# QA Pilot Task at CLEF 2004

Jesús Herrera
**Anselmo Peñas**
Felisa Verdejo

**UNED NLP Group**

Cross-Language Evaluation Forum
Bath, UK - September 2004

# Objectives of the Pilot Task

- Difficult questions

  Inspect systems performance in answering :

  - Conjunctive lists
  - Disjunctive lists
  - Questions with temporal restrictions
  - Make some inference
- Self-scoring

  Study systems capability to give accurate confidence score

  - New evaluation measures
  - Correlation between confidence score and correctness
- Generate evaluation and training resources
- Feedback for the QA@CLEF Main Tasks

# Pilot Task definition

- Usual methodology
- Carried out simultaneously with Main Track
- Same guidelines except:
  - Source and target language: Spanish
  - Number and type of questions:
    - Factoid: 18 (2 NIL)
    - Definition: 2
    - Conjunctive list: 20
    - Temporally restricted: 60
  - Number of answers per question: unlimited
    - Several correct and distinct answers per question (disjunctive list)
    - Context dependant or evolving in time
    - Reward correct and distinct answers and “punish” incorrect ones
  - Evaluation Measures

# Temporally restricted questions

- 3 Moments with regard to the restriction
  - Before
  - During
  - After
- 3 Types of restrictions:
  - Restricted by Date (20 questions): day, month, year, etc.
    - *¿Qué sistema de gobierno tenía Andorra hasta mayo de 1993?*
  - Restricted by Period (20 questions)
    - *¿Quién gobernó en Bolivia entre julio de 1980 y agosto de 1981?*
  - Restricted by Event (nested question) (20 questions)
    - *¿Quién fue el rey de Bélgica inmediatamente antes de la coronación de Alberto II?*
- Inspect several documents to answer a question

# Evaluation measures

- Considerations
  - “Do it better” versus “How to get better results?”
  - Systems are tuned according the evaluation measure
- Criteria. Reward systems that give:
  - Answer to more questions
  - More different correct answers to each question
  - Less incorrect answers to each question
  - Higher confidence score to correct answers
  - Lower confidence score to incorrect answers
  - Answer to questions with less known answers
- “Punish” incorrect answers
  - Users prefer void answers rather than incorrect ones
  - Promote answer validation and accurate self-scoring
  - Unlimited number of answers is permited $\rightarrow$ self-regulation

# Evaluation measure

## *K-measure*

$$K(sys) = \frac{1}{\#questions} \cdot \sum_{i \in questions} \frac{\sum_{r \in answers(sys,i)} score(r) \cdot eval(r)}{max\{R(i), answered(sys,i)\}}$$

$$eval(r) = \begin{cases} 1 & \text{if } r \text{ is judged as correct} \\ 0 & \text{if } r \text{ is a repeated answer} \\ -1 & \text{if } r \text{ is judged as incorrect} \end{cases}$$

*score(r):* confidence self-scoring [0,1]

*R(i):* number of different known answers to question *i*

*answered(sys,i):* number of answers given by *sys* to question *i*

*K(sys)*$\in$ [-1,1]

Baseline: *K(sys)=0* $\approx \forall r.score(r)=0$ *(System without knowledge)*

# Self-scoring and correctness

- Correlation coefficient (*r*)
  - Correctness (human assessment):
    - assess(sys,r) $\in$ {0,1}
    - 0: incorrect
    - 1: correct
  - Self-scoring
    - score(sys,r)$\in$[0,1]
  - r$\in$[-1,1]
    - 0: no correlation
    - 1: perfect correlation
    - -1: inverse correlation

$$r(sys) = \frac{\sigma_{assess(sys)score(sys)}}{\sigma_{assess(sys)} \cdot \sigma_{score(sys)}}$$

# Results at the Pilot Task

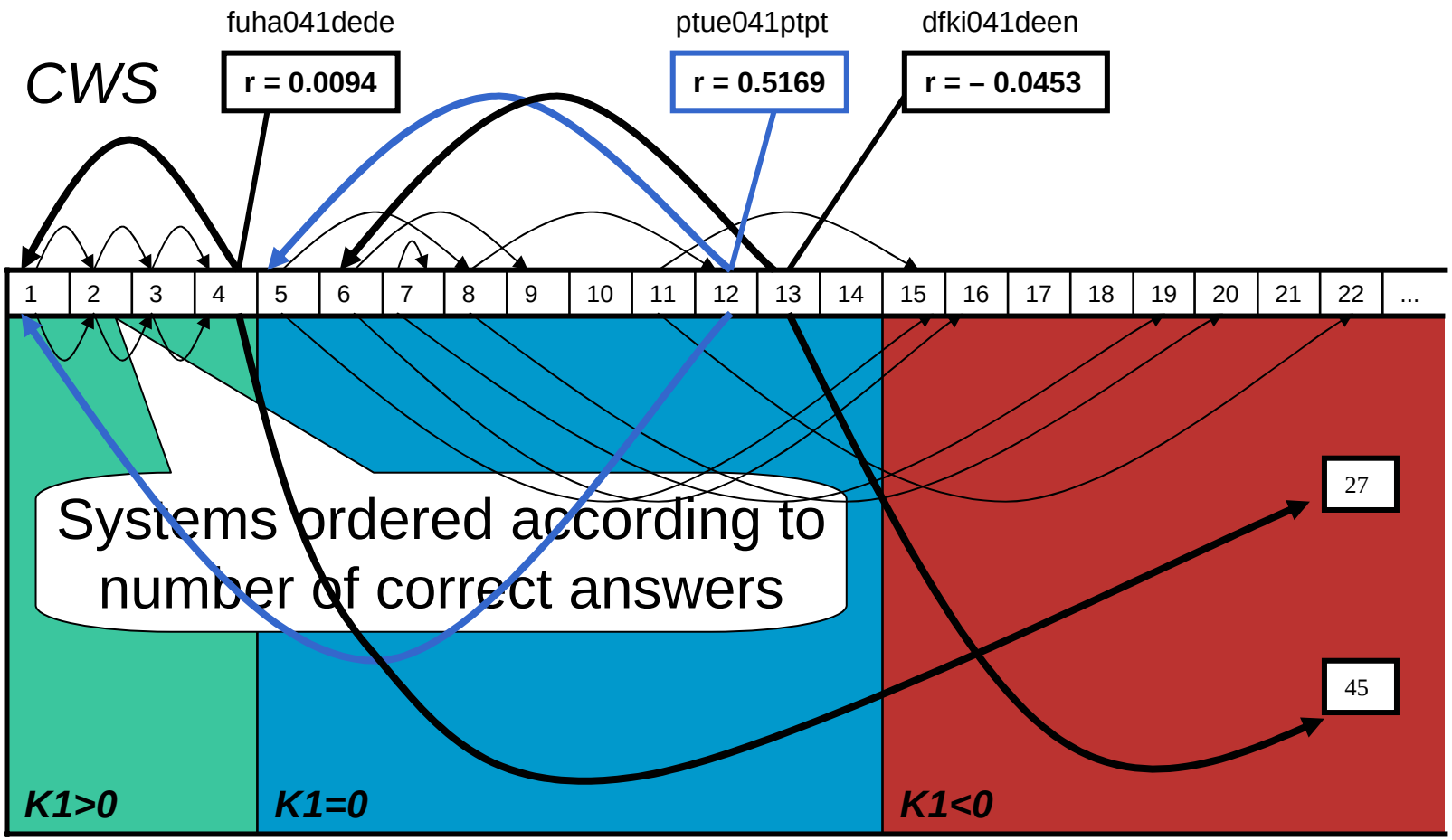
- Only one participant: U. Alicante
  - Splitting of nested questions (Saquete et al., 2004)
- Correctly answered: 15% (factoid: 22%)
  - Correctly answered in Main Track: 32%
  - Evaluated over TERQAS obtain better results
  - Questions too difficult
- Correlation between assessment and self-scoring: 0.246
  - Further work on improving self-scoring
- K= –0.086
  - $k < 0$

# Case of study

- Are systems able to give an accurate confidence score?
- Do *K-measure* reward it better than others?
- Study the ranking of the 48 participant systems at the Main Track
  - *number of correct answers*
  - *CWS*
  - *K1*, variant of *K-measure* when just 1 answer per question is requested

$$K1(sys) = \frac{\sum_{r \in answers(sys)} score(r) \cdot eval(r)}{\#questions}$$

# Re-ranking with *K1* vs. *CWS*



# Case of study

- CWS reward some systems with very bad confidence self-scoring
  - For example: fuha041dede, *r=0.0094*
    - *CWS*: 1st position
    - *K1*: 27th position
  - Strategy oriented to obtain better CWS
    - Convert answers with low confidence to NIL with *score=1* ensures 20 correct answers in the top of the ranking (the 20 NIL questions)
  - However, it shows very good self-knowledge
    - Giving *score=0* to its NIL answers: *r=0.7385*
    - *K1*: 1st position

# Conclusions

- Some systems are able to give an accurate self-scoring: *r* up to 0.7
- *K-measures* reward good confidence self-scoring better than CWS
- But not only good self-scoring (high r)
  - A system with a perfect score (r=1) would need to answer correctly more than 40 questions to reach 1st position
  - Find a good balance
- Promote answer validation and accurate self-scoring

# Conclusions

- *"Difficult"* questions still remain a challenge
- Some specialisation should be expected
  - QA Main Track shows that different systems answer correctly different subsets of questions
- *K-measures* permit
  - Some specialisation
  - Pose new types of questions
  - Leave the door open to new teams

"Just give *score=0* to the things you don't *K-now*"


# And, What about Multilinguality?

- Start thinking about promoting fully multilingual systems
  - Too soon for a unique task with several target languages? (Multilingual collection)
  - Join Bilingual Subtasks with the same target language into a Multilingual task? (Multilingual set of questions)
    - Allow bilingual, promote multilingual (help transition)
    - ~50 questions in each different language
    - Systems could answer with *score=0* to the questions in source languages they don't manage
    - Systems that manage several source languages would be rewarded (transition could be expected)

# Thanks!

# QA Pilot Task at CLEF 2004

Jesús Herrera
**Anselmo Peñas**
Felisa Verdejo

**UNED NLP Group**

Cross-Language Evaluation Forum
Bath, UK - September 2004