A Web-based Demonstrator of 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

Abstract

This paper describes a multi-lingual phrase-based Statistical Machine Translation system accessible by means of a Web page. The user can issue translation requests from Arabic, Chinese or Spanish into English. The same phrase-based statistical technology is employed to realize the three supported language-pairs. New language-pairs can be easily added to the demonstrator. The Web-based interface allows the use of the translation system to any computer connected to the Internet.

1 Introduction

At this time, Statistical Machine Translation (SMT) has empirically proven to be the most competitive approach in international competitions like the NIST Evaluation Campaigns[1] and the International Workshops on Spoken Language Translation (IWSLT-2004[2] and IWSLT-2005[3]).

In this paper we describe our multi-lingual phrase-based Statistical Machine Translation system which can be accessed by means of a Web page. Section 2 presents the general log-linear framework to SMT and gives an overview of our phrase-based SMT system. In section 3 the software architecture of the demo is outlined. Section 4 focuses on the currently supported language-pairs: Arabic-to-English, Chinese-to-English and Spanish-to-English. In section 5 the Web-based interface of the demo is described.

[1] http://www.nist.gov/speech/tests/mt/
[2] http://www.slt.atr.jp/IWSLT2004/
[3] http://www.is.cs.cmu.edu/iwslt2005/

2 SMT System Description

2.1 Log-Linear Model

Given a string $\mathbf{f}$ in the source language, the goal of the statistical machine translation is to select the string $\mathbf{e}$ in the target language which maximizes the posterior distribution $\Pr(\mathbf{e} \mid \mathbf{f})$. By introducing the hidden word *alignment* variable $\mathbf{a}$, the following approximate optimization criterion can be applied for that purpose:

$$\begin{aligned} \mathbf{e}^* &= \arg\max_{\mathbf{e}} \Pr(\mathbf{e} \mid \mathbf{f}) \\ &= \arg\max_{\mathbf{e}} \sum_{\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \\ &\approx \arg\max_{\mathbf{e},\mathbf{a}} \Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \end{aligned}$$

Exploiting the maximum entropy (Berger et al., 1996) framework, the conditional distribution $\Pr(\mathbf{e}, \mathbf{a} \mid \mathbf{f})$ can be determined through suitable real valued functions (called *features*) $h_r(\mathbf{e}, \mathbf{f}, \mathbf{a}), r = 1 \ldots R$, and takes the parametric form:

$$p_{\boldsymbol{\lambda}}(\mathbf{e}, \mathbf{a} \mid \mathbf{f}) \propto \exp\{\sum_{r=1}^{R} \lambda_r h_r(\mathbf{e}, \mathbf{f}, \mathbf{a})\}$$

The ITC-irst system (Chen et al., 2005) is based on a log-linear model which extends the original IBM Model 4 (Brown et al., 1993) to *phrases* (Koehn et al., 2003; Federico and Bertoldi, 2005). In particular, target strings $\mathbf{e}$ are built from sequences of phrases $\tilde{e}_1 \ldots \tilde{e}_l$. For each target phrase $\tilde{e}$ the corresponding source phrase within the source string is identified through three random quantities: the *fertility* ϕ, which establishes its length; the *permutation* π_i, which sets its first position; the *tablet* $\tilde{f}$, which tells its word string. Notice that target phrases might have fertility equal to zero, hence they do not translate any

source word. Moreover, uncovered source positions are associated to a special target word (*null*) according to specific fertility and permutation random variables.
The resulting log-linear model applies eight feature functions whose parameters are either estimated from data (e.g. target language models, phrase-based lexicon models) or empirically fixed (e.g. permutation models). While feature functions exploit statistics extracted from monolingual or word-aligned texts from the training data, the scaling factors λ of the log-linear model are estimated on the development data by applying a *minimum error training* procedure (Och, 2004).

2.2 Decoding Strategy

The translation of an input string is performed by the SMT system in two steps. In the first pass a beam search algorithm (decoder) computes a word graph of translation hypotheses. Hence, either the best translation hypothesis is directly extracted from the word graph and output, or an N-best list of translations is computed (Tran et al., 1996). The N-best translations are then re-ranked by applying additional features and the top ranking translation is finally output.

The decoder exploits dynamic programming, that is the optimal solution is computed by expanding and recombining previously computed partial theories. A theory is described by its *state* which is the only information needed for its expansion. Expanded theories sharing the same state are recombined, that is only the best scoring one is stored for further expansions. In order to output a word graph of translations, backpointers to all expanded theories are mantained, too.

To cope with the large number of generated theories some approximations are introduced during the search: less promising theories are pruned off (*beam search*) and a new source position is selected by limiting the number of vacant positions on the left-hand and the distance from the left most vacant position (*re-ordering constraints*).

2.3 Phrase extraction and model training

Training of the phrase-based translation model requires a parallel corpus provided with word-alignments in both directions, i.e. from source to target positions, and viceversa. This pre-processing step can be accomplished by applying the GIZA++ toolkit (Och and Ney, 2003) that provides Viterbi alignments based on IBM Model-4.

Starting from the parallel training corpus, provided with direct and inverted alignments, the so-called *union alignment* (Och and Ney, 2003) is computed.
Phrase-pairs are extracted from each sentence pair which correspond to sub-intervals of the source and target positions, J and I, such that the union alignment links all positions of J into I and all positions of I into J. In general, phrases are extracted with maximum length in the source and target defined by the parameters J_{max} and I_{max}. All such phrase-pairs are efficiently computed by an algorithm with complexity $\mathcal{O}(lI_{max}J_{max}^2)$ (Cettolo et al., 2005).

Given all phrase-pairs extracted from the training corpus, lexicon probabilities and fertility probabilities are estimated.

Target language models (LMs) used by the decoder and rescoring modules are, respectively, estimated from 3-gram and 4-gram statistics by applying the *modified Kneser-Ney* smoothing method (Goodman and Chen, 1998). LMs are estimated with an in-house software toolkit which also provides a compact binary representation of the LM which is used by the decoder.

3 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.

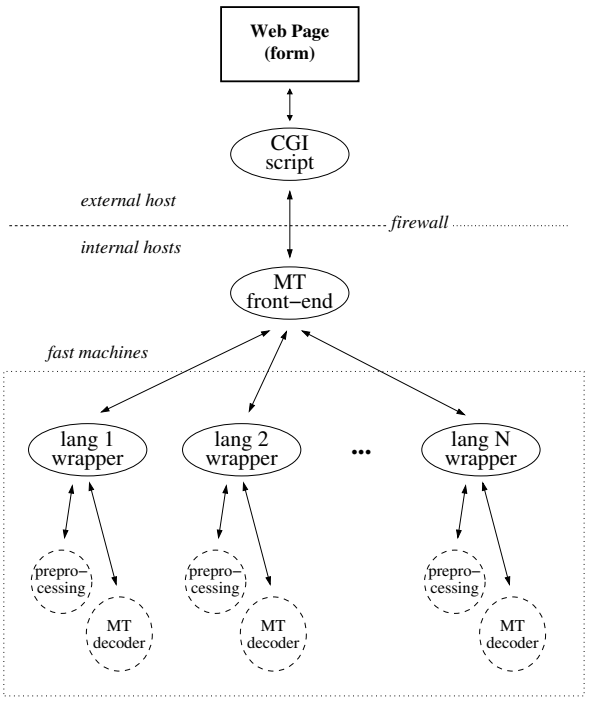

Figure 1: Architecture of the demo. For each language-pair a set of programs (in particular the MT decoder) provides the translation service. The request issued by the user on the Web page is sent by the cgi script to the MT front-end. The translation is then performed on the appropriate language-pair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

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.

4 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 three source languages: Arabic, Chinese and Spanish. For demonstration purpose, three different application domains are covered too.

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.

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[4]. 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.

5 The Web-based Interface

Figure 2 shows a snapshot of the Web-based interface of the demo – the URL has been removed to make this submission anonymous. In the upper part of the 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.

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

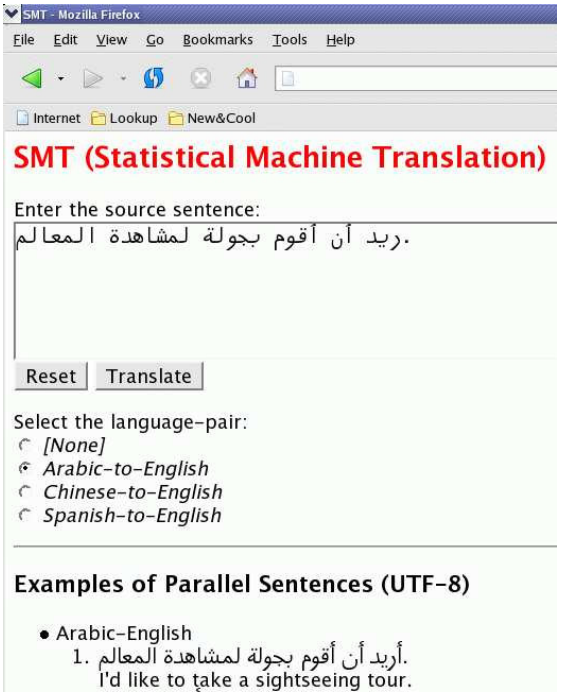

Figure 2: A snapshot of the Web-based interface. The user provides the sentence to be translated in the desired language-pair. Some examples of bilingual sentences are also available to the user.

The output of a translation request is simple: the requested source sentence, the translation in the target language and the selected language-pair are presented to the user. Figure 3 shows an example of an Arabic sentence translated into English.

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.

6 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

A.L. Berger, S.A. Della Pietra, and V.J. Della Pietra. 1996. A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics*, 22(1):39–71.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–313.

Figure 3: Example of an Arabic sentence translated into English.

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.

B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico. 2005. The ITC-irst SMT System for IWSLT-2005. In *Proceedings of the IWSLT 2005*, Pittsburgh, USA.

M. Federico and N. Bertoldi. 2005. A Word-to-Phrase Statistical Translation Model. *ACM Transactions on Speech and Language Processing*. to appear.

J. Goodman and S. Chen. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Harvard University, August.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of HLT-NAACl 2003*, pages 127–133, Edmonton, Canada.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F.J. Och. 2004. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, Sapporo, Japan.

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.

B. H. Tran, F. Seide, and V. Steinbiss. 1996. A Word Graph based N-Best Search in Continuous Speech Recognition. In *Proceedings of ICLSP*, Philadelphia, PA, USA.