An Interactive Machine Translation System with Online Learning

Daniel Ortiz-Martínez, Luis A. Leiva, Vicent Alabau,
Ismael García-Varea†, Francisco Casacuberta
ITI - Institut Tecnològic d'Informàtica, Universitat Politècnica de València
† Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha
{dortiz,luileito,valabau,fcn}@iti.upv.es, †ismael.garcia@uclm.es

Abstract

State-of-the-art Machine Translation (MT) systems are still far from being perfect. An alternative is the so-called Interactive Machine Translation (IMT) framework, where the knowledge of a human translator is combined with the MT system. We present a statistical IMT system able to learn from user feedback by means of the application of online learning techniques. These techniques allow the MT system to update the parameters of the underlying models in real time. According to empirical results, our system outperforms the results of conventional IMT systems. To the best of our knowledge, this online learning capability has never been provided by previous IMT systems. Our IMT system is implemented in C++, JavaScript, and ActionScript; and is publicly available on the Web.

1 Introduction

The research in the field of machine translation (MT) aims to develop computer systems which are able to translate text or speech without human intervention. However, current translation technology has not been able to deliver full automated high-quality translations. Typical solutions to improve the quality of the translations supplied by an MT system require manual post-editing. This serial process prevents the MT system from integrating the knowledge of the human translator.

An alternative way to take advantage of the existing MT technologies is to use them in collaboration with human translators within a computer-assisted translation (CAT) or interactive framework (Isabelle and Church, 1997). Interactivity in CAT has been explored for a long time. Systems have been designed to interact with linguists to solve ambiguities or update user dictionaries.

An important contribution to CAT technology was pioneered by the *TransType* project (Foster et al., 1997; Langlais et al., 2002). The idea proposed in that work was to embed data driven MT techniques within the interactive translation environment. Following the TransType ideas, Barrachina et al. (2009) proposed the so-called IMT framework, in which fully-fledged statistical MT (SMT) systems are used to produce full target sentences hypotheses, or portions thereof, which can be accepted or amended by a human translator. Each corrected text segment is then used by the MT system as additional information to achieve improved suggestions. Figure 1 shows an example of a typical IMT session.

The vast majority of the existing work on IMT makes use of the well-known *batch learning* paradigm. In the batch learning paradigm, the training of the IMT system and the interactive translation process are carried out in separate stages. This paradigm is not able to take advantage of the new knowledge produced by the user of the IMT system. In this paper, we present an application of the *online learning* paradigm to the IMT framework. In the online learning paradigm, the training and prediction stages are no longer separated. This feature is particularly useful in IMT since it allows to take into account the user feedback. Specifically, our proposed IMT system can be extended with the new training samples that are generated each time the user validates the translation of a given source sentence. The online learning techniques implemented in our IMT system incrementally update the statistical models involved in the translation process.

2 Related work

There are some works on IMT in the literature that try to take advantage of user feedback. One example is the work by Nepveu et al. (2004), where dynamic adaptation of an IMT system via cache-based model extensions to language and translation models is proposed. One major drawback of such proposal is its inability to learn new words.

Proceedings of the ACL-HLT 2011 System Demonstrations, pages 68–73,
Portland, Oregon, USA, 21 June 2011. ©2011 Association for Computational Linguistics

source($\mathbf{f}$):		Para ver la lista de recursos
reference($\hat{\mathbf{e}}$):		To view a listing of resources
interaction-0	$\mathbf{e}_p$	
	$\mathbf{e}_s$	*To view the resources list*
interaction-1	$\mathbf{e}_p$	To view
	k	**a**
	$\mathbf{e}_s$	*list of resources*
interaction-2	$\mathbf{e}_p$	To view a list
	k	**i**
	$\mathbf{e}_s$	*ng resources*
interaction-3	$\mathbf{e}_p$	To view a listing
	k	**o**
	$\mathbf{e}_s$	*f resources*
accept	$\mathbf{e}_p$	To view a listing of resources

Figure 1: IMT session to translate a Spanish sentence into English. In interaction-0, the system suggests a translation ($\mathbf{e}_s$). In interaction-1, the user moves the mouse to accept the first eight characters "To view " and presses the **a** key (k), then the system suggests completing the sentence with "*list of resources*" (a new $\mathbf{e}_s$). Interactions 2 and 3 are similar. In the final interaction, the user accepts the current suggestion.

Recent research on IMT has proposed the use of online learning as one possible way to successfully incorporate user feedback in IMT systems (Ortiz-Martínez et al., 2010). In the online learning setting, models are trained sample by sample. For this reason, such learning paradigm is appropriate for its use in the IMT framework. The work by Ortiz-Martínez et al. (2010) implements online learning as incremental learning. Specifically, an IMT system able to incrementally update the parameters of all of the different models involved in the interactive translation process is proposed. One previous attempt to implement online learning in IMT is the work by Cesa-Bianchi et al. (2008). In that work, the authors present a very constrained version of online learning, which is not able to extend the translation models due to the high time cost of the learning process.

We have adopted the online learning techniques proposed in (Ortiz-Martínez et al., 2010) to implement our IMT system. We are not aware of other IMT tools that include such functionality. For instance, a prototype system for text prediction to help translators is shown in (Foster et al., 2002). Additionally, Koehn (2009) presents the *Caitra* translation tool. *Caitra* aids linguists suggesting sentence completions, alternative words or allowing users to post-edit machine translation output. However, neither of these systems are able to take advantage of the user validated translations.

3 Interactive Machine Translation

IMT can be seen as an evolution of the statistical machine translation (SMT) framework. In SMT, given source string $\mathbf{f}$, we seek for the target string $\mathbf{e}$ which maximizes the posterior probability:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}}\, Pr(\mathbf{e}|\mathbf{f}) \tag{1}$$

Within the IMT framework, a state-of-the-art SMT system is employed in the following way. For a given source sentence, the SMT system automatically generates an initial translation. A human translator checks this translation from left to right, correcting the first error. The SMT system then proposes a new extension, taking the correct prefix $\mathbf{e}_p$ into account. These steps are repeated until the whole input sentence has been correctly translated. In the resulting decision rule, we maximize over all possible extensions $\mathbf{e}_s$ of $\mathbf{e}_p$:

$$\hat{\mathbf{e}}_s = \underset{\mathbf{e}_s}{\operatorname{argmax}}\, Pr(\mathbf{e}_s|\mathbf{e}_p, \mathbf{f}) \tag{2}$$

It is worth to note that the user interactions are at character level, that is, for each submitted keystroke the system provides a new extension (or suffix) to the current hypothesis. A typical IMT session for a given source sentence is depicted in Figure 1.

State-of-the-art SMT systems follow a log-linear approach (Och and Ney, 2002), where the posterior

probability $Pr(\mathbf{e}\,|\,\mathbf{f})$ of Eq. (1) is used. Such log-linear approach can be easily adapted for its use in the IMT framework as follows:

$$\hat{\mathbf{e}}_s = \underset{\mathbf{e}_s}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}_p, \mathbf{e}_s, \mathbf{f}) \right\} \quad (3)$$

where each $h_m(\mathbf{e}_p, \mathbf{e}_s, \mathbf{f})$ is a feature function representing a statistical model and λ_m its corresponding weight. Typically, a set of statistical generative models are used as feature functions. Among this feature functions, the most relevant are the language and translation models. The language model is implemented using statistical n-gram language models and the translation model is implemented using phrase-based models.

The IMT system proposed here is based on a log-linear SMT system which includes a total of seven feature functions: an n-gram language model, a target sentence length model, inverse and direct phrase-based models, source and target phrase length models and a reordering model.

4 Online Learning

In the online learning paradigm, *learning* proceeds as a sequence of trials. In each trial, a sample is presented to the learning algorithm to be classified. Once the sample is classified, its correct label is told to the learning algorithm.

The online learning paradigm fits nicely in the IMT framework, since the interactive translation of the source sentences generates new user-validated training samples that can be used to extend the statistical models involved in the translation process.

One key aspect in online learning is the time required by the learning algorithm to process the new training samples. One way to satisfy this constraint is to obtain incrementally updateable versions of the algorithms that are executed to train the statistical models involved in the translation process. We have adopted this approach to implement our IMT system. Specifically, our proposed IMT system implements the set of training algorithms that are required to incrementally update each component of the log-linear model. Such log-linear model is composed of seven components (see section 3). One key aspect of the required training algorithms is the necessity to replace the conventional expectation-maximization (EM) algorithm by its incremental version (Neal and Hinton, 1998). The complete details can be found in (Ortiz-Martínez et al., 2010).

5 System Overview

In this section the main features of our prototype are shown, including prototype design, interaction protocol, prototype functionalities and demo usage.

5.1 Prototype Design

Prototype architecture has been built on two main aspects, namely, accessibility and flexibility. The former is necessary to reach a larger number of potential users. The latter allows researchers to test different techniques and interaction protocols.

For that reason, we developed an CAT Application Programming Interface (API) between the client and the actual translation engine, by using a network communication protocol and exposing a well-defined set of functions.

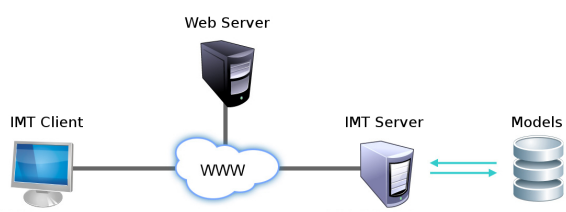


Figure 2: IMT system architecture.

A diagram of the architecture is shown in Figure 2. On the one hand, the IMT client provides a User Interface (UI) which uses the API to communicate with the IMT server through the Web. The hardware requirements in the client are very low, as the translation process is carried out remotely on the server, so virtually any computer (including netbooks, tablets or 3G mobile phones) should be fairly enough. On the other hand, the server, which is unaware of the implementation details of the IMT client, uses and adapts the statistical models that are used to perform the translation.

5.2 User Interaction Protocol

The protocol that rules the IMT process has the following steps:

1. The system proposes a full translation of the selected text segment.

Figure 3: Demo interface. The source text segments are automatically extracted from source document. Such segments are marked as pending (light blue), validated (dark green), partially translated (light green), and locked (light red). The translation engine can work either at full-word or character level.

2. The user validates the longest prefix of the translation which is error-free and/or corrects the first error in the suffix. Corrections are entered by amendment keystrokes or mouse clicks/wheel operations.
3. In this way, a new extended consolidated prefix is produced based on the previous validated prefix and the interaction amendments. Using this new prefix, the system suggests a suitable continuation of it.
4. Steps 2 and 3 are iterated until the user-desired translation is produced.
5. The system adapts the models to the new validated pair of sentences.

5.3 Prototype Functionality

The following is a list of the main features that the prototype supports:

- When the user corrects the solution proposed by the system, a new improved suffix is presented to the user.
- The system is able to *learn* from user-validated translations.
- The user is able to perform *actions* by means of keyboard shortcuts or mouse gestures. The supported actions on the proposed suffix are:

  Substitution Substitute the first word or character of the suffix.

  Deletion Delete the first word of the suffix.

  Insertion Insert a word before the suffix.

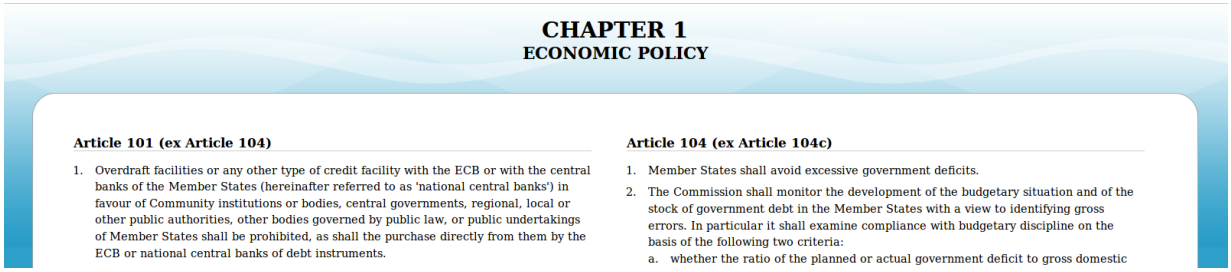
  Rejection The rejected word will not appear in the following proposals.

  Acceptance Assume that the current translation is correct and *adapt* the models.
- At any time, the user is able to visualize the original document (Figure 4(a)), as well as a properly formated draft of the current translation (Figure 4(b)).
- Users can select the document to be translated from a list or upload their own documents.

5.4 Demo Description and Usage

This demo exploits the WWW to enable the connection of simultaneous accesses across the globe, coordinating client-side scripting with server-side technologies. The interface uses web technologies such as XHTML, JavaScript, and ActionScript; while the IMT engine is written in C++.

The prototype is publicly available on the Web (`http://cat.iti.upv.es/imt/`). To begin with, the UI loads an index of all available translation corpora. Currently, the prototype can be tested with the well-known Europarl corpora (Koehn, 2005). The user chooses a corpus and navigates to the main interface page (Figure 3), where she interactively translates the text segments one by one. User's feedback is then processed by the IMT server.



CHAPTER 1
ECONOMIC POLICY

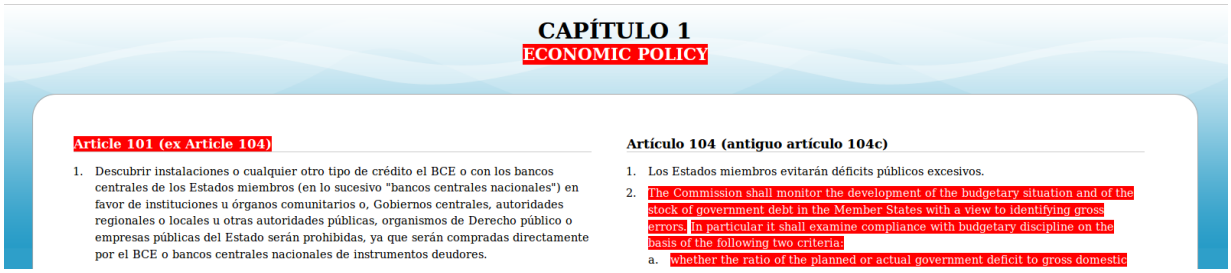
Article 101 (ex Article 104)

1. Overdraft facilities or any other type of credit facility with the ECB or with the central banks of the Member States (hereinafter referred to as 'national central banks') in favour of Community institutions or bodies, central governments, regional, local or other public authorities, other bodies governed by public law, or public undertakings of Member States shall be prohibited, as shall the purchase directly from them by the ECB or national central banks of debt instruments.

Article 104 (ex Article 104c)

1. Member States shall avoid excessive government deficits.
2. The Commission shall monitor the development of the budgetary situation and of the stock of government debt in the Member States with a view to identifying gross errors. In particular it shall examine compliance with budgetary discipline on the basis of the following two criteria:
   a. whether the ratio of the planned or actual government deficit to gross domestic

(a) Source document example, created from EuroParl corpus.



CAPÍTULO 1
ECONOMIC POLICY

Article 101 (ex Article 104)

1. Descubrir instalaciones o cualquier otro tipo de crédito el BCE o con los bancos centrales de los Estados miembros (en lo sucesivo "bancos centrales nacionales") en favor de instituciones u órganos comunitarios o, Gobiernos centrales, autoridades regionales o locales u otras autoridades públicas, organismos de Derecho público o empresas públicas del Estado serán prohibidas, ya que serán compradas directamente por el BCE o bancos centrales nacionales de instrumentos deudores.

Artículo 104 (antiguo artículo 104c)

1. Los Estados miembros evitarán déficits públicos excesivos.
2. The Commission shall monitor the development of the budgetary situation and of the stock of government debt in the Member States with a view to identifying gross errors. In particular it shall examine compliance with budgetary discipline on the basis of the following two criteria:
   a. whether the ratio of the planned or actual government deficit to gross domestic

(b) Translated example document, preserving original format and highlighting non-translated sentences.

Figure 4: Translating documents with the proposed system.

All corrections are stored in plain text logs on the server, so the user can retake them in any moment, also allowing collaborative translations between users. On the other hand, this prototype allows uploading custom documents in text format.

Since the users operate within a web browser, the system also provides crossplatform compatibility and requires neither computational power nor disk space on the client's machine. The communication between application and web server is based on asynchronous HTTP connections, providing thus a richer interactive experience (no page refreshes are required.) Moreover, the Web server communicates with the IMT engine through binary TCP sockets, ensuring really fast response times.

6 Experimental Results

Experimental results were carried out using the Xerox corpus (Barrachina et al., 2009), which consists of translation of Xerox printer manual involving three different language pairs: French-English, Spanish-English, and German-English. This corpus has been extensively used in the literature to report IMT results. The corpus consists of approximately 50,000 sentences pairs for training, 1,000 for development, and 1,000 for test.

The evaluation criteria used in the experiments are the *key-stroke and mouse-action ratio* (KSMR) metric (Barrachina et al., 2009), which measures the user effort required to generate error-free translations, and the well-known BLEU score, which constitutes a measure of the translation quality.

The test corpora were interactively translated from English to the other three languages, comparing the performance of a batch IMT (baseline) and the online IMT systems. The batch IMT system is a conventional IMT system which is not able to take advantage of user feedback after each translation is performed. The online IMT system uses the translations validated by the user to adapt the translation models at runtime. Both systems were initialized with a log-linear model trained in batch mode using the training corpus. Table 1 shows the BLEU score and the KSMR for the batch and the online IMT systems (95% confidence intervals are shown). The BLEU score was calculated from the first translation hypothesis produced by the IMT system for each source sentence. All the obtained improvements with the online IMT system were statistically significant. The average online training time for each new sample presented to the system, and the average response time for each user interaction

(that is, time that the system uses to propose new extensions for corrected prefixes) are also shown in Table 1, which are less than a tenth of a second and around two tenths of a second respectively[1]. According to the reported response and online training times, we can argue that the system proposed here is able to be used on real time scenarios.

	System	BLEU	KSMR	LT/RT (s)
En-Sp	batch	55.1± 2.3	18.2± 1.1	– /0.09
	online	60.6± 2.3	15.8± 1.0	0.04 /0.09
En-Fr	batch	33.7± 2.0	33.9± 1.3	– /0.14
	online	42.2± 2.2	27.9± 1.3	0.09 /0.14
En-Ge	batch	20.4± 1.8	40.3± 1.2	– /0.15
	online	28.0± 2.0	35.0± 1.3	0.07 /0.15

Table 1: BLEU and KSMR results for the XEROX test corpora using the batch and the online IMT systems, reporting the average online learning (LT) and the interaction response times (RP) in seconds.

It is worth mentioning that the results presented here significantly improve those presented in (Barrachina et al., 2009) for other state-of-the-art IMT systems using the same corpora.

7 Conclusions

We have described an IMT system with online learning which is able to learn from user feedback in real time. As far as we know, to our knowledge, this feature have never been provided by previously presented IMT prototypes.

The proposed IMT tool is publicly available through the Web (`http://cat.iti.upv.es/imt/`). Currently, the system can be used to interactively translate the well-known Europarl corpus. We have also carried out experiments with simulated users. According to such experiments, our IMT system is able to outperform the results obtained by conventional IMT systems implementing batch learning. Future work includes researching further on the benefits provided by our online learning techniques with experiments involving real users.

Acknowledgments

Work supported by the EC (FEDER/FSE), the Spanish Government (MEC, MICINN, MITyC, MAEC, "Plan E", under grants MIPRCV "Consolider Ingenio 2010" CSD2007-00018, iTrans2 TIN2009-14511, erudito.com TSI-020110-2009-439), the Generalitat Valenciana (grant Prometeo/2009/014, grant GV/2010/067), the Universitat Politècnica de València (grant 20091027), and the Spanish JCCM (grant PBI08-0210-7127).

[1]All the experiments were executed in a PC with 2.40 GHz Intel Xeon processor and 1GB of memory.

References

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. Lagarda, H. Ney, J. Tomás, and E. Vidal. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

N. Cesa-Bianchi, G. Reverberi, and S. Szedmak. 2008. Online learning algorithms for computer-assisted translation. Deliverable D4.2, SMART: Stat. Multilingual Analysis for Retrieval and Translation.

G. Foster, P. Isabelle, and P. Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194.

G. Foster, P. Langlais, and G. Lapalme. 2002. Transtype: text prediction for translators. In *Proc. HLT*, pages 372–374.

P. Isabelle and K. Church. 1997. Special issue on new tools for human translators. *Machine Translation*, 12(1–2).

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the MT Summit X*, pages 79–86, September.

P. Koehn. 2009. A web-based interactive computer aided translation tool. In *Proc. ACL-IJCNLP*, ACLDemos, pages 17–20.

P. Langlais, G. Lapalme, and M. Loranger. 2002. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 15(4):77–98.

R.M. Neal and G.E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Proc. of the NATO-ASI on Learning in graphical models*, pages 355–368, Norwell, MA, USA.

L. Nepveu, G. Lapalme, P. Langlais, and G. Foster. 2004. Adaptive language and translation models for interactive machine translation. In *Proc. EMNLP*, pages 190–197.

F. J. Och and H. Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. ACL*, pages 295–302.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2010. Online learning for interactive statistical machine translation. In *Proc. NAACL/HLT*, pages 546–554.