Contextualized PACRR for Complex Answer Retrieval

Sean MacAvaney\textsuperscript{1}\* Andrew Yates\textsuperscript{2} Kai Hui\textsuperscript{2}

\textsuperscript{1} IRLab, Georgetown University
\textsuperscript{2} Max Planck Institute for Informatics

\* Joint work with IRLab and MPII
Contribution: Contextualized PACRR

• We start with the PACRR neural ranking architecture (Hui et al. 2017)
• Include task-specific contextual vectors

PACRR overview

1. Retrieve candidate documents to re-rank using basic unsupervised ranking approach (e.g. BM25)

Query: green sea turtle ecology and behavior life cycle

Results:
1. As one of the first sea turtle species studied much of what is known of sea turtle ecology...
2. Reptiles such as sea turtles spend most of their lives in the ocean. However their life cycle...
3. Jesús A. Rivas (born in Caracas Venezuela) is a Venezuelan herpetologist tropical ecologist...
4. As of 2015 Mote employs over 200 staff members conducting research on 25 different...
5. Galápagos green turtles lifestyle is similar to other populations of Chelonians. The behavior...
6. Within the sea turtles E. imbricata has several unique anatomical and ecological traits. It is...
PACRR overview

2. Build similarity matrix for each document

\[ \text{sim}_{|q| \times |d|} \]

<table>
<thead>
<tr>
<th>green</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>turtle</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

reptiles such as sea turtles spend most of ...
PACRR overview

3. Convolution over similarity matrix to find matching patterns (e.g. ngrams)
3. Convolution over similarity matrix to find matching patterns (e.g. ngrams)
PACRR overview

4. Perform max pooling over filters
PACRR overview

4. Perform max pooling over filters

<table>
<thead>
<tr>
<th>green</th>
<th>sea</th>
<th>turtle ...</th>
<th>reptiles such</th>
<th>as</th>
<th>sea</th>
<th>turtles spend</th>
<th>most</th>
<th>of</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>f1</td>
<td>f2</td>
<td>f3</td>
<td>f4</td>
<td>f5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PACRR overview

5. Perform k-max pooling over query terms
PACRR overview

5. Perform k-max pooling over query terms

```
green
sea
turtle
...```
PACRR overview

6. Concatenation with query term contextual vectors
7. Combination (dense layers) to produce relevance score
Challenge: Discourse context

Well-written articles will often avoid overusing terms within a document because the context is clear.
Challenge: Query term utility

Some query components exhibit a general structure.

Other query components exhibit a specificity to the particular topic.

Who are the main characters in *The Lord of the Rings*?

What influences did *The Lord of the Rings* have on the fantasy genre?

Who are the main characters in *House*?
Heading position vectors

• Indicator of whether query term came from heading, intermediate, or main heading

<table>
<thead>
<tr>
<th></th>
<th>green</th>
<th>sea</th>
<th>turtle</th>
<th>ecology</th>
<th>and</th>
<th>behavior</th>
<th>life</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>title</td>
<td>1</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>intermediate</td>
<td>0</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>main</td>
<td>0</td>
<td></td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

• Intuition: Title is important to keep intact
Our approach performs well
Heading usage frequency

Content; does not occur frequently in training data

Structural; occurs frequently in training data
Heading usage frequency vector

• Calculate how frequently each heading occurs in training data
• Stratify by percentile (60\textsuperscript{th}=1, 90\textsuperscript{th}=2; 99\textsuperscript{th}=3)

<table>
<thead>
<tr>
<th>green</th>
<th>sea</th>
<th>turtle</th>
<th>ecology</th>
<th>and</th>
<th>behavior</th>
<th>life</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

• Intuition: **Common headings are less likely to match verbatim**
Heading usage frequency performs well, too
Term occurrence vector

• For each term in each heading, calculate the frequency that it appears in relevant paragraphs

<table>
<thead>
<tr>
<th></th>
<th>green</th>
<th>sea</th>
<th>turtle</th>
<th>ecology</th>
<th>and</th>
<th>behavior</th>
<th>life</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

• Intuition: Inform the model directly about which terms are likely to match
Term occurrence vector - expansion

• Since the PACRR model uses word similarity scores (not exact term matches), find words that are similar and include these as matches as well
• Perform relevance feedback on most common headings (e.g. history, life cycle, etc.)

<table>
<thead>
<tr>
<th></th>
<th>green</th>
<th>sea</th>
<th>turtle</th>
<th>ecology</th>
<th>and</th>
<th>behavior</th>
<th>life</th>
<th>cycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without expansion</td>
<td>0.6</td>
<td>0.5</td>
<td>0.6</td>
<td>0.1</td>
<td>0.8</td>
<td>0.2</td>
<td>0.1</td>
<td>0.3</td>
</tr>
<tr>
<td>With expansion</td>
<td>0.6</td>
<td>0.5</td>
<td>0.7</td>
<td>0.1</td>
<td>0.8</td>
<td>0.3</td>
<td>0.9</td>
<td>0.3</td>
</tr>
</tbody>
</table>
Expanded term occurrence is better for lenient
Antioxidant » Health effects » Relation to diet

As with the minerals discussed above, some vitamins are recognized as organic essential nutrients necessary in the diet for good health... Moreover, thousands of different phytochemicals have recently been discovered in food (particularly in fresh vegetables) which may have desirable properties including antioxidant activity...
Failure case – No mention of diet

**Query:** Antioxidant » Health effects » Relation to diet

According to a March 2015 Scientific American Special Report on Aging article laboratory mice at a University of Washington laboratory who produced more catalase which is an antioxidant lived longer. Research on the topic both supports and cautions against the benefit of antioxidants for health effects on aging.
Positive case – Patterns look similar

**Antioxidant** » Health effects » Relation to diet

...As a result of animal experimental studies antioxidant and anti-inflammation are expected to be effective countermeasures for CNS risks from space radiation. Diets of blueberries and strawberries were shown to reduce CNS risks after heavy-ion exposure. Estimating the effects of diet and nutritional supplementation will be a primary goal of CNS research on countermeasures...

<table>
<thead>
<tr>
<th></th>
<th>Position</th>
<th>Pos. + Heading Usage</th>
<th>Pos. + Term Occ.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank</td>
<td>19</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
All three configurations outperform “vanilla” PACRR

(Vanilla PACRR was not an official run. These values are based on training automatic judgments.)
Summary

• Adding contextual vectors to the PACRR architecture performs well on CAR

• Simply informing the model of the information hierarchy does well. Adding more specialized contextual vectors doesn’t help much.

Special thanks to the members of the IRLab for constructive comments.