ICT for EU-India Cross-Cultural Dissemination

University of Genova

Department of Computer and Information Sciences

PROJECT CO-FUNDED BY
THE EUROPEAN COMMISSION

Natural language processing: turning barriers into bridges

Stefano Rovetta

DISI – University of Genova, Italy

Outline

- The role of language in digital communication
- What is “natural language processing”?
- The activities of University of Genova – DISI

The focus of the project: Creating Communities

Community is made of communication

>> Digital communication:

information and communication technologies

at the heart of knowledge exchange (cultural and technical)

ability to keep people tightly in touch,

irrespective of their actual location

is it enough?

- digital libraries
- distance learning
- cultural heritage preservation
- documentation

Sophisticated tools to assist users and maintainers

for automatic organization / fast access to digital contents

>> all based on the use of human natural languages

The role of natural language

- Documents are written in natural language(s)
  – not digital codes
- Humans interact in natural language(s)
- Users of digital services often prefer natural language for interacting with computers and programs

Two issues

1. How can computers understand human language?
   - instructing the machine
   - storing, organizing and accessing written documents

2. How can we deal with cross-cultural environments?
   - different languages
   - different conventions

About cross-cultural environments

India and Europe together are a cross-cultural community

...but India and Europe by themselves are already cross-cultural communities

Communicating in a neutral language is not always a solution

Diversity is a wealth, not a problem and should be preserved!

Natural language processing

The discipline which studies the interaction between man and machine using natural (human) language

A wide and complex area!

The activities of the University of Genova – DISI

Two workgroups
(cooperation with Universidad Politècnica de Valencia)

>> Topic of WG4:

Clustering techniques for document organisation and retrieval

>> Topic of WG8:

Semantic Information Retrieval:
A Natural Language Processing Task

Workgroup 4: Clustering techniques for document organisation and retrieval

Organizing collections of documents written in natural language

with the aim of speeding up and optimizing access to contents

Organizing documents

**When searching a document collection,
it is often useful to provide it with a structure**

>> Example: categories in a library

**Digital document collections:
structure should be provided automatically**

Clustering

An often-used automatic organization scheme

Items to be organized are grouped in clusters

A cluster is analogous to a category,
but arises from similarities and dissimilarities
between items

— whereas a category is defined externally

Clustering document collections

Documents can be clustered
on the basis of different similarity criteria

In accessing information, clustering can be used
in two ways:

>> Preliminarily

Documents are organized in a way
that speeds up and guides the search process

>> A posteriori

The documents to be accessed are presented to the user in groups with similar content

The activity

Focused on the research on novel clustering techniques and their application to the information retrieval task

Some original clustering algorithms have been proposed by the participants

- one is especially suitable for the organisation of a document collection
- the second is useful in the presentation of search results

Comments on WG4 activity

- Clustering is a technique based only on the available collection of items (documents) and not on detailed external knowledge

- Document clustering only needs a way to assess similarity or dissimilarity

- It is easy to use in collections written in any language; may also be used for cross-language organization (clustering documents written in different languages)

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

Introducing semantic knowledge into document retrieval methods

for providing the user with answers which better fulfill the requirements (both expressed and implied)

Information Retrieval

**Objective: given a query by the user,
the system should return all and only the documents
which fulfill the user's request**

**Often user's requests are expressed as keywords;
documents are represented by the words they contain**

**>> Such systems cannot retrieve documents
on the basis of similarity in meaning**

Adding semantics

For inexperienced users, it is crucial to take also into account document meaning

>> add word semantics to word-based access

We obtain methods for the search of conceptual information

Objective of the work

The aim is applying semantic information retrieval to the search of Web documents

The main problem in semantic information processing is eliminating ambiguities in the meaning of words

>> Word Sense Disambiguation

The activity

Knowledge-based approaches to word sense disambiguation will be merged with keyword-based information retrieval methods

Knowledge-based approaches rely on external knowledge sources

>> For instance: an ontology

Ontologies

An ontology is a collection of pieces of knowledge

A semantic ontology is a collection of relationships between word meanings in a vocabulary

Example: apple IS A fruit

Comments on WG8 activity

- This approach to conceptual information retrieval can be tuned to different languages by switching to a different ontology for each language

- Multilingual information retrieval is obtained by automatic language recognition and activation of the relevant ontology

Comments

The activities of WG4 and WG8 aim at providing new tools for accessing collections of documents in a way that is:

- **more time-efficient**
- **more manageable by the user**
- **accessible to non-technically literate people**
- **accessible to people from different cultures and languages**
- **usable on collections written in different languages**

Conclusions

We believe that the technologies addressed in our activity are one of the key factors for effective building of cross-cultural communities of

- **scholars**
- **students**
- **professionals**

needed in the close future for cooperative development of the Indian and European countries