A Continuously Growing Dataset of Sentential Paraphrases

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu

The Ohio State University, University of Pennsylvania, University of Maryland

Introduction

- Rich application usage of paraphrase
- A simple but powerful URL based data collection method
- Collecting >30k sentential paraphrases with ~70% precision per month
- Largest annotated paraphrase corpus to date

Paraphrase Corpus

Comparison between our data and two existing corpus (MSRP [Dolan et al. 2005] and PIT-2015 [Xu et al. 2015]).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>MSRP</th>
<th>PIT-2015</th>
<th>Our work</th>
</tr>
</thead>
<tbody>
<tr>
<td>corpora size</td>
<td>6k</td>
<td>14k</td>
<td>51k</td>
</tr>
<tr>
<td>+/- balance</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>manual inspection</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>over-identification</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>high precision</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>dynamic growing</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Our Method

Step 1: Collecting tweets that refer to the same URL for account in [CNN, BBC, New York Times et al.]: for each tweet in [account.tweets_of_yesterday]:

- if tweet contains URL:
  - search_tweet = Twitter.search(key=URL)
  - search_tweet.clean()

- if NotOverlap(tweet, search_tweet):
  - candidate.append(search_tweet)

Step 2: Labelling for gold standard corpus

Sampling ~51k candidate sentence pairs Labeled by Amazon Mechanical Turk

Step 3: Training

Splitting 42k for training and 9k for testing

Run various paraphrase identification models

Get best classifier M

Step 4: Predicting

M.predict(unlabeled data)

Analysis

1) Distributional lexical dissimilarity (PINC [Chen et al. 2011])

- Lexically divergent in our URL dataset

2) Distributional semantic similarity (OrMF [Guo et al. 2014])

- Semantically coherent in our URL dataset

Twitter URL Corpus

- Original Tweet: Singer-songwriter Bob Dylan awarded the 2016 Nobel Prize in Literature.

- Paraphrase: The Nobel Prize for Literature goes to American singer-songwriter Bob Dylan.

- Non-Paraphrase: Bob Dylan wins 2016 Nobel Prize for Literature.

Analysis

1) Distributional lexical dissimilarity (PINC [Chen et al. 2011])

- Lexically divergent in our URL dataset

2) Distributional semantic similarity (OrMF [Guo et al. 2014])

- Semantically coherent in our URL dataset

Phrasal Paraphrase

- a 15-year-old girl, a 15yr old, a 15 y/o girl
- fetuses, fetal tissue, miscarried fetuses

Data and Code

- https://github.com/lanwuwei/paraphrase-dataset

Well balanced in our URL dataset