Improved Topic-based Semantic Title Evaluation and Recommendation techniques and systems

Xiang Zhang, u5199740@anu.edu.au
Supervisor: Dr. Warren Jin

Comp8780 Information and Human centered computing project
Department of Computer Science
Australian National University
outline

• Background
• Motivations
• Challenges
• What we did and achieved
• Conclusions
• Future researches
• This thesis is mainly developing, revisiting and experimenting the previous work - "Semantic Title Evaluation and Recommendation Based on Topic Models" by Warren Jin et al 2013.

• Background:
• Basically, the previous work introduced a system that is able to do semantic title evaluation (STE) and semantic title recommendation (STR).
• The system can work with either topic model - latent dirichlet allocation (LDA) or segmented topic model (STM). So, STESTM, STELDA, STRSTM and STRLDA.
• What does topic model do? – generate latent topics, which is a probability distribution over words. Document/sentence/title presented as a mixture of latent topics.
• Difference between LDA & STM?

Figure 2.2: Graphical model representation of LDA

Figure 2.3: Graphical model representation of STM
• **Motivations:**
• Title evaluation and recommendation are real needs: how many of you have been attracted by a title, and then find out the content is so irrelevant?

an autobiography for Victoria Beckham                      a piloting & flight instruction

• The system in previous work has made a big step forward: it does not require human reference summary for title evaluation and also considers polysemy and synonymy.
• Challenges:
  • Getting familiar with the system in previous work
  • Trying to develop a new method for the STE part of system. The semantic similarity and GEV steps are replaced by a Chi-square goodness of fit test. Also a equation is empirically developed to complement the hypothesis test.
  • How to prove that the new method of STE is valid? A number of measures are used to suggest the new method is valid at some level.
  • A series of experiments on different parameter settings of new STE and STR were conducted. Tons of data (about 10 GB in total) from each of the intermediate stages need to be managed.
• What we did and achieved

• **For STR part:** (STR’s performance is evaluated by ROUGE (used for evaluating auto-summarization techniques), in terms of precision, recall and F-measure)

• **STR experiment 1** varied the number of topics parameter in topic model training. K=30, 50, 80, 110, 140, 170, 200.

• Finding: there is no real impact on the quality of recommended sentence. STRSTM and STRLDA reacted similarly to the variation
• What we did and achieved
• **STR experiment 2**: varied pre-processing parameter settings. Condition 1, 2 and 3 varied the number of top common words. Condition 4, 5 and 6 varied the minimum document frequency and minimum sentence frequency.
• Finding: no real impact. STRSTM and STRLDA reacted similarly.
• What we did and achieved

• **STR experiment 3**: varying the size of corpus used in topic model training. Three conditions, single document, 30 documents, 487 documents.

• Finding: STRSTM performed better than STRLDA with the variation. For STRSTM, P-value from T test between performances of single and 30 is **0.56**. Between single and 487 is **0.13**. For STRLDA, P-value from T test between performances of single and 30 is **0.016**. Between single and 487 is **0.0002**.

<table>
<thead>
<tr>
<th>condition</th>
<th>Rouge-1 F</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRLDA</td>
<td></td>
</tr>
<tr>
<td>singleSTRLDA</td>
<td>0.1068623</td>
</tr>
<tr>
<td>group30STRLDA</td>
<td>0.1238643</td>
</tr>
<tr>
<td>group487STRLDA</td>
<td>0.165328</td>
</tr>
<tr>
<td>STRSTM</td>
<td></td>
</tr>
<tr>
<td>singleSTRSTM</td>
<td>0.202821</td>
</tr>
<tr>
<td>group30STRSTM</td>
<td>0.1951953</td>
</tr>
<tr>
<td>group487STRSTM</td>
<td>0.1810087</td>
</tr>
</tbody>
</table>

*Table 4.6*: The average Rouge-1 F-measure score of different conditions (F for F measure, these measures are averages over the 30 documents)
• What we did and achieved
• Comparing STR’s performance with other methods participated in DUC-2002.

• Finding: STR is quite competitive.
  In terms of precision, LDA is better than STM. In terms of F-measure, STM is better.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-L precision scores</th>
<th>ROUGE-1 precision scores</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STRSTM</td>
<td>STRLDA</td>
</tr>
<tr>
<td>Barzilay</td>
<td>0.309</td>
<td>0.356</td>
</tr>
<tr>
<td>Ercan</td>
<td>0.285</td>
<td>0.365</td>
</tr>
<tr>
<td>GongLiu</td>
<td>0.356</td>
<td>0.430</td>
</tr>
<tr>
<td>SteinbergerJezek</td>
<td>0.307</td>
<td>0.356</td>
</tr>
<tr>
<td>Murray et al.</td>
<td>0.347</td>
<td>0.347</td>
</tr>
<tr>
<td>Cross</td>
<td>0.333</td>
<td>0.333</td>
</tr>
<tr>
<td>Topic</td>
<td>0.347</td>
<td>0.347</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GL</th>
<th>SJ</th>
<th>MRC</th>
<th>Cross</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>STRLDA</td>
<td>0.094</td>
<td>0.098</td>
<td>0.109</td>
<td>0.177</td>
<td>0.108</td>
</tr>
<tr>
<td>STRSTBM</td>
<td>0.103</td>
<td>0.104</td>
<td>0.176</td>
<td>0.104</td>
<td>0.104</td>
</tr>
<tr>
<td>freq.</td>
<td>0.234</td>
<td>0.224</td>
<td>0.230</td>
<td>0.196</td>
<td>0.230</td>
</tr>
<tr>
<td>binary</td>
<td>0.149</td>
<td>0.172</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tf-idf</td>
<td>0.093</td>
<td>0.095</td>
<td>0.098</td>
<td>0.179</td>
<td>0.097</td>
</tr>
<tr>
<td>logent.</td>
<td>0.103</td>
<td>0.098</td>
<td>0.104</td>
<td>0.176</td>
<td>0.104</td>
</tr>
<tr>
<td>root</td>
<td>0.103</td>
<td>0.098</td>
<td>0.104</td>
<td>0.176</td>
<td>0.104</td>
</tr>
<tr>
<td>mod.tf-idf</td>
<td>0.094</td>
<td>0.098</td>
<td>0.119</td>
<td>0.177</td>
<td>0.118</td>
</tr>
</tbody>
</table>

Table 4.7: Comparison of Rouge-L F-measure scores produced by different techniques with different way of creating input matrix on duc2002 task 1 dataset
• What we did and achieved

• For STE part:

• In old STE, the GEV stage cannot guarantee to properly fit the semantic similarity values every time. Hence, the new STE replaced the semantic similarity and GEV stages with goodness of fit test directly comparing sentence/title to document.

• A equation for calculating sample size was developed to complement the test based on empirical analysis.

\[ n = \log(nd \times ns) \times \log(nt) \]

- nd: number of words in document
- ns: number of words in the sentence
- nt: number of topics set in topic model training

why log(nd \times ns) ?
why log(nt)?
• What we did and achieved

• **STE experiment 1**: used STM and two versions of LDA as topic model in STE. LDA1 produces output in two phase, sentence latent topics in model training, document latent topics in model inferring/testing. LDA2 produces output in model training phase.

• Finding: Only STESTM produced sensible evaluations on technical documents’ title. As for technical documents, their titles ought to be descriptive.

![Diagram showing p-values for title evaluation by STELDA1, STELDA2, and STESTM](image-url)

*Figure 6.2:* Overview of the p-values produced for title evaluation by STELDA1, STELDA2, and STESTM
• What we did and achieved
  
  **STE experiment 2**: tried to prove the validity of the new STE by transitivity in logic. Linking new STE to STR, as STR is already proved to be competitive among many other methods. A correlation test is also conducted for new STE and old STE.

  
  **Findings:**
  
  (1) great overlap between the performances of STR using cosine similarity and Chi-square distance.
• What we did and achieved

(2) A level of overlap found between cosine similarity and p-value. When the cosine similarity for a title is low, then the p-value for this title is always low.

(3) the correlation coefficient from the correlation test carried between the p-values produced by new STE and p-values produced by old STE, is 0.675
• What we did and achieved
• **STE experiment 3**: STESSTM with different size of the corpus used in topic model training. Single document, 21 document and 99 document training conditions.
• Findings: some titles are consistently rated as unfavorable across conditions.
• **What we did and achieved**

  • **STE experiment 4**: editing on the unfavorable title by using the system’s byproduct. Step 1, comparing sentence level topic vector with document level topic vector. Step 2, mark the dimensions where the difference between these the vectors is greater than 0.1. Step 3, make some human decisions to choose some words from the top 50 words list of the dimensions. Step 4, add or remove the words accordingly.

  • Finding: The titles are made better. Also, the improvements are quantified by the p-values produced for the new title.

  • Example, “web page performance scoring” -> “web page scoring by collecting statistics”.

Example’s abstract:

A browser-based tool is provided that loads a Webpage, accesses the document object model (DOM) of the page, collects information about the page structure and parses the page, determines through the use of heuristics such factors as how much text is found on the page and the like, produces statistical breakdown of the page, and calculates a score based on performance of the page. Key to the operation of the invention is the ability to observe operation of the Webpage as it actually loads in real time, scoring the page for several of various performance factors, and producing a combined score for the various factors.
• Conclusions:

• The newly developed STE works, suggested by experimental results.

• STM is a better candidate of topic model for the system LDA. Only in terms of precision that LDA is better than STM.

• When using STM, the system is able to work well with single document used in topic model training.

• STM and LDA reacted similarly to most of the parameter settings.
• **Future researches:**

  • More experiments should be conducted for STR and STE with single document training, to prove its validity and more importantly its consistency. From my point of view, STR with single document training is quite promising. Moreover, STE with different number of documents training may be the key to true title evaluation, not just sensible any more.

  • All the experimental results suggest that the equation of sample size in new STE works, but it should be revisited with more empirical data and careful analysis. As illustrated, the sample size really affects the p-value. In other words, tuning on the sample size is similar to tuning on the whole system.

  • External data should be added into the system to increase the intelligence of the system, things like WordNet. Ideally, by adding corpus independent knowledge like WordNet, the system may be able to independently edit titles by follow the instructions stated in STE experiment 4.