Multilingual Content Encoding and Translation

Panel Session at LREC-2000

Antonio Sanfilippo

European Commission,
Information Society DG, Luxembourg,
Antonio.Sanfilippo@cec.eu.int

1. Topic

Text encoding protocols and language description standards for online translation and localisation services.

2. Panel Participants

2.1. Moderator

Antonio Sanfilippo, European Commission, Information Society DG, Luxembourg, Antonio.Sanfilippo@cec.eu.int

2.2. Panelists

- Roberto Cencioni, European Commission, Information Society DG, Luxembourg, Roberto.Cencioni@cec.eu.int
- Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR, Italy, glottolo@ilc.pi.cnr.it
- Bente Maegard, Center for Sprogteknologi, Denmark, bente@cst.ku.dk
- Yuji Matsumoto, Graduate School of Information Science, Nara Institute of Science and Technology, Japan, matsu@is.aist-nara.ac.jp
- Dimitri Theologitis, European Commission, Translation Service, Luxembourg, Dimitri.Theologitis@cec.eu.int
- Gregor Thurmair, SAIL Labs, Germany, gregor.thurmair@sail-labs.de
- Jun-ichi Tsujii, University of Tokyo, Japan and University of Manchester Institute of Science and Technology, UK, tsujii@is.s.u-tokyo.ac.jp
- Antonio Zampolli, UPI and Istituto di Linguistica Computazionale del CNR, Italy, pisa@ilc.pi.cnr.it

3. Introduction

Augmenting Web technologies with real-time automated translation can reduce the cost of doing business and increase international co-operation by allowing users to cope effectively with global communication tasks. Web-based machine translation services such as Systran on Altavista provide useful and interesting ways of tackling the issue of global communication. However, current machine translation technologies fall short of achieving an adequate answer to global communication needs because of limitations in dealing with general unrestricted text. The goal of this panel is to envisage ways in which such limitations can be addressed in the face of emerging communication technologies and the evolving behaviour of the Web community.

4. Multilingual Content Encoding and Translation

One of the major bottlenecks for current machine translation systems resides in the analysis and understanding of the source language text. Natural occurring language exhibits such a variety in content representation and communication strategies that the provision of a correct analysis solution for any arbitrary input text is still below acceptable levels. This is a major hindrance for machine translation systems since a proper understanding of the source text is a sine qua non for attaining a successful translation.

Nevertheless, current work shows that the language understanding problem in Automated Translation can be successfully addressed reducing the analysis space for the source text through techniques such as adaptation to sub-language domains, and the use of translation memories and controlled languages. A further measure consists in enabling users to generate linguistic representations where all/most analysis conflicts and uncertainties in the source text are solved interactively so as to facilitate generation into the target language(s).

The combined effect of these techniques has yielded encouraging results in sector specific localisation activities, but has been of little avail for Web-based online translation services so far. This is largely due to the fact that there are still no widely accepted operative text encoding protocols and language description standards available which might enable the provision of linguistic content in a format more amenable to localisation. The absence of such protocols and standards pre-empts the creation of Web authoring tools which would allow users to take advantage of techniques geared to render Automated Translation more robust and portable --- e.g. those techniques described above.

5. Intended Audience and Discussion Themes

The panel is intended to provide a forum in which to exchange information, voice opinions and share visions on issues related to the creation and deployment of language technology standards with reference to the requirements of multilingual communication technologies and applications. Some of the issues which tabled for discussion include:

Current translation technologies: data encoding problems and limitations for online usage

Text encoding protocols and language description standards for translation and multilingual authoring tools: thematic priorities and promotion strategies.

Inter-lingual structured data encoding for multilingual generation.

6. Take-Home Message

The current combined growth of internet penetration and socio-economic globalisation requires further immediate development of Web-based technologies which provide users with seamless information access and dissemination across languages, e.g. robust and reliable translation and multilingual authoring tools. The availability of widely accepted operative text encoding protocols and language description standards is essential for the creation of such technologies.