LIA/LINA at the INEX 2012 Tweet Contextualization track

Romain Deveaud* and Florian Boudin**

* LIA – University of Avignon

** LINA – University of Nantes

UNIVERSITÉ D'AVIGNON

UNIVERSITÉ DE NANTES

Introduction

- definition of a « context » ?
  - informational context
  - set of informative sentences related to an expression (e.g. a tweet, a query...)

- between focused IR and automatic summarization

Introduction (2)

- using the best retrieved documents as a pool of candidate sentences
- scoring these sentences with various metrics
- also used the URLs in the tweets

Outline

- Introduction
- Finding relevant candidate sentences
- Sentence scoring
- Results
- Conclusions and future work

Finding relevant candidate sentences

- relevant Wikipedia articles must contain relevant sentences
- yet another ad-hoc task, where all sentences in the top 5 articles are our candidates
- but...

#HashtagSplitting and Tweet cleaning

- tweets are dirty

... actually this one is pretty clean

#HashtagSplitting and Tweet cleaning (2)

- hashtags are very important pieces of information, just like user tags
- #whitneyhouston ? => [‘whitney’, ‘houston’]
- we also removed all RT, @, urls and stopwords (INQUERY stoplist)
- the final output is like a TREC topic
  - hashtags words as `<title>`, rest of the tweet as `<desc>`

Wikipedia articles retrieval

$$s(H_{\mathcal{T}}, Q_{\mathcal{T}}, D) = \lambda \times score_{MRF}(H_{\mathcal{T}}, D) + (1 - \lambda) score_{MRF}(Q_{\mathcal{T}}, D)$$

- empirically, λ = 0.8
- Markov Random Field (MRF) [Metzler & Croft, 2005]
  - modeling query unigram and bigram occurencies in the documents
- retrieving the top 5 articles and considering all the sentences as a set of candidates

Outline

- Introduction
- Finding relevant candidate sentences
- Sentence scoring
- Results
- Conclusions and future work

Sentence scoring

- we need to score each sentence in order to generate the « context »
- several metrics, mostly based on similarities between the tweet and the candidate sentences
- also used tweeted URLs for one run... but realised to late that it was counted as manual

Summarization run

- given S a candidate sentence, $H_{\mathcal{T}}$ a set of hashtags words and $Q_{\mathcal{T}}$ the « clean » tweet

$$m_1(S) = \frac{|S \cap \mathcal{T}|}{|min(S, \mathcal{T})|} \qquad m_2(S) = cos(S, \mathcal{T})$$

- $m_3(S)$ is the TextRank of sentence S [Mihalcea & Tarau, 2004]

$$m_1(S) \times m_2(S) \times m_3(S) \times p(D|\mathcal{T})$$

Using tweeted URLs

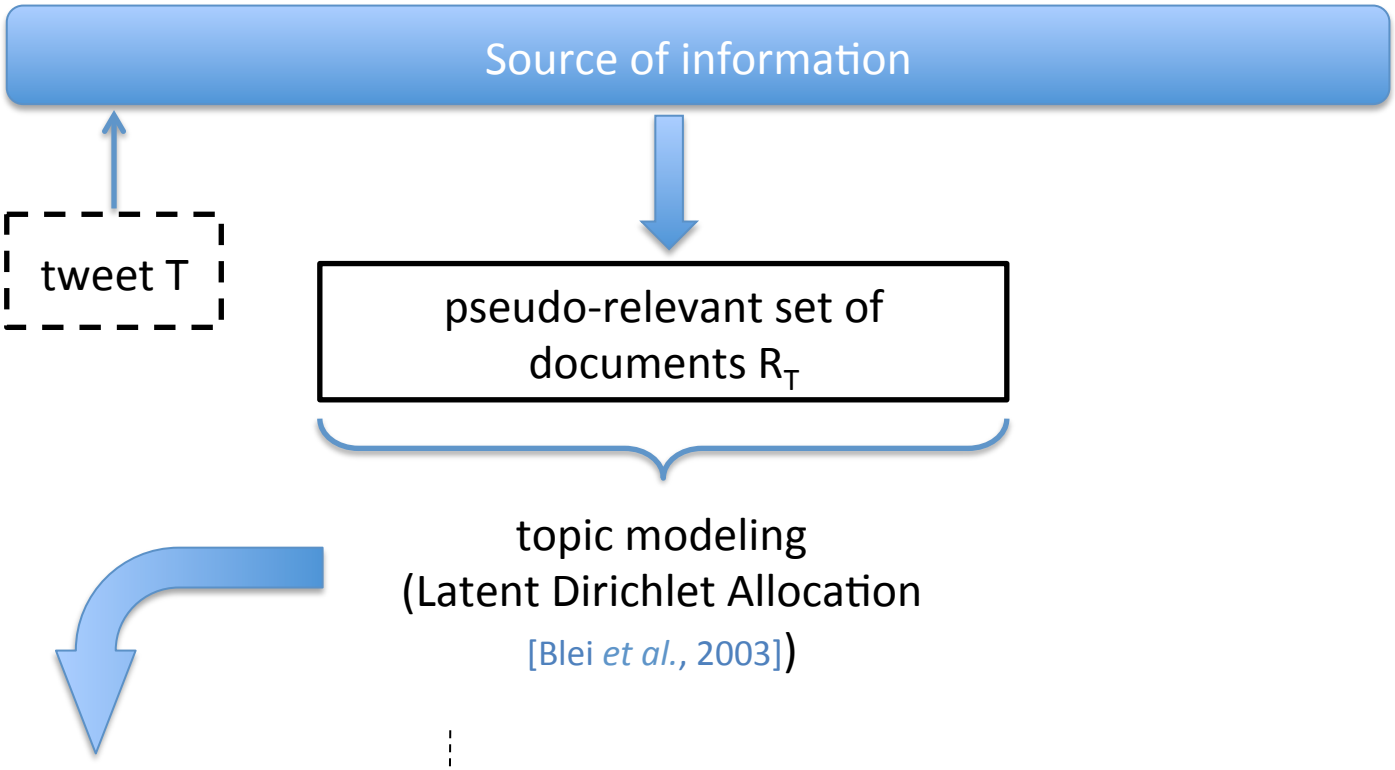
- some tweets have embedded URLs
  - we fetch the title and the body of the web page
  - considered as manual run

$$m_4(S) = \frac{|S \cap U_{title}|}{|min(S, U_{title})|} \qquad m_5(S) = \frac{|S \cap U_{body}|}{|min(S, U_{body})|}$$

$$m_6(S) = cos(S, U_{title}) \qquad m_7(S) = cos(S, U_{body})$$

$$m_1(S) \times \underbrace{m_4(S) \times m_5(S) \times m_6(S) \times m_7(S)}_{\text{URLs only}} \times p(D|\mathcal{T})$$

Modeling tweet concepts



- set of LDA topics as informative concepts
- can be viewed as weighted sentences

Modeling tweet concepts (2)

$P(w\|k_1)$	w
0.17647058823528927	boots
0.13235294117646695	ice
0.11764705882353424	aircraft
0.11764705882352615	deicing
0.1029411764705973	air
0.10294117647058541	systems
0.05882352941177073	wings
0.058823529411763095	electrothermal
0.058823529411763095	bleed

$P(w\|k_2)$	w
0.2500000000000005	jet
0.14285714285714315	engines
0.10714285714285737	airliner
0.09523809523808895	aircraft
0.08333333333333677	airliners
0.07142857142857154	passenger
0.07142857142857154	boeing
0.059523809523809645	powered
0.059523809523809645	airlines

$\sum_d P(k_1|d)P(d|\mathcal{T})$ = 0.37196633186504946

$\sum_d P(k_2|d)P(d|\mathcal{T})$ = 0.6280336681349504

Modeling tweet concepts (3)

- two sources of information for modeling concepts :
  - Wikipedia as encyclopedic source (recent may 2012 dump)
  - ClueWeb09 category B (without spam [Cormack *et al.*, 2011]*) as web source

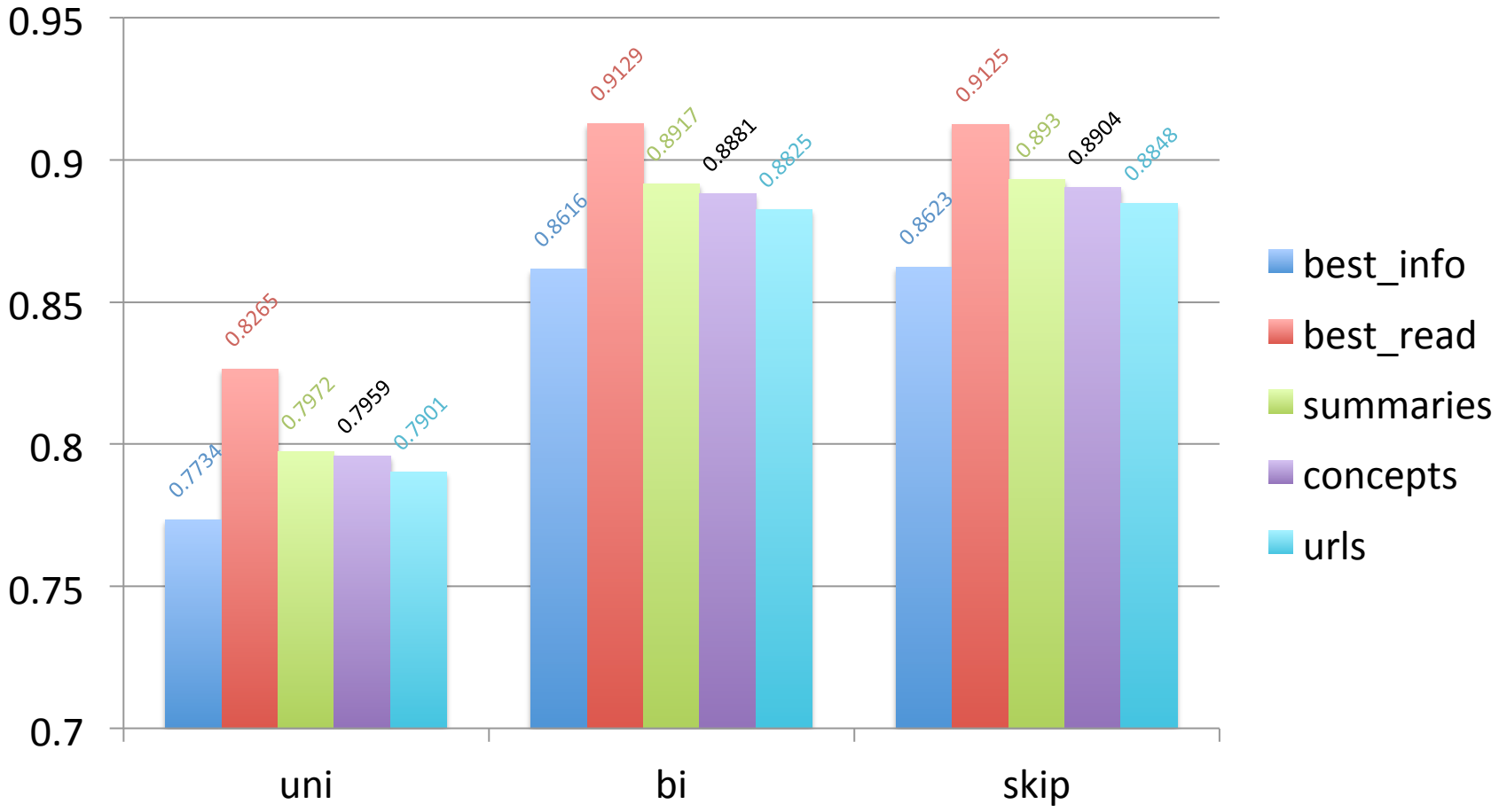
$$\sigma(S) = \frac{1}{|K_{\mathcal{T}}|} \sum_{k \in K_{\mathcal{T}}} \left(\sum_{d} P(k|d) P(d|\mathcal{T}) \sum_{w \in S} p(w|k) \log \frac{N}{df_w} \right)$$

* **http://plg.uwaterloo.ca/~gvcormac/clueweb09spam/**

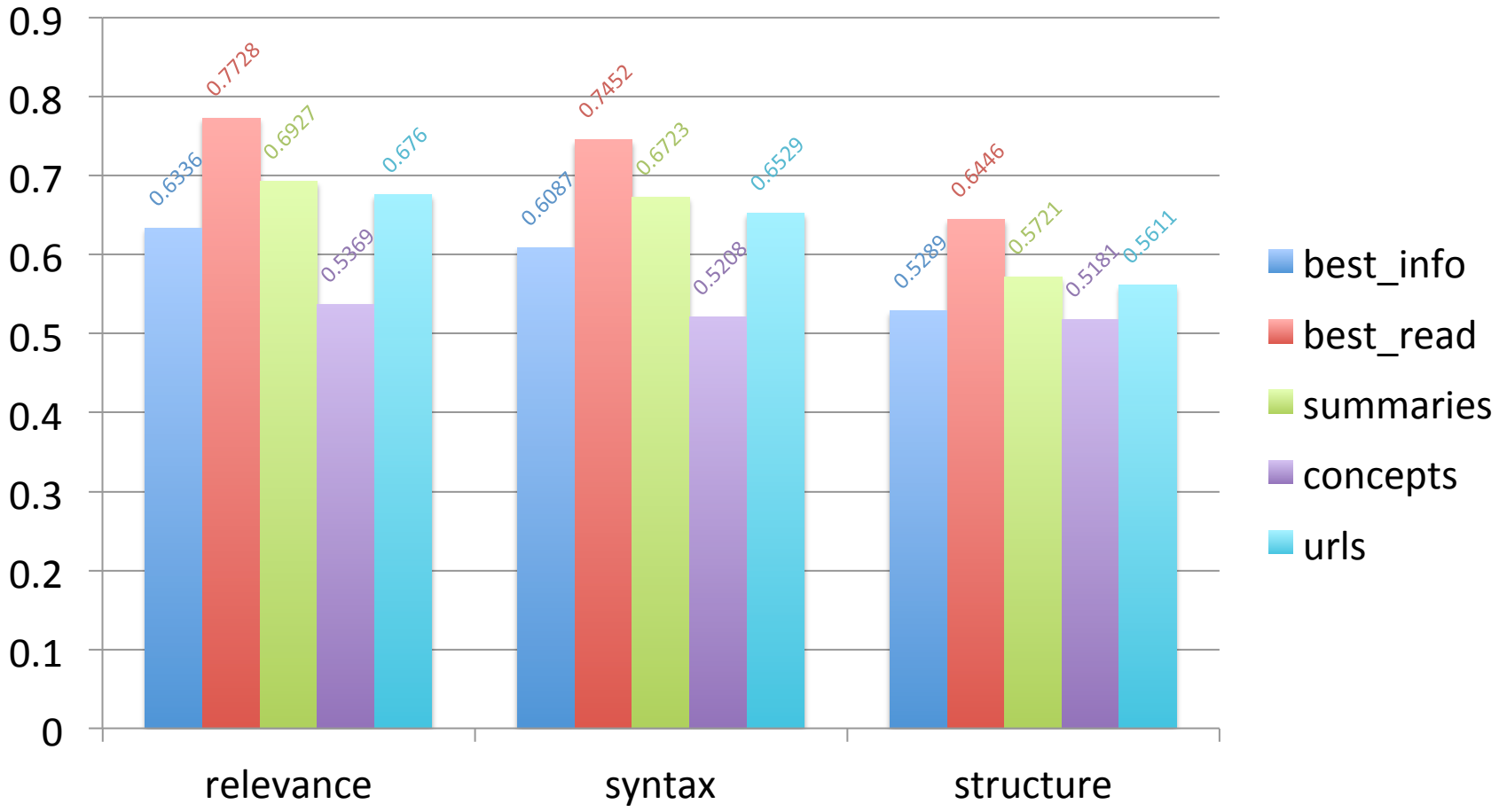
Outline

- Introduction
- Finding relevant candidate sentences
- Sentence scoring
- Results
- Conclusions and future work

Informativeness



Readability



Outline

- Introduction
- Finding relevant candidate sentences
- Sentence scoring
- Results
- Conclusion

Conclusion

- all approaches performed well
- both informativeness and readability were good
  - improvements needed for the conceptual metrics
- further exploration of focused IR methods to effectively retrieve sentences before applying a scoring function

Conclusion (2)

- another evaluation with more difficult tweets?

- future of the track?
  - contextual diversity
  - exploring tweet facets
  - using the RT and reply network of each tweet
  - …

may I contextualize your questions?

The "Airbus A380" is a double-deck, wide-body, four-engine jet airliner manufactured by the European corporation Airbus, a subsidiary of EADS.

The aircraft was known as the "Airbus A3XX" during much of its development, before receiving the A380 designation.

Airbus considered several designs, including an odd side-by-side combination of two fuselages from the A340, which was Airbus' largest jet at the time.

On 19 December 2000, the supervisory board of newly restructured Airbus voted to launch an €8.8-billion programme to build the A3XX, re-christened as the A380, with 50 firm orders from six launch customers.

Airbus obtained type certificates for the A380-841 and A380-842 model from the EASA and FAA on 12 December 2006 in a joint ceremony at the company's French headquarters.

The A380's wing is sized for a maximum take-off weight (MTOW) over 650 tonnes in order to accommodate these future versions, albeit with some strengthening required.

The stronger wing (and structure) would be used on the A380-800F freighter.

Another A380 following an A380 should maintain a separation of .

Improved A380-800 From 2013, Airbus will introduce a new A380 build standard incorporating a strengthened airframe structure and a 1.5° increase in wing twist.

The "Airbus A380" is a double-deck, wide-body, four-engine jet airliner manufactured by the European corporation Airbus, a subsidiary of EADS.

On 3 October 2006, upon completion of a review of the A380 program, Airbus CEO Christian Streiff announced a third delay, pushing the first delivery to October 2007, to be followed by 13 deliveries in 2008, 25 in 2009, and the full production rate of 45 aircraft per year in 2010.

A380-800 freighter Airbus originally accepted orders for the freighter version, offering the second largest payload capacity of any cargo aircraft, exceeded only by the Antonov An-225.

In July 2010 its orders comprise of 212 aircraft from Airbus, including 80 Airbus A380-800s (15 delivered as of March 2011), and 63 Boeing aircraft.

On 16 November 2003, Emirates ordered 41 Airbus aircraft, comprising two A340-500s, 18 A340-600s and 21 A380-800s.

Emirates also announced it had signed a contract for Engine Alliance GP7200 engines to power the 32 Airbus A380 aircraft it ordered in June at the Berlin Air Show.

The airline has orders for 90 Airbus A380-800 aircraft and was the second airline to receive the aircraft, after Singapore Airlines, the launch customer.