Using Active Learning and Semantic Clustering for Noise Reduction in Distant Supervision

Lucas Sterckx, Thomas Demeester, Johannes Deleu and Chris Develder
Email: lucas.sterckx@intec.ugent.be

Noise Reduction in Distantly Supervised Data

- **Distant Supervision**:
  - State-of-the-art for training Knowledge Base Population
  - Inherently suffers from noise
  - We propose noise reduction using:
    - Discriminative classifier trained on a small set of labeled examples
    - Active learning strategy and Semantic Similarity between the contexts of the training examples

- Combination facilitates the creation of a clean training set for relation extraction, at a reduced manual labeling cost.

Semantic Clustering

- Distributional Hypothesis
- Summed Average of Embeddings of Words in context of relations
- Use semantic vector representations in cluster based active learning

Cluster-based Active Learning

- Exploit hierarchical cluster structure in data
- Efficient search through hypothesis space

Experimental Setup

- Part of a participation in the TAC-KBP slot filing competition, precision after noise reduction + 8%
- 2,000 training samples assigned with a True or False label with respect to the 2014 TAC-annotation guidelines for a selection of 12 relations with a large quantity of training data

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distant Supervision (Baseline)</td>
<td>51.9</td>
<td>100.0</td>
<td>60.3</td>
</tr>
<tr>
<td>Random Sampling</td>
<td>72.0</td>
<td>72.8</td>
<td>72.4</td>
</tr>
<tr>
<td>Bag-of-Words Clustering</td>
<td>75.4</td>
<td>65.2</td>
<td>69.6</td>
</tr>
<tr>
<td>Latent Semantic Indexing Clustering</td>
<td>73.7</td>
<td>68.5</td>
<td>68.3</td>
</tr>
<tr>
<td>Semantic Vector Space Clustering</td>
<td>74.6</td>
<td>71.4</td>
<td>72.9</td>
</tr>
</tbody>
</table>

^ Macro-average filter performance using 70 labeled distantly supervised training examples

Conclusion

Novel approach for filtering a distantly supervised training set by building a binary classifier to detect true relation mentions. The classifier is trained using a cluster based active learning strategy. Clustering of relation mentions and adding semantic information reduces human effort and makes this a promising approach more feasible to filter a wide variety of relations.

Future work: we suggest the use of more compositional methods for transforming to a semantic vector space from the field of paraphrase detection.