The Portuguese Language in CLEF

Paulo Alexandre Rocha
& Diana Santos

The Portuguese language at CLEF

- What is CLEF?
- Portuguese in CLEF
  - IR: information retrieval
  - Q&A: question answering
- What does it take to add a new language?

What is CLEF?

http://www.clef-campaign.org

- Cross-Language Evaluation Forum
  - The Cross-Language Evaluation Forum (CLEF) supports global digital library applications by
    - (i) developing an infrastructure for the testing, tuning and evaluation of information retrieval systems operating on European languages in both monolingual and cross-language contexts, and
    - (ii) creating test-suites of reusable data which can be employed by system developers for benchmarking purposes.
  - The final goal is to assist and stimulate the development of European cross-language retrieval systems in order to guarantee their competitiveness on the global marketplace.

Project details
Project Acronym: CLEF
Project Reference: IST-2000-31002
Start Date: 2001-10-01
Duration: 30 months
Project Cost: 585927.00 euro
Contract Type: Preparatory, accompanying and support measures
End Date: 2004-03-31
Project Status: Execution
Project Funding: 387450.00 euro

Why take part in CLEF?

Linguateca's mission

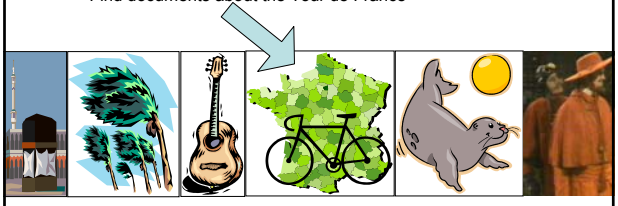
- Our mission is to raise the quality of Portuguese language processing, through the removal of difficulties for the researchers and developers involved. This is done by
  - providing resources that enable sophisticated processing of Portuguese.
  - monitoring and cataloguing the area
  - organizing evaluation activities

CLEF tracks

- Information Retrieval on News Collections
  - Multilingual Information Retrieval
  - Bilingual Information Retrieval
  - Monolingual (non-English) Information Retrieval
- GIRT Mono- and Cross-Language Information Retrieval on Structured Scientific Data
- iCLEF Interactive Cross-Language Information Retrieval
- QA@CLEF Multiple Language Question Answering
  - Monolingual Q&A
  - Crosslingual Q&A
- ImageCLEF Cross-Language Retrieval in Image Collections
- CL-SDR Cross-Language Spoken Document Retrieval

Monolingual Information Retrieval

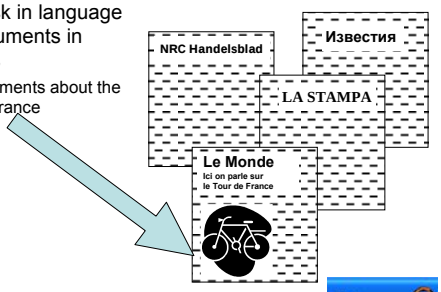
- To return documents related to select topics:
  - Find documents about the Tour de France



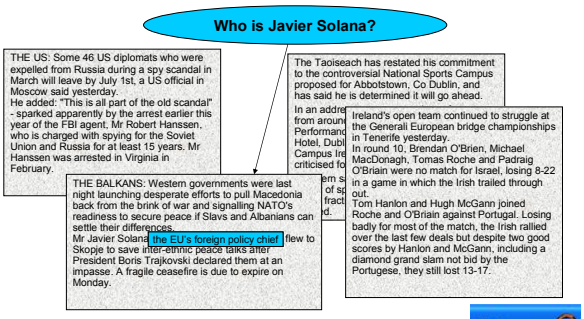
Cross-lingual information retrieval

(bilingual and multilingual)

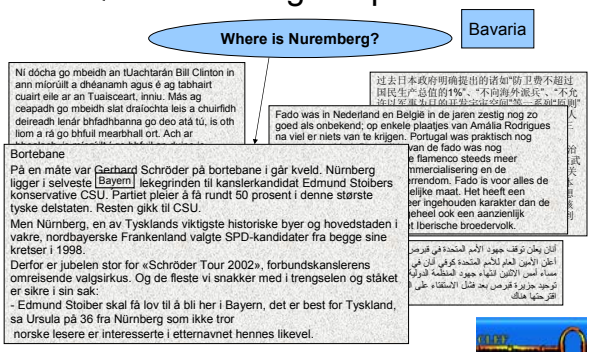
- Given a task in language A, find documents in language B
  - Find documents about the Tour de France



Q&A: question answering



Q&A: crosslingual questions



Q&A: which questions?

- *I keep six honest serving men,*
  They taught me all I knew
  Their names are What and Where and When
  And How and Why and Who.

  Rudyard Kipling
- DEFINITION: *What* is Nike?
  - A sport material company
- FACTOID: *Where* is Alcatraz Island?
  - San Francisco
  - California

Q&A: which answers?

- PERSON: *Who* killed Roger Ackroyd?
- LOCATION: *Where* was Roger Ackroyd killed?
- TIME: *When* was Roger Ackroyd killed?
- OBJECT: *What weapon* was used in the murder of Roger Ackroyd?
- MANNER: *How* was Roger Ackroyd killed?
- MEASURE: *How long* did Hercule Poirot take to solve the murder?
- OTHER: *What* was Hercule Poirot's job?

Organization

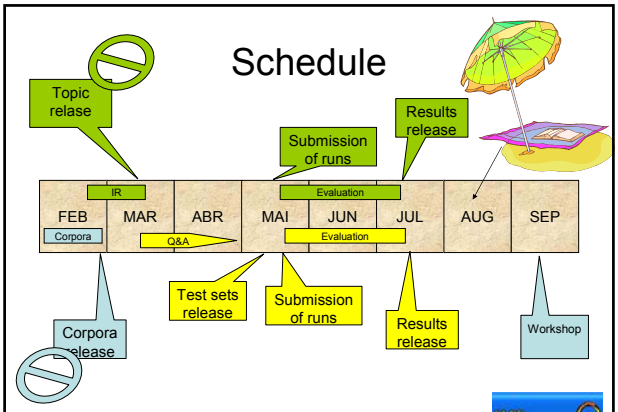
Information retrieval

- Leader
  - Carol Peters (ITC-irst)
- Organiser
  - CNR-ISTI (IT)
  - ELRA/ELDA (FR)
  - Eurospider (RU)
  - IZ-Bonn (DE)
  - Univ. Tammerfors (FI)
  - Linguateca (PT)

QA@CLEF

- Leader
  - Alessandro Vallin (ITC-irst)
- Organisers
  - ITC-irst (IT,EN)
  - UNED (ES)
  - ILLC, UvA (NL)
  - DFKI (DE)
  - ELDA/ELRA (FR)
  - Univ. Limerick
  - BulTreeBank Project (BG)
  - Linguateca (PT)

Schedule

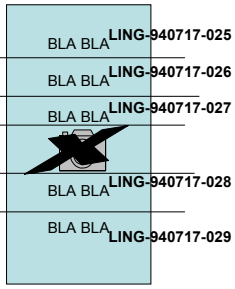


Tasks for adding Portuguese

- Create a collection of texts
- Choose topics for information retrieval
- Choose questions for Q&A
- Evaluate result

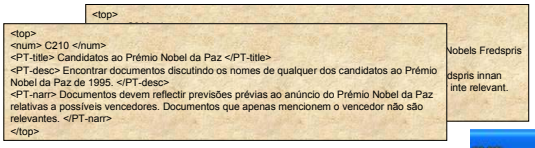
Creating a collection

- Create a text collection
  - Newspaper text 1994/1995
    - Público
- Divide text into documents
- Eliminate garbage
  - 106,821 documents
  - 348 MB
- Add ID to documents

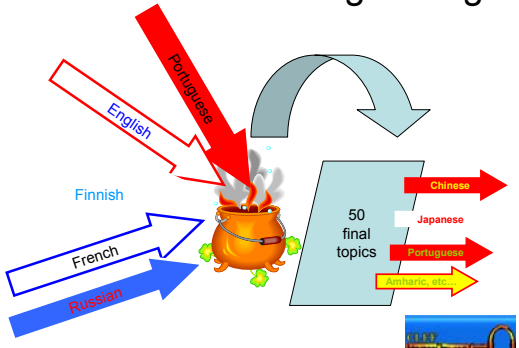


Preparing material for IR tracks

- Choose ±15 topics present in our collection (1995 only)
- Translate into English these topics
- Check number of hits of topics chosen by the other groups
- *Select 50 topics from the general pool of 98 topics*
  - *collective task among the six participating groups*
- Translate these 50 topics into Portuguese



Information retrieval: organizing



Information retrieval: guidelines

- Choose only topics relevant to 1995
- Choose topics
  - General topics (earthquakes)
  - Non-European topics (an earthquake in Botswana)
  - European topics (a minor earthquake in Nice)
  - Local topics (an earthquake at Vinderen)
- Avoid too frequent topics (10+ hits)
  - "find documents on any football game"
  - "find documents on any legislative election"
- Avoid topics used in previous years

Information retrieval: our methodology

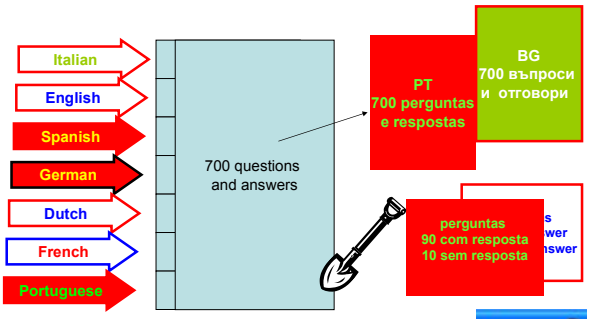
- Events topics (from Wikipedia's 1995 chronology)
  - Specific events of 1995
    - Portuguese legislative election, Alexander the Great's tomb
  - Recurring yearly events
    - Tour de France, Ig Nobel prizes
- General topics (personal tastes and chronology inspired)
  - Iranian cinema, Pope's travels
- Avoid topics unlikely to appear in other collections
  - "find documents on the films showing at Klinkenberg 1"
- Try to cover alternative ways of describing the topic

<top>
<num> C249 </num>
<PT-title> Campeã dos 10.000 metros femininos </PT-title>
<PT-desc> Quem venceu os **10.000 metros *femininos*** nos Mundiais de Atletismo em Gotemburgo? </PT-desc>
<PT-narr> Documentos relevantes devem nomear a ***vencedora*** da final dos **dez mil metros** nos Mundiais de Atletismo em Gotemburgo. </PT-narr>
</top>

QA@CLEF: preparing material

- Choose 100 questions answered by our collection
- Translate into English these 100 questions and their answers as present in the collection
- Translate into Portuguese the 600 questions proposed by the other six groups
- From these 600 questions, select 90 answered by our collection, and 10 not answered

Q&A@CLEF: collectively gathering questions



QA@CLEF: guidelines

- Questions must have an answer within 1994-1995 texts
- Questions not acceptable
  - Subjective questions
    - Who was the greatest Norwegian writer of the 19th century?
  - Lists
    - Mention works by Ibsen,
  - Closed questions
    - Did Ibsen write *Peer Gynt*?
  - Nested questions
    - When did Gro's sucessor take office?
  - Why-questions
    - Why are why-questions excluded from CLEF?
- Definition questions must address only persons and organizations

QA@CLEF: our methodology

- Balance categories
  - 11 DEFINITION (8 PERSON, 3 ORGANIZATION)
  - 89 FACTOID (22 LOCATION, 20 PERSON, 11 MEASURE, 9 TIME, 6 OBJECT, 2 MANNER, 10 OTHER)
- Choose questions on Portuguese matters
  - 37 Portugal, 12 other Portuguese-speaking countries
- Avoid
  - Questions too difficult to parse
  - Questions with too complex answers
  - Artificial or uninteresting questions

Formulating questions: syntactic variation

- What's X's age?
- How old is X?

- When did A and B marry?
- When did A and B get married?

- *primeiro ministro* | *primeiro-ministro*
- *Ministro da Economia* | *ministro da Economia*
- *Yeltsin* | *Ieltsine* | *Ieltsin*

Formulating questions: semantic variation

- Simple unambiguous
  - What is the capital of Norway?
- Tricky
  - answer depending on gender
    - Quem é o Ministro da Educação da Noruega?
  - answer depending on article
    - "a EUA" (=European University Association)
    - "os EUA" (=Estados Unidos da América)
  - answer depending on compound
    - Where is Charleston? (West Virginia)
  - answer depending on complex reasoning
    - Who is the king of Finland?

QA@CLEF: semantic variation of answers

- Granularity
  - *Hvor ligger Nord-Ossetia?*
    - » Nord-Ossetia ligger *i **Kaukasus-regionen** helt sør i Russland*, ved grensen til Georgia.
- disagreement between hits in the collections
  - Who wrote "*Hunger*"?
    - » Knut Hamsun
    - » Karen Blixen
- further specification
  - Who was Karen Blixen?
    - » Danish writer
- different currency
- subtle differences:
  - "secret service" or "secret police"

QA@CLEF: syntactic variation of answers

- What is the right answer?
  - *in 1876* | *1876*
  - *her father* | *father*
  - *Reykjavik* | *Reiquevique* | *Reikjavik* | *Reikyavik*

QA@CLEF: translating questions

Important rules:

- Sticking as much as possible to the text found in the corpus, rather than to a literal translation
- Formulating the question in a natural way in Portuguese
- Sticking as much as possible to the original question, rather to its translation into English

Future work

- Evaluate results
- Publish