OPINION MINING OF MICROBLOG EVENTS

Zhou Xiao
1. Background

2. Opinion Mining Track on Chinese Microblog
   - 2.1 Introduction
   - 2.2 Opinionated Sentence Identification
   - 2.3 Opinion Target Extraction

3. Discussion & Summary
1. Background (1/3)

- Social media is becoming a major trend\(^1\)
  - Over 270 million microblog users in China
  - Rapid growth on mobile devices

- Opinion mining on user generated result has been studied extensively in the past decades
  - Lots of work in both academic and industry world

1. CNICC: http://www.cnnic.cn
1. Background (2/3)
1. Background (3/3)

- Most existing methods work well in specific fields, such as product
- Microblog environment raise new challenge
  - Lots of data
  - Informal writing style
  - Open field
Contents

1. Background

2. Opinion Mining Track on Chinese Microblog
   - 2.1 Introduction
   - 2.2 Opinionated Sentence Identification
   - 2.3 Opinion Target Extraction

3. Discussion & Summary
2.1 Introduction

- **Opinion Mining Track on Chinese Microblog**
  - 34 Teams Participated, Including high schools and companies
  - 3 tasks
    - Task 1: Opinionated Sentence Identification
    - Task 2: Sentence Polarity
    - Task 3: Opinion Target Extraction
  - 20 events, range from entertainment to society
  - For every event, about 1000 microblogs are tested
2.1 Introduction

- We participated in two tasks:
  - Task 1
    | Precision | Recall | F-Score |
    |-----------|--------|---------|
    | 0.671     | 0.944  | 0.784   |
  - Task 3
    | Precision | Recall | F-Score |
    |-----------|--------|---------|
    | Restrict  | 0.160  | 0.160   | 0.160   |
    | Loos      | 0.290  | 0.220   | 0.250   |
2.2 Opinionate Sentence Identification

- **Preprocess**
  - Identify different kinds of text
    - URL
    - Retweet
    - User name(@somebody)
    - Hashtag(“#”)
    - Normal text
  - Word segmentation, POS tagging, dependency parsing on normal text
  - Opinion words identification
    - Construct Opinion word dictionary
    - Add Web words and slangs
    - Use patterns to match phrase
2.2 Opinionated Sentence Identification

- Use Classification

  - Unbalance distribution data: VFI (Voting Feature Interval) classifier doesn’t use prior probability

- Various Features

  - POS tag: number of pronoun, number of conjunction, etc.
  - Punctuation: “!”, “?”
  - Sentiment word: number of sentiment words
2.3 Opinion Target Extraction

- Also based on VFl classifier

- Extract all the candidate targets in microblog $m$
- For every candidate and every sentence in $m$, construct an instance
- Label is true: the candidate is correct target for that sentence

- Candidate targets
  - Noun phrases in microblog
  - Nouns in hashtag
2.3 Opinion Target Extraction

- Features
  - For an instance \(<t, s>, \) where \(t\) is a candidate target, \(s\) is a sentence
    - \(t\)'s syntactic feature
    - Distance between \(t\) and \(s\)
    - Position of \(t\): whether it appears between a hashtag, etc.
    - Frequencies of \(t\)
Contents

- 1. Background
- 2. Opinion Mining Track on Chinese Microblog
  - 2.1 Introduction
  - 2.2 Opinionated Sentence Identification
  - 2.3 Opinion Target Extraction
- 3. Discussion & Summary
3. Discussion & Summary

- However, by using larger dataset, we can use simpler method to get more readable result.
## 3. Discussion & Summary

<table>
<thead>
<tr>
<th>Aspect</th>
<th>Opinions</th>
</tr>
</thead>
<tbody>
<tr>
<td>日本 鬼子  (“Japanese” “devil”)</td>
<td>痛恨 讨厌 仇恨 不是人 变态  (“abhor” “hate” “hatred” “inhuman” “sick”)</td>
</tr>
<tr>
<td>电影  (“movie”)</td>
<td>便宜 不错 最好 精彩 别离  (“inexpensive” “not bad” “best” “excellent” “leave”)</td>
</tr>
<tr>
<td>南京 日军  (“Nanjing” “Japanese army”)</td>
<td>审判 强奸 灾难 好听 熟悉的  (“judge” “rape” “disaster” “sound” “familiar”)</td>
</tr>
<tr>
<td>张艺谋  (“Yimou Zhang”)</td>
<td>最好 最新 战争 感谢 不愧  (“best” “new” “war” “grateful” “deserve”)</td>
</tr>
<tr>
<td>影院 电影院  (“movie theater” “cinema”)</td>
<td>值得 没去 光明 提高 支持  (“worthy” “haven’t been there” “bright” “elevation” “support”)</td>
</tr>
<tr>
<td>敌人 金陵  (“enemy” “NanKing”)</td>
<td>完了 好看 刚刚 准备 很好  (“doom” “interesting” “just now” “preparing” “very good”)</td>
</tr>
<tr>
<td>影片 片子  (“movie” “picture”)</td>
<td>不错 最佳 沉重的 最好 热门  (“not bad” “peak” “heavy” “best” “hot”)</td>
</tr>
</tbody>
</table>
3. Discussion & Summary

Experience

• Include WEB words and slangs in your dictionary
• Use simple method
• Combine different strategies (machine learning, rule based method, statistical information)

Shortage

• The result is still not desirable
• It’s hard to manually maintain a WEB dictionary
• High level NLP tools (such as parsing) are slow and inefficient
Thanks!