Weighting of Passages in Question Answering

An Empirical Evaluation of the Godwin’s Law

Vít Novotný, witiko@mail.muni.cz
Petr Sojka, sojka@fi.muni.cz

December 7, 2018
Introduction

The vector space model (tf-idf) is well-understood and scalable, let’s keep it.

\[ d_1 \text{ Player with 350 or more goals in a domestic league?} \]

\[ \text{football} \quad \text{coding} \quad \text{cooking} \]
Introduction

The vector space model (tf-idf) is well-understood and scalable, let’s keep it.

\[ d_1 \text{ Player with 350 or more goals in a domestic league?} \]

\[ d_2 \text{ Why is using pixels as a unit of measure bad in web design?} \]
The vector space model (tf-idf) is well-understood and scalable, let’s keep it.

- **d₁** Player with 350 or more goals in a domestic league?
- **d₂** Why is using pixels as a unit of measure bad in web design?
- **d₃** Can I use red wine instead of white wine for chicken scallops?
The vector space model (tf-idf) is well-understood and scalable, let’s keep it.

- **d₁** Player with 350 or more goals in a domestic league?
- **d₂** Why is using pixels as a unit of measure bad in web design?
- **d₃** Can I use red wine instead of white wine for chicken scallops?
The vector space model (tf-idf) is well-understood and scalable, let’s keep it.

Problem: Long documents that range many topics are almost never retrieved!

Novotný and Sojka · Weighting of Passages in Question Answering · December 7, 2018
Introduction

- The vector space model (tf-idf) is well-understood and scalable, let’s keep it.
- Problem: Long documents that range many topics are almost never retrieved!
- Solution: Segmentation to semantically coherent passages.

But what if we need to retrieve full documents?
Our system **segments both queries and indexed documents.**

<table>
<thead>
<tr>
<th>$u_1$</th>
<th>I did enact Julius Caesar</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_2$</td>
<td>I was killed in the Capitol</td>
</tr>
<tr>
<td>$u_3$</td>
<td>Brutus killed me</td>
</tr>
<tr>
<td>$v_1$</td>
<td>So let it be with Caesar</td>
</tr>
<tr>
<td>$v_2$</td>
<td>The noble Brutus hath told you</td>
</tr>
<tr>
<td>$v_3$</td>
<td>Caesar was ambitious</td>
</tr>
</tbody>
</table>
Our system segments both queries and indexed documents. At query time, we construct a matrix $M_{uv}$ of similarities between the segments of the query $u$ (rows) and the segments of a document $v$ (columns).

<table>
<thead>
<tr>
<th>$u_1$</th>
<th>I did enact Julius Caesar</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_2$</td>
<td>I was killed i’ the Capitol</td>
</tr>
<tr>
<td>$u_3$</td>
<td>Brutus killed me</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$v_1$</th>
<th>So let it be with Caesar</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_2$</td>
<td>The noble Brutus hath told you</td>
</tr>
<tr>
<td>$v_3$</td>
<td>Caesar was ambitious</td>
</tr>
</tbody>
</table>

$$\implies M_{uv} = \begin{pmatrix} 0.18 & 0 & 0.26 \\ 0 & 0.16 & 0.24 \\ 0 & 0.24 & 0 \end{pmatrix}$$
Our system segments both queries and indexed documents. At query time, we construct a matrix $M_{uv}$ of similarities between the segments of the query $u$ (rows) and the segments of a document $v$ (columns). Then, we reduce $M_{uv}$ to an aggregate similarity.

\[
M_{uv} = \begin{pmatrix}
0.18 & 0 & 0.26 \\
0 & 0.16 & 0.24 \\
0 & 0.24 & 0 \\
\end{pmatrix}
\leadsto \quad \bigcirc_k \bigcirc_l m_{kl} = 0.205
\]
Our system segments both queries and indexed documents. At query time, we construct a matrix $M_{uv}$ of similarities between the segments of the query $u$ (rows) and the segments of a document $v$ (columns). Then, we reduce $M_{uv}$ to an aggregate similarity.

The devil is in the detail. How exactly do we aggregate $M_{uv}$?
Datasets

- We evaluated our system on SemEval 2016 and 2017 question answering datasets.
Datasets

- We evaluated our system on SemEval 2016 and 2017 question answering datasets.
- We focused on the following two subtasks of the SemEval question answering task:
  
  A) Given a question, rank the first ten comments by relevance.
Datasets

- We evaluated our system on [SemEval](https://www.aclweb.org/anthology/S16-1047) 2016 and 2017 question answering datasets.
- We focused on the following two subtasks of the SemEval question answering task:
  A) Given a question, rank the first ten comments by relevance.
  B) Given an original question, rank ten related questions by relevance.
Datasets

- We evaluated our system on SemEval 2016 and 2017 question answering datasets.
- We focused on the following two subtasks of the SemEval question answering task:
  A) Given a question, rank the first ten comments by relevance.
  B) Given an original question, rank ten related questions by relevance.
- Subtask A comes with a training dataset of 2,654 questions. We analyzed these to learn how to aggregate $M_{UV}$. 
Datasets

- We evaluated our system on SemEval 2016 and 2017 question answering datasets.
- We focused on the following two subtasks of the SemEval question answering task:
  
  A) Given a question, rank the first ten comments by relevance.
  
  B) Given an original question, rank ten related questions by relevance.

- Subtask A comes with a training dataset of 2,654 questions. We analyzed these to learn how to aggregate $M_{UV}$.

- Subtask B comes with a validation dataset of 50 original questions and 244 related questions (2016), and 88 original questions and 239 related questions (2017). We used these to evaluate our system. Question text and comments are the segments.
Analysis

- In 1991, Mike Godwin posited that later comments are likely to be less relevant.

![Graph showing the probability of comment position](image)
Analysis

- In 1991, Mike Godwin posited that later comments are likely to be less relevant.
- Godwin’s rule applies to the subtask A dataset with statistical significance.
In 1991, Mike Godwin posited that later comments are likely to be less relevant. Godwin’s rule applies to the subtask A dataset with statistical significance. We aggregate $M_{uv}$ using weighted average, where the weight of a segment is proportional to the inverse of the segment’s position. We also evaluated weighting tokens in unsegmented documents.
## Results

- Our system (primary) is **on-par with winners** of SemEval 2016 and 2017.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Segm.</th>
<th>Text summ.</th>
<th>S. f. S</th>
<th>Aggregate s. f. $S'$</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td>Yes</td>
<td></td>
<td>$bfx.tfx$</td>
<td>$\oplus = \text{wavg}_{\text{length}}$</td>
<td><strong>76.77</strong></td>
</tr>
<tr>
<td><strong>SemEval-2016 task 3 subtask B winner (UH-PRHLT-primary)</strong></td>
<td></td>
<td></td>
<td>$bfx.tfx$</td>
<td>$\ominus = \text{wavg}_{\text{Godwin}}$</td>
<td><strong>76.70</strong></td>
</tr>
<tr>
<td>Third contrastive</td>
<td>No</td>
<td>FirstTwoPara</td>
<td>$bfx.tfx$</td>
<td></td>
<td><strong>75.21</strong></td>
</tr>
<tr>
<td><strong>SemEval-2016 task 3 subtask B IR baseline</strong></td>
<td></td>
<td></td>
<td>$bfx.tfx$</td>
<td></td>
<td><strong>74.75</strong></td>
</tr>
<tr>
<td>First contrastive</td>
<td>No</td>
<td></td>
<td>$bfx.tfx$, Godwin</td>
<td></td>
<td><strong>73.94</strong></td>
</tr>
<tr>
<td>Second contrastive</td>
<td>No</td>
<td></td>
<td>$bfx.tfx$, Godwin</td>
<td></td>
<td><strong>70.28</strong></td>
</tr>
</tbody>
</table>
Results

Our system (primary) is on-par with winners of SemEval 2016 and 2017.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Segm.</th>
<th>Text summ.</th>
<th>S. f. S</th>
<th>Aggregate s. f. S’</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>Yes</td>
<td></td>
<td>bfx.tfx</td>
<td></td>
<td>47.45</td>
</tr>
<tr>
<td><strong>SemEval-2017 task 3 subtask B winner (SimBow-primary)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.22</td>
</tr>
<tr>
<td>Third contrastive</td>
<td>No</td>
<td>FirstTwoPara</td>
<td>bfx.tfx</td>
<td></td>
<td>44.67</td>
</tr>
<tr>
<td><strong>SemEval-2017 task 3 subtask B IR baseline</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.85</td>
</tr>
<tr>
<td>Second contrastive</td>
<td>No</td>
<td></td>
<td>bfx.tfx,</td>
<td></td>
<td>37.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Godwin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First contrastive</td>
<td>No</td>
<td></td>
<td>bfx.tfx</td>
<td></td>
<td>36.82</td>
</tr>
</tbody>
</table>

Novotný and Sojka • Weighting of Passages in Question Answering • December 7, 2018
**Results**

- Our system (primary) is *on-par with winners* of SemEval 2016 and 2017.
- *Weighting tokens* (second contrastive) does *worse than baseline*, which shows that *segments* are the *correct level of analysis*.

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Segm.</th>
<th>Text summ.</th>
<th>S. f. S</th>
<th>Aggregate s. f. S’</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary</strong></td>
<td>Yes</td>
<td></td>
<td>bfx.tfX</td>
<td></td>
<td>47.45</td>
</tr>
<tr>
<td><em>SemEval-2017 task 3 subtask B winner (SimBow-primary)</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>47.22</td>
</tr>
<tr>
<td>Third contrastive</td>
<td>No</td>
<td>FirstTwoPara</td>
<td>bfx.tfX</td>
<td></td>
<td>44.67</td>
</tr>
<tr>
<td><em>SemEval-2017 task 3 subtask B IR baseline</em></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.85</td>
</tr>
<tr>
<td>Second contrastive</td>
<td>No</td>
<td></td>
<td>bfx.tfX,</td>
<td></td>
<td>37.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Godwin</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First contrastive</td>
<td>No</td>
<td></td>
<td>bfx.tfX</td>
<td></td>
<td>36.82</td>
</tr>
</tbody>
</table>
Conclusion

We have accomplished the following:

- We showed that **Godwin’s rule applies** in question answering.
Conclusion

We have accomplished the following:

- We showed that Godwin’s rule applies in question answering.
- We achieved state-of-the-art results on a SemEval task with no semantic modeling.
Conclusion

We have accomplished the following:

- We showed that Godwin’s rule applies in question answering.
- We achieved state-of-the-art results on a SemEval task with no semantic modeling.

In our future work, we will focus on the following:

- We will test the hypothesis on other datasets.
We have accomplished the following:

- We showed that Godwin’s rule applies in question answering.
- We achieved state-of-the-art results on a SemEval task with no semantic modeling.

In our future work, we will focus on the following:

- We will test the hypothesis on other datasets.
- We will investigate how this generalizes outside question answering.
Thank you for your attention!