A Web-based Interface to a Multi-lingual Phrase-based Translation System

Roldano Cattoni, Nicola Bertoldi, Mauro Cettolo, Boxing Chen and Marcello Federico
ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
38050 Povo - Trento, Italy
{surname}@itc.it

1 Introduction

In this demonstration we present our multi-lingual phrase-based Statistical Machine Translation system (Cettolo et al., 2005) which can be accessed by means of a Web page. In section 2 the software architecture of the demonstrator is outlined. Section 3 focuses on the currently supported language-pairs: Arabic-to-English, Chinese-to-English, Spanish-to-English, Italian-to-English and English-to-Italian. In section 4 the Web-based interface of the demo is described.

2 Demo Architecture

Figure 1 shows the two-layer architecture of the demo. At the bottom lie the programs that provide the actual translation services: for each language-pair a wrapper coordinates the activity of a specialized pre-processing tool and a MT decoder. The translation programs run on a grid-based cluster of high-end PCs to optimize the processing speed. All the wrappers communicate with the MT front-end whose main task is to forward translation requests to the appropriate language-pair wrapper and to report an error in case of wrong requests (e.g. unsupported language-pair). It is worth noticing here that a new language-pair can be easily added to the system with a minimal intervention on the code of the MT front-end.

At the top of the architecture are the programs that provide the interface with the user. This layer is separated from the translation layer (hosted by internal machines only) by means of a firewall. The user interface is implemented as a Web page in which a translation request (a source sentence and a language-pair) is input by means of an HTML form. The cgi script invocated by the form manages the interaction with the MT front-end.

When a user issues a translation request after filling the form fields, the cgi script sends the request to the MT front-end and waits for its reply. The input sentence is then forwarded to the wrapper of the appropriate language-pair. After a preprocessing step, the actual translation is performed by the specific MT decoder. The output in the target language is then sent back to the user's Web browser through the chain in the reverse order.

From a technical point of view, the inter-process communication is realized by means of standard TCP-IP sockets. As far as the encoding of texts is concerned, all the languages are encoded in UTF-8: this allows to manage the processing phase in an uniform way and to render graphically different character sets.

3 The supported language-pairs

Although there is no theoretical limit to the number of supported language-pairs, the current version of the demo provides translations to English from four source languages: Arabic, Chinese, Spanish and Italian. Moreover translation from English to Italian is also supported.

Arabic-to-English (Tourism)

The Arabic-to-English system has been trained with the data provided by the International Workshop on Spoken Language Translation 2005 The context is that of the Basic Traveling Expression Corpus (BTEC) task (Takezawa et al., 2002). BTEC is a multilingual speech corpus which contains sentences coming from phrase books for tourists. Training set includes 20k sentences containing 159K Arabic and 182K English running words; vocabulary size is 18K for Arabic, 7K for English.

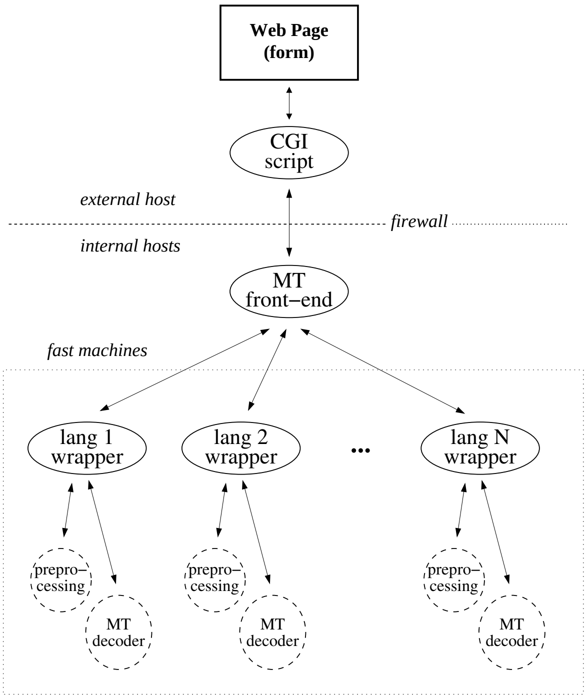

Figure 1: Architecture of the demo. For each language-pair a set of programs (in particular the MT decoder) provides the translation service. The request issued by the user on the Web page is sent by the cgi script to the MT front-end. The translation is then performed on the appropriate language-pair service and the output sent back to the Web browser.

Chinese-to-English (Newswire)

The Chinese-to-English system has been trained with the data provided by the NIST MT Evaluation Campaign 2005 , large-data condition. In this case parallel data are mainly news-wires provided by news agencies. Training set includes 71M Chinese and 77M English running words; vocabulary size is 157K for Chinese, 214K for English.

Spanish-to-English (European Parliament)

The Spanish-to-English system has been trained with the data provided by the Evaluation Campaign 2005 of the European integrated project TC-STAR[1]. The context is that of the speeches of the European Parliament Plenary sessions (EPPS) from April 1996 to October 2004. Training set for the Final Text Edition transcriptions includes 31M Spanish and 30M English running words; vocabulary size is 140K for Spanish, 94K for English.

Italian-to-English and English-to-Italian (European Parliament + Tourism)

The Italian-to-English system has been trained with two bilingual corpora (1) the *Europarl* corpus (Koehn, 2005), extracted from the proceedings of the European Parliament in the 1996-2003 period and (2) a larger version of the above described BTEC corpus. The joined training set includes 30M Italian and 30M English running words; vocabulary size is 141K for Italian, 98K for English.

4 The Web-based Interface

As highlighted by the demo storyboard, in the upper part of the Web-page the user provides the two information required for the translation: the source sentence can be input in a 80x5 *textarea* html structure, while the language-pair can be selected by means of a set a *radio-buttons*. The user can reset the input area or send the translation request by means of standard reset and submit buttons. Some examples of bilingual sentences are provided in the lower part of the page.

The output of a translation request is presented in a box in the middle of the page.

We plan to extend the interface with the possibility for the user to ask additional information about the translation – e.g. the number of explored theories or the score of the first-best translation.

5 Acknowledgements

This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

References

Mauro Cettolo, Marcello Federico, Nicola Bertoldi, Roldano Cattoni, and Boxing Chen. 2005. A look inside the itc-irst smt system. In *Proceedings of the 10th Machine Translation Summit*, pages 451–457, Phuket, Thailand, September.

Philipp Koehn. 2005. Europarl: A Multilingual Corpus for Evaluation of Machine Translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, September.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a Broad-Coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World. In *Proceedings of 3rd LREC*, pages 147–152, Las Palmas, Spain.

[1] http://www.tc-star.org