ICT for Eu-India cross-cultural dissemination

Co-financed by the European Commission

Stefano Rovetta

University of Genova

Department of Computer
and Information Sciences

ICT for Eu-India cross-cultural dissemination

Workgroup 8 – Semantic Information Retrieval: A Natural Language Processing Task

Multi-Language Communication: Two Sides of a Golden Coin

Outline

- Multi-Language Communication as an ICT task
- Multi-Language Communication as a challenge
- Multi-Language Communication as an opportunity
- Preview: Genoa contribution to Workgroup 8

Multi-Language communication

Communication

- Communicating and community making:
  by necessity goes through computers
- Language is still an issue
- Access to digital documents:
  — search
  — organize and group
  — present
  — answer questions directly
  — suggest interesting items
  — . . .

June 2005 WG4 Workshop

- The **2005 Cross-Language Information Processing** Workshop was held in Genoa (http://www.disi.unige.it/clip2005)
- Participants from WG4 countries (Italy and Spain) and from Russia
- Topics discussed:
  - Cross-language question answering
  - Document organization and clustering
  - Structural analysis of documents
  - Content personalization
- There was also a panel discussion about more general pattern recognition topics

Workshop conclusions

- Electronic documents form the basis
  of many everyday tasks,
  both for personal productivity and for group work
- Automatic document organization is
  of **vital importance** in this regard
- Despite its advancement, further work is needed
- Structural and simple content-based analysis are
  the basic tools
- Significant improvements need also
  an approach based on **semantic analysis**

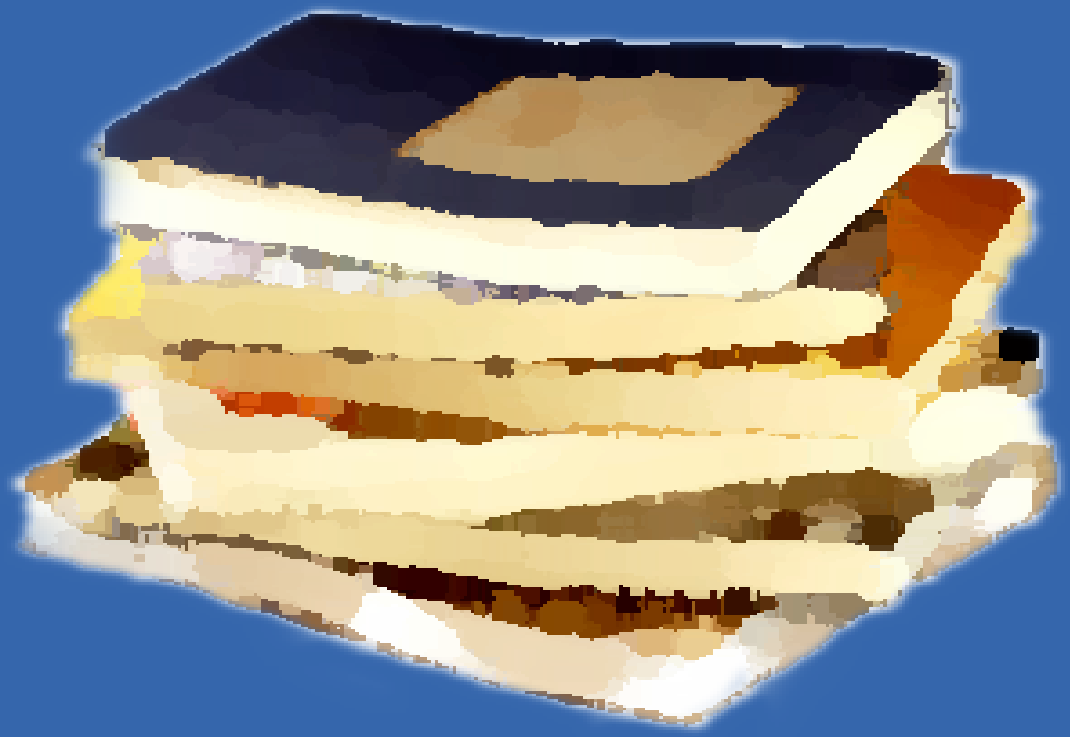
More workshop conclusions

- Cross-language document processing is possible:

  — either by using knowledge encoded into language-dependent resources, such as **ontologies** and **automatic translators** (intensive methods)

  — or by using **trainable systems** that learn from examples of different languages (extensive methods)

Side I: The challenge



Organizing and searching documents

- Traditional area for computers
- In the past 10 years it has developed exponentially:
  - the Web
  - desktop document production and processing
  - powerful aids for digitization (scanners, OCR)

The status of multi-language methods research

- Typical cross-language task:
  retrieve documents from a collection in more than one target language

- Usually target languages are **known in advance**

- This helps in the **preliminary processing** steps:
  — eliminating uninformative terms
  — extracting the stem
  — part-of-speech tagging
  — . . .

CLEF

- The **Cross-Language Evaluation Forum** (http://www.clef-campaign.org/) is the most representative international initiative in this field

- Periodically **poses challenges** and **gathers results** in annual workshops

- Typical methods presented are based on **translation software** or on **ontologies** (which are ready-made knowledge repositories)

Some remarks

- Multi-language communities from Europe and India have to face much more complex situations

- Although there are widespread languages both across India and across Europe, the effective number of languages used is at least of the order of 100

- There is also the issue of **different scripts**

Solutions to the multi-script problem

- European languages are widely studied and standard encodings for all significant scripts are available
- Indian languages are receiving attention (e.g. the ISCII code)
- The multi-script problem may be tackled with tools which are becoming **standard** such as **Unicode**

Language independence

- For a **universal** multi-language approach, language-specific facts should be **learned from examples**
- Methods should be based as much as possible on **statistical approaches** rather than a-priori knowledge
- Methods based on plug-in **knowledge repositories** are also useful — but limited to those language for which **translators** or **ontologies** exist

The contribution from Genoa

- WG4 – A task that has been studied:
  organizing documents in coherent clusters
  both for **efficient indexing**
  and for **meaningful presentation**

- WG8 – A technical problem to be solved:
  finding the best keywords for document indexing

Side II: The opportunities

The language-independent approach

- In many instances the proposed approach has already been implemented or prepared
- A prominent example: Google (http://www.google.com) is **not** based on language-dependent preprocessing (stemming)

Benefits of this activity

- The results of these studies are likely to impact on important areas of interest:
  - the EU priorities to bring ICT to the citizen ("e-inclusion")
  - the Indian Minister of Communications and Information Technology agenda, point 9 ("Language Computing")

- However, the **fact itself** of working on these topics has **already had an impact** over creation of multi-language communities

Widening the network

As a result of the Project's activities,
more initiatives and new partnerships have been launched by WG4/WG8 participants:

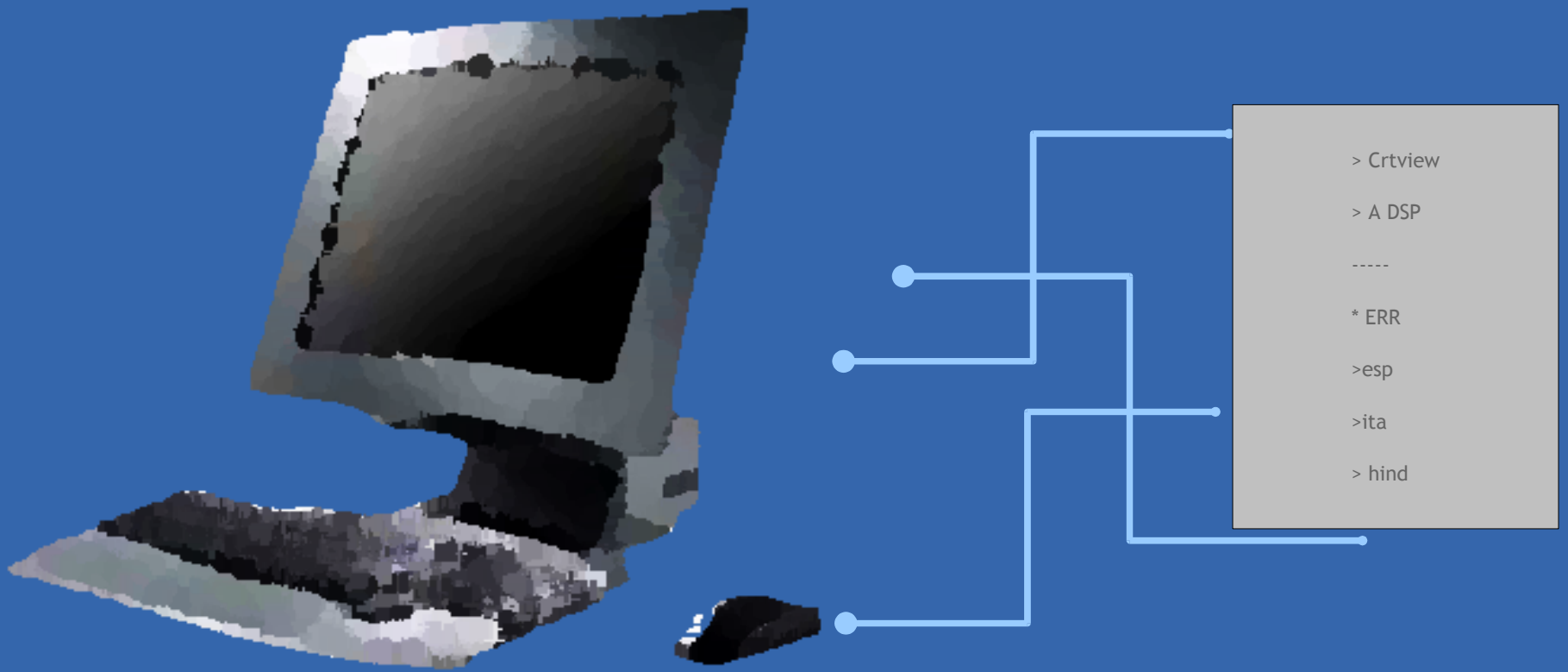
- Research cooperation with Indian Statistical Institute, Kolkata
- Partnership and cooperation with **other European research centres** on document and language technology (from Greece and Switzerland)
- **Hosting more young Indian researchers** with support from the Italian Ministry of University

A golden coin

- We believe that the expected benefits,
  are of **great** importance in building and supporting
  multi-language communities

- The benefits **already achieved** are a confirmation

Preview: WG8 contribution



Workgroup 8

- WG8 is dedicated to the following topic

  “Semantic Information Retrieval:
  A Natural Language Processing Task”

- Start: September 2005 — End: April 2006

- The Genoa contribution is focused on **automatic keyword extraction**

The Vector Space model

- It is the main approach of the field
- **Represents a document as a list of keywords**
- Keywords are extensive
  i.e. Take all terms as keywords – Exclude only some
- **How do we know what keywords are important?**
- Knowledge of the topic and the language is necessary

Natural language processing

- Alternative, powerful approach
- **The content of documents is analyzed at the grammatical and semantic levels**
- We need to store the knowledge about languages in resources such as
  - **a corpus (or training collection)**
  - **an ontology (or semantic network)**

Language independence

- The approach with methods **learning from examples** is a third way

- Combines implicit **semantic informations** with **language independence**

Automatic keyword selection

- All terms in a document are possible keywords
- But not all would make for **good** keywords
- A method has been developed to identify the most relevant terms
- The method is **fully automatic**
  and focused on the task of **document clustering**

Expected results

- WG8 is focused on
  taking into account the meaning of documents
  (semantic analysis)
- The keyword selection method provides
  an automatic evaluation of which terms are interesting (useful)
- This is learned from examples and therefore
  independently from the specific language
- The method works also for
  multi-language documents

Final remarks

The approach

- Accessing collections of documents is **one of the key points** for cooperation in teams and communities
- The main requirement in multilingual communications is **language independent methods**
- We try not to rely only only on pre-existing resources
  - **methods based on learning from data**

Summary of Genoa contribution to WG 4 and WG 8

- Workgroup 4 provided tools for **automatic organization** of collections of documents
- Workgroup 8 is working on techniques to exploit **the content of documents and their meaning**
- The Genova group is studying **techniques to automatically find relevant keywords** from documents **in a language-independent setting**
- Community building is being widened **outside the project consortium**

— the end —