Predicting Hierarchical Relationship in Job Title Taxonomy
Shuang Jin sjin1@stanford.edu https://youtu.be/yb6EjHsbJ8

Problem Statement

**Goal**
- predict the relationship between job titles in the taxonomy.

**Motivation**
- build a well-structured taxonomy to organize job market knowledge.

**Problem Definition**
- given a job title entity pair \((x_{\text{source}}, x_{\text{target}})\) predict their relationship \(y\)

\[
x_{\text{source}}: \text{Machine Learning Engineer} \quad x_{\text{target}}: \text{Computer Software Engineer}
\]

\(x_{\text{target}}\) is a Broader Term of \(x_{\text{source}}\).

Data

**Source**
- an established title taxonomy curated by a dedicated taxonomy team.

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>labels &amp; Size</td>
<td>290683</td>
<td>16723</td>
<td>3106</td>
</tr>
<tr>
<td>Label</td>
<td>Meaning</td>
<td>Data Size</td>
<td></td>
</tr>
<tr>
<td>BT</td>
<td>Broader Term</td>
<td>100K</td>
<td></td>
</tr>
<tr>
<td>NT</td>
<td>Narrower Term</td>
<td>100K</td>
<td></td>
</tr>
<tr>
<td>P1</td>
<td>Preferred Term</td>
<td>20K</td>
<td></td>
</tr>
<tr>
<td>PNT</td>
<td>Non-Preferred Term</td>
<td>20K</td>
<td></td>
</tr>
<tr>
<td>UKN</td>
<td>Unknown</td>
<td>60K</td>
<td></td>
</tr>
</tbody>
</table>

Features

<table>
<thead>
<tr>
<th>Source</th>
<th>Raw</th>
<th>Tokenized</th>
</tr>
</thead>
<tbody>
<tr>
<td>supply chain specialist</td>
<td>0 0 0 0 0 65</td>
<td>73 18</td>
</tr>
<tr>
<td>supplier quality specialist</td>
<td>0 0 0 0 1097 79 10</td>
<td></td>
</tr>
<tr>
<td>Label</td>
<td>NT</td>
<td>0 1 0 0 0</td>
</tr>
</tbody>
</table>

Models & Results

\[
J = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{C} \lambda_j \cdot y_j^i \cdot \log \hat{y}_j^i
\]

**Challenges**
- Embedding trained from this project outperformed Glove embedding?
  - Large percentage of spelling errors that cannot be recognized by Glove
  - High repetitiveness of words in training data making embedding training less difficult
  - Solution: Removing spelling mistakes
- Large fluctuation on validation set performance after a few epochs?
  - Learning rate set too large
  - Solution: Calibrating on the learning rate
- One label performance significantly worse than others?
  - Data imbalance
  - Solution: adding class weights to loss

Selected Model

**Input**

- Embedding w/ Glove
- LSTM
- Softmax

**Param #**

- (16) 351700
- (16, 100) 42240
- (64) 325

**Models**

- Base: 77.52% 72.11% 73.66%
- Simple LSTM 64d: 94.63% 91.90% 93.14%
- Simple LSTM 128d: 93.98% 93.04% 93.44%
- Glove: 77.33% 71.75% 73.27%
- Glove LSTM 64d: 95.84% 94.70% 95.21%
- Glove LSTM 128d: 95.31% 94.56% 94.90%

Selected Model

**Glove LSTM 64d**

**Selected Model**

**Glove LSTM 64d**

**Selected Model**

**Glove LSTM 64d**

Discussion

**Observation**

**Label Difficulty for Machine vs. Human:**
- Machine: NPT > PT > UKN > BT > NT
- Human: PT > NPT > BT > NT > UKN

**Why is UKN easy for human but not so easy for machine?**
- Humans use knowledge such as related industries, skills, etc. to make the judgement
- Future Work: Mine title-related data as additional features

**What do humans do to get better on PT vs. NPT?**
- Humans use popularity data to compare which title is used more often
- Future Work: adding counts as new features

**What do humans do to get better on PT vs. NPT?**
- Humans use popularity data to compare which title is used more often
- Future Work: adding counts as new features

**What could be the pitfalls for the model in production?**
- Labeled data has sector bias
- Future Work: balancing sector data

References

