

The Portuguese Language in CLEF

Paulo Alexandre Rocha
& Diana Santos

Oslo, 27. april 2004

Portuguese in CLEF



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?

Oslo, 27. april 2004

Portuguese in CLEF



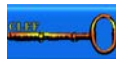
What is CLEF?

<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

Oslo, 27. april 2004



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

Oslo, 27. april 2004

Portuguese in CLEF



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

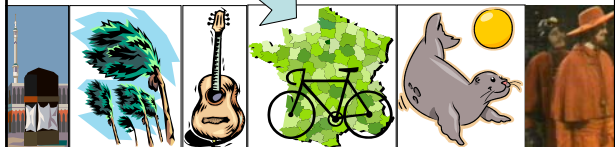
Oslo, 27. april 2004

Portuguese in CLEF



Monolingual Information Retrieval

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



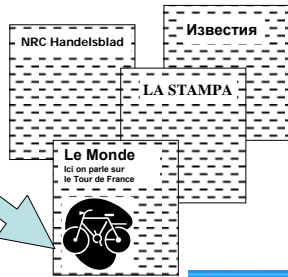
Oslo, 27. april 2004

Portuguese in CLEF



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



Oslo, 27. april 2004

Portuguese in CLEF



Q&A: question answering

Who is Javier Solana?

THE US: Some 46 US diplomats who were expelled from Russia during a spy scandal in March will leave by July 1st, a US official in Moscow said yesterday. He added: "This is all part of the old scandal" - sparked apparently by the arrest earlier this year of the FBI agent, Mr Robert Hanssen, who is charged with spying for the Soviet Union and Russia for at least 15 years. Mr Hanssen was arrested in Virginia in February.

The Taoiseach has restated his commitment to the controversial National Sports Campus proposed for Abbotstown. Co Dublin, and has said he is determined it will go ahead. In an address from around Performa Hotel, Dublin Campus Ireland's open team continued to struggle at the Generali European bridge championships in Tenerife yesterday. In round 10, Brendan O'Brien, Michael MacDonald, Tomas Roche and Padraig O'Brien were no match for Israel, losing 8-22 in a game in which the Irish trailed through out. Tom Hanlon and Hugh McGann joined Roche and O'Brien against Portugal. Losing badly for most of the match, the Irish rallied over the last few deals but despite two good scores by Hanlon and McGann, including a diamond grand slam not bid by the Portuguese, they still lost 13-17.

THE BALKANS: Western governments were last night launching desperate efforts to pull Macedonia back from the brink of war and signalling NATO's readiness to secure peace if Slavs and Albanians can settle their differences. Mr Javier Solana, the EU's foreign policy chief, flew to Skopje to save inter-ethnic peace talks after President Boris Trajkovski declared them at an impasse. A fragile ceasefire is due to expire on Monday.

Oslo, 27. april 2004

Portuguese in CLEF



Q&A: crosslingual questions

Where is Nuremberg?

Bavaria

Ni dócha go mbeidh an tUachtarán Bill Clinton in ann miorúilt a dhéanamh agus é ag tabhairt cuairt eile ar an Tuaisceart, inniu. Más ag ceapadh go mbeidh siad draoichte leis a chuirfidh díreadh lenár bhfadhbanna go deo atá tu, is oth lom a rá go bhfuil mearbhall ort. Ach ar

过去日本政府明确提出的诸如“防卫费不超过国民生产总值的1%”、“不同海外派兵”、“不向海外派兵”等主张，在野党一直予以反对。

Fado was in Nederland en België in de jaren zestig nog zo goed als onbekend, op enkele plaatsen van Amália Rodrigues na viel er niets van te krijgen. Portugal was praktisch nog van de fado was nog le fado was steeds meer nmercialisering en de irendom. Fado is voor alles de elijke maal. Het heeft een eer ingehouden karakter dan de eheel ook een aanzienlijk t Iberische broedervolk.

Bortebane
På en måte var Gerhard Schröder på bortebane i går kveld. Nürnberg ligger i selveste Bayern lekegrinden til kanslerkandidat Edmund Stoibers konservative CSU. Partiet pleier å få rundt 50 prosent i denne største tyske delstaten. Resten gikk til CSU.
Men Nürnberg, en av Tysklands viktigste historiske byer og hovedstaden i vakre, nordbayerske Frankentland valgte SPD-kandidater fra begge sine kretser i 1998.
Derfor er jubelen stor for «Schröder Tour 2002», forbundskanslerens omreisende valgsirkus. Og de fleste vi snakker med i trenselen og staket er sikre i sin sak:
- Edmund Stoiber skal få lov til å bli her i Bayern, det er best for Tyskland, sa Ursula på 30 fra Nürnberg som ikke tror norske lesere er interesserte i etternavnet hennes likevel.

Oslo, 27. april 2004

Portuguese in CLEF

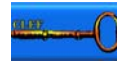


Q&A: which questions?

- 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

Oslo, 27. april 2004

Portuguese in CLEF



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?

Oslo, 27. april 2004

Portuguese in CLEF



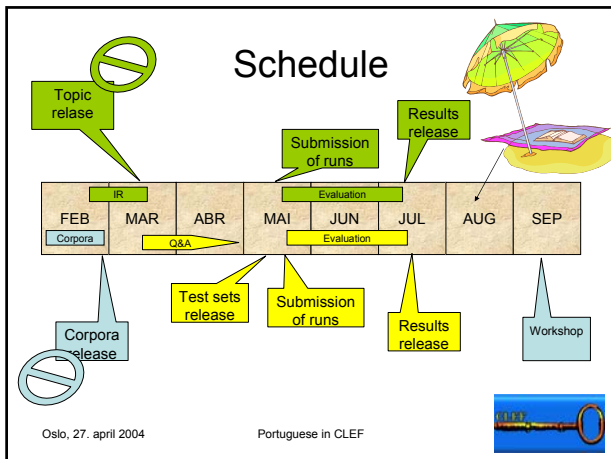
Organization

- | | |
|---|--|
| Information retrieval | QA@CLEF |
| <ul style="list-style-type: none"> Leader <ul style="list-style-type: none"> Carol Peters (ITC-irst) Organiser <ul style="list-style-type: none"> CNR-ISTI (IT) ELRA/ELDA (FR) Eurospider (RU) IZ-Bonn (DE) Univ. Tammerfors (FI) Linguatca (PT) | <ul style="list-style-type: none"> Leader <ul style="list-style-type: none"> Alessandro Vallin (ITC-irst) Organisers <ul style="list-style-type: none"> ITC-irst (IT,EN) UNED (ES) ILLC, UvA (NL) DFKI (DE) ELDA/ELRA (FR) Univ. Limerick BulTreeBank Project (BG) Linguatca (PT) |

Oslo, 27. april 2004

Portuguese in CLEF





- ### Tasks for adding Portuguese
- Create a collection of texts
 - Choose topics for information retrieval
 - Choose questions for Q&A
 - Evaluate result
- Oslo, 27. april 2004
- Portuguese in CLEF

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

Oslo, 27. april 2004

Portuguese in CLEF

Preparing material for IR tracks

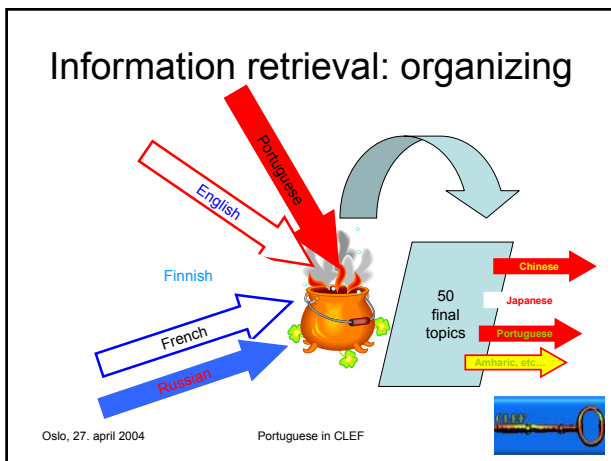
- Choose ±15 topics present in our collection (1995 only)
- Translate into English these topics
- Check number of hits of topics chosen by the other groups
- Select 50 topics from the general pool of 98 topics
 - collective task among the six participating groups
- Translate these 50 topics into Portuguese

```

<top>
<num> C210 </num>
<PT-title> Candidatos ao Prémio Nobel da Paz </PT-title>
<PT-desc> Encontrar documentos discutindo os nomes de qualquer dos candidatos ao Prémio Nobel da Paz de 1995. </PT-desc>
<PT-narr> Documentos devem reflectir previsões prévias ao anúncio do Prémio Nobel da Paz relativas a possíveis vencedores. Documentos que apenas mencionem o vencedor não são relevantes. </PT-narr>
</top>
  
```

Oslo, 27. april 2004

Portuguese in CLEF



- ### 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
- Oslo, 27. april 2004
- Portuguese in CLEF

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 Klinckenberg 1"
- Try to cover alternative ways of describing the topic

```
<!--
<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>
-->
```

Oslo, 27. april 2004

Portuguese in CLEF



QA@CLEF: preparing material

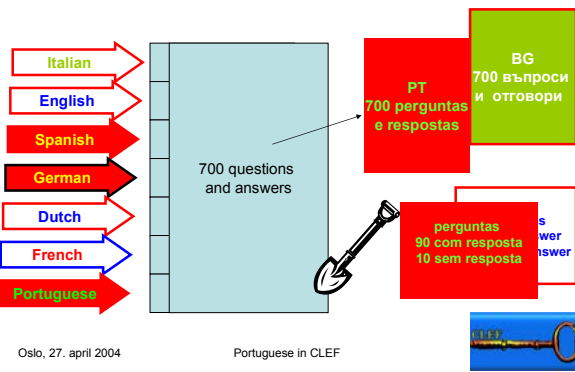
- Choose 100 questions answered by our collection
- Translate into English these 100 questions and their answers as present in the collection
- Translate into Portuguese the 600 questions proposed by the other six groups
- From these 600 questions, select 90 answered by our collection, and 10 not answered

Oslo, 27. april 2004

Portuguese in CLEF



Q&A@CLEF: collectively gathering questions



Oslo, 27. april 2004

Portuguese in CLEF



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 successor take office?
 - Why-questions
 - Why are why-questions excluded from CLEF?
- Definition questions must address only persons and organizations

Oslo, 27. april 2004

Portuguese in CLEF



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

Oslo, 27. april 2004

Portuguese in CLEF



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* | *leltsine* | *leltsin*

Oslo, 27. april 2004

Portuguese in CLEF



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?

Oslo, 27. april 2004

Portuguese in CLEF



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"

Oslo, 27. april 2004

Portuguese in CLEF



QA@CLEF: syntactic variation of answers

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

Oslo, 27. april 2004

Portuguese in CLEF



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

Oslo, 27. april 2004

Portuguese in CLEF



Future work

- Evaluate results
- Publish



Oslo, 27. april 2004

Portuguese in CLEF

