Aspect Term Extraction with History Attention and Selective Transformation

Xin Li¹, Lidong Bing², Piji Li¹, Wai Lam¹, Zhimou Yang³
Presenter: Lin Ma²

¹ The Chinese University of Hong Kong
² Tencent AI Lab
³ Northeastern University

IJCAI 2018

¹ Joint work with Tencent AI Lab
Outline

1 Aspect Term Extraction
   - What is “Aspect Term”? 
   - Problem Formulation

2 The Proposed Model
   - Motivation
   - Our Model

3 Comparative Study
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
Outline

1 Aspect Term Extraction
   - What is “Aspect Term”?
   - Problem Formulation

2 The Proposed Model
   - Motivation
   - Our Model

3 Comparative Study
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
What is “Aspect Term”? 

**Definition:** Explicitly mentioned entities / product attributes in the review sentences where the users express their opinions.

- Also called “Aspect Phrase” or “Opinion Target” in the existing works [4].

**Examples**

- *Its size is ideal and the weight is acceptable.*
- *The pizza is overpriced and soggy.*
Outline

1 Aspect Term Extraction
   - What is “Aspect Term”?
   - Problem Formulation

2 The Proposed Model
   - Motivation
   - Our Model

3 Comparative Study
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
Aspect Term Extraction is to automatically extract the aspect term from user reviews.

As a natural information extraction problem, it can be formulated as a sequence labeling problem or a token-level classification problem.

Examples

<table>
<thead>
<tr>
<th>I</th>
<th>love</th>
<th>the</th>
<th>operating</th>
<th>system</th>
<th>and</th>
<th>the</th>
<th>preloaded</th>
<th>software</th>
</tr>
</thead>
<tbody>
<tr>
<td>O-T</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>T</td>
<td>T</td>
<td>O</td>
<td>O</td>
<td>T</td>
</tr>
<tr>
<td>B-I-O</td>
<td>O</td>
<td>O</td>
<td>O</td>
<td>B</td>
<td>I</td>
<td>O</td>
<td>O</td>
<td>B</td>
</tr>
</tbody>
</table>
1. **Aspect Term Extraction**
   - What is “Aspect Term”? 
   - Problem Formulation

2. **The Proposed Model**
   - Motivation
   - Our Model

3. **Comparative Study**
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
Motivation

1. We still adopt the aspect-opinion joint modeling strategy \[3, 5, 11, 12\] in our model.
   - The existence of opinion (aspect) term can provide indicative clues for finding the collocated / correlated aspect (opinion) term.

2. Local attention and global soft attention have some limitations.
   - Local attention \[3\] can NOT capture the long term dependency between the aspect term and the opinion words.
     - Example: We ordered the special, grilled branzino, that was so infused with bone, it was difficult to eat.
   - Global soft attention \[12\] may introduce some irrelevant information.
The previous predictions can help the current prediction to reduce the error space.

- If the previous prediction is “O”, then current prediction cannot be “I”.
- Some previously predicted commonly-used aspect terms can guide the model to find the co-occurred infrequent aspect terms.
  - Example: Apple is unmatched in **product quality**, **aesthetics**, **craftmanship**, and **customer service**.
  - If we know “**product quality**” is an aspect, then “**aesthetics**” and “**craftmanship**” which belong to the same co-ordinate structure with “**product quality**” are very likely to be aspect terms.
1 Aspect Term Extraction
   • What is “Aspect Term”?
   • Problem Formulation

2 The Proposed Model
   • Motivation
   • Our Model

3 Comparative Study
   • Baselines
   • Main Results
   • Effectiveness of “History Attention” and “Selective Transformation”
Model Overview

Figure: The proposed architecture for Aspect Term Extraction

\[ \tilde{h}_t^A \quad \text{previous history-aware aspect representation} \]
\[ h_t^A \quad \text{history-aware aspect representation} \]
\[ h_{1:t-1}^A \quad \text{previous aspect representation} \]
\[ h_t^A \quad \text{aspect representation} \]
\[ h_t^O \quad \text{opinion representation} \]
\[ \tilde{h}_t^O \quad \text{opinion summary} \]

\[ \mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_t \end{bmatrix} \]

\[ \mathbf{y}_t^A \]

\[ \mathbf{h}_t^A \]

\[ \mathbf{h}_t^O \]

\[ y_t^A \]

\[ \mathbf{h}_i^A \quad \text{aspect representation} \]
\[ \mathbf{h}_i^O \quad \text{opinion summary} \]

THA

Bi-Linear Attention

STN

FC Layer
Core components of the proposed model

- Long Short-Term Memory Networks (LSTMs)
  - Learning word-level representations.
- Truncated History Attention (THA) component
  - Explicitly modeling the aspect-aspect relation based on self-attention.
- Selective Transformation Networks (STN)
  - Making use of global opinion information without introducing too much noise.
The primary goal of THA is to explicitly model the relation between the previous predictions and the current prediction.

- Adding more constraints on the current prediction.
  - E.g., if the previous hidden vector $h_{t-1}$ was predicted as tag “O” then the current tag cannot be “I”.

- Providing more information for the current predictions based on the collocated aspects.
  - Example: Apple is unmatched in product quality, aesthetics, craftsmanship, and customer service.
  - Given the current input “aesthetics”, modeling the relation between it and “product quality” implicitly captures the co-ordinate structure.
Truncated History Attention (THA)

Solutions provided By THA:

1. Calculate the association scores between the previous representations \((h^A_i & \tilde{h}^A_i)\) and the current representation \(h^A_t\) (self-attention):

\[
a^t_i = v^\top \tanh(W_1 h^A_i + W_2 h^A_t + W_3 \tilde{h}^A_i),
\]

\[
s^t_i = \text{Softmax}(a^t_i).
\]

2. Incorporate the aspect history \(\hat{h}^A_t\) into the aspect representation \(h^A_t\):

\[
\hat{h}^A_t = \sum_{i=t-N^A}^{t-1} s^t_i \times \tilde{h}^A_i.
\]

\[
\tilde{h}^A_t = h^A_t + \text{ReLU}(\hat{h}^A_t),
\]
This component tries to make use of the global information without introducing too much noises.

- **Global soft attention [10]:**
  1. Computing association scores between aspect and opinion representations
  2. Aggregating the opinion features based on association scores

- **Local attention [3]:**
  - Assume the aspect is close to its opinion modifier.
  - Only paying attention to a few surrounding words (i.e., opinion representations)
Selective Transformation Networks (STN)

- Our STN:
  - Capture long-term aspect-opinion dependency: make use of the global opinion information.
  - Reduce noises: add more constraint on the opinion representation $h^O_i$ with current aspect representation $\tilde{h}^A_t$.
  - Refine opinion representations $h^O_i$: introduce a residual block [2] to combine the original and the transformed opinion representations.
  - The produced opinion features $\hat{h}^O_t$ is aspect-dependent or time-dependent.

\[
\hat{h}^O_{i,t} = h^O_i + \text{ReLU}(W^A \tilde{h}^A_t + W^O h^O_i),
\]

\[
w_{i,t} = \text{Softmax}(\tanh(\tilde{h}^A_t W_{bi} \hat{h}^O_{i,t} + b_{bi})),
\]

\[
\hat{h}^O_t = \sum_{i=1}^T w_{i,t} \times \hat{h}^O_{i,t}.
\]
1. **Aspect Term Extraction**
   - What is “Aspect Term”?
   - Problem Formulation

2. **The Proposed Model**
   - Motivation
   - Our Model

3. **Comparative Study**
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
Baselines

- CRF and Semi-CRF [7]
- SemEval ABSA winning systems [1, 9, 6, 8]
- LSTMs
- WDEmb [13]
- Memory Interaction Networks (MIN) [3]
- Recursive Neural Conditional Random Fields (RNCRF) [11]
- Coupled Multi-Layer Attention (CMLA) [12]
Outline

1 Aspect Term Extraction
   - What is “Aspect Term”? 
   - Problem Formulation 

2 The Proposed Model
   - Motivation 
   - Our Model 

3 Comparative Study
   - Baselines 
   - Main Results 
   - Effectiveness of “History Attention” and “Selective Transformation”
## Main Results

<table>
<thead>
<tr>
<th>Models</th>
<th>$D_1$ (LAPTOP14)</th>
<th>$D_2$ (REST14)</th>
<th>$D_3$ (REST15)</th>
<th>$D_4$ (REST16)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF-1</td>
<td>72.77</td>
<td>79.72</td>
<td>62.67</td>
<td>66.96</td>
</tr>
<tr>
<td>CRF-2</td>
<td>74.01</td>
<td>82.33</td>
<td>67.54</td>
<td>69.56</td>
</tr>
<tr>
<td>Semi-CRF</td>
<td>68.75</td>
<td>79.60</td>
<td>62.69</td>
<td>66.35</td>
</tr>
<tr>
<td>LSTM</td>
<td>75.71</td>
<td>82.01</td>
<td>68.26</td>
<td>70.35</td>
</tr>
<tr>
<td>IHS_RD ($D_1$ winner)</td>
<td>74.55</td>
<td>79.62</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DLIREC ($D_2$ winner)</td>
<td>73.78</td>
<td>84.01</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EliXa ($D_3$ winner)</td>
<td>-</td>
<td>-</td>
<td>70.04</td>
<td>-</td>
</tr>
<tr>
<td>NLANGP ($D_4$ winner)</td>
<td>-</td>
<td>-</td>
<td>67.12</td>
<td>72.34</td>
</tr>
<tr>
<td>WDEmb (IJCAI 2016)</td>
<td>75.16</td>
<td>84.97</td>
<td>69.73</td>
<td>-</td>
</tr>
<tr>
<td>MIN (EMNLP 2017)</td>
<td>77.58</td>
<td>-</td>
<td>-</td>
<td>73.44</td>
</tr>
<tr>
<td>RNCRF (EMNLP 2016)</td>
<td>78.42</td>
<td>84.93</td>
<td>67.74</td>
<td>69.72*</td>
</tr>
<tr>
<td>CMLA (AAAI 2017)</td>
<td>77.80</td>
<td>85.29</td>
<td>70.73</td>
<td>72.77*</td>
</tr>
<tr>
<td><strong>OURS w/o THA</strong></td>
<td>77.64</td>
<td>84.30</td>
<td>70.89</td>
<td>72.62</td>
</tr>
<tr>
<td><strong>OURS w/o STN</strong></td>
<td>77.45</td>
<td>83.88</td>
<td>70.09</td>
<td>72.18</td>
</tr>
<tr>
<td><strong>OURS w/o THA &amp; STN</strong></td>
<td>76.95</td>
<td>83.48</td>
<td>69.77</td>
<td>71.87</td>
</tr>
<tr>
<td><strong>OURS</strong></td>
<td><strong>79.52</strong></td>
<td><strong>85.61</strong></td>
<td><strong>71.46</strong></td>
<td><strong>73.61</strong></td>
</tr>
</tbody>
</table>

*Table: Experimental results ($F_1$ score, %).*
1 Aspect Term Extraction
   - What is “Aspect Term”?
   - Problem Formulation

2 The Proposed Model
   - Motivation
   - Our Model

3 Comparative Study
   - Baselines
   - Main Results
   - Effectiveness of “History Attention” and “Selective Transformation”
Effectiveness of “History Attention” and “Selective Transformation”

The generated attention scores of our model and our model w/o STN:

(a) OURS

(b) OURS w/o STN.

(c) OURS

(d) OURS w/o STN.
Effectiveness of “History Attention” and “Selective Transformation”

We also compare the output of our model and its variants:

<table>
<thead>
<tr>
<th>Input sentences</th>
<th>Output of LSTM</th>
<th>Output of OURS w/o THA &amp; STN</th>
<th>Output of OURS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. <em>the device speaks about it self</em></td>
<td>device</td>
<td>NONE</td>
<td></td>
</tr>
<tr>
<td>2. <em>Great service!</em></td>
<td>NONE</td>
<td>service</td>
<td></td>
</tr>
<tr>
<td>3. <em>Apple is unmatched in product quality, aesthetics, craftsmanship, and customer service</em></td>
<td>quality, aesthetics, customer service</td>
<td>quality, customer service</td>
<td>product quality, aesthetics, craftsmanship, customer service</td>
</tr>
<tr>
<td>4. <em>I am pleased with the fast log on, speedy WiFi connection and the long battery life</em></td>
<td>WiFi connection, battery life</td>
<td>log, WiFi connection, battery life</td>
<td>log on, WiFi connection, battery life</td>
</tr>
<tr>
<td>5. <em>Also, I personally wasn’t a fan of the portobello and asparagus mole</em></td>
<td>asparagus mole</td>
<td>asparagus mole</td>
<td>portobello and asparagus mole</td>
</tr>
</tbody>
</table>

**Table:** The gold standard aspect terms are underlined and in red.
In this paper, we design a convolution-based framework for Aspect Term Extraction, which achieves state-of-the-art results on four SemEval ABSA datasets.

The proposed THA component explicitly models the aspect-aspect relation for more accurate extraction.

The proposed STN component makes full use of the opinion information without introducing too much noises.
References:


