LIA participation at INEX Tweet contextualization

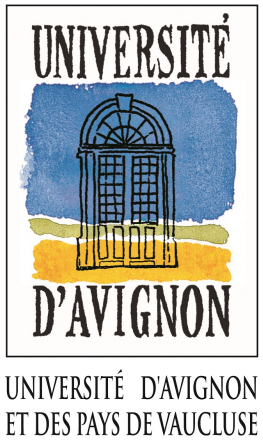
J.V. Cossu
University of Avignon, France

CLEF - INEX'2014

16 September 2014

Authors

- Juan-Manuel Torres-Moreno, **Jean-Valère Cossu**



Plan

- LIA Systems participating to INEX
- Experiments about tweet contextualization and ORM

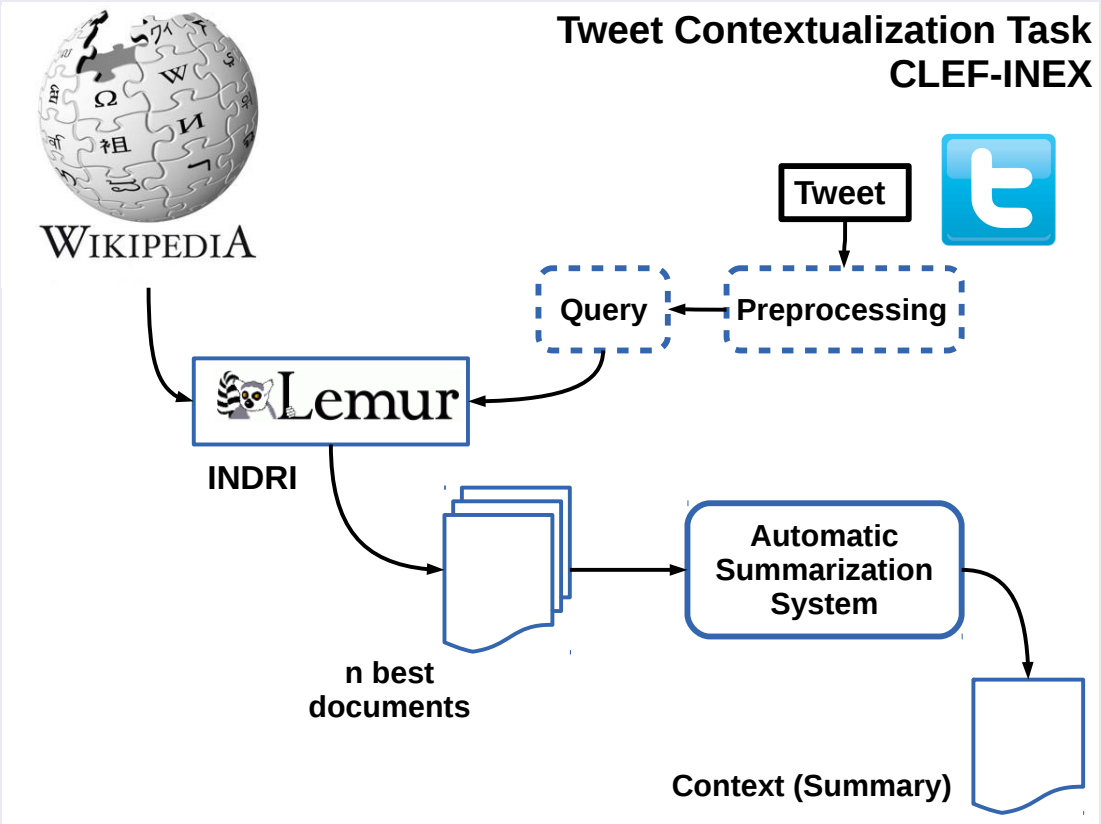
Systems participating to INEX

- Cortex
- Artex
- REG

LIA strategy for INEX

- Classical approach coming from IR
- Vector-Space Model based
- Without linguistics resources

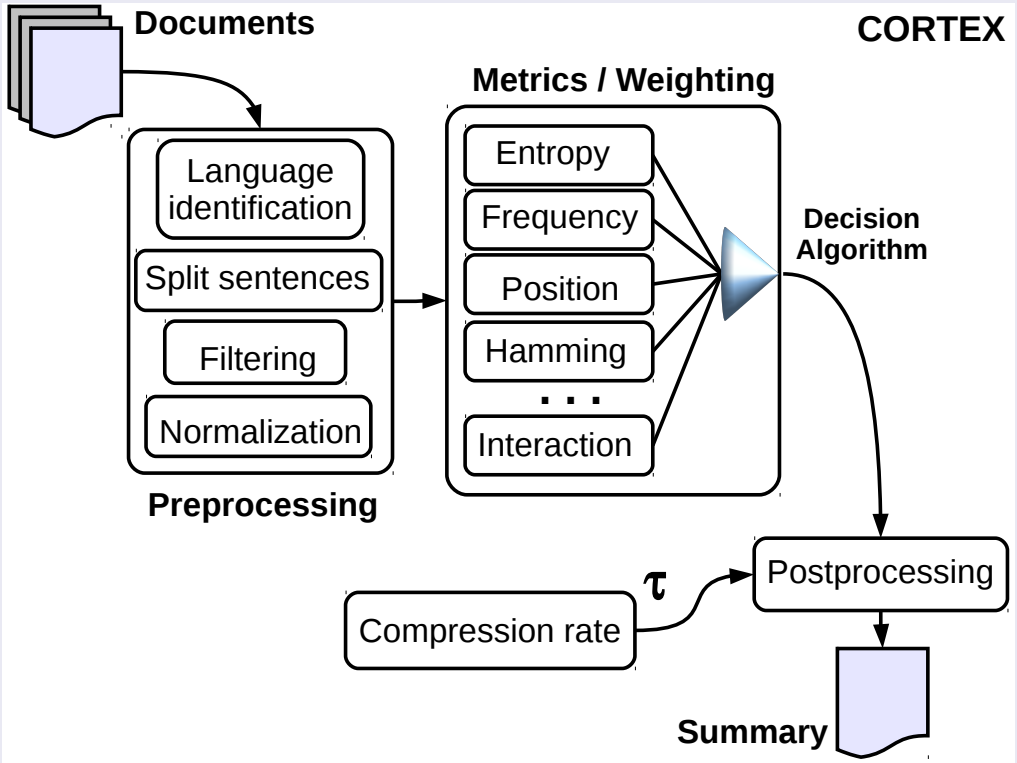
Architecture



Ideas

- Plug-in metrics (systems) coming from IR domain
- Vector-Space Model based
- Optimal Decision Algorithm merging the metrics

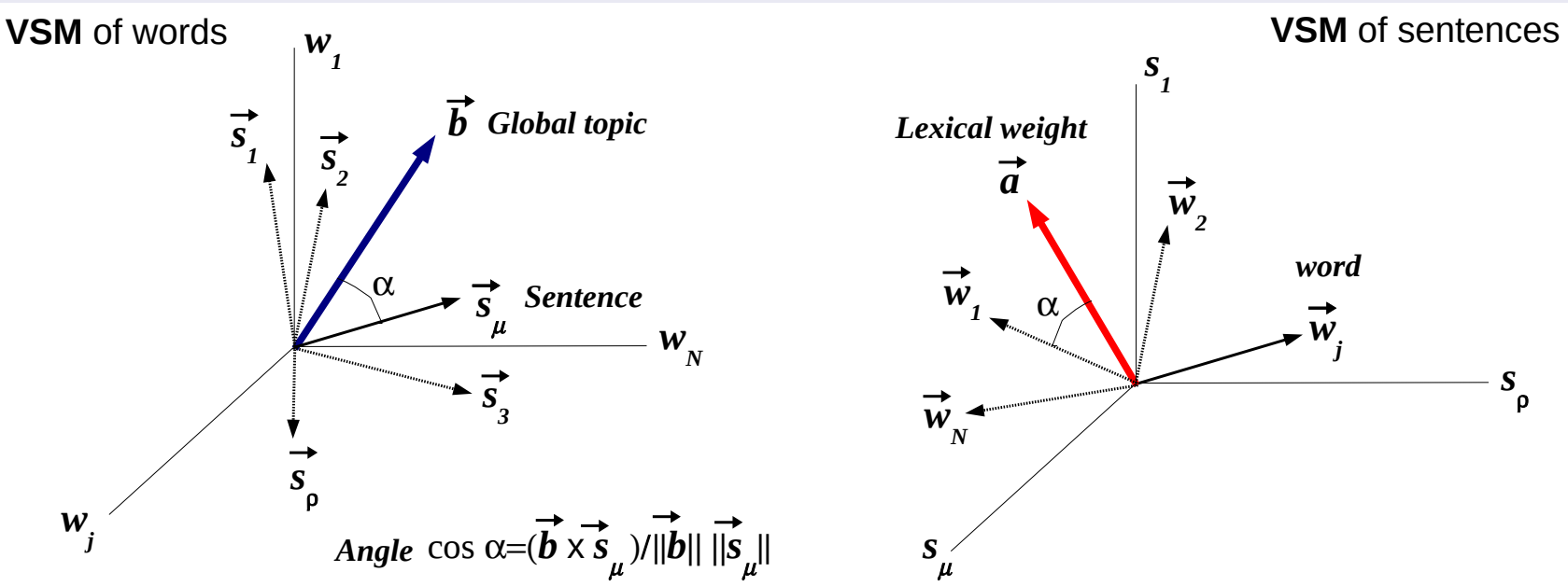
Architecture



Dual computation of:

- Pseudo Sentence-topic
- Pseudo Word-topic
- Vector-Space Model based

Idea



Graph-based Algorithm

- Compute the weight of the vertices from adjacency matrix
- Compute the degree of each vertex: shared words

Weighting sentences

- Algorithm inspired from Kruskal's algorithm
- Vertex represent the sentences
- Solution based on a calculation of greedy search paths

To make it easier

- Graph structure
- Each node is a sentence, links are shared words
- Path is computed with common words

Readability Evaluation (Official Results)

- ID Systems: REG 358, Cortex: 356, Artex: 357

Readability (3 first places)

Run	Readable	Syntax	Diversity	Structure	Avg
358	94.82%	87.31%	72.17%	93.10%	86.85%
356	95.24%	85.19%	70.31%	92.40%	85.78%
357	94.88%	82.53%	71.34%	91.58%	85.08%
364	88.05%	69.94%	63.91%	86.92%	77.20%
360	92.60%	70.35%	58.84%	86.33%	77.03%
ref2013	91.74%	69.82%	60.52%	85.80%	76.97%
ref2012	91.39%	69.58%	60.67%	85.56%	76.80%
359	93.03%	70.64%	53.53%	86.34%	75.88%
363	83.68%	67.92%	61.13%	87.55%	75.07%
362	83.67%	68.00%	60.81%	87.59%	75.02%
361	93.23%	70.41%	50.12%	85.97%	74.93%
368	90.88%	68.89%	56.59%	80.88%	74.31%
369	91.23%	69.47%	54.93%	81.56%	74.30%
370	90.10%	68.30%	53.83%	80.70%	73.23%

Legend:

Readable:	% of passages considered as readable (Non trash)
Syntax	% of passages without syntax or grammatical errors
Diversity	% of non redundant passages
Structure	% of non breaking anaphora passages

LIA participation to INEX

Common sentences pre-processing

- Minimal clean of tweets (punctuation)
- Ultra-stemming of Corpus (five characters stem)
- Break the long sentences (than the mean length on the corpus)

Statistical summarizers

- Without linguistic rules
- Without external knowledge

Results

- Top results on Readability Evaluation
- 9-12 places on Informativeness (sentences and NPs)

Can Tweet Contextualization be helpful for e-reputation management ?

J.V. Cossu
University of Avignon, France

CLEF - INEX'2014

16 September 2014

Author

- **J.-V. Cossu**

Task

Subset of the Replab 2013 data-set Classification according to the standard reputation dimension
(1) Products/Services (2) Innovation (3) Workplace (4) Citizenship (5) Governance (6) Leadership (7) Performance (8) Undefined

- Products/Services represent more than 50% of the data-set

Official results

- LIA systems only using tweet content and author performed bad
- Glasgow run using tweet enrichment obtained the best score

System	F-Score	Accuracy
Glasgow 4	48.9	69.5
Baseline	38	62.2
LIA 4	16	54.9

How far without enrichment ?

A new system

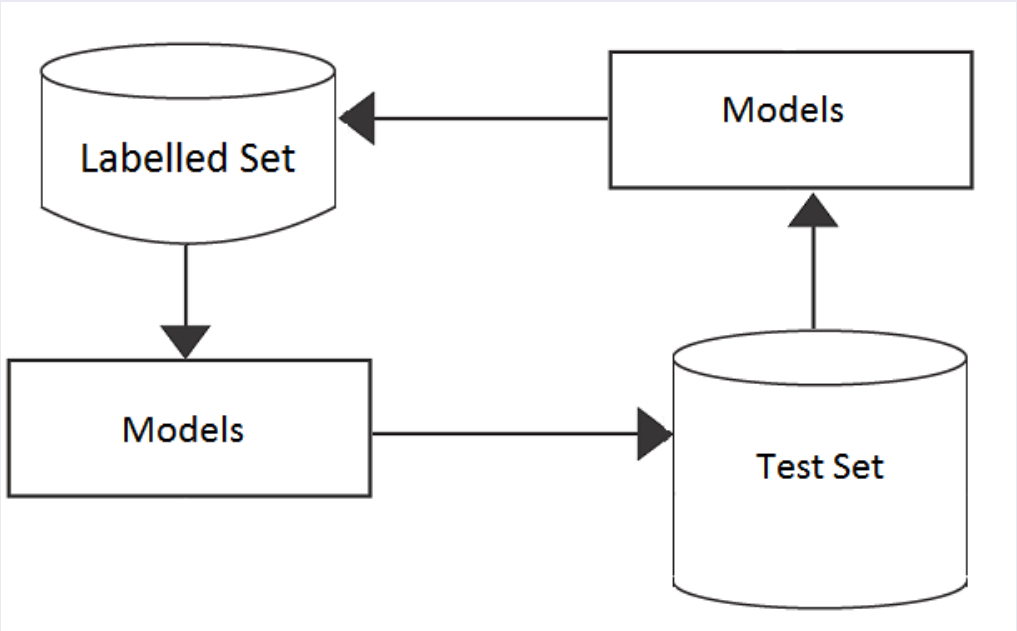
- Cosinus classifier using bag of words representation
  - A basic version
  - An improved version using a better features selection
- Results near the best system and outperforming our official run

Results over the complete Replab set

System	F-Score	Accuracy
Cosinus with FS	52	71
Glasgow 4	48.9	69.5
Basic Cosinus	48	66
LIA 4	16	54.9

Flip-flop process

- Selecting the best weighted values of TF, IDF and Purity index
- Look for the best automatic annotation of the test set that will allow us to retrieve the train annotation



Can Tweet Contextualization be helpful for e reputation management ?

Motivations in link with the work previously presented

- Use the context information to improve a reputation management system
- Measure the improvements over RepLab 2014 Reputation Dimensions task
- Introduce a way to automatically rank contextualization systems

INEX 2014 Tweets

- 240 long tweets without URLs manually selected for their readability
- Link with Wikipedia
- Easier to handle from a Machine Learning PoV ?

RepLab 2014 Reputation Dimensions task

- Accuracy as main measure and F-Score computed with precision/recall
- 77 tweets from the RepLab Test-set match the INEX ones

Can Tweet Contextualization be helpful for ORM ?

Protocol

- For a each tweet different contexts are added in the bag of words
- We perform Dimension classification with this new document
- Looking for improvement of the classification results ?

Experiences

For each tweet, we selected from each INEX run

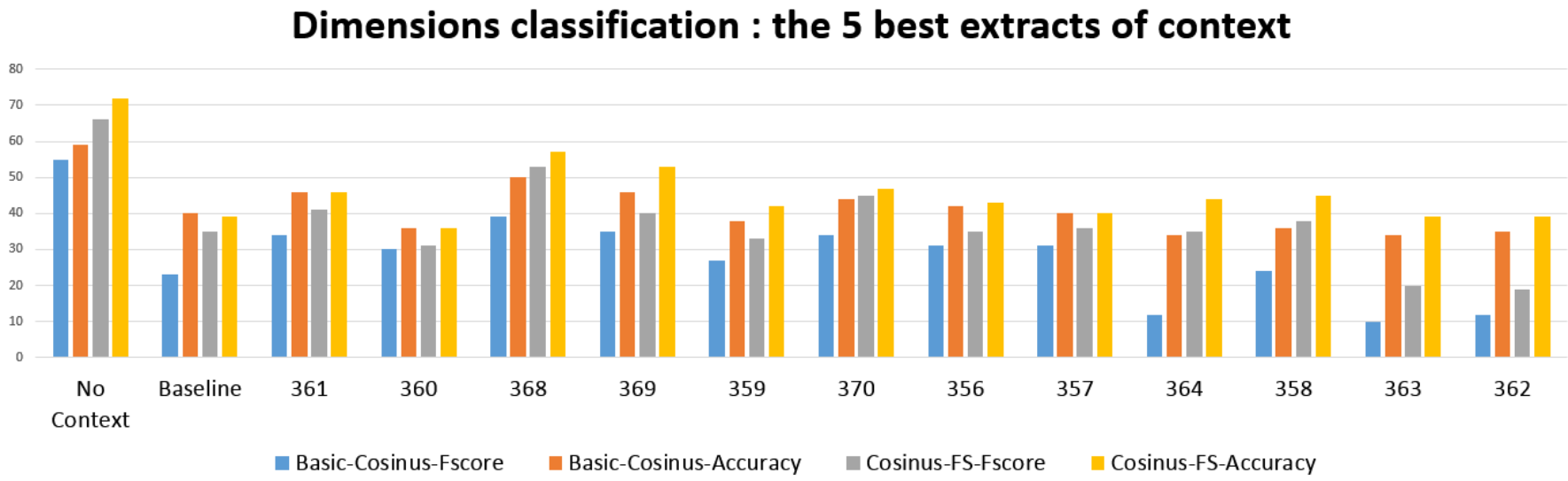
- 5 best context extracts
- 5 best context extracts in addition to the tweet
- Only the best context extract in addition to the tweet

Results over the INEX subset

System	F-Score	Accuracy
Cosinus with FS	66	72
Basic Cosinus	55	59

INEX tweets seem the be easier to classify
Label balance changed Products/Services $< 30\%$

Can Tweet Contextualization be helpful for ORM ?

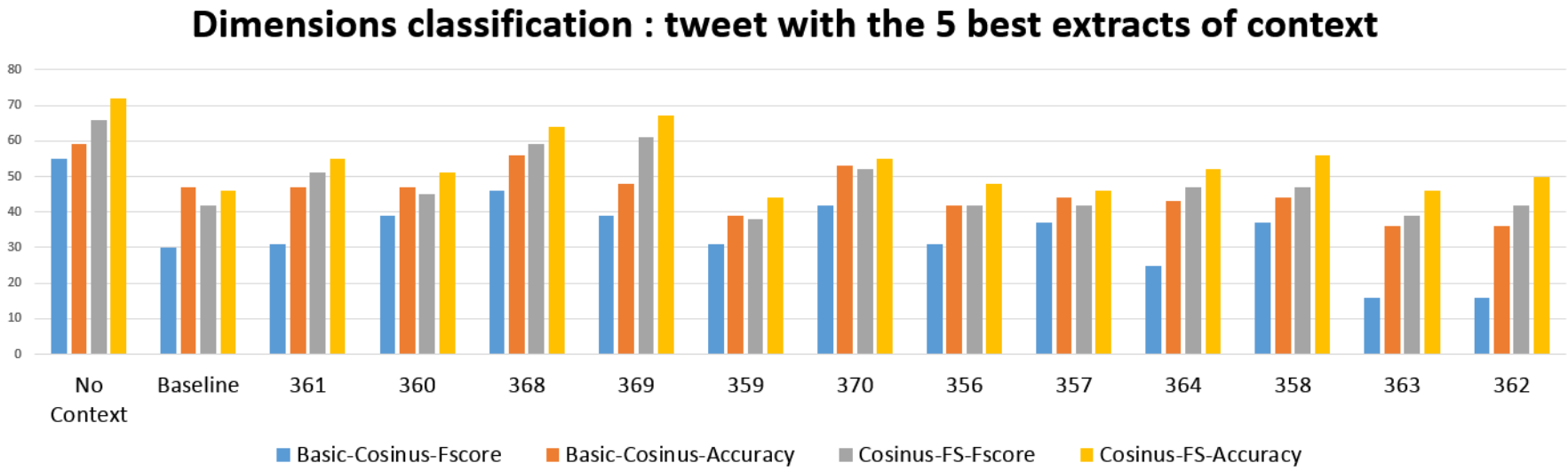


Experiment 1

Is it better to classify a tweet or his large context ?

- The baseline (tweet content) outperform each context
- Some contexts don't provide efficient reputation dimension information

Can Tweet Contextualization be helpful for ORM ?

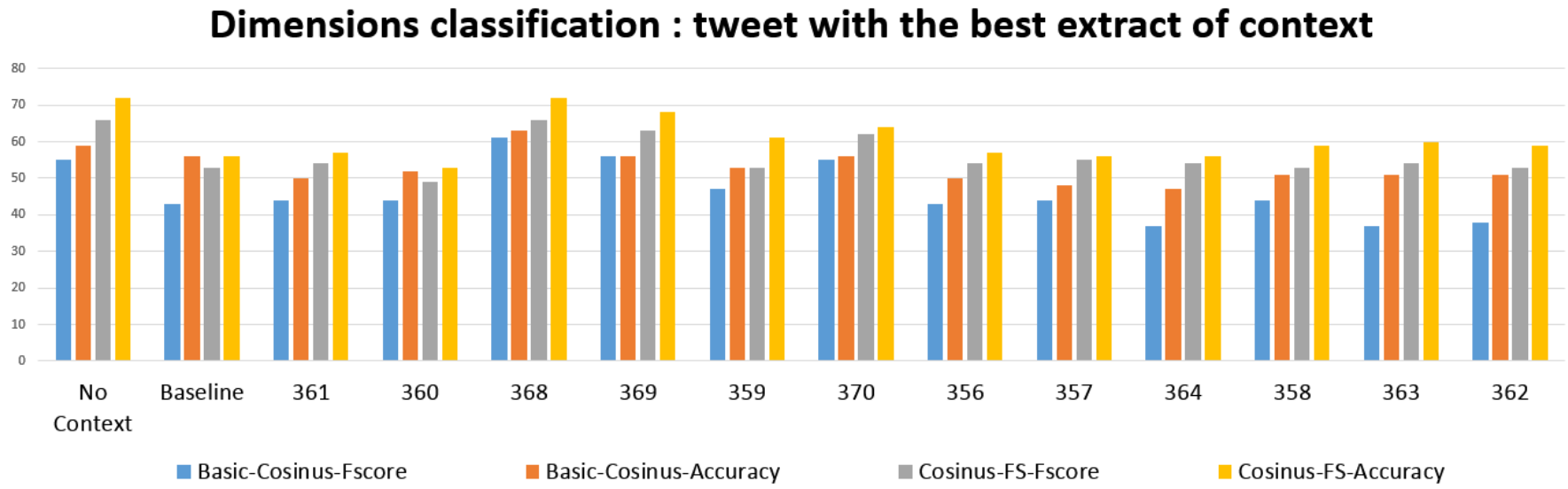


Experiment 2

Any improvement considering a tweet with his large context ?

- The baseline (tweet content) still outperforms each context
- There are even more differences between contexts

Can Tweet Contextualization be helpful for ORM ?



Experiment 3

Finally what can we expect from a short context ?

- With INEX 368 we improve the basic cosinus
- With INEX 368 we acheive the same results as the advanced cosinus

Conclusion

- Considering the short context seems to be better than a larger one
- Probably less noise is introduced regarding the original content

Ranking comparison INEX/Classification

Based on the last experiment

INEX Run	Readability	Informativeness	Classification
358	1	11	6
356	2	8	8
357	3	9	9
364	4	10	13
360	5	3	12
Baseline	6	1	5
359	7	6	4
363	8	12	10
362	9	13	11
361	10	2	7
368	11	4	1
369	12	5	2
370	13	7	3

Original system ranking based on readability is totally modified
Readability does not provide valuable information for the classification

Example

Tweet 281209605940449280 stopbeingfamous RL2013D02E057 Wells-Fargo

Here we have Wells-Fargo funding Mexican drug cartels, the ATF FBI allowing class 3 weapons across the US-Mexico border, HSBC laundering
Citizenship Dimension

Exemple of context (system 368)

Mexican officials denied any corruption in the government 's treatment of drug cartels.

Exemple of context (system 364)

ATF announced a goal to deploy eTrace software to all thirty-one states within the Republic of Mexico. As part of eTrace expansion, ATF continues to provide training to Mexican and Central American countries to ensure that the technology is utilized to a greater extent. Sauceda was killed in a gun battle with the Federal Police on February 17, 2009, and Heriberto Lazcano formed his own rival cartel, Los Zetas Costilla is one of a number of high ranking members of the Gulf Cartel who has been indicted in the United States for drug trafficking activities. Chambers finished serving his sentence in November 2006. In February 2007, the Associated Press reported that Chambers has chosen to describe himself as a retired criminal, and has begun employment at an undisclosed location. Chambers had pleaded guilty to 16 charges related to the theft. In August 1997, Campbell informed Ghantt about an old high school friend of hers named Steve Chambers, who could assist Ghantt to execute a massive cash robbery of the Loomis Fargo vault in one night. Chambers had broached the possibility of a robbery to Campbell earlier in the summer. On July 16, 2008, the Mexican Navy intercepted a 10-meter long narco submarine travelling about 200 kilometers off the southwest of Oaxaca ; in a raid, Special Forces rappelled from a helicopter onto the deck of the submarine and arrested four smugglers before they could scuttle their vessel.

Conclusion

What has been showed ?

- A new method to rank summaries
- A good system without additional information can be as good as a lower one using external ressources
  - M-H. Peetz, D. Spina, J. Gonzalo and M. de Rijke - Towards an Active Learning System for Company Name Disambiguation in Microblog Streams (CLEF 2013)
  - M. Qureshi, C. O'Riordan and G. Pasi - Exploiting Wikipedia for Entity Name Disambiguation in Tweets (NLDB 2014)
  - G. McDonald, R. Deveaud, C. Macdonald, R. McCreadie, T. Gollins and I. Ounis - University of Glasgow terrier team / project abaca at RepLab 2014: reputation dimensions task (CLEF 2014)
- The interest of using the context information
- Large contexts are too generic wtr to the tweet content

Coming soon - Short future perspectives

- Confirm these results on the complete Replab tasks (more than 70,000 tweets in test)
- Find a new constraint for automatic summarizers

Perspectives

Tweet expansion more than tweet contextualization ?

Any questions or suggestions ?