Multi-Granular Aspect Aggregation in Aspect-Based Sentiment Analysis

John Pavlopoulos and Ion Androutsopoulos
NLP Group, Department of Informatics
Athens University of Economics and Business, Greece
http://nlp.cs.aueb.gr/
A laptop with great design, but the service was horrible!

Good food and wine, few beers, and very polite service...

- User generated reviews
- Aspect terms carry sentiment
- Aspect Based Sentiment Analysis
  - Aspect Term Extraction
  - Aspect (Term) Aggregation
  - Aspect Sentiment Estimation
- We focus on Aspect Aggregation
What is aspect aggregation with multiple granularities?

Top Aspect Terms

1. Food
2. Wine
3. Beers
4. Service
Approaches to aspect aggregation with multiple granularities

- Synonym Grouping (e.g., group “cost” and “price”)
  - Only aggregates aspect terms at the lowest granularity
  - E.g., “wine” and “beers” are not synonyms
  - But they could be aggregated along with “drinks” if a coarser granularity (fewer groups of aspect terms) is desirable

- Predefined Taxonomies
  - Hard to find and to manually construct and maintain

- Flat Clustering aiming at fewer or more clusters of aspect terms
  - E.g., k-means with smaller or larger k
  - Does not satisfy consistency constraint: If “wine” and “beers” are grouped together for 5 clusters, they should remain grouped together for 4, 3, and 2 clusters (consistent sense of “zoom out”)

- Hierarchical Agglomerative Clustering (HAC)
  - Dissecting the resulting hierarchy at different depths produces clusterings at different granularities satisfying the consistency constraint
Aspect aggregation via hierarchical agglomerative clustering

ABSA

pizza, food, fish, ... ***
prices, price **
ambience, atmosphere, ... *
waiter, staff, service ***

Restaurant aspects hierarchy

pizza
food
fish
sushi
meal
dishes
menu
portions
wine
drinks
prices
price
ambience
atmosphere
decor
place
table
waiter
staff
service
Datasets: agreement problems

Benchmark datasets presuppose *inter-annotator agreement*

- Humans **agree** when asked to cluster near-synonyms, but **not** when asked to produce coarser clusters of aspect terms
- Humans **don’t agree** when judging given aspect term hierarchies
- Humans **don’t agree** when asked to create aspect term hierarchies

**But!**

- Humans **agree** when asked to fill in an aspect term similarity matrix

<table>
<thead>
<tr>
<th></th>
<th>food</th>
<th>fish</th>
<th>sushi</th>
<th>dishes</th>
<th>wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>4</td>
<td></td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>fish</td>
<td></td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>sushi</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>dishes</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>wine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A two phase decomposition

We propose decomposing aspect aggregation into 2 phases:

- **Phase A:** Systems try to produce (fill in) a similarity matrix as close as possible to the gold similarity matrix
- **Phase B:** The similarity matrix of Phase A is used as a distance measure in hierarchical agglomerative clustering (along with a linkage criterion) to produce an aspect term hierarchy, from which clusterings of different granularities can be obtained.
Phase A data

- Customer review data (Restaurants and Laptops)
- Subjective sentences, manually annotated aspect terms
- 20 most frequently annotated aspect terms per domain
- 3 human judges asked to fill in a similarity matrix (1-5)
- Pearson’s rho: \( \rho(\text{restaurants}) = 0.81, \rho(\text{laptops}) = 0.74 \)
- Absolute agreement: \( \alpha(\text{restaurants}) = 0.90, \alpha(\text{laptops}) = 0.91 \)
- Gold similarity matrix: average scores of the 3 judges

<table>
<thead>
<tr>
<th></th>
<th>food</th>
<th>fish</th>
<th>sushi</th>
<th>dishes</th>
<th>wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>food</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>fish</td>
<td></td>
<td>4</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>sushi</td>
<td></td>
<td></td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>dishes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>wine</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phase A methods

Systems fill in the similarity matrix. The similarity matrix of each system is evaluated by comparing it to the gold similarity matrix.

**WordNet-based**: Wu & Palmer, Lin, Jiang & Conrath, Shortest Path
- No word-sense disambiguation. Greedy approach instead:
  - For aspect terms $a_1, a_2$: $\text{sim}(a_1, a_2) = \max \text{ sense similarity}$

**Distributional (DS)**: Cosine similarity between $v(a_1)$ and $v(a_2)$
- $v(a) = <\text{PMI}(a, w_1), ..., \text{PMI}(a, w_n)>$

**AVG**: Average of all measures
**WN**: Average of WordNet-based measures
**WNDS**: Average of WN and DS
Phase A methods: sense pruning

Sense pruning applied to WordNet-based methods only

WordNet-based: Wu & Palmer, Lin, Jiang & Conrath, Shortest Path

- Greedy approach: for aspect terms $a_1, a_2$ : $\text{sim}(a_1, a_2) = \max \text{ sense similarity}$
- Sense Pruning: For each aspect term $a_i$ discard some senses $s_{ij}$ before the greedy approach!
- For each sense $s_{ij}$ of aspect term $a_i$ we compute the relevance of $s_{ij}$ to all the other aspect terms $a_i'$
  \[
  \text{rel}(s_{ij}, a_i') = \max_{s_{i'j'} \in \text{senses}(a_i')} \text{sim}(s_{ij}, s_{i'j'})
  \]
- We take the average relevance of each sense $s_{ij}$ of aspect term $a_i$ to all the other aspect terms $a_i'$
- For each aspect term $a_i$ we keep its top-5 senses, i.e., the 5 senses with the highest average relevance to the other aspect terms
- The discarded senses are considered to be domain irrelevant
Phase A results: Pearson correlation to gold similarity matrix

<table>
<thead>
<tr>
<th>Method</th>
<th>without SP</th>
<th>with SP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rest.</td>
<td>Lapt.</td>
</tr>
<tr>
<td>WP</td>
<td>0.475</td>
<td>0.216</td>
</tr>
<tr>
<td>PATH</td>
<td>0.524</td>
<td>0.301</td>
</tr>
<tr>
<td>LIN@domain</td>
<td>0.390</td>
<td>0.256</td>
</tr>
<tr>
<td>LIN@Brown</td>
<td>0.434</td>
<td>0.329</td>
</tr>
<tr>
<td>JCN@domain</td>
<td>0.467</td>
<td>0.348</td>
</tr>
<tr>
<td>JCN@Brown</td>
<td>0.403</td>
<td>0.469</td>
</tr>
<tr>
<td>DS</td>
<td>0.283</td>
<td>0.517</td>
</tr>
<tr>
<td>AVG</td>
<td>0.499</td>
<td>0.352</td>
</tr>
<tr>
<td>WN</td>
<td>0.490</td>
<td>0.328</td>
</tr>
<tr>
<td>WNDS</td>
<td>0.523</td>
<td>0.453</td>
</tr>
</tbody>
</table>

A paired t test indicates that the differences (with and without pruning) are statistically significant (p < 0.05).
Phase A results: Pearson correlation to gold similarity matrix

Now comparing our best system (WNDS with SP) to two state of the art term similarity methods and human judges

Mikolov et. al., 2013

<table>
<thead>
<tr>
<th>Method</th>
<th>Restaurants</th>
<th>Laptops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Han et al. (2013)</td>
<td>0.450</td>
<td>0.471</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>0.434</td>
<td>0.485</td>
</tr>
<tr>
<td>WNDS with SP</td>
<td>0.545</td>
<td>0.546</td>
</tr>
<tr>
<td>Judge 1</td>
<td>0.913</td>
<td>0.875</td>
</tr>
<tr>
<td>Judge 2</td>
<td>0.914</td>
<td>0.894</td>
</tr>
<tr>
<td>Judge 3</td>
<td>0.888</td>
<td>0.924</td>
</tr>
</tbody>
</table>
Phase B

- Get a similarity matrix (e.g., from a Phase A method or humans)
- Use the similarity matrix to compute the distance between any two aspect terms
- Choose a linkage criterion in effect to compute the distance between any two clusters of aspect terms:
  - *Single*: min distance of any two terms of the clusters
  - *Complete*: max distance of any two terms of the clusters
  - *Average*: average distance between the terms of the clusters
  - *Ward’s*: minimum variance criterion (this is not a distance function)
- Use Hierarchical Agglomerative Clustering to build an aspect term hierarchy
- Dissect the aspect term hierarchy at different depths, to obtain fewer or more clusters.
Phase B: evaluation

- **Silhouette Index (Rousseeuw, 1987)**
  - Considers both inter and intra cluster coherence
  - Ranges from -1.0 to 1.0
  - Requires the correct distances between cluster elements (aspect terms) to be known when evaluating clusters
  - We use the correct distances provided by the gold Phase A similarity matrix

- **Different indices produce similar results**
  - Dunn Index (Dunn, 1974)
  - Davies-Bouldin Index (Davies and Bouldin, 1979)
Phase B results: gold similarity matrix, different linkage criteria

We use the gold similarity matrix from Phase A and Hierarchical Agglomerative Clustering with 4 different linkage criteria

- No linkage criterion clearly outperforms the others
- All four criteria perform reasonably well

![Silhouette Index](https://via.placeholder.com/150)

<table>
<thead>
<tr>
<th>Restaurants</th>
<th>Laptops</th>
</tr>
</thead>
<tbody>
<tr>
<td>single</td>
<td>single</td>
</tr>
<tr>
<td>complete</td>
<td>complete</td>
</tr>
<tr>
<td>average</td>
<td>average</td>
</tr>
<tr>
<td>Ward</td>
<td>Ward</td>
</tr>
</tbody>
</table>
Phase B results: similarity matrix of WNDS+SP, different linkage criteria

We now use the similarity matrix of the best Phase A method (WNDS+SP) and Hierarchical Agglomerative Clustering with 4 linkage criteria:

- All the scores deteriorate significantly
- Again, no clear winner among the linkage criteria
Phase B results: human evaluation

We asked 4 human judges to evaluate (1-5 scale) clusterings of varying granularities (fewer or more clusters)

- **System 1**: gold similarity matrix of Phase A plus Hierarchical Agglomerative Clustering (HAC) with average linkage
- **System 2**: WNDS+SP similarity matrix plus HAC with average linkage
- **Absolute inter-annotator agreement**: greater than 0.8 in all cases

![Restaurants](chart.png)

![Laptops](chart.png)
Conclusions

- We introduced aspect aggregation at multiple granularities and a two-phase decomposition
  - Phase A fills in a pairwise aspect term similarity matrix
  - Phase B uses the similarity matrix of Phase A, a linkage criterion, and hierarchical agglomerative clustering to produce an aspect hierarchy

- Dissecting the generated aspect hierarchy at different depths produces consistent clusterings at different granularities

- Our decomposition leads to high inter-annotator agreement and allows previous work on term similarity and HAC to be reused

- We introduced a sense-pruning mechanism that improves WordNet-based similarity measures and leads to the best performing method in Phase A, but large scope for improvement

- With the gold Phase A similarity matrix, the quality (perceived and measured with SI) of the clusters of Phase B is high, but much lower quality with the similarity matrix of the best Phase A method

- We also provide publicly available datasets
Thank you!

Questions?