1. Introduction

The problem of representing document in short form is very important in different areas including news aggregation and providing search results. Last years text summarization algorithms have been significantly improved due to applying of novel learning techniques. The scope of the project includes applying various RNN-LSTM based summarization approaches to the text and prepare summaries for text documents.

2. Dataset description

We use 2 different datasets with very different structures of text and headlines:

1) Amazon food reviews

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mean article length (words)</th>
<th>Mean summary length (words)</th>
<th>Total vocabulary size (words)</th>
<th>Occuring 10+ times (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>680,654</td>
<td>40</td>
<td>32,384</td>
<td>64,183</td>
</tr>
</tbody>
</table>

2) Cornell NEWSROOM Summarization dataset

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Mean article length (words)</th>
<th>Mean summary length (words)</th>
<th>Total vocabulary size (words)</th>
<th>Occuring 10+ times (words)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cornell</td>
<td>1,201,995</td>
<td>655.6</td>
<td>62,937</td>
<td>78,488</td>
</tr>
</tbody>
</table>

While Amazon food reviews is a much more simple data set, it allows to make quick verification of the ideas for different models. Text preprocessing:

1) convert text to lowercase
2) tokenization and separate punctuation from words
3) English language contractions (don’t -> do not)
4) Remove infrequent words (less than 10 per text corpus)
5) Use <UNK> instead of unknown words and add <EOS> to the end of string

4. Results

We trained the model (using training set) and defined optimal hyper parameters for each data set (using the dev set).

The results were evaluated using test set.

<table>
<thead>
<tr>
<th>Method</th>
<th>ROUGE 1 F1</th>
<th>ROUGE 2 F1</th>
<th>ROUGE 3 F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>20</td>
<td>17</td>
<td>13</td>
</tr>
<tr>
<td>Cornell</td>
<td>14</td>
<td>11</td>
<td>7</td>
</tr>
</tbody>
</table>

Example:

SUMMARY: A young woman has been arrested after allegedly glassing another woman during a wild brawl at a Sydney train station.

ARTICLE: A wild brawl between two women at a Sydney train station has left one with head injuries after she was struck with a glass bottle. The women were fighting at Redfern train station just before 4am on Friday before police used pepper spray to break them up. One of the women, aged 27, was taken to hospital after being hit by a bottle to her head...

5. Conclusions

In presented project we trained an encoder-decoder neural network with LSTM units and attention for text summarization problem. We used 2 completely different data sets and got much better result on data set with short summaries and much more simple language structure, which we can expect a proxy. Approach demonstrated the feasibility and definitely it will be interesting to apply for this problem the latest techniques which were proposed in the latest articles.

We observed that text summarization is very sensitive to the text cleaning method and vocabulary. It was unexpected that we didn’t get significant dependency on embedding dimension (we compare 50 and 300 sin G2Ve embeddings).

References