Leveraging Just a Few Keywords for Fine-Grained Aspect Detection Through Weakly Supervised Co-Training

Giannis Karamanolakis  
Columbia University  
gkaraman@cs.columbia.edu

Daniel Hsu  
Columbia University  
djhsu@cs.columbia.edu

Luis Gravano  
Columbia University  
gravano@cs.columbia.edu

**Aspect Detection in Online Reviews**

**Goal:** Identify which product aspects (e.g., price, quality) are discussed in individual segments (e.g., sentences, clauses) of a review.

**Key task in:** Sentiment Analysis, Opinion Mining, Review Summarization.

1. Supervised
   - Easier to collect
2. Unsupervised (Neural Topic Models)
   - Leverage descriptive seed words as a “weak” source of supervision.

**Issue:** Ground-truth aspect labels are not readily available.
- Manual segment annotation is expensive and not scalable.

**Neural Networks For Aspect Detection**

1. **Supervised**
   - Segment $\rightarrow$ NN $\rightarrow$ $p = \langle p^1, \ldots, p^K \rangle$

2. **Unsupervised (Neural Topic Models)**
   - Topic 1
   - Topic 2
   - Topic 100

**Issue:** May not capture the K aspects of interest.

3. **Weakly Supervised: Learning with Seed Words**
   - Idea: Leverage descriptive seed words as a “weak” source of supervision.

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Seed Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td>price, value, money, ...</td>
</tr>
<tr>
<td>Image</td>
<td>picture, color, bright, ...</td>
</tr>
<tr>
<td>Sound</td>
<td>sound, speaker, noise, ...</td>
</tr>
</tbody>
</table>

**Idea:** Easier to collect seed words than ground-truth aspect labels.
- manually (humans) or automatically (small number of labeled segments).

**Issue:** Previous approaches use seed words just for initialization.

**Datasets**
- Amazon Product Reviews
- Previous year: 6 product domains
- Amazon reviews from 6 product domains (9 aspects)
- Restaurant reviews from 6 languages (12 aspects)
- Multi-seed Aspect Extractor (MATE) [Angelidis & Lapata '18]

**Baselines**
- LDA-Anchors [Lund et al. '17]
- Aspect-Based Auto-Encoder (ABAE) [He et al. '17]
- Teacher (BoSW)
- Student (BoSW)
- Teacher (BERT)

**Results**
- Teacher and Student-BoW outperform previous approaches.
- Student-* outperforms MATE across all domains and languages (best model: Student-BERT).
- Most improvement achieved after one round of co-training.

**Experiments**

- Amazon Reviews from 6 Product Domains
  - 1M unlabeled segments
  - 30 seed words per aspect (same across models)

- Initial Aspect Embeddings
  - Pre-Trained Word Embeddings
  - Unsupervised Model

Acknowledgements: This material is based upon work supported by the National Science Foundation under Grant No. IIS-15-63785.