Language Independent Answer Prediction from the Web

in

HS Current topics in Information Extraction
PD Dr Günter Neumann

Michael Wirth
Universität des Saarlandes
Fakultät 4.7 Allgemeine Linguistik
miwirth@coli.uni-sb.de
Outline

- Motivation
- WebQA approaches
- Language Answer Prediction formal
- Experiments
- Summary
Motivation

What's the problem?

textQA

natural language query

processing large unstructured document collections

precise answer

webQA

conversion of NL into search engine specific queries

IR via interface to public search engine

extraction of answers

- no NLP-oriented preprocessing possible
- growing multilinguality
A different webQA approach

webQA
conversion of NL into search engine specific queries

IR via interface to public search engine

extraction of answers

webQA
IR via interface to public search engine

find the n-best query results and improve the query

before document retrieval
pseudo relevance feedback

webQA

after document retrieval

extraction of answers
The answer candidate prediction approach
- a PRF like approach

- no preprocessing

- IR via interface to public search engine
  - list of ranked document links
  - snippets

- list of ranked document links
- snippets

- best-n snippets

- m-best ranked predicted answer strings

- the best predicted answer strings for each snippet
Ranking

\[ R = \{(s_1, l_1), (s_2, l_2), \ldots, (s_\sigma, l_\sigma)\} \]

\[ \text{rank} : S \rightarrow L \]

- how to represent documents?
- how to rank sentences?
- how to extract predicted answers?
- how to rank predicted answers?
How to represent a document?

Definition “document“ :

A document is a multi-set of all sentences extracted from the n-best snippets returned by the search engine.

Representation: Vector-space document representation
How to represent a document?

Properties of the vector-space representation

\[ X_{sik} = \begin{cases} 
1 & \text{if the word } w_i \text{ is in the sentence } s_s \text{ at position } k \\
0 & \text{otherwise.} 
\end{cases} \]

\[ \text{freq}(w_i) = \sum_{s=1}^{\sigma} \sum_{k=1}^{\text{len}(s_s)} X_{sik}, \quad \forall w, \quad 1 \leq i \leq \omega \]

\[ \text{freq}(w_i, w_j, \epsilon) = \sum_{s=1}^{\sigma} \sum_{k=\epsilon+1}^{\text{len}(s_s)} X_{si(k-\epsilon)}X_{skj} \]

- \( \sigma \) = number of sentences in the document
- \( \text{len}(s_s) \) = length of sentence \( s \)
- \( \omega \) = size of Dictionary \( W \)
- \( \epsilon \) = absolute distance between \( w_i \) and \( w_j \)
How to represent a document?

Representation of a document

\[ D = \{ \langle w_i, w_j, \epsilon, \text{freq}(w_i, w_j, \epsilon) \rangle, \ \forall i, j, \epsilon, 0 \leq \epsilon \leq \tau \ \land \ \text{freq}(w_i, w_j, \epsilon) > 0 \} \]

\[ T = \text{length of the longest sentence in the document} \]
How to rank an sentence?

We rank a sentence $s_s$ in a document by means of a specially designed matrix $M$.

(Figueroa and Neumann 2006, p.426)

$$M_{ij}(s_s) = \begin{cases} 
freq(w_i, w_j, \varepsilon) & \text{if } i < j; \\
freq(w_j, w_i, \varepsilon) & \text{if } i > j; \\
0 & \text{otherwise.}
\end{cases}$$

Filtering Rule:

$$\forall i, j \ M_{ij} \leq \zeta \Rightarrow M_{ij} = 0$$

In this formula $i$ and $j$ are indices of the position of word $w$ in sentence $s$.

$\zeta = \text{empirical determined threshold}$
How to rank an sentence?

**Definition:** Rank of a sentence $s_s$

$$\text{rank}(s_s) = \lambda_{\text{max}}(M(s_s))$$

$\lambda_{\text{max}}(M(s_s))$ = the greatest eigenvalue of Matrix M constructed from sentence $s_s$

1. $\forall i \ M_{ii} = 0 \Rightarrow \sum_{\forall i} M_{ii} = 0 \Rightarrow \sum_{\forall f} \lambda_f = 0$.
2. $\forall i, j \ M_{ij} = M_{ji}$, the spectral theorem implies that $\forall f \ \lambda_f \in \mathbb{R}$, and all the eigenvectors are orthogonal.
How to extract predicted answers?

Algorithm 1: extractPredictedAnswers

\begin{algorithm}
\begin{algorithmic}[1]
  \Require $M, s_s$
  \Begin 
  \State predictedAnswers = $s_s$;
  \State \If {$\text{numberOfWords}(w_i) > 3$} \Then 
  \ForAll {$w_i \in s_s$} 
  \State flag = true;
  \ForAll {$w_j \in s_s$} 
  \State \If {$M_{ij} > 0$} \Then flag = false; \End 
  \End
  \If {flag} \Then replace $w_i$ with "*"; \End 
  \End
  \State predictedAnswers = split($s_s,$"*"$);
  \End
  \Return predictedAnswers;
\End
\end{algorithmic}
\end{algorithm}
How to rank predicted answers?

1) 

\[ \text{rank}(v) = \text{rank}(s_s) \times \sum_{b=2}^{\beta} P(B_b | B_{b-1}) \]

\[ P(B_b | B_{b-1}) = \frac{\log(freq(B_{b-1}, B_b, 1))}{\log(freq(B_{b-1}))} \]

\[ \nu = \text{predicted answer} \]

\[ B_b = \text{the words in } \nu \]

\[ \beta = \text{length of } \nu \]

2) Redundant predicted answers are removed

Definition redundant predicted answer:

1. If there exists another predicted answer \( \nu' \), such that \( \text{rank}(\nu) < \text{rank}(\nu') \).
2. If \( \nu \) is a substring of \( \nu' \).
Answer Extraction

When- Answer Extraction
Who- Answer Extraction
Where- Answer Extraction

main task: removing predicted answer candidates with the wrong expected answer type (EAT)

Table 1. Some sample Wh-question keywords for the covered languages

<table>
<thead>
<tr>
<th></th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Wann, When, Cuándo, Qué año, Welchem Jahr, Que ano</td>
</tr>
<tr>
<td>Location</td>
<td>Wo, Where, Dónde, Onde</td>
</tr>
<tr>
<td>Person</td>
<td>Wer, Who, Quién, Quem</td>
</tr>
</tbody>
</table>
Experiments

Procedure:

- sending an unmodified question to Google
- extract the first 30 snippets
- normalize each snippet (removing HTML encoding/ upercasing text)
- extract correct answers via the presented system

Test set: 889 questions in four languages (EN, DE, ES, PT) taken as Q/A pairs from the multilingual CLEF 2004 corpus
Experiments

<table>
<thead>
<tr>
<th>CA</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
</table>

CA : correct answer (exact answer, inexact answer)

MRR: mean reciprocal rank

NAG: no answer in the snippets; the system returns NIL

WAG: no answer in the snippets; the system gave three wrong answers

NAF: no answer found, although there is one in the snippets

1,2,3: the three best returned candidates
Experiments

Results

Table 2. Results for each question type over all languages

<table>
<thead>
<tr>
<th>CA</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEN</td>
<td>218</td>
<td>0.60</td>
<td>25.11</td>
<td>10.96</td>
<td>21.46</td>
<td>35.16</td>
<td>5.02</td>
</tr>
<tr>
<td>WHERE</td>
<td>232</td>
<td>0.57</td>
<td>10.77</td>
<td>24.14</td>
<td>20.68</td>
<td>30.60</td>
<td>9.91</td>
</tr>
<tr>
<td>WHO</td>
<td>439</td>
<td>0.38</td>
<td>11.39</td>
<td>27.56</td>
<td>32.57</td>
<td>18.90</td>
<td>6.83</td>
</tr>
</tbody>
</table>

combined MRR = 0.52

Table 3. Distribution of answer candidates (all languages)

<table>
<thead>
<tr>
<th>CA</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHEN</td>
<td>33.82</td>
<td>55.42</td>
<td>7.91</td>
<td>2.84</td>
</tr>
<tr>
<td>WHERE</td>
<td>31.86</td>
<td>47.00</td>
<td>15.23</td>
<td>5.95</td>
</tr>
<tr>
<td>WHO</td>
<td>53.37</td>
<td>30.97</td>
<td>11.19</td>
<td>4.47</td>
</tr>
</tbody>
</table>

(Figueroa and Neumann 2006, p. 431)
Experiments

Results

<table>
<thead>
<tr>
<th></th>
<th>CA(EN)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>when</td>
<td>69</td>
<td>0.69</td>
<td>15.69</td>
<td>15.69</td>
<td>17.65</td>
<td>45.10</td>
<td>3.92</td>
<td>1.96</td>
</tr>
<tr>
<td>where</td>
<td>64</td>
<td>0.74</td>
<td>7.81</td>
<td>12.5</td>
<td>15.62</td>
<td>53.12</td>
<td>10.93</td>
<td>0</td>
</tr>
<tr>
<td>who</td>
<td>148</td>
<td>0.50</td>
<td>7.43</td>
<td>12.83</td>
<td>32.43</td>
<td>33.78</td>
<td>10.14</td>
<td>3.38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CA(DE)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wann</td>
<td>58</td>
<td>0.45</td>
<td>36.20</td>
<td>12.07</td>
<td>27.59</td>
<td>22.03</td>
<td>1.17</td>
<td>0</td>
</tr>
<tr>
<td>Wo</td>
<td>58</td>
<td>0.46</td>
<td>9.37</td>
<td>18.75</td>
<td>23.43</td>
<td>20.31</td>
<td>12.5</td>
<td>6.25</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CA(ES)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuándo</td>
<td>59</td>
<td>0.55</td>
<td>16.64</td>
<td>11.86</td>
<td>23.73</td>
<td>32.20</td>
<td>10.17</td>
<td>11.86</td>
</tr>
<tr>
<td>Dónde</td>
<td>63</td>
<td>0.59</td>
<td>10.93</td>
<td>31.25</td>
<td>15.62</td>
<td>26.56</td>
<td>10.93</td>
<td>3.21</td>
</tr>
<tr>
<td>Quién</td>
<td>86</td>
<td>0.27</td>
<td>9.65</td>
<td>40.68</td>
<td>28.96</td>
<td>11.72</td>
<td>6.21</td>
<td>2.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CA(PT)</th>
<th>Total MRR</th>
<th>NAG(%)</th>
<th>WAG(%)</th>
<th>NAF(%)</th>
<th>1(%)</th>
<th>2(%)</th>
<th>3(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quando</td>
<td>56</td>
<td>0.04</td>
<td>30.76</td>
<td>12.30</td>
<td>42.45</td>
<td>3.08</td>
<td>1.54</td>
<td>0</td>
</tr>
<tr>
<td>Onde</td>
<td>47</td>
<td>0.18</td>
<td>10.93</td>
<td>25</td>
<td>20.31</td>
<td>10.93</td>
<td>1.56</td>
<td>4.68</td>
</tr>
<tr>
<td>Quem</td>
<td>146</td>
<td>0.14</td>
<td>17.12</td>
<td>29.45</td>
<td>36.30</td>
<td>10.95</td>
<td>4.11</td>
<td>2.05</td>
</tr>
</tbody>
</table>
Experiments

Results

• system performance comparable to the sixth highest score at each TREC 9-12

• different behavior of the answer prediction strategy among the different question types and languages
Experiments

Results

• Stronger relations occur much fewer than weaker relations

<table>
<thead>
<tr>
<th>Frequency</th>
<th>$\tilde{G}(v)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.999925</td>
</tr>
<tr>
<td>1</td>
<td>0.000685</td>
</tr>
<tr>
<td>2</td>
<td>0.00005</td>
</tr>
<tr>
<td>3 and more</td>
<td>0.000015</td>
</tr>
</tbody>
</table>
Questions ?
Literature


Thank you!