Aspect Based Sentiment Analysis

Ioannis Pavlopoulos
PhD student

Natural Language Processing Group
Department of Informatics
Athens University of Economics and Business
The food was delicious! Nice food but horrible wine and beers. Excellent service! Thank you 😊 ...

Aspect term extraction
food, wine, beers, service, ...

Aspect term aggregation
food, wine, beers, service, ...

Aspect term polarity
food, wine, beers, service, ...

See our paper for more details (Pontiki et al., 2014)
1. Aspect term extraction
2. Multi-granular aspect aggregation
3. Message-level sentiment estimation
4. Aspect term sentiment estimation
## Previous datasets vs. our datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Inter- Annotator Agreement</th>
<th># Domains</th>
<th>Gold Aspect Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hu &amp; Liu 2004</td>
<td>✗</td>
<td>1</td>
<td>✓</td>
</tr>
<tr>
<td>Ganu et al. 2009</td>
<td>✓</td>
<td>1</td>
<td>✗</td>
</tr>
<tr>
<td>Blitzer et al. 2007</td>
<td>✗</td>
<td>4</td>
<td>✗</td>
</tr>
<tr>
<td>Pavlopoulos &amp; Androutsopoulos 2014</td>
<td>✓</td>
<td>3</td>
<td>✓</td>
</tr>
</tbody>
</table>
## Aspect term extraction

### # distinct aspect terms with $n$ occurrences

<table>
<thead>
<tr>
<th>Domain</th>
<th>$n = 0$</th>
<th>$n &gt; 0$</th>
<th>$n &gt; 1$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>1,590</td>
<td>2,120</td>
<td>872</td>
<td>3,710</td>
</tr>
<tr>
<td>Hotels</td>
<td>1,622</td>
<td>1,978</td>
<td>652</td>
<td>3,600</td>
</tr>
<tr>
<td>Laptops</td>
<td>1,760</td>
<td>1,325</td>
<td>416</td>
<td>3,085</td>
</tr>
</tbody>
</table>

### # sentences with $n$ aspect term occurrences

<table>
<thead>
<tr>
<th>Domain</th>
<th>$n = 0$</th>
<th>$n &gt; 0$</th>
<th>$n &gt; 1$</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>1,590</td>
<td>2,120</td>
<td>872</td>
<td>3,710</td>
</tr>
<tr>
<td>Hotels</td>
<td>1,622</td>
<td>1,978</td>
<td>652</td>
<td>3,600</td>
</tr>
<tr>
<td>Laptops</td>
<td>1,760</td>
<td>1,325</td>
<td>416</td>
<td>3,085</td>
</tr>
</tbody>
</table>

### Inter-Annotator Agreement:
- Dice: ~70% in all domains

### Our new datasets

#### battery life

<table>
<thead>
<tr>
<th>Domain</th>
<th>multi-word</th>
<th>single-word</th>
<th>multi-word</th>
<th>single-word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>593</td>
<td>452</td>
<td>67</td>
<td>195</td>
</tr>
<tr>
<td>Hotels</td>
<td>199</td>
<td>262</td>
<td>24</td>
<td>120</td>
</tr>
<tr>
<td>Laptops</td>
<td>350</td>
<td>289</td>
<td>67</td>
<td>137</td>
</tr>
</tbody>
</table>

---

5/70
Multi-word vs. single-word distinct aspect terms per domain

Electronics (Hu & Liu, 2004) & laptops reviews contain more multi-word distinct aspect terms
Aspect term extraction

Precision, Recall, F-measure

Gold: “design” (94), “service” (3), “screen” (2)

Predicted: “design” (92/94), “service” (1/3 + 1), “screen” (0/2), “foo” (+3)

Computed on **tokens** (aspect term occurrences):

\[
P = \frac{|\text{true predicted tokens}|}{|\text{predicted tokens}|} = \frac{92+1+0+0}{92+2+0+3} = 0.96
\]

\[
R = \frac{|\text{true predicted tokens}|}{|\text{true tokens}|} = \frac{92+1+0+0}{94+3+2+0} = 0.94
\]

Computed on **types** (distinct aspect terms):

\[
P = \frac{|\text{true predicted types}|}{|\text{predicted types}|} = \frac{|\text{design,service}|}{|\text{design,service,foo}|} = \frac{2}{3} = 0.66
\]

\[
R = \frac{|\text{true predicted types}|}{|\text{true types}|} = \frac{|\text{design,service}|}{|\text{design,service,screen}|} = \frac{2}{3} = 0.66
\]

Frequent distinct aspect term are treated as rare ones

Over sensitive to high-frequent aspect terms
The users care only about the top $m$ (e.g., 10-20) most frequently discussed distinct aspect terms.
- The value of $m$ depends on screen size, available time etc.

Finding or missing a truly more frequent distinct aspect term should be rewarded or penalized more.

Placing a truly high-frequency distinct aspect term towards the beginning of the returned list should be rewarded more.
Order all the correct distinct aspect terms by human annotation frequency (G list).

Each method returns a list of distinct aspect terms, ordered by predicted frequency (A list).
- Given an $m$ value, use the first $m$ elements of the A list ($A_m$).
- Compare $G$ and $A_m$ for different $m$ values.
Weighted precision and recall

By varying $m$, we obtain $WP_m - WR_m$ curves.

Also, average weighted precision at 11 weighted recall levels.

$WP_m$ is similar to $nDCG@m$, but no counter-part for $WR_m$. 

$WP_m = \frac{\sum_{i=1}^{m} \frac{1}{i} \cdot 1 \cdot \{a_i \in G\}}{\sum_{i=1}^{m} \frac{1}{i}} \rightarrow m=3 \frac{1 + 0 + \frac{1}{3}}{1 + \frac{1}{2} + \frac{1}{3}} = 0.73$

$WR_m = \frac{\sum_{i=1}^{m} \frac{1}{r(a_i)} \cdot 1 \cdot \{a_i \in G\}}{\sum_{j=1}^{|G|} \frac{1}{j}} \rightarrow m=3 \frac{1 + 0 + \frac{1}{2}}{1 + \frac{1}{2} + \frac{1}{3}} = 0.82$
**Freq baseline**
- Considered effective & popular *unsupervised* baseline (Liu, 2012)
- Returns the **most frequent nouns and noun phrases**, ordered by decreasing sentence frequency

**H&L** *(Hu & Liu 2004)*
- Also *unsupervised*, finds frequent nouns and noun phrases, plus…
- Discards candidate aspect terms that are subparts of other candidate aspect terms
- Finds *adjectives that modify candidate aspect terms*, uses them to detect additional candidate aspect terms
- Details previously unclear, **full pseudo-code** published
We use word vectors (Mikolov, 2013) computed using Word2Vec:

\[ v('king') - v('man') + v('woman') \approx v('queen') \]

\[
\begin{align*}
\nu_{\text{domain}} &= \frac{\sum_{w \in \text{aspect terms}} v(w)}{|\text{aspect terms}|} \\
\nu_{\text{common}} &= \frac{\sum_{w \in \text{common words}} v(w)}{|\text{common words}|}
\end{align*}
\]

For each candidate aspect term \( a \), measure its similarity (cosine) with the two centroids

Prune \( a \), if \( \cos(a, v_{\text{aspect}}) < \cos(a, v_{\text{common}}) \)

Applicable to both \textit{Freq} and \textit{H\&L}
### Results:

**average weighted precision**

<table>
<thead>
<tr>
<th>Method</th>
<th>Restaurants</th>
<th>Hotels</th>
<th>Laptops</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freq</td>
<td>43.40</td>
<td>30.11</td>
<td>9.09</td>
</tr>
<tr>
<td>Freq+w2v pruning</td>
<td><strong>45.17</strong></td>
<td><strong>30.54</strong></td>
<td>7.18</td>
</tr>
<tr>
<td>Hu&amp;Liu</td>
<td>52.23</td>
<td>49.73</td>
<td>34.34</td>
</tr>
<tr>
<td>H&amp;L+w2v pruning</td>
<td><strong>66.80</strong></td>
<td><strong>53.37</strong></td>
<td><strong>38.93</strong></td>
</tr>
</tbody>
</table>

All differences are statistically significant (p<0.01)
Aspect term extraction

Results:

average weighted precision
Results:

average weighted precision
Aspect term extraction

Results:

average weighted precision

[Graph showing weighted precision vs. weighted recall for Laptops, with different methods represented by different markers.]
Summary & contribution of this section

- Introduced 3 new aspect term extraction datasets
  - Laptops/Restaurants/Hotels
  - Domain variety is important

- New evaluation measures
  - Weighted precision, weighted recall, average weighted precision

- Improved the popular unsupervised method of Hu & Liu
  - Additional pruning step based on continuous space word vectors (using Word2Vec)

- The ‘Aspect term extraction’ subtask of ABSA SemEval 2014 & 2015 was based on the work of this section

See our paper for more details (Pavlopoulos and Androutsopoulos, 2014a)
1. Aspect term extraction
2. Multi-granular aspect aggregation
3. Message-level sentiment estimation
4. Aspect term sentiment estimation
The food was delicious! Nice food but horrible wine and beers. Excellent service! Thank you 😊...
Multi-granular aspect aggregation

Aspect aggregation with multiple granularities

Top Aspect Terms
1. Food
2. Wine
3. Beers
4. Service

food  *****
wine  ***
beers  ***
service  *****
Multi-granular aspect aggregation

Approaches to aspect aggregation with multiple granularities

- (Near-) 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”)

---

- food
- **wine**, beers
- service

- **service**, wine
- food, beers
Multi-granular aspect aggregation via agglomerative clustering
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*.

### Datasets: agreement problems

<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>4</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>sushi</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dishes</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>wine</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
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.
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>
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, but greedy approach instead, for aspect terms $a_1, a_2$: $\text{sim}(a_1, a_2) = \text{max sense similarity}$

**Distributional (DS):** Cosine similarity between $\nu(a_1)$ and $\nu(a_2)$
- $\nu(a) = \langle \text{PMI}(a, w_1), \ldots, \text{PMI}(a, w_n) \rangle$

**AVG:** Average of all measures

**WN:** Average of WordNet-based measures

**WNDS:** Average of WN and DS
Sense pruning applied to WordNet-based methods only

- **Greedy approach**: for aspect terms $a_1, a_2 : \text{sim}(a_1, a_2) = \max$ 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
### Multi-granular aspect aggregation

**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><strong>Method</strong></td>
<td></td>
<td></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 ($p < 0.05$).
Now comparing our best system (WNDS with SP) to two state of the art term similarity methods and human judges

<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>
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.
Multi-granular aspect aggregation

Phase B: evaluation

- **Silhouette Index (Rousseeuw, 1987)**
  - Considers both inter and intra cluster coherence
  - Ranges from -1.0 to 1.0
  - Requires the 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)
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.

**Phase B results:**

- Multi-granular aspect aggregation

![Silhouette Index charts for Restaurants and Laptops]
We now use the similarity matrix of the best Phase A method (WNDS+SP) and Hierarchical Agglomerative Clustering with 4 linkage criteria.

- Again, no clear winner among the linkage criteria.
- All the scores deteriorate significantly.
Multi-granular aspect aggregation

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
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 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.

See our paper for more details (Pavlopoulos and Androutsopoulos, 2014b).
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.

Summary & contribution of section (2/2)

See our paper for more details (Pavlopoulos and Androutsopoulos, 2014b)
1. Aspect term extraction
2. Multi-granular aspect aggregation
3. Message-level sentiment estimation
4. Aspect term sentiment estimation
Message-level sentiment estimation

- Messages: sentences, social media updates, SMS
- Our message-level sentiment estimation system
  - Androutsopoulos I.
  - Karampatsis M.
  - Makrynioti N. (2013)
  - Malakasiotis P.
  - Pavlopoulos J.

I like the new charsets of Word 😊 !!!

I hate technology!!!
Goal: Classify each message to positive, negative, or neutral

Train: 8730 positive, negative, and neutral messages from

- Originally 9728, but privacy issues…

Dev: 1654 positive, negative, and neutral messages from Twitter

Test: 3814 messages from Twitter & 2094 SMS messages
Goal: Classify each message to positive, negative, or neutral

Train: 8730 train + 1654 dev messages from Twitter (2013)

Dev: 3814 + 2094 test messages from Twitter and SMS (2013)

Test: 8987 tweets, tweets with sarcasm, SMS, messages from blog posts (Live Journal)
Message-level sentiment estimation

Our system’s architecture

Subjectivity detection

SVM

Objective messages

Subjective messages

Polarity detection

SVM

Positive messages

Negative messages

LIBLINEAR

LIBLINEAR
Message-level sentiment estimation

Data preprocessing

- Twitter-specific tokeniser & POS tagger (Owoputi 2013)

- Text normalization and slang removal
  - edit distance to replace unknown words with their most similar word in an English dictionary (see also Karampatsis 2012)
  - e.g., flames → angry comments, xmpl → example, etc.
Morphological (e.g., #elongated_words ‘gooood’, or #capitalized_tokens ‘I WANT MORE’)

POS-tags based (e.g., #nouns, etc.)

Sentiment lexicons (e.g., AFINN, SentiWordNet, NRC, MPQA)
- For lexicons with no scores (e.g., MPQA) we compute our own

Miscellaneous (e.g., existence of negation, Twitter clusters)
For each lexicon we compute the following:

- Sum, max, min, average of scores of the message’s words: (7, 4, 3, 3.5)
- Count of words with scores (2)
- Score of the last word (e.g., ‘happy’ yields 3)
- All features normalized to [0, 1]
Message-level sentiment estimation

Results: Average F1(±)

<table>
<thead>
<tr>
<th>Test set</th>
<th>Best</th>
<th>AUEB</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>LJ (2014)</td>
<td>74.84</td>
<td>70.75 (9\textsuperscript{th}/50)</td>
<td>65.48</td>
</tr>
<tr>
<td>SMS (2013)</td>
<td>70.28</td>
<td>64.32 (8\textsuperscript{th}/50)</td>
<td>57.53</td>
</tr>
<tr>
<td>TW (2013)</td>
<td>72.12</td>
<td>63.92 (21\textsuperscript{st}/50)</td>
<td>62.88</td>
</tr>
<tr>
<td>TW (2014)</td>
<td>70.96</td>
<td>66.38 (14\textsuperscript{th}/50)</td>
<td>63.03</td>
</tr>
<tr>
<td>TWSarc (2014)</td>
<td>58.16</td>
<td>56.16 (4\textsuperscript{th}/50)</td>
<td>45.77</td>
</tr>
<tr>
<td>AVGall</td>
<td>68.78</td>
<td>64.31 (6\textsuperscript{th}/50)</td>
<td>56.56</td>
</tr>
<tr>
<td>AVG (2014)</td>
<td>67.62</td>
<td>64.43 (5\textsuperscript{th}/50)</td>
<td>57.97</td>
</tr>
</tbody>
</table>
Message-level sentiment estimation

- Message-level sentiment estimation system
- ‘Sentiment Analysis in Twitter’ SemEval task
  - 2013: good rank
  - 2014: better rank
- 2-stage pipeline approach
  - Handles well class imbalance
- Good generalization ability

Summary & contribution of section

Work of this section has been published in (Malakasiotis, 2013), (Karampatsis, 2014)
1. Aspect term extraction
2. Multi-granular aspect aggregation
3. Message-level sentiment estimation
4. Aspect term sentiment estimation
The food was delicious! Nice food but horrible wine and beers. Excellent service! Thank you 😊 …

Aspect term extraction
food, wine, beers, service, …

Aspect term aggregation
food, wine, beers, service, …

Aspect term polarity
food, wine, beers, service, …
“Estimate the sentiment polarities of the aspect term occurrences in a sentence”

“I hated their fajitas, but their salads were great”
“The fajitas were their starters”
“The fajitas were great to taste, but not to see”

- Sentiment of the aspect term, not the sentence per se
- Positive, negative, neutral, or conflict
- Subtask of ABSA in SemEval 2014
Aspect-term sentiment estimation

Our aspect term polarity datasets

<table>
<thead>
<tr>
<th>Domain</th>
<th>Train</th>
<th>Test</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurants</td>
<td>3041</td>
<td>800</td>
<td>3841</td>
</tr>
<tr>
<td>Laptops</td>
<td>3045</td>
<td>800</td>
<td>3845</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>6086</td>
<td>1600</td>
<td>7686</td>
</tr>
</tbody>
</table>

Human annotations of aspect term occurrences and their polarities

Inter-annotator agreement: Kappa ≥ 75%

Our waiter was friendly and it is a shame that he didn’t have a supportive staff to work with.
### Aspect-term sentiment estimation

#### Our aspect term polarity datasets

<table>
<thead>
<tr>
<th></th>
<th>Restaurants</th>
<th>Laptops</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>2164</td>
<td>805</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>728</td>
<td>196</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2892</td>
<td>1001</td>
</tr>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td><strong>Train</strong></td>
<td>987</td>
<td>866</td>
</tr>
<tr>
<td><strong>Test</strong></td>
<td>341</td>
<td>128</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1328</td>
<td>994</td>
</tr>
</tbody>
</table>
Aspect-term sentiment estimation

Aspect term polarity evaluation: common measures

\[
\begin{align*}
\text{Acc} &= \frac{\text{correctly classified aspect term occurrences}}{|\text{aspect term occurrences}|} \\
\text{Pre}_c &= \frac{\text{correctly classified aspect term occurrences of class } c}{|\text{aspect term occurrences classified as } c|} \\
\text{Rec}_c &= \frac{\text{correctly classified aspect term occurrences of class } c}{|\text{aspect term occurrences of class } c|} \\
F_1_c &= 2 \frac{\text{Pre}_c \text{Rec}_c}{\text{Pre}_c + \text{Rec}_c}, \text{ where } c: +, -, 0, \pm
\end{align*}
\]
The *fajitas* were rather good!!!

The *fajitas* were rather good!!!

The *fajitas* were rather good!!!

Aspect term polarity evaluation in ABSA
We care more about frequently discussed aspect terms, in ABSA

Aspect term polarity evaluation in ABSA

- **food**: *****
- **wine, beers**: **
- **service**: *****
I. For each distinct aspect term $a_i$ we measure its average polarity in all texts
   • E.g., if $a_i$ has 3 positive occurrences, 4 negative and 3 neutral and conflict ones, then: $v_i = \frac{3(1) + 4(-1) + 3(0)}{10} = -0.1$
   • We compute both the predicted $v_i$ and the true $v_i^*$ average polarity.

II. Then, for the $m$ most frequent (distinct) aspect terms:

$$ MAE_m = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} |v_i - v_i^*| $$
Aspect-term sentiment estimation

Aspect term polarity evaluation: mean absolute error

Restaurants

Laptops

Food

wine, beers

service

56/70
Aspect-term sentiment estimation

Aspect term polarity: cumulative frequency

Restaurants

Cumulative Frequency

Number of distinct aspects

Laptops

Cumulative Frequency

Number of distinct aspects
“I hated their fajitas, and their salads”

Problem for sentences containing aspect terms with multiple polarities

“I hated their fajitas, but their salads were great”

Same label for all aspect term occurrences
Aspect-term sentiment estimation

Multi-polar and single-polar sentences

7-14% accuracy loss

6-10% accuracy loss

Restaurants

- #sentences with single polarity: 86%
- #sentences with multiple polarities: 14%

Laptops

- #sentences with single polarity: 91%
- #sentences with multiple polarities: 9%
# Evaluation

## Aspect-term sentiment estimation

<table>
<thead>
<tr>
<th>Teams</th>
<th>Error rate: Restaurants</th>
<th>Error rate: Laptops</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCU</td>
<td>0.191 (1&lt;sup&gt;st&lt;/sup&gt;)</td>
<td>0.295 (1&lt;sup&gt;st&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Median</td>
<td>0.292 (14&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>0.414 (14&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>AUEB</td>
<td>0.318 (16&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>0.427 (16&lt;sup&gt;th&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.357 (21&lt;sup&gt;st&lt;/sup&gt;)</td>
<td>0.486 (22&lt;sup&gt;nd&lt;/sup&gt;)</td>
</tr>
<tr>
<td>Worst</td>
<td>0.583 (24&lt;sup&gt;th&lt;/sup&gt;)</td>
<td>0.635 (23&lt;sup&gt;rd&lt;/sup&gt;)</td>
</tr>
</tbody>
</table>
## Aspect-term sentiment estimation

### Evaluation:

**Restaurants test set**

<table>
<thead>
<tr>
<th>Teams</th>
<th>Error Rate (1-Acc)</th>
<th>$\text{MAE}_{m=50}$</th>
<th>$\text{MAE}_{m=500}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCU</td>
<td>0.191 (1st)/26</td>
<td>0.076 (2nd)/26</td>
<td>0.126 (1st)/26</td>
</tr>
<tr>
<td>NRC</td>
<td>0.199 (2nd)/26</td>
<td>0.062 (1st)/26</td>
<td>0.141 (2nd)/26</td>
</tr>
<tr>
<td>XRCE</td>
<td>0.223 (3rd)/26</td>
<td>0.088 (4th)/26</td>
<td>0.156 (4th)/26</td>
</tr>
<tr>
<td>UWB</td>
<td>0.223 (4th)/26</td>
<td>0.090 (5th)/26</td>
<td>0.143 (3rd)/26</td>
</tr>
<tr>
<td>SZTENLP</td>
<td>0.248 (5th)/26</td>
<td>0.120 (10th)/26</td>
<td>0.164 (5th)/26</td>
</tr>
<tr>
<td>AUEB</td>
<td>0.318 (16th)/26</td>
<td>0.097 (6th)/26</td>
<td>0.194 (8th)/26</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.357 (22nd/26)</td>
<td>0.769 (26th/26)</td>
<td>0.737 (24th/26)</td>
</tr>
</tbody>
</table>

Teams with multiple submissions are shown under one name.
### Evaluation: Laptops test set

<table>
<thead>
<tr>
<th>Teams</th>
<th>Error Rate (1-Acc)</th>
<th>MAE_{m=50}</th>
<th>MAE_{m=500}</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCU</td>
<td>0.295 (1^{st}/26)</td>
<td>0.118 (3^{rd}/26)</td>
<td>0.165 (3^{rd}/26)</td>
</tr>
<tr>
<td>NRC</td>
<td>0.295 (2^{nd}/26)</td>
<td>0.141 (12^{th}/26)</td>
<td>0.160 (2^{nd}/26)</td>
</tr>
<tr>
<td>IITPatan</td>
<td>0.330 (3^{rd}/26)</td>
<td>0.111 (1^{st}/26)</td>
<td>0.178 (4^{th}/26)</td>
</tr>
<tr>
<td>SZTENLP</td>
<td>0.330 (4^{th}/26)</td>
<td>0.124 (4^{th}/26)</td>
<td>0.190 (7^{th}/26)</td>
</tr>
<tr>
<td>UWB</td>
<td>0.333 (5^{th}/26)</td>
<td>0.118 (2^{nd}/26)</td>
<td>0.182 (5^{th}/26)</td>
</tr>
<tr>
<td>AUEB</td>
<td>0.427 (16^{th}/26)</td>
<td>0.147 (13^{th}/26)</td>
<td>0.201 (11^{th}/26)</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.486 (25^{th}/26)</td>
<td>0.704 (25^{th}/26)</td>
<td>0.687 (25^{th}/26)</td>
</tr>
</tbody>
</table>
Aspect-term sentiment estimation

Evaluation: with an ensemble (restaurants)

<table>
<thead>
<tr>
<th>Ensembles</th>
<th>Error Rate</th>
<th>$\text{MAE}_{m=50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1-AUEB</td>
<td>0.202</td>
<td>0.070</td>
</tr>
<tr>
<td>EC2-AUEB</td>
<td>0.196</td>
<td>0.058*</td>
</tr>
<tr>
<td>EC1-UWB</td>
<td>0.198</td>
<td>0.118</td>
</tr>
<tr>
<td>EC2-UWB</td>
<td><strong>0.184</strong></td>
<td><strong>0.075</strong></td>
</tr>
<tr>
<td>Best</td>
<td>0.190</td>
<td>0.076</td>
</tr>
</tbody>
</table>

Same as above, but instead of AUEB, the system uses UWB.
Aspect-term sentiment estimation

Evaluation: with an ensemble (laptops)

<table>
<thead>
<tr>
<th>Ensembles</th>
<th>Error Rate</th>
<th>$\text{MAE}_{m=50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC1-AUEB</td>
<td>0.269★</td>
<td>0.114</td>
</tr>
<tr>
<td>EC2-AUEB</td>
<td>0.282</td>
<td>0.108</td>
</tr>
<tr>
<td>EC1-UWB</td>
<td>0.283</td>
<td>0.100</td>
</tr>
<tr>
<td>EC2-UWB</td>
<td>0.282</td>
<td>0.120</td>
</tr>
<tr>
<td>Best</td>
<td>0.295</td>
<td>0.118</td>
</tr>
</tbody>
</table>
New datasets
New evaluation measure
Message-level sentiment estimation system applied to aspect term sentiment estimation

- Good performance on the aspect term polarity task, especially with MAE or when integrated in an ensemble
- The ‘Aspect term polarity’ subtask of ABSA SemEval 2014 & 2015 was based on the work of this section

Summary & contribution of this section
Contributions of this thesis (1/4)

- Clear decomposition of Aspect Based Sentiment Analysis (ABSA)
  - Systems may compare to each other

- ABSA SemEval task (2014, 2015) based on the work of this thesis
Contributions of this thesis (2/4)

- Introduced 3 new aspect term extraction datasets
  - Laptops/Restaurants/Hotels
  - Showed that domain variety is important

- New aspect term extraction evaluation measures
  - Weighted precision, weighted recall, average weighted precision

- The ‘Aspect term extraction’ subtask of ABSA SemEval 2014 & 2015 was based on the work of this section
Contributions of this thesis (3/4)

- Introduction of a Multi-granular Aspect Aggregation ABSA step
- Two-phase methodology for Multi-granular Aspect Aggregation
- Sense pruning mechanism which improves WordNet-based measures and leads to best performing method
- Publicly available datasets
Contributions of this thesis (4/4)

- 2-stage sentiment estimation system
  - Good generalization ability
  - High rank in Sentiment tasks of SemEval 13/14

- Ensemble of classifiers
  - Best results in ‘Aspect Polarity’ subtask of ABSA task in SemEval ‘14

- Mean Absolute Error evaluation measure
Publications


Thank you!

Questions?