Mining What Developers Are Talking About Deep Learning

LYU1801
JIN Fenglei
Supervisor: Michael R. Lyu
Motivation

- Deep learning is popular!
• Lots of engineers and researchers are jumping into this area.

  - More and more papers about deep learning
  - 36 FYP about deep learning this year!
- Many new developers tend to enter this field and ask some basic questions.

- It is significant and necessary for the “newbies” to have a brief understanding about this field.
Motivation

- Questions posted by developers directly reflect the focus of the deep leaning field
  - In October 2017, "Sophia", which is an AI robot and the first robot to receive citizenship at that time, is popular.

- For experienced developers, knowing the newest information gives them inspiration.

Questions about “Sophia”
- We propose a framework called IEDL to automatically track topic changes and identify emerging topics from Deep Learning-related posts in Q&A forum effectively.
- We propose a novel topic interpretation method, which improve the topic coherence dramatically.
- We visualize the variations of the captured (emerging) topics along with time slices, with the emerging ones highlighted.
Related work
• Previous works for aspect extraction can be categorized into three approaches: **rule-based, supervised, and unsupervised**

- LDA (Blei et al., 2003) and its variants are the most popular unsupervised approaches
- Attention-based Aspect Extraction (ABAE) model (He et al., 2017)
- On-line Latent Dirichlet Allocation (OLDA)
- IDEA: with adaptively online latent Dirichlet allocation approach (AOLDA)
- Didn’t extract phrases
- Didn’t reconstruct sentences
- Didn’t detect emerging topics
- No comparison

Illustration of ABAE of last semester
Overview

- Part A: preprocesses the raw posts
- Part B: extract aspects
- Part C: interpret the topic
- Part D: visualization

Framework of IEDL
Data Crawling

- Over 7,000 questions provided by StackExchange
- Over 9,000 questions under the tag of deep-learning in StackOverflow
- Use a python package called scrapy to crawl the data in StackOverflow
- Enter the website of every question to crawl the detailed information
Data Analysis

StackExchange deep-learning questions

- **Votes** and **Views** are important attributes!
• **We manually label 507 posts**

  • Categories: Image, NLP, Game-ai, Self-driving, Programming-languages, and Reinforcement-learning.
  
  • The labels are determined based on the tags provided by Stack Exchange and to maximize their distinguishability.
Preprocessing

• massive noisy words
• codes, terminologies and websites
• HTML tags

Massive question
Preprocessing

- Word Formatting:
  - lowercase
  - lemmatization
- Word Filtering:
  - reduce the non-informative words
- Word Replacement:

<table>
<thead>
<tr>
<th>Non-informative parts</th>
<th>Replacing words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Websites (eg: http://..., https://...)</td>
<td>url</td>
</tr>
<tr>
<td>All numbers</td>
<td>&lt;digit&gt;</td>
</tr>
<tr>
<td>Image html tag</td>
<td>img</td>
</tr>
<tr>
<td>Code, pseudocode</td>
<td>code</td>
</tr>
<tr>
<td>Unknown words in dictionary</td>
<td>&lt;unk&gt;</td>
</tr>
</tbody>
</table>
**Preprocessing**

- **HTML Tags Summarization:**

<table>
<thead>
<tr>
<th>Tags</th>
<th>Description</th>
<th>Tags</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;br&gt;</td>
<td>new line</td>
<td>&lt;ol&gt;</td>
<td>ordered list</td>
</tr>
<tr>
<td>&lt;hr&gt;</td>
<td>thematic change in the content</td>
<td>&lt;blockquote&gt;</td>
<td>a section that is quoted from another source</td>
</tr>
<tr>
<td>&lt;em&gt;</td>
<td>stress emphasis</td>
<td>&lt;pre&gt;</td>
<td>a preformatted text</td>
</tr>
<tr>
<td>&lt;strong&gt;</td>
<td>important text</td>
<td>&lt;code&gt;</td>
<td>a code or pseudocode (handled before)</td>
</tr>
<tr>
<td>&lt;h1&gt;, &lt;h2&gt;, &lt;h3&gt;</td>
<td>define HTML headings</td>
<td>&lt;img src=...&gt;</td>
<td>image (handled before)</td>
</tr>
<tr>
<td>&lt;ul&gt;</td>
<td>unordered (bulleted) list</td>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Preprocessing

Phrase Extraction:

\[ PMI(w_i, w_j) = \log \frac{p(w_i w_j)}{p(w_i)p(w_j)} \]

Extracted phrases: output_layer, dot_product, neural_network, initial_state, hide_layer, cross_entropy, cross_validation, computer_science, tic_tac_toe, activation_function.
Model

- Goal: learn a topic distribution and detect emerging topics
Model

• Normal LDA

Illustration of IEDL
• OLDA

Illustration of IEDL
Model

$$\beta_k^t = \sum_{i=1}^{w} \mu_i^i \gamma_k^i \phi_k^{t-i}$$

Illustration of IEDL
\[ \beta_k^t = \sum_{i=1}^{w} \mu^i \gamma^i_k \phi^i_k \]

\[ \gamma^i_k = \frac{\exp(\phi^i_k \cdot \beta_k^{t-1})}{\sum_{j=1}^{w} \phi^j_k \cdot \beta_k^{t-1}}, \]

\[ \mu^i = \exp(-\lambda^i), \]
Anomaly Detection:

\[ D_{JS}(\phi^t_k || \phi^{t-1}_k) = \frac{1}{2} D_{KL}(\phi^t_k || M) + \frac{1}{2} D_{KL}(\phi^{t-1}_k || M) \]

\[ M = \frac{1}{2} (\phi^t_k + \phi^{t-1}_k) \]

\[ D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \]
Automatic Topic Interpretation:

$$SCORE_{qua}(l) = \exp\left(-\frac{1}{\ln(v_l + 1) \ln(r_l + 1)} - \eta \times \frac{1}{\ln(h_l + 1)}\right)$$

- \(l\) is the post, and \(v_l\), \(r_l\), \(h_l\) are the votes, views, and length of the post respectively.
Experiment
• **StackExchange: 7,067**
• **507 labeled data**
• **Divided dataset in 2017 into 12 months**

<table>
<thead>
<tr>
<th>Month</th>
<th>Question No.</th>
<th>Month</th>
<th>Question No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2017-01</td>
<td>294 questions</td>
<td>2017-07</td>
<td>358 questions</td>
</tr>
<tr>
<td>2017-02</td>
<td>226 questions</td>
<td>2017-08</td>
<td>458 questions</td>
</tr>
<tr>
<td>2017-03</td>
<td>288 questions</td>
<td>2017-09</td>
<td>358 questions</td>
</tr>
<tr>
<td>2017-04</td>
<td>306 questions</td>
<td>2017-10</td>
<td>374 questions</td>
</tr>
<tr>
<td>2017-05</td>
<td>272 questions</td>
<td>2017-11</td>
<td>350 questions</td>
</tr>
<tr>
<td>2017-06</td>
<td>228 questions</td>
<td>2017-12</td>
<td>378 questions</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td></td>
<td>3,890 questions</td>
</tr>
</tbody>
</table>
We use the topic distribution of each post as features, and classify the 507 labeled posts by SVM.

IEDL outperforms the baseline model by 5% for average precision.
### Topic Coherence

<table>
<thead>
<tr>
<th></th>
<th>OLDA</th>
<th>IDEA</th>
<th>IDEA+Quality Score</th>
<th>IEDL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.133</td>
<td>0.166</td>
<td>0.217</td>
<td>0.222</td>
</tr>
</tbody>
</table>

- IEDL improves the topic coherence greatly!
Visualization

- https://github.com/Alle
  nFenglei/IEDL
- ./run normal
- ./run test
- ./run views
- http://appsrv.cse.cuhk.e
du.hk/~fljin7/fyp_term2
  /index.html
Case Study

How will morality questions be settled in the domain of self-driving cars?

For example...

1) If a dog is crossing the road, I'd expect the car to try to avoid it. But what if this leads to 0.00001% more risk for the driver? What is the 'risk cut-off'?

2) What if a cockroach is crossing the road? Will the car have a list of animals okay to run over?

3) What if a kid is crossing the street and avoiding it would kill the driver?

These questions seem to not really have an answer, yet self-driving cars are almost ready. What are they doing about all of this?

Tesla plans to unveil its all-electric semi truck on October 26th

Month later than Musk originally announced

By Zac Estrada | @zacestrada | Sep 13, 2017, 7:55pm EDT
Acknowledgement

• My deeper gratitude goes to my supervisor Michael and PhD mentor Cuiyun Gao

• Submitted to ACL workshop & ICML workshop
THANK YOU