Contribution

- Introduce a novel task of discovering emerging entities (EEs) in microblogs
- Revisit the definition of EEs so that it does not depend on temporal resource (KB etc.)
- Propose a method that can discover EEs accurately, and quickly

Introduction

- Emerging entities (EEs) are appearing ceaselessly in microblogs real-time
- Recognizing those EEs is important for applications such as social listening

Task definition

Find EEs as many and early as possible from microblog posts

Challenges:

- How to solidly define EEs without relying on temporal resources?
- How to discover diverse EEs including long-tail ones?
- How to discover EEs early while their frequencies are low?

Resource-independent definition of EEs

Define EEs only in terms of how people describe their contexts (emerging contexts (EC)):

- Emerging Contexts: Contexts in which the writer assumed the reader does not know the existence of the entities
- Emerging Entities: Entities in the state of being still observed in EC

By capturing those ECs that are independent of any resources, we can discover diverse EEs early since EEs are common whether they are long-tail or not, and appear in the early-stage of their appearance

(LSTM-)CRF based on timely distant supervision

Collect early-stage microblog posts where EEs are likely to appear by using distant supervision on time-series posts

1. Retrieve first N reposts where the titles of Wikipedia articles appear as EEs with EEs (Pos. ex.)
2. Retrieve last N posts one year after the time of collecting ECs with collected EEs (Neg. ex.)

Experiments

- Training data:
  - 222,092 posts collected using proposed method from the period of 2012 to 2015 in our Twitter archive
- Methods:
  - Proposed: (LSTM-)CRF using the training data
  - Baseline: Output NEs found by NER that are absent in a dictionary or Wikipedia (Baseline1), or in the past tweets (Baseline2)
- Evaluation:
  - Precision: Evaluate top-500 discovered entities from daily retweets, which are sorted using confidence scores of each method
  - (Relative) Recall: Evaluate how many target entities registered in Wikipedia from Jan. ’17 to Jun. ’18 could be discovered from all the retweets that include those entity surfaces

Examples of discovered EEs and ECs

<table>
<thead>
<tr>
<th>Discovered EE</th>
<th>Example contexts of entities (input tweet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAGON BALL</td>
<td>New fighting game DRAGON BALL FighterZ is announced! Trailer is become public. Expected to appear early 2018.</td>
</tr>
<tr>
<td>Kayakku LIVING</td>
<td>Kayakku established WEB company by business acquisition, which is related to housing named “Kayakku LIVING”</td>
</tr>
</tbody>
</table>

Resource-independent definition of EEs

Define EEs only in terms of how people describe their contexts (emerging contexts (EC)):

- Emerging Contexts: Contexts in which the writer assumed the reader does not know the existence of the entities
- Emerging Entities: Entities in the state of being still observed in EC

By capturing those ECs that are independent of any resources, we can discover diverse EEs early since EEs are common whether they are long-tail or not, and appear in the early-stage of their appearance

(LSTM-)CRF based on timely distant supervision

Collect early-stage microblog posts where EEs are likely to appear by using distant supervision on time-series posts

1. Retrieve first N reposts where the titles of Wikipedia articles appear as EEs with EEs (Pos. ex.)
2. Retrieve last N posts one year after the time of collecting ECs with collected EEs (Neg. ex.)

Experiments

- Training data:
  - 222,092 posts collected using proposed method from the period of 2012 to 2015 in our Twitter archive
- Methods:
  - Proposed: (LSTM-)CRF using the training data
  - Baseline: Output NEs found by NER that are absent in a dictionary or Wikipedia (Baseline1), or in the past tweets (Baseline2)
- Evaluation:
  - Precision: Evaluate top-500 discovered entities from daily retweets, which are sorted using confidence scores of each method
  - (Relative) Recall: Evaluate how many target entities registered in Wikipedia from Jan. ’17 to Jun. ’18 could be discovered from all the retweets that include those entity surfaces

Examples of discovered EEs and ECs

<table>
<thead>
<tr>
<th>Discovered EE</th>
<th>Example contexts of entities (input tweet)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAGON BALL</td>
<td>New fighting game DRAGON BALL FighterZ is announced! Trailer is become public. Expected to appear early 2018.</td>
</tr>
<tr>
<td>Kayakku LIVING</td>
<td>Kayakku established WEB company by business acquisition, which is related to housing named “Kayakku LIVING”</td>
</tr>
</tbody>
</table>