# Human judgment as a parameter in evaluation campaigns

Jean-Baptiste Berthelin
Cyril Grouin
Martine Hurault-Plantet
Patrick Paroubek

LIMSI-CNRS, Orsay, France

Coling 2008 workshop on human judgements in Computational Linguistics, Manchester, 23 August 2008

# DEFT
## (Défi Fouille de Textes, Text-Mining Challenge)
## deft.limsi.fr

- 2005-2008 and beyond
- French corpora
- from 6 to 14 competitors...
- ... academic and industrial
- design and organisation...
- ...involving human judges

# Contents

- Managing a campaign
- Feasibility of tasks
- Adjusting parameters
- Validation of choices
- Conclusion and prospects

# Managing a campaign: classical steps

- Preparation
  - potential topics
  - task and corpora
  - measurements
  - test by humans
- Execution
  - launching the task
  - training period
  - adjudication
  - final workshop

# Feasibility of tasks

- 07: discarding a book review corpus
- 08: examining “text puzzles”

# Adjusting parameters

- depends on task
  - Opinion analysis (07)
    - scales of marks
  - Identification of genre and topic (08)
    - sets of topical categories

# Adjusting, 1: corpora, original scales

- parliamentary debates, “for” vs “against”
- scientific paper reviews, 4 values
- film and book reviews, 5 values
- video game reviews, 20 values

# Adjusting, 2: kappa on original scales

| Judge | Ref. | 1 | 2 | 3 |
|---|---|---|---|---|
| **Ref.** | | 0.17 | 0.12 | 0.07 |
| **1** | 0.17 | | 0.03 | 0.05 |
| **2** | 0.12 | 0.03 | | 0.07 |
| **3** | 0.07 | 0.05 | 0.07 | |

# Adjusting, 3: from original to restricted scales

- debates, no adjustment needed
- scientific
- film and book reviews
- video game reviews

# kappa on restricted scales

| **Judge** | **Ref.** | **1** | **2** | **3** |
|---|---|---|---|---|
| **Ref.** |  | 0.74 | 0.79 | 0.69 |
| **1** | 0.74 |  | 0.74 | 0.54 |
| **2** | 0.79 | 0.74 |  | 0.69 |
| **3** | 0.69 | 0.54 | 0.69 |  |

# Adjusting, 08-1: initial set

- Le Monde
  - notebook
  - economy
  - France
  - international
  - science
  - society
  - sport
  - television

- Wikipedia
  - people
  - economy
  - French politics
  - politics minus French
  - science
  - society (minus some)
  - sport
  - television

# Adjusting, 08-2: grouping categories

- discarding category “Notebook”
- balancing easy and difficult ones to recognise
  - Art, Economy, Sport, Television (genre and topic)
  - France, International, Literature, Science, Society (topic)

# Validation

- 07, scientific vs parliamentary
- 08,
  - genre easy,
  - category, competitors better than judges

# Conclusion

- relevance of human judgments
  - checking the feasibility of tasks
    - matching results by different judges
  - adjusting parameters
  - final results happen to validate initial choices
  - planning future campaigns

# Prospects for DEFT-09

- continue to rely upon human judgment
- more scientific paper reviews?
- multilingual corpora?
- ...