CL-QA @ CLEF 2004

Julio Gonzalo (UNED)

Douglas W. Oard (UMD)

The iCLEF approach to CLIR research

- Look for retrieval scenarios
  - Realistic
  - Challenging for CLIR research
  - Susceptible of comparative evaluation.
- Study them from a user-inclusive perspective.
- Promote comparative incremental research via TREC-like evaluations.

Why Do People Use IR Systems?

- Learning about a topic iCLEF 2001, 2002, 2003
- Finding a known item
- Substantiating a claim
- Finding a person/organization/service
- Answering a question

2001-2003: some results

- Interactive features make a difference!
- Interactive features are more important than CLIR performance.
- Cross-language document selection is harder than monolingual selection even for searchers trained in the target language.
- Word-by-word T < MT < Cross-Language summaries
- Automatic translation < Assisted user translation
- Users prefer monolingual search interfaces for CLIR

Why Do People Use IR Systems?

- Learning about a topic
- Finding a known item
- Substantiating a claim
- Finding a person/organization/service
- Answering a question

iCLEF 2004

iCLEF 2004: Interactive Cross-Language Question Answering

How can a system help a user to find, recognize and use the answer to a particular question, even if the answer is expressed in some foreign language?

- Realistic (even more than plain QA?)
- Challenging
- Comparative evaluation feasible
- Potentially wide research community.
- assessment support in CLEF

CL-QA Evaluation Design

- Standard set of documents
  - – CLEF uses news text
- Standard set of 200 “factoid” questions
  - – In some other language
- System finds a single best answer
  - – Correct: exact, with a pointer to the doc
  - – Unsupported: exact, but no correct pointed
  - – Inexact: too much or too little

Some Differences for iCLEF

- People know some of the answers
  - Which answers depends on cultural factors
- People can draw inferences
  - Answers may draw from more than one doc
- People answer in the question language
  - But assessors work in the document language
- Assessors hold people to a higher standard

iCLEF 2004 User Study Design

- 8 users (native query language)
- 16 evaluation questions (+ 4 for training)
  - 5 Measure, 4 Time, 4 Person, 3 Organization
  - All with available answers (for Spanish+English)
- 5 minutes per search (~3 hours/session)
- Independent variable: CLIR system design
  - 8 questions per system
- Dependent variable: accuracy (exact)
- Latin square to block user/question effects

The iCLEF 2004 Questions

1 What year was Thomas Mann awarded the Nobel Prize?
2 How many human genes are there?
3 Who is the German Minister for Economic Affairs?
4 Who committed the terrorist attack in the Tokyo underground?
5 How much did the Channel Tunnel cost?
6 When did Latvia gain independence?
7 How many people were declared missing in the Philippines after the typhoon "Angela"?
8 Who is the managing director of the International Monetary Fund?
9 When did Lenin die?
10 How many people died of asphyxia in the Baku underground?
11 Who is the president of Burundi?
13 Of what team is Bobby Robson coach?
12 What is Charles Millon's political party?
14 When did the attack at the Saint-Michel underground station in Paris occur?
15 How many people live in Bombay?
16 Who won the Nobel Prize for Literature in 1994?

Participants

- UMD: KWIC vs passages
- UNED: docs. vs (filtered) passages
- SICS: SICS: with or without topic-tailored term expansion built from external parallel corpora.
- U. Salamanca: docs. vs passages
- U. Alicante: concepts vs syntactic-semantic patterns
- ALL: strong baseline for the task

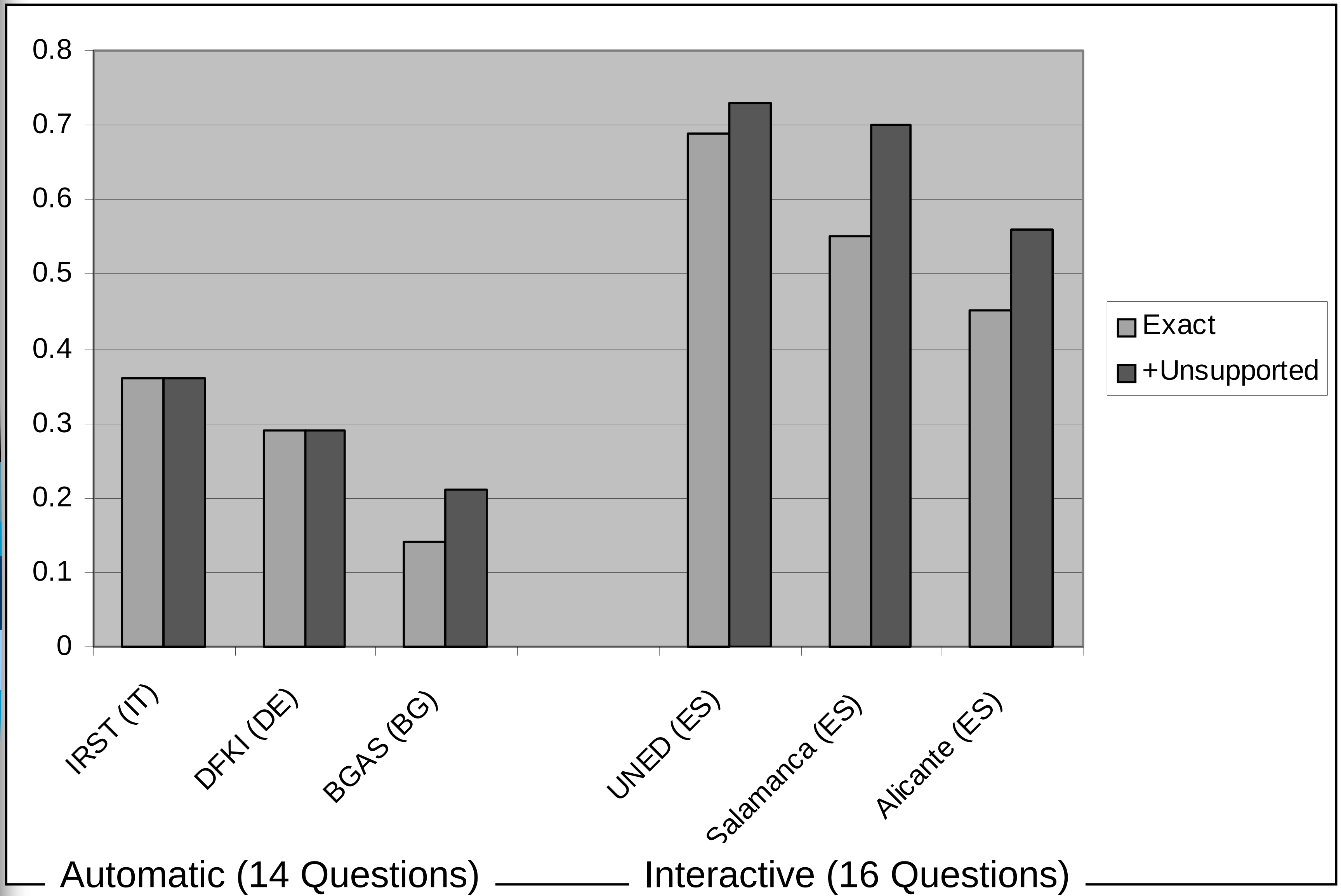
Main results

- ALL: strong baseline for the task:

50% accuracy (average), 69% (best, IR+MT)

- **plus several insights into search behavior**
- UMD: KWIC vs passages
- UNED: docs. vs (filtered) passages
- SICS: with or **without** topic-tailored term expansion built from external parallel corpora
- U. Salamanca: **docs.** vs passages
- U. Alicante: concepts vs **syntactic-semantic patterns**

Users vs QA machines



Best submitted systems,, English documents

Next Steps

- Need to measure inter-annotator agreement
  - – Not so important for (bad) automated systems!
- What lessons can we learn from searchers?
  - – Might help with automated system design
- How can we get QA teams involved?
  - – Which parts of a QA system would be useful?

Conclusions

- Interactive CLIR works
  - Real task, real systems, representative users
- iCLEF is where the action is!
- Automatic CL-QA has a long way to go
  - Half the accuracy in twice the clock time!