Generating Links to Background Knowledge:

A Case Study Using Narrative Radiology Reports

Jiyin He\textsuperscript{1}, Maarten de Rijke\textsuperscript{2}, Merlijn Sevenster\textsuperscript{3},
Rob van Ommering\textsuperscript{3}, Yuechen Qian\textsuperscript{3}

\textsuperscript{1} CWI; \textsuperscript{2} University of Amsterdam; \textsuperscript{3} Philips Research
Medical content on the Web

Yahoo! Answers

Resolved Question

What is lacunar infarct it showed up on a CT but i can not find anything about it on the medical sites thanks?

6 years ago

Best Answer - Chosen by Asker

Lissacal

Lacunar infarct
small deep cerebral infarcts located mainly in the basal ganglia in the brain stem, internal and external capsules and periventricular regions. They are most likely due to atherosclerotic occlusion of penetrating arteries less likely due to small emboli. Clinically they may present with pure motor or pure sensory dysarthria, symptoms and signs that usually clear in a relat
Automatically generate explanatory links to background resources

**FINDINGS:**

There is a small curvilinear area of restricted diffusion involving the **posterior limb of the right internal capsule** extending to the adjacent **periventricular white matter** along the posterior body of the **caudate nucleus**. There is an old **punctate lacunar infarct** at the **genu of the right internal capsule**.

There is no **acute hemorrhage**, no...
A case study

- Narrative neuroradiology reports
  - Gives narrative descriptions of the radiologist’s findings, diagnoses and recommendations for followup actions

- Wikipedia as background knowledge resource
  - Much work has been done in automatic link generation with Wikipedia in general domain
  - Rich interlinking structure provides valuable training data
  - Covers many medical thesauri and ontologies, e.g., MeSH, ICD-9, ICD-10
A solved problem?

- State-of-the art linking systems
  - E.g., Wikify! (Mihalcea and Csomai, 2007), Wikipedia Miner (Milne and Witten 2008)
  - Exploit Wikipedia link structure
  - Domain independent

- How do they perform in generating links for medical content?
  - An empirical evaluation of existing linking systems on a manually annotated test collection
Two state-of-the-art linking systems

• Wikify!
  • Step 1 - Anchor detection:
    • Keyphraseness score - the more often a phrase occurs in WP as an anchor text, the more likely it will be used as an anchor text again.
  • Step 2 - Target finding:
    • Lesk algorithm - Measuring the similarity between the context of an anchor text and the target page
    • Machine learning based approach

• Wikipedia Miner
  • Step - 1: For each phrase in the current text, finding candidate target pages by measuring the relatedness of a WP page and the context of the phrase
  • Step - 2: Classification to determine the target page for a phrase
  • Step - 3: Classification on anchor - target pairs for anchor detection
Test collection

- 860 anonymized narrative neuroradiology reports
  - 29,256 anchor - target pairs; 6,440 unique links
  - Anchors are body locations, findings and diagnosis
- Annotated by 3 medical informatics specialists
  - Stage 1: Manually select anchor texts
  - Stage 2: Search for target pages with Wikipedia search engine
    - If no direct matched Wikipedia page was found, a more general concept that reasonably covers the topic was sought
    - If no such page was found, no target was assigned
  - Disagreements were resolved through communication (~5% cases)
Experimental setup

• System setup
  • Re-implemented Wikify! ; two versions for target finding - Lesk and machine learning based approach
  • Use Wikipedia miner as a blackbox
• Evaluation metrics: precision, recall and F-measure
• Evaluation on
  • anchor detection
  • target finding - only on correctly identified anchors
  • and overall performance
## Results

<table>
<thead>
<tr>
<th>System</th>
<th>Anchor detection</th>
<th>Target finding</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>Wikify! (Lesk)</td>
<td>0.35</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>Wikify! (ML)</td>
<td>0.35</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>WM</td>
<td>0.35</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

- Generally not satisfactory
- only 30% of the links were correctly identified
- Low performance for anchor detection
- Relatively OK performance for target finding
Some observations

- Two properties of the medical anchor texts
  - Regular syntactic structure
    - 70% are noun phrases, where 38 % are single nouns, 32% are nouns with one or more modifiers
    - Can be useful features for anchor detection
  - Complicated semantic structure
    - e.g. “acute cerebral and cerebellar infarction”
    - May cause problems: Wikipedia concepts are usually short and with less complicated structure

<table>
<thead>
<tr>
<th></th>
<th>Occurrences in WP links</th>
<th>Coverage</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exact match</strong></td>
<td>923</td>
<td>14.3</td>
<td>“brain” (Report) &amp; “brain” (WP)</td>
</tr>
<tr>
<td><strong>Partial match</strong></td>
<td>1,038</td>
<td>16.1</td>
<td>“infarction” (Report) &amp; “cerebellar infarction” (WP)</td>
</tr>
<tr>
<td><strong>Sub-exact match</strong></td>
<td>5,257</td>
<td>81.6</td>
<td>“acute cerebral infarction” (Report) &amp; “cerebral infarction” (WP)</td>
</tr>
</tbody>
</table>
Link generation revisited

- The observed structural mismatching between the medical anchor texts and Wikipedia anchor texts causes problems.
- Both state-of-the-art systems highly rely on the existing Wikipedia links.
- e.g., keyphraseness equals to 0 when a phrase does not occur in WP anchors.
Our approach part 1: anchor detection

• Exploiting the syntactic regularity of medical anchor texts

• A sequential labeling problem: annotate each word of a report with one of the following labels:
  • Begin-of-anchor (BOA); In-anchor (IA); End-of-anchor (EA); Outside-anchor (OA); Single-word-anchor (SWA)

• Conditional random field models (CRFs) with syntactic features
  • The word itself, its POS tag, its syntactic chunk tag
Our approach part II: target candidate identification

- Exploiting existing Wikipedia links with a sub-anchor based approach

- For a given anchor $a$, we decompose it into a set of sub-sequences $S_a$

white matter disease- {white, matter, disease, white matter, matter disease, white matter disease}

- For each sub-anchor $s_i$, we retrieve top 10 Wikipedia pages as candidates $c$ based on their target probability:

The more often a page is linked to a phrase, the more likely it should be linked to it again.

$$p(c_i|s) = \frac{|L_{s,c_i}|}{\sum_{j=1}^{n} |L_{s,c_j}|}$$
Our approach part III: target detection

• A classification problem: classify each anchor-candidate pair \((a, c)\) as “link” or “non-link”

• Three types of features
  
  • Title matching - Whether a sub-anchor matches the title of the candidate page; weighted by the similarity of the sub-anchor to the original anchor

  \[
  tm(s, c) = f_{tm}(s, c) \frac{len(s)}{len(a)}, \quad f_{tm}(s, c) = \begin{cases} 
  1 & \text{if } s \text{ equals title of } c \\
  0 & \text{otherwise}.
\end{cases}
\]

  • Language model comparison - how likely is the candidate page about neuroradiology?

  \[
  LM-LLR(T) = \log \left( \frac{p(T|\theta_R)}{p(T|\theta_W)} \right) = \sum_{i=1}^{n} \log p(t_i|\theta_R) - \sum_{i=1}^{n} \log p(t_i|\theta_W)
\]

  • Target probability
    • Pre-calculated at candidate identification stage
    • Aggregate from sub-anchor level to anchor level: Max, Min, Avg
Experiment setup

• 3-fold cross-validation

• Classifiers for target detection:
  • SVM, NB and Random Forest*

• A post-processing step for target detection
  • If all candidates are classified as “non-link”, the one with the lowest confidence score is chosen
  • If multiple candidates are classified as “link”, the one with the highest confidence score is chosen
Evaluation

- Anchor detection

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiRa</td>
<td>0.9</td>
<td>0.8</td>
<td>0.85</td>
</tr>
<tr>
<td>Wikify!</td>
<td>0.35</td>
<td>0.16</td>
<td>0.22</td>
</tr>
<tr>
<td>WM</td>
<td>0.35</td>
<td>0.36</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Results of anchor detection
LiRa: system using our proposed approach
Evaluation

• Target finding

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiRa</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>Wikify! (Lesk)</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
<tr>
<td>Wikify! (ML)</td>
<td>0.69</td>
<td>0.69</td>
<td>0.69</td>
</tr>
</tbody>
</table>

Results of target finding for anchors identified by Wikify!

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiRa</td>
<td>0.68</td>
<td>0.68</td>
<td>0.68</td>
</tr>
<tr>
<td>Wikify! (Lesk)</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Wikify! (ML)</td>
<td>0.26</td>
<td>0.26</td>
<td>0.26</td>
</tr>
</tbody>
</table>

Results of target finding for annotated anchors

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiRa</td>
<td>0.89</td>
<td>0.89</td>
<td>0.89</td>
</tr>
<tr>
<td>WM</td>
<td>0.84</td>
<td>0.84</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Results of target finding for annotated anchors
Evaluation

- **Overall performance**

<table>
<thead>
<tr>
<th>System</th>
<th>P</th>
<th>R</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>LiRa</td>
<td>0.65</td>
<td>0.58</td>
<td>0.61</td>
</tr>
<tr>
<td>Wikify! (Lesk)</td>
<td>0.14</td>
<td>0.07</td>
<td>0.09</td>
</tr>
<tr>
<td>Wikify! (ML)</td>
<td>0.25</td>
<td>0.12</td>
<td>0.16</td>
</tr>
<tr>
<td>WM</td>
<td>0.29</td>
<td>0.3</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Impact of anchor frequencies

- Some anchors occur more frequent than others
  - Frequent anchors are likely to be general concepts
  - More likely to occur in Wikipedia
  - Large amount of infrequent anchors, few frequent anchors

<table>
<thead>
<tr>
<th>Top 5</th>
<th>Bottom 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>mass</td>
<td>vestibular nerves</td>
</tr>
<tr>
<td>brain</td>
<td>Virchow-Robin space</td>
</tr>
<tr>
<td>meningioma</td>
<td>Warthin’s tumor</td>
</tr>
<tr>
<td>frontal</td>
<td>Wegner’s granulomatosis</td>
</tr>
<tr>
<td>white matter</td>
<td>xanthogranulomas</td>
</tr>
</tbody>
</table>
Impact of anchor frequencies

How does this influence the performance of linking systems?

<table>
<thead>
<tr>
<th>Group</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq. range</td>
<td>&gt;100</td>
<td>51-100</td>
<td>11-50</td>
<td>6-10</td>
<td>2-5</td>
<td>1</td>
</tr>
<tr>
<td>#Anchors</td>
<td>116</td>
<td>108</td>
<td>527</td>
<td>482</td>
<td>1,399</td>
<td>2,149</td>
</tr>
</tbody>
</table>

![Graph showing recall vs segments of anchor frequency](image1)

![Graph showing recall vs segments of anchor frequency](image2)
Conclusions

• Existing link generation systems trained on general domain corpora do not provide a satisfactory solution to linking radiology reports

• Structural mismatch between medical phrases and Wikipedia concepts is a major problem

• Our proposed approach was shown to be effective

• Frequent anchor texts tend to be “easier” than anchor texts with a low frequency
Conclusions

• Existing link generation systems trained on general domain corpora do not provide a satisfactory solution to linking radiology reports

• Structural mismatch between medical phrases and Wikipedia is a major problem

• Our proposed approach was shown to be effective

• Frequent anchor texts tend to be “easier” than anchor texts with a low frequency

Questions?