Towards Real-Time Summarization of Scheduled Events from Twitter Streams

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, Julio Gonzalo

arkeitz.zubiaga@qc.cuny.edu, {damiano,enrique,julio}@lsi.uned.es

23rd ACM Conference on Hypertext and Social Media (HT2012)

Introduction

• Users live-tweet about events they are following in real-time.
• The overwhelming amount of tweets makes it difficult for the user:
  – to follow the full stream while finding out about new sub-events
  – to retrieve from Twitter the main, summarized information about which are the key things happening at the event.
• Real-time summarization of events enables to follow streams and to find out about main sub-events happening during the event.

Dataset

• Soccer competition: #copaamerica 2011.
• 26 soccer games - 1.4M+ tweets.
• Multilingual (majority of es, pt & en).
• Reference summaries from Y! Sports:
  – Key sub-events (goals, red cards, penalties,...)
  – Timestamps.

System

Sub-event detection

By looking at the histogram in real-time, are we observing an outstanding activity right now?
1. Increase (based on recent activity).
2. Outliers (based on full histogram).

Tweet selection

If a sub-event happened right now, what tweet describes it?
1. TF (representative tweet right in the moment).
2. KLD (representative tweet comparing to whole game).

Results

Sub-event detection

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
<th>Redundancy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase</td>
<td>0.29</td>
<td>0.81</td>
<td>0.41</td>
<td>45.4</td>
</tr>
<tr>
<td>Outliers</td>
<td>0.51</td>
<td>0.84</td>
<td>0.63</td>
<td>25.6</td>
</tr>
</tbody>
</table>

Tweet selection

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF</td>
<td>es 0.79</td>
<td>en 0.74</td>
</tr>
<tr>
<td>KLD</td>
<td>es 0.84</td>
<td>en 0.77</td>
</tr>
</tbody>
</table>

Conclusions

• Real-time summaries with precision and recall values above 80%.
• The fact that users tweet at the same time with overlapping vocabulary helps not only detect that a sub-event occurs, but also select a tweet to describe it.
• Considering previous information seen during the event is helpful to this end, rather than only consider the most recent activity.

Acknowledgements: This work has been partly-funded by MA2VICMR (S-2009/TIC-1542), the Holopedia project (TIN2010-21126-C02-01), the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 288024 (LiMoSINe project) and the Spanish Ministry of Education for a doctoral grant (AP2009-0507).