Sentiment Classification towards Question-Answering with Hierarchical Matching Network

Min Zhang
Soochow University

2019 NExT++ Workshop, NUS
NLP in Soochow University

30 Years  21 Faculties  150+ Students

Application: Platform, System and Application

Apply Research: Machine Translation, Knowledge Graph, Question Answering, Dialogue, Information Extraction, Sentiment Analysis

Fundamental Research (NLU): Lexical, Syntax, Semantic and Discourse (Monolingual and Bilingual)
Outline

• Introduction
• Related Work
• Data Collection & Annotation
• Approach
• Experiment
• Conclusion
Outline

- Introduction
- Related Work
- Data Collection & Annotation
- Approach
- Experiment
- Conclusion
Introduction

**Question 1:** Is the screen clear? How is the battery?

**Answer 1:** It’s a nice phone with high-quality screen. But the battery is not durable.

**Question 2:** Summer is coming, I’m afraid of getting darker. Is the sun cream really effective?

**Answer 2:** No, just depending on my own experience.

Figure 1: Two examples of QA text pairs from “customer questions & answers” section in Amazon.
Introduction

**Three differences** (compared with traditional sentiment classification):

- From sequence to parallel unit
- Both question text and answer text are equally important to carry sentiment tips
- In each QA text pair, the importance degrees of different units/sentences are different
Introduction

Q1: Is the screen clear?  How is the battery?
A1: A nice phone with high-quality screen. But the battery is not durable.

Problem 1:
- Semantic distance
- Conflict instance

Solution 1:
- Segmentation and combination
Introduction

Q1: Is the screen clear? How is the battery?  
A1: It is a nice phone with high-quality screen. But the battery is not durable.

[Q-sentence, A-sentence] units:

- [Is the screen clear? , It is a nice phone with high-quality screen. ]
- [Is the screen clear? , But the battery is not durable.]
- [How is the battery? , It is a nice phone with high-quality screen.]
- [How is the battery? , But the battery is not durable.]
Introduction

Q1: Summer is coming, I’m afraid of getting darker. Is the sun cream really effective?
A1: No, just depending on my own experience.

Problem 2:
- Both question and answer text carry important sentiment tips to predict the polarity.

Solution 2:
- Bidirectional matching mechanism to capture both sentiment information in question and answer text.
Introduction

Q1: Summer is coming, I’m afraid of getting darker. Is the sun cream really effective?
A1: No, just depending on my own experience.

Problem 3:

➤ Importance of different [Q-sentence, A-sentence] units can be different.

Solution 3:

➤ Self-matching attention mechanism to capture the importance of different [Q-sentence, A-sentence] units.
Introduction

Three contributions:

- **Define** a new research problem: QA-style sentiment analysis (This is a real request from real industry)

- **Build** a large-scale annotated corpus for this task (for the first time in research and industry communities, data share)

- **Propose** a hierarchical matching network to address the challenges of this task
Outline

• Introduction
• Related Work
• Data Collection & Annotation
• Approach
• Experiment
• Conclusion
Related Work

- **Word level** (Turney, 2002; Hassan et al., 2011; Tang et al., 2014…)
- **Sentence level** (Choi and Cardie, 2008; Nakagawa et al., 2010; Zhou et al., 2011)
- **Document level** (Li et al., 2010; Blitzer et al., 2007; Xu et al., 2016…)
- **Aspect level** (Wang et al., 2016; Tang et al., 2016; Wang et al., 2018…)

To the best of our knowledge, first attempt to perform sentiment classification on QA-style text level
Outline

• Introduction
• Related Work
• Data Collection & Annotation
• Approach
• Experiment
• Conclusion
Data Collection & Annotation

- **Source**: “Asking All” in Taobao
- **Domain**: Beauty, Shoe and Electronics
- **Categories**
  - Positive
  - Negative
  - Neutral (neither positive nor negative)
  - Conflict (both positive and negative)
Data Collection & Annotation

Several special annotation guidelines of Neutral:

1) Q and A **do not match**:  
   **Q:** *Is the screen clear?*  
   **A:** *The battery life is decent.*

2) A is with an unknown or uncertain answer:  
   **Q:** *What about these sneakers?*  
   **A:** *I do not know, I bought it for my dad.*

3) A is **objective description**:  
   **Q:** *What is the OS of this phone?*  
   **A:** *Android.*
Data Collection & Annotation

Several special annotation guidelines of Non-neutral:

1) Negative Sentimental Expression:
   Q: How is the rock climbing shoe?  
   A: I am so 
   **disappointed**, my feet felt hurt.

2) Positive Sentimental Expression:
   Q: How about the fragrance?  
   A: I am so 
   **satisfied**, it smells distinctive.

3) Depending on both Q&A:
   Q: Will the phone get hot when gaming?  
   A: No.  
   Q: Is the sun cream really economic?  
   A: No.
Data Collection & Annotation

Category distribution of the annotated data in three domains

<table>
<thead>
<tr>
<th></th>
<th>Positive</th>
<th>Negative</th>
<th>Conflict</th>
<th>Neutral</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beauty</td>
<td>3,676</td>
<td>981</td>
<td>318</td>
<td>5,025</td>
<td>10,000</td>
</tr>
<tr>
<td>Shoe</td>
<td>4,025</td>
<td>819</td>
<td>412</td>
<td>4,744</td>
<td>10,000</td>
</tr>
<tr>
<td>Electronic</td>
<td>3,807</td>
<td>1,017</td>
<td>528</td>
<td>4,648</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Kappa consistency check value: 0.84

Data Release: https://github.com/clshenNLP/QASC/
Outline

• Introduction
• Related Work
• Data Collection & Annotation
• Approach
• Experiment
• Conclusion
Overall Architecture of Our Hierarchical Matching Network
Approach

**Step 1**: Segment question and answer text into sentences respectively, and construct [Q-sentence, A-sentence] units.
Step 2: QA bidirectional matching layer: encode each Q&A unit into a vector which contains both sentiment information of question and answer.
Step 3: A self-matching attention layer captures the importance of all [Q-sentence, A-sentence] units.
Approach

We use Bi-LSTM to get contextual representations of two sentences individually.

**Step 2-1**: QA Bidirectional Matching Mechanism
Approach

Bidirectional matching matrix (dot product)

Step 2-2: QA Bidirectional Matching Mechanism
Approach

Answer-to-Question attention (row-wise)

Step 2-3: QA Bidirectional Matching Mechanism
Approach

**Step 2-4**: QA Bidirectional Matching Mechanism

*Question-to-Answer* attention (column-wise)
Approach

Bidirectional matching vector

Step 2-5: QA Bidirectional Matching Mechanism
Step 3: Self-Matching Attention Mechanism
Outline

- Introduction
- Related Work
- Data Collection & Annotation
- Approach
- Experiment
- Conclusion
Experiment

- **Dataset:** 70% training set, 10% development set and 20% test set.
- **Word Embedding:** word2vec
- **Sentence Segmentation:** CoreNLP
- **Evaluation Metric:** Accuracy and Macro-F1
- **Others:**
  - 128 for hidden state
  - 32 for batch size
  - 0.2 for dropout rate
  - 100 for word embedding dimensionality
## Overall performance comparison

<table>
<thead>
<tr>
<th></th>
<th>Beauty</th>
<th></th>
<th>Shoe</th>
<th></th>
<th>Electronic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
</tr>
<tr>
<td>SVM</td>
<td>0.362</td>
<td>0.684</td>
<td>0.381</td>
<td>0.718</td>
<td>0.435</td>
<td>0.691</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.499</td>
<td>0.712</td>
<td>0.520</td>
<td>0.754</td>
<td>0.562</td>
<td>0.715</td>
</tr>
<tr>
<td>Bi-LSTM</td>
<td>0.527</td>
<td>0.719</td>
<td>0.531</td>
<td>0.759</td>
<td>0.574</td>
<td>0.723</td>
</tr>
<tr>
<td>Bidirectional-Match</td>
<td>0.526</td>
<td>0.747</td>
<td>0.557</td>
<td>0.796</td>
<td>0.582</td>
<td>0.741</td>
</tr>
<tr>
<td>AtoQ-Match</td>
<td>0.543</td>
<td>0.745</td>
<td>0.602</td>
<td>0.792</td>
<td>0.567</td>
<td>0.754</td>
</tr>
<tr>
<td>QtoA-Match</td>
<td>0.573</td>
<td>0.751</td>
<td>0.647</td>
<td>0.807</td>
<td>0.608</td>
<td>0.752</td>
</tr>
<tr>
<td>Bidirectional-Match QA</td>
<td>0.583</td>
<td>0.760</td>
<td>0.666</td>
<td>0.815</td>
<td>0.617</td>
<td>0.764</td>
</tr>
<tr>
<td>HMM</td>
<td>0.598</td>
<td>0.776</td>
<td>0.683</td>
<td>0.827</td>
<td>0.640</td>
<td>0.779</td>
</tr>
</tbody>
</table>

Overall Performance Comparison

---
Experiment

<table>
<thead>
<tr>
<th></th>
<th>Beauty</th>
<th></th>
<th>Shoe</th>
<th></th>
<th>Electronic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
</tr>
<tr>
<td>Bidirectional-Match</td>
<td>0.526</td>
<td>0.747</td>
<td>0.557</td>
<td>0.796</td>
<td>0.582</td>
<td>0.741</td>
</tr>
<tr>
<td>Bidirectional-Match QA</td>
<td>0.583</td>
<td>0.760</td>
<td>0.666</td>
<td>0.815</td>
<td>0.617</td>
<td>0.764</td>
</tr>
</tbody>
</table>

- **Bidirectional-Match**: QA bidirectional matching mechanism
- **Bidirectional-Match QA**: sentence segmentation strategy & QA bidirectional matching mechanism

- Sentence segmentation strategy helps to extract the sentiment matching information between the question and answer text
### Experiment

<table>
<thead>
<tr>
<th></th>
<th>Beauty</th>
<th></th>
<th>Shoe</th>
<th></th>
<th>Electronic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
</tr>
<tr>
<td>AtoQ-Match</td>
<td>0.543</td>
<td>0.745</td>
<td>0.602</td>
<td>0.792</td>
<td>0.567</td>
<td>0.754</td>
</tr>
<tr>
<td>QtoA-Match</td>
<td>0.573</td>
<td>0.751</td>
<td>0.647</td>
<td>0.807</td>
<td>0.608</td>
<td>0.752</td>
</tr>
<tr>
<td>Bidirectional-Match QA</td>
<td>0.583</td>
<td>0.760</td>
<td>0.666</td>
<td>0.815</td>
<td>0.617</td>
<td>0.764</td>
</tr>
</tbody>
</table>

- **A to Q-Match**: `Answer-to-Question` attention & sentence segment
- **Q to A-Match**: `Question-to-Answer` attention & sentence segment
- **Bidirectional-Match QA**: QA bidirectional matching mechanism & sentence segmentation strategy

- **Both the question and answer information contribute to sentiment polarity of the QA text pair**
Experiment

<table>
<thead>
<tr>
<th></th>
<th>Beauty</th>
<th>Shoe</th>
<th>Electronic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
</tr>
<tr>
<td>Bidirectional-Match QA</td>
<td>0.583</td>
<td>0.760</td>
<td>0.666</td>
</tr>
<tr>
<td>HMN</td>
<td>0.598</td>
<td>0.776</td>
<td>0.683</td>
</tr>
</tbody>
</table>

- **Bidirectional-Match QA**: QA bidirectional matching mechanism & sentence segmentation strategy
- **HMN**: QA bidirectional matching mechanism & self-matching attention mechanism & sentence segmentation strategy

- **HMN performs best**: both QA bidirectional matching mechanism and self-matching attention mechanism
# Experiment

<table>
<thead>
<tr>
<th></th>
<th>Beauty</th>
<th></th>
<th>Shoo</th>
<th></th>
<th>Electronic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
<td>Macro-F1</td>
<td>Accuracy</td>
</tr>
<tr>
<td>CNN-Tensor</td>
<td>0.500</td>
<td>0.731</td>
<td>0.535</td>
<td>0.765</td>
<td>0.576</td>
<td>0.734</td>
</tr>
<tr>
<td>Attention-LSTM</td>
<td>0.509</td>
<td>0.725</td>
<td>0.571</td>
<td>0.755</td>
<td>0.576</td>
<td>0.721</td>
</tr>
<tr>
<td>BiMPM</td>
<td>0.553</td>
<td>0.745</td>
<td>0.587</td>
<td>0.766</td>
<td>0.584</td>
<td>0.746</td>
</tr>
<tr>
<td>HMN</td>
<td>0.598</td>
<td>0.776</td>
<td>0.683</td>
<td>0.827</td>
<td>0.640</td>
<td>0.779</td>
</tr>
</tbody>
</table>

- **CNN-Tensor** (Lei et al., 2015): sentence level
- **Attention-LSTM** (Wang et al., 2016): aspect level
- **BiMPM** (Wang et al., 2017): QA matching
- **HMN**: our proposed approach

- Matching strategy is important for QA-style sentiment classification, which is the unique challenge in sentiment classification task.
Outline

• Introduction
• Related Work
• Data Collection & Annotation
• Approach
• Experiment
• Conclusion
Conclusion

- **Define** a new research problem: QA-style sentiment analysis (This is a real request from real industry)

- **Build** a large-scale annotated corpus for this task (for the first time in research and industry communities, data share)

- **Propose** a hierarchical matching network to address the challenges of this task
Conclusion

Next Step:

Q&A Style SA ➡️ Dialogue-style SA
Supervised ➡️ Semi-/Un-Supervised
Sentence Sentiment ➡️ Aspect and Stance
Domain Adaptation (Cross-Domain/Unsupervised)
Acknowledgement

- Chenlin SHEN, Jingjing WANG, Shoushan LI, Min Zhang, Guodong ZHOU from Soochow University

- Luo SI, Xiaozhong LIU, Changlong SUN, Yangyang KANG from Alibaba Group
Thanks for your attention!

Data and Code Release: https://github.com/clshenNLP/QASC/
Aspect Data Annotation

- **Data Collection:**
  - Collected QA style reviews consist of three domains: *Bags, Cosmetics and Electronics*.
  - Each domain contains 10k QA text pairs.

- **Data Annotation:**
  - Each QA style review is annotated with **two tuples**:

  **Tuple (Aspect Term, Polarity)**
  
  **Example 1:**
  - Question: How is **operating speed** of this phone?
  - Answer: Quite **obtuse** and **slow**.

  - Annotate:
    - (operating speed, negative)
    - (System Performance, negative)

  **Tuple (Aspect Category, Polarity)**
  
  **Example 2:**
  - Question: How is **battery life**?
  - Answer: Very **durable**.

  - Annotate:
    - (battery life, negative)
    - (Battery, negative)
Select **10k QA text pairs** from each domain to perform annotation.

**Totally annotate 1415 aspect terms.**

**Define 15, 16, 10 aspect categories** for three domains, i.e., **Bags, Cosmetics and Electronics** respectively.

---

### Table 1: The defined aspect categories in each domain.

<table>
<thead>
<tr>
<th>Domains</th>
<th>Aspect Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bags</strong></td>
<td>Size, Price, Appearance, Quality, Weight, Certified Products, Smell, Accessories, Material, Life Timer, Style, Workmanship, Color, Stain Resistant, Practicality</td>
</tr>
<tr>
<td><strong>Cosmetics</strong></td>
<td>Price, Efficacy, Moisturizing Performance, Certified Products, Adverse Reaction, Exfoliator, Texture, Long Lasting, Smell, Material, Noticeable Color, Quality, Colour, Touch, Skin Whitening, Acne</td>
</tr>
<tr>
<td><strong>Electronics</strong></td>
<td>System Performance, Appearance, Battery, Computing (e.g., cpu, gpu, tpu etc.), Certified Products, Quality, IO (e.g., keyboard, screen, etc.), Price, Storage, Function (e.g., touch id, waterproof etc.)</td>
</tr>
</tbody>
</table>

---

### Table 2: Statistics of the corpus (Pos., Neg. and Neu. denote the number of positive, negative and neutral for aspect term; #Term denotes the number of aspect term).

<table>
<thead>
<tr>
<th>Domains</th>
<th>Pos.</th>
<th>Neg.</th>
<th>Neu.</th>
<th>All</th>
<th>#Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bags</td>
<td>2503</td>
<td>724</td>
<td>453</td>
<td>3680</td>
<td>451</td>
</tr>
<tr>
<td>Cosmetics</td>
<td>2834</td>
<td>956</td>
<td>503</td>
<td>4293</td>
<td>621</td>
</tr>
<tr>
<td>Electronics</td>
<td>2742</td>
<td>821</td>
<td>531</td>
<td>4094</td>
<td>343</td>
</tr>
</tbody>
</table>
Reinforced Bidirectional Attention Network

Step 1. **Word Encoder:** to encode the question words and answer words.

Two *Word Selection Model (RAWS)* are used to discard the noisy words inside both question and answer. *(Introduced Next)*
Step 2. Reinforced Bidirectional Attention: to capture the semantic matching information between question and answer.
Reinforced Bidirectional Attention Network

Step 3. Softmax Decoder:
To predict the sentiment polarity
The Word Selection Model: Reinforced Aspect-relevant Word Selector (RAWS)

RAWS: aims to discard noisy words and only select aspect-relevant words inside a word sequence for a specific aspect.

Problem: Word selection functions as "hard" attention mechanism. Thus, cannot be directly optimized through back-propagation due to the non-differentiable problem.

Solution: Model the word selection model via Reinforcement Learning algorithm, i.e., Policy Gradient.
Experimental Results

<table>
<thead>
<tr>
<th>Approaches</th>
<th>Term-level ASC-QA</th>
<th>Category-level ASC-QA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Bags</td>
<td>Cosmetics</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>Acc.</td>
</tr>
<tr>
<td>LSTM (Wang et al., 2016)</td>
<td>0.571</td>
<td>0.757</td>
</tr>
<tr>
<td>RAM (Chen et al., 2017)</td>
<td>0.605</td>
<td>0.782</td>
</tr>
<tr>
<td>GCAE (Xue and Li, 2018)</td>
<td>0.617</td>
<td>0.779</td>
</tr>
<tr>
<td>S-LSTM (Wang and Lu, 2018)</td>
<td>0.615</td>
<td>0.824</td>
</tr>
<tr>
<td>BIDAF (Seo et al., 2016)</td>
<td>0.613</td>
<td>0.815</td>
</tr>
<tr>
<td>HMN (Shen et al., 2018a)</td>
<td>0.607</td>
<td>0.817</td>
</tr>
<tr>
<td>MAMC (Yin et al., 2017)</td>
<td>0.621</td>
<td>0.825</td>
</tr>
<tr>
<td>RBAN w/o RAWS</td>
<td>0.623</td>
<td>0.826</td>
</tr>
<tr>
<td>RBAN w/o Q2A</td>
<td>0.595</td>
<td>0.788</td>
</tr>
<tr>
<td>RBAN w/o A2Q</td>
<td>0.623</td>
<td>0.837</td>
</tr>
<tr>
<td><strong>RBAN</strong></td>
<td><strong>0.648</strong></td>
<td><strong>0.856</strong></td>
</tr>
</tbody>
</table>

Table 3: Performances of all the approaches to two sub-tasks, i.e., **Term-level** and **Category-level** ASC-QA. In each sub-task, all approaches are evaluated in three different domains, i.e., **Bags**, **Cosmetics** and **Electronics**.