Towards Real-Time Summarization of Scheduled Events from Twitter Streams

Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, Julio Gonzalo

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

City University of New York
New York, NY, USA

UNED NLP & IR Group
Madrid, Spain

23rd ACM Conference on Hypertext and Social Media (HT2012)
Milwaukee, WI, USA. June 25-28, 2012.

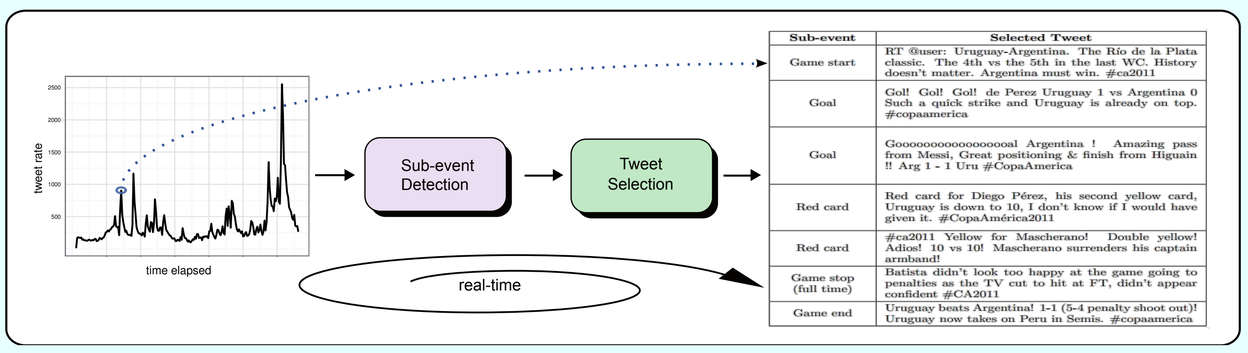
Introduction

- **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	Selected Tweet
Game start	RT @user: Uruguay-Argentina. The Río de la Plata classic. The 4th vs the 5th in the last WC. History doesn't matter. Argentina must win. #ca2011
Goal	Gol! Gol! Gol! de Perez Uruguay 1 vs Argentina 0 Such a quick strike and Uruguay is already on top. #copaamerica
Goal	Goooooooooooooooal Argentina ! Amazing pass from Messi, Great positioning & finish from Higuain !! Arg 1 - 1 Uru #CopaAmerica
Red card	Red card for Diego Pérez, his second yellow card, Uruguay is down to 10, I don't know if I would have given it. #CopaAmérica2011
Red card	#ca2011 Yellow for Mascherano! Double yellow! Adios! 10 vs 10! Mascherano surrenders his captain armband!
Game stop (full time)	Batista didn't look too happy at the game going to penalties as the TV cut to hit at FT, didn't appear confident #CA2011
Game end	Uruguay beats Argentina! 1-1 (5-4 penalty shoot out)! Uruguay now takes on Peru in Semis. #copaamerica

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

	Precision	Recall	F1	Redundancy
Increase	0.29	0.81	0.41	45.4
Outliers	**0.51**	**0.84**	**0.63**	**25.6**

Tweet selection

Recall			
	es	**en**	**pt**
TF	0.79	0.74	0.78
KLD	**0.84**	**0.77**	**0.82**

$$\text{Recall} = \frac{\text{sub-events in reference}}{\text{correct tweets in summary}}$$

Precision			
	es	**en**	**pt**
TF	0.79	0.74	0.79
KLD	**0.84**	**0.79**	**0.83**

$$\text{Precision} = \frac{\text{correct + novel tweets in summary}}{\text{total tweets in summary}}$$

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.
- **Regardless of the audience** tweeting about an event, **our method effectively reports** the key sub-events occurred during a game.
- The **most important sub-events**, such as goals and game ends, are **reported almost perfectly**.

Acknowledgements: This work has been part-funded by MA2VICMR (S-2009/TIC-1542), the Holopedia project (TIN2010-21128-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).