided exploratory search for handling topic ambiguity
von Willebrand Disease

Type 2 von Willebrand disease is characterised by qualitative defects in von Willebrand factor (VWF). Von Willebrand disease (VWD) is caused by a deficiency or dysfunction of von Willebrand factor (VWF).

Intracellular storage and regulated secretion of von Willebrand factor... quantitative von Willebrand disease.

Acquired von Willebrand syndrome (AVWS) usually mimics von Willebrand disease (VWD) type 1 or 2A.

Porcine and canine von Willebrand factor and von Willebrand disease... hemostasis, thrombosis, and atherosclerosis.

Pregnancy and delivery in women with von Willebrand's disease.... different von Willebrand factor mutations.

Investigation of von Willebrand factor gene.... mutations in Korean von Willebrand disease patients....

Multiple von Willebrand factor mutations in patients with recessive type 1 von Willebrand disease.

Oligosaccharide structures of von Willebrand factor and their potential role in von Willebrand disease.

bleeding disorder caused by decrease

is characterized

platelet glycoprotein

missense

coincides

von willebrand factor pseudogene
Topic Graphs

Main data structure

- A graphical summary of relevant text fragments in form of a graph
- Nodes and edges are text fragments
  - Nodes: entities phrases
  - Edges: relation phrases
- Content of a node: set of snippets it has been extracted from, and the documents retrievable via the snippets’ web links.

Properties

- Open domain
- Dynamic index structure
- Weight-based filtering/construction
Construction of Topic graphs

- Identification of relevant text fragments
  - A document consisting of topic-query related text fragments

- Identification of nodes and edges
  - Distance-aware collocation
  - Clustering-based labels for filtering

- Technology
  - Shallow Open relation Extraction (ORE) for snippets
  - Deeper ORE for more regular text

For each chunk $c_i$ do:

\[(c_i, c_{i+1}, d_{(i+1)})$, $(c_i, c_{i+2}, d_{(i+2)})$, ...\]

\[(c_i, c_j, \#c_i, \#c_j, D_{ij})\] with

\[D_{ij} = \{(freq_1, dist_1), (freq_2, dist_2), ...\}\]

\[PMI(cp) = \log_2 \left( \frac{p(c_i, c_j)}{p(c_i) \cdot p(c_j)} \right)\]

\[= \log_2(p(c_i)) - \log_2(p(c_i) \cdot p(c_j))\]

The visualized topic graph $TG$ is then computed from a subset $CPD'_M \subseteq CPD_M$ using the $m$ highest ranked $cpd$ for fixed $c_i$. In other words, we restrict the complexity of a TG by restricting the number of edges connected to a node.
Evaluation of Mobile Touchable User Interface

- 20 testers
  - 7 from our lab
  - 13 “normal” people

- 10 topic queries
  - Definitions: EEUU, NLF
  - Person names: Bieber, David Beckham, Pete Best, Clark Kent, Wendy Carlos
  - General: Brisbane, Balancity, Adidas.

- Average answer time for a query: ~0.5 seconds

<table>
<thead>
<tr>
<th></th>
<th>v.good</th>
<th>good</th>
<th>avg.</th>
<th>poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>results first sight</td>
<td>55%</td>
<td>40%</td>
<td>15%</td>
<td>-</td>
</tr>
<tr>
<td>query answered</td>
<td>71%</td>
<td>29%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>interesting facts</td>
<td>33%</td>
<td>33%</td>
<td>33%</td>
<td>-</td>
</tr>
<tr>
<td>suprising facts</td>
<td>33%</td>
<td>-</td>
<td>-</td>
<td>66%</td>
</tr>
<tr>
<td>overall feeling</td>
<td>33%</td>
<td>50%</td>
<td>17%</td>
<td>4%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>v.good</th>
<th>good</th>
<th>avg.</th>
<th>poor</th>
</tr>
</thead>
<tbody>
<tr>
<td>results first sight</td>
<td>43%</td>
<td>38%</td>
<td>20%</td>
<td>-</td>
</tr>
<tr>
<td>query answered</td>
<td>65%</td>
<td>20%</td>
<td>15%</td>
<td>-</td>
</tr>
<tr>
<td>interesting facts</td>
<td>62%</td>
<td>24%</td>
<td>10%</td>
<td>4%</td>
</tr>
<tr>
<td>suprising facts</td>
<td>66%</td>
<td>15%</td>
<td>13%</td>
<td>6%</td>
</tr>
<tr>
<td>overall feeling</td>
<td>54%</td>
<td>28%</td>
<td>14%</td>
<td>4%</td>
</tr>
</tbody>
</table>
Guided Exploratory Search

- Problem: a topic graph might merge information from different topics/concepts

- Solution:
  - Guided exploratory search
  - Using an external KB (e.g., Wikipedia)

- Strategy
  - Compute topic graph $TD_q$ for query $q$
  - Ask KB (Wikipedia or any other KB) if $q$ is ambiguous
  - Let user select reading $r$, and use selected Wikipedia article for expanding $q$ to $q'$
  - Compute new topic graph $TD_{q'}$
Evaluation

List of celebrity guest stars in Sesame Street:

209 different queries

List of film and television directors:

229 different queries
Evaluation

- **Goal:**
  - We want to analyze whether our approach helps building topic graphs which express a preference for the selected reading.

- **Automatic evaluation:**
  - **Method**
    - For each reading article $r$, compute topic graph $TD_r$ using expanded query
    - Compare $TD_r$ with all readings and check whether best reading equals $r$
  - **Advantage:** No manual checking necessary
  - **Disadvantage:** Correctness of $TD_R$ needs to be proven

- **Manual evaluation:**
  - Double-check the results of the automatic evaluation
  - Prove the results at least for the examples used in evaluation
## Results

<table>
<thead>
<tr>
<th>set</th>
<th>#queries</th>
<th>good</th>
<th>bad</th>
<th>acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sesame + Colloc.</td>
<td>209</td>
<td>375</td>
<td>54</td>
<td>87.41 %</td>
</tr>
<tr>
<td>Sesame + Colloc.+ SemLabel</td>
<td>209</td>
<td>378</td>
<td>51</td>
<td>88.11 %</td>
</tr>
<tr>
<td>Hollywood + Colloc.+ SemLabel</td>
<td>229</td>
<td>472</td>
<td>28</td>
<td>94.40 %</td>
</tr>
<tr>
<td>Hollywood + Colloc.+ SemLabel</td>
<td>229</td>
<td>481</td>
<td>19</td>
<td>96.20 %</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1st task</th>
<th>2nd task</th>
</tr>
</thead>
<tbody>
<tr>
<td>set</td>
<td>guidance</td>
</tr>
<tr>
<td>Sesame</td>
<td>ca. 95 %</td>
</tr>
<tr>
<td>Hollywood</td>
<td>ca. 95 %</td>
</tr>
<tr>
<td>Sesame</td>
<td>&gt; 97 %</td>
</tr>
<tr>
<td>Hollywood</td>
<td>&gt; 97 %</td>
</tr>
</tbody>
</table>

- **Automatic**
  - Colloc. – empirical collocations for topic graph computation
  - SemLabel – Filtering of nodes using semantic labels computed via SVD (Carrot2)

- **Manual**
  - 2 test persons
  - 20 randomly chosen celebrities and 20 randomly chosen directors
  - 1st task: Exploratory search and personal judgments of the Guidance by the system
  - 2nd task: Check all associated nodes after choosing a meaning in the list
Summary and Discussion

- Interactive topic graph exploration
  - Unsupervised open information extraction
  - On-demand computation of topic graphs
  - Strategies for guided exploratory search
  - Effective for Web snippet like text fragments
  - Implemented for EN and DE on mobile touchable device

- Drawback
  - Problems in processing text fragments from large-scale text directly
  - Especially Open Relation Extraction for German is challenging

- Solution:
  - Nemex - A new multilingual Open Relation Extraction approach
Nemex – A Multilingual Open Relation Extraction Approach

- Uniform multilingual core ORE
  - N-ary extraction
  - Clause-level

- Multi-lingual
  - Very few language-specific constraints over dependency trees
  - Current: English and German

- Efficiency
  - Complete pipeline (form sentence splitting, to POS-tagging, to NER, to dependency parsing, to relation extraction)
  - About 800 sentences/sec
  - Streaming based – small memory footprint
German ORE is Challenging

- Challenging properties of German
  - Morphology/Compounding*
  - No strict word ordering (especially between phrases)
  - Discontinuous elements, e.g., verb groups

- Simple, pattern-based ORE approach difficult to realize (e.g., ReVerb)

- Deep sentence analysis helpful
  - Current multilingual dependency parsers provide very good robustness!
  - DFKI’s MDParser is very efficient: 1000 sentences/second (but see also Chen&Manning, 2014)

- Challenge:
  - Can we design a core uniform ORE approach for English, German, ... ?

*Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz
"the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef"
Multilingual ORE – Our Approach

- Multi-lingual open relation extraction
  - Only few Language-specific constraints necessary (constraints over direct dependency relations (head, label, modifier))
  - Few language-independent constraints in case of uniform dependency annotations, e.g., McDonald et al., 2013

- Processing strategy
  - Head-Driven Phrase Extraction
  - Top-down head-driven traversal of dependency tree
Mammalian NMD was mostly studied in cultured cells so far and there was no direct evidence yet that NMD could operate in the brain.
(Mammalian NMD, was mostly studied so far, in cultured cells)
(no direct evidence, was yet, there)
(NMD, could operate, in the brain)

**Annotated sentence:**
[[[Arg11 Mammalian NMD Arg11]]] --->Rel1 was mostly studied
[[[Arg13 in cultured cells Arg13]]] so far Rel1<--- and [[[Arg23 there Arg23]]] --->Rel2 was [[[Arg21 no direct evidence Arg21]]] yet
Rel2<--- that [[[Arg31 NMD Arg31]]] --->Rel3 could operate Rel3<---
[[[Arg33 in the brain Arg33]]].

*Details omitted

**Extension of the annotation scheme introduced by Mesquita et al., 2013**
Zuvor hatte Asmussen mitgeteilt, dass er sein Amt als EZB-Direktor in Kürze aufgeben will:

*Earlier had Asmussen informed, that he his position as EZB-director in the_near_future quit will:*

Earlier Asmussen has informed that he will quit his position as EZB-director in the_near_future:
Example German – Cont.

(Asmussen, Zuvor hatte mitgeteilt)
(er, aufgeben will, sein Amt, als EZB-Direktor, in Kürze)

Annotation:
--->Rel1 Zuvor hatte [[[Arg11 Asmussen Arg11]]] mitgeteilt Rel1<--- , dass [[[Arg21 er Arg21]]] [[[Arg22 sein Amt Arg22]]] [[[Arg23 als EZB-Direktor Arg23]]] [[[Arg24 in Kürze Arg24]]] ---&gt;Rel2 aufgeben will Rel2<--- :
Nemex – Current Status

- **Properties**
  - Efficient text stream for EN and DE implemented
  - Uniform POS and Dependency labels
  - Small set of uniform constraints over dependency relations

- **Very fast & Domain independent**
  - About 800 sentences per second for complete pipeline

- **Current /near future work**
  - Improve cross-clausal resolution
  - Extensive evaluation, intrinsic and extrinsic
  - Adaptation to other languages
    - Conll based dependency treebanks (uniform and specific)
Future action points

- Cross-sentence open information extraction
  - **Goal:** co-reference resolution, integration of more fine-grained information to dependency parsers (morphology), text inference

- Beyond isolated topic graphs
  - **Goal:** share topic graphs, compare topic graphs, monitor topic graphs

- Interactive text data mining and knowledge discovery
  - **Goal:** support abstract interactions, e.g., “more like this”, “less like this”, “what is this”, ...
DONE

Thank you for Your Attention!