Product-Patent Matching

**Task:** given a Wiki article describing a product, find its matched entities from target domain.

**Dataset:**
- 13,085 Wiki articles;
- 2,000 Chinese articles from Baidu Baike;
- 15,000 patents from USPTO.

**Method**

- CS+LDA
- RW+LDA
- LFG:
  - Title Only
  - SVM-S
- LFG+CST

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Measure</th>
<th>F2-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS+LDA</td>
<td>0.805</td>
<td>0.774</td>
<td>0.794</td>
<td>0.397</td>
</tr>
<tr>
<td>RW+LDA</td>
<td>0.853</td>
<td>0.801</td>
<td>0.823</td>
<td>0.756</td>
</tr>
<tr>
<td>RTM</td>
<td>0.907</td>
<td>0.907</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>RW+RTM</td>
<td>0.867</td>
<td>0.849</td>
<td>0.856</td>
<td>0.926</td>
</tr>
<tr>
<td>CST</td>
<td>0.861</td>
<td>0.850</td>
<td>0.855</td>
<td>0.889</td>
</tr>
</tbody>
</table>

**Cross-lingual Matching**

**Task:** given an English Wiki article, find all Chinese article reporting the same content.

**Dataset:**
- 2,000 English articles from Wikipedia;
- 2,000 Chinese articles from Baidu Baike;
- Each English article reports to one Chinese article.

**Method**

- SVM-S: famous cross-lingual Wikipedia matching toolkit.
- LFG: mainly considers the structural information of Wiki articles.
- LFG+LDA: adds content feature (topic distributions) to LFG by LDA.
- LFG+LDF: adds content feature to LFG by employing CST.

<table>
<thead>
<tr>
<th>Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Measure</th>
<th>F2-Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title Only:</td>
<td>0.563</td>
<td>0.563</td>
<td>0.769</td>
<td>0.171</td>
</tr>
<tr>
<td>SVM-S:</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
<td>0.805</td>
</tr>
<tr>
<td>LFG:</td>
<td>0.697</td>
<td>0.697</td>
<td>0.697</td>
<td>0.697</td>
</tr>
<tr>
<td>LFG+LDA:</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
</tr>
<tr>
<td>LFG+LDF:</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
</tr>
<tr>
<td>LFG+LDF:</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
<td>0.662</td>
</tr>
</tbody>
</table>

**Experimental Results**

**Product-Patent Matching**

**Cross-lingual Matching**

**Proposed Model**

- Learning LDA on the source domain and target domain respectively.
- Given an entity, ranking others as candidates according to the topic similarity.

**Cross-Sampling-Based Entity Generation**

**Matching Relation Generation**

**Cross-Source Topic Model**

- E-Step: update variational parameters:
- M-Step: update model parameters:

**Integrate topic extraction and entity matching into a unified framework**

**Proposed Model (CST)**

- Random Walk based on CST (RW+CST):
  - Search engine: object relevant area
  - Ranking module: ranking module rank candidate

- Relational Topic Model (RTM): used to model links between documents.
- Random Walk based on CST (RW+CST): uses CST instead of LDA.

**Baseline method (CS+LDA, RW+LDA)**

- \( P@3 \) : 0.250
- \( R#20 \) : 0.217
- \( MAP / F1 \) : 0.025

**Performance (MAP / F1)**

- LFG+CST: 0.662
- LFG+LDA: 0.652
- SVM-S: 0.805
- RTM: 0.907
- RW+LDA: 0.853
- CS+LDA: 0.805

**Content Similarity based on LDA (CS+LDA): cosine similarity between two articles’ topic distribution extracted by LDA. Random Walk based on LDA (RW+LDA): random walk on a graph where edges indicate the topic similarity between articles. Relational Topic Model (RTM): used to model links between documents. Random Walk based on CST (RW+CST): uses CST instead of LDA comparing with RW+LDA.**

**Parameter Analysis**

- (a) Number of topics
- (b) Cross-sampling ratio
- (c) Precision
- (d) Convergence