Online Learning for Interactive Statistical Machine Translation

Daniel Ortiz-Martínez
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46071 Valencia, Spain
dortiz@dsic.upv.es

Ismael García-Varea
Dpto. de Informática
Univ. de Castilla-La Mancha
02071 Albacete, Spain
ivarea@info-ab.uclm.es

Francisco Casacuberta
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46071 Valencia, Spain
fcn@dsic.upv.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. In this framework, the knowledge of a human translator is combined with a MT system. 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 the user feedback to be taken into account. The online learning techniques proposed here incrementally update the statistical models involved in the translation process. Empirical results show the great potential of online learning in the IMT framework.

1 Introduction

Information technology advances have led to the need for more efficient translation methods. Current MT systems are not able to produce ready-to-use texts. Indeed, MT systems usually require human post-editing to achieve high-quality translations.

One way of taking advantage of MT systems is to combine them with the knowledge of a human translator in the IMT paradigm, which is a special type of the computer-assisted translation paradigm (Isabelle and Church, 1997). An important contribution to IMT technology was pioneered by the *TransType* project (Foster et al., 1997; Langlais et al., 2002) where data driven MT techniques were adapted for their use in an interactive translation environment.

Following the TransType ideas, Barrachina et al. (2009) proposed a new approach to IMT, in which fully-fledged statistical MT (SMT) systems are used to produce full target sentence hypotheses, or portions thereof, which can be partially or completely accepted and amended by a human translator. Each partial, correct text segment is then used by the SMT system as additional information to achieve improved suggestions. Figure 1 illustrates a typical IMT session.

source(**f**): Para ver la lista de recursos
reference($\hat{\mathbf{e}}$): To view a listing of resources

inter.-0	$\mathbf{e}_p$	
	$\mathbf{e}_s$	*To view the resources list*
inter.-1	$\mathbf{e}_p$	To view
	k	**a**
	$\mathbf{e}_s$	*list of resources*
inter.-2	$\mathbf{e}_p$	To view a list
	k	**i**
	$\mathbf{e}_s$	*ng resources*
inter.-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.

In this paper, we also focus on the IMT framework. Specifically, we present an IMT system that is able to learn from user feedback. For this purpose, we apply the *online learning* paradigm to the IMT framework. The online learning techniques that we propose here allow the statistical models involved in the translation process to be incrementally updated.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 546–554,
Los Angeles, California, June 2010.

Figure 2 (inspired from (Vidal et al., 2007)) shows a schematic view of these ideas. Here, $\mathbf{f}$ is the input sentence and $\mathbf{e}$ is the output derived by the IMT system from $\mathbf{f}$. By observing $\mathbf{f}$ and $\mathbf{e}$, the user interacts with the IMT system until the desired output $\hat{\mathbf{e}}$ is produced. The input sentence $\mathbf{f}$ and its desired translation $\hat{\mathbf{e}}$ can be used to refine the models used by the system. In general, the model is initially obtained through a classical batch training process from a previously given training sequence of pairs ($\mathbf{f}_i$,$\mathbf{e}_i$) from the task being considered. Now, the models can be extended with the use of valuable user feedback.

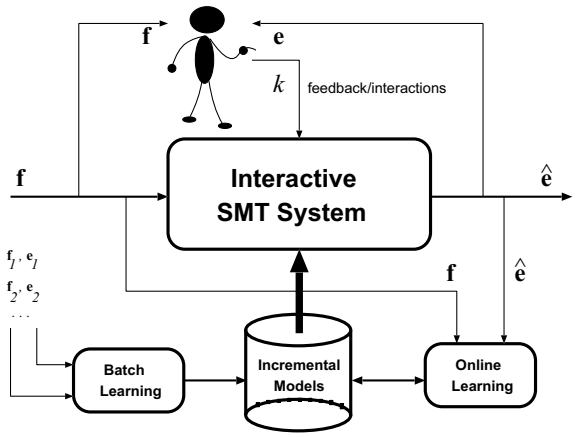


Figure 2: An Online Interactive SMT system

2 Interactive machine translation

IMT can be seen as an evolution of the SMT framework. Given a sentence $\mathbf{f}$ from a source language $\mathcal{F}$ to be translated into a target sentence $\mathbf{e}$ of a target language $\mathcal{E}$, the fundamental equation of SMT (Brown et al., 1993) is the following:

$$\hat{\mathbf{e}} = \operatorname*{argmax}_{\mathbf{e}} \{Pr(\mathbf{e} \mid \mathbf{f})\} \quad (1)$$

$$= \operatorname*{argmax}_{\mathbf{e}} \{Pr(\mathbf{f} \mid \mathbf{e})\, Pr(\mathbf{e})\} \quad (2)$$

where $Pr(\mathbf{f} \mid \mathbf{e})$ is approximated by a *translation model* that represents the correlation between the source and the target sentence and where $Pr(\mathbf{e})$ is approximated by a *language model* representing the well-formedness of the candidate translation $\mathbf{e}$.

State-of-the-art statistical machine translation systems follow a loglinear approach (Och and Ney, 2002), where direct modelling of the posterior probability $Pr(\mathbf{e} \mid \mathbf{f})$ of Equation (1) is used. In this case, the decision rule is given by the expression:

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

where each $h_m(\mathbf{e}, \mathbf{f})$ is a feature function representing a statistical model and λ_m its weight.

Current MT systems are based on the use of phrase-based models (Koehn et al., 2003) as translation models. The basic idea of Phrase-based Translation (PBT) is to segment the source sentence into phrases, then to translate each source phrase into a target phrase, and finally to reorder the translated target phrases in order to compose the target sentence. If we summarize all the decisions made during the phrase-based translation process by means of the hidden variable $\tilde{a}_1^K$, we obtain the expression:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{K,\tilde{a}_1^K} Pr(\tilde{f}_1^K, \tilde{a}_1^K \mid \tilde{e}_1^K) \quad (4)$$

where each $\tilde{a}_k \in \{1 \ldots K\}$ denotes the index of the target phrase $\tilde{e}$ that is aligned with the k-th source phrase $\tilde{f}_k$, assuming a segmentation of length K.

According to Equation (4), and following a maximum approximation, the problem stated in Equation (2) can be reframed as:

$$\hat{\mathbf{e}} \approx \arg\max_{\mathbf{e},\mathbf{a}} \{p(\mathbf{e}) \cdot p(\mathbf{f}, \mathbf{a} \mid \mathbf{e})\} \quad (5)$$

In the IMT scenario, we have to find an extension $\mathbf{e}_s$ for a given prefix $\mathbf{e}_p$. To do this we reformulate Equation (5) as follows:

$$\hat{\mathbf{e}}_s \approx \arg\max_{\mathbf{e}_s,\mathbf{a}} \{p(\mathbf{e}_s \mid \mathbf{e}_p) \cdot p(\mathbf{f}, \mathbf{a} \mid \mathbf{e}_p, \mathbf{e}_s)\} \quad (6)$$

where the term $p(\mathbf{e}_p)$ has been dropped since it does depend neither on $\mathbf{e}_s$ nor on $\mathbf{a}$.

Thus, the search is restricted to those sentences $\mathbf{e}$ which contain $\mathbf{e}_p$ as prefix. It is also worth mentioning that the similarities between Equation (6) and Equation (5) (note that $\mathbf{e}_p\mathbf{e}_s \equiv \mathbf{e}$) allow us to use the same models whenever the search procedures are adequately modified (Barrachina et al., 2009).

Following the loglinear approach stated in Equation (3), Equation (6) can be rewriten as:

$$\hat{\mathbf{e}}_s = \operatorname*{argmax}_{\mathbf{e}_s,\mathbf{a}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}, \mathbf{a}, \mathbf{f}) \right\} \quad (7)$$

which is the approach that we follow in this work.

A common problem in IMT arises when the user sets a prefix ($\mathbf{e}_p$) which cannot be found in the phrase-based statistical translation model. Different solutions have been proposed to deal with this problem. The use of word translation graphs, as a compact representation of all possible translations of a source sentence, is proposed in (Barrachina et al., 2009). In (Ortiz-Martínez et al., 2009), a technique based on the generation of partial phrase-based alignments is described. This last proposal has also been adopted in this work.

3 Related work

In this paper we present an application of the online learning paradigm to the IMT framework. In the online learning setting, models are trained sample by sample. Our work is also related to model adaptation, although model adaptation and online learning are not exactly the same thing.

The online learning paradigm has been previously applied to train discriminative models in SMT (Liang et al., 2006; Arun and Koehn, 2007; Watanabe et al., 2007; Chiang et al., 2008). These works differ from the one presented here in that we apply online learning techniques to train generative models instead of discriminative models.

In (Nepveu et al., 2004), dynamic adaptation of an IMT system via cache-based model extensions to language and translation models is proposed. The work by Nepveu et al. (2004) constitutes a domain adaptation technique and not an online learning technique, since the proposed cache components require pre-existent models estimated in batch mode. In addition to this, their IMT system does not use state-of-the-art models.

To our knowledge, the only previous work on online learning for IMT is (Cesa-Bianchi et al., 2008), where a very constrained version of online learning is presented. This constrained version of online learning is not able to extend the translation models due to technical problems with the efficiency of the learning process. In this paper, we present a purely statistical IMT system which is able to incrementally update the parameters of all of the different models that are used in the system, including the translation model, breaking with the above mentioned constraints. What is more, our system is able to learn from scratch, that is, without any preexisting model stored in the system. This is demonstrated empirically in section 5.

4 Online IMT

In this section we propose an online IMT system. First, we describe the basic IMT system involved in the interactive translation process. Then we introduce the required techniques to incrementally update the statistical models used by the system.

4.1 Basic IMT system

The basic IMT system that we propose uses a log-linear model to generate its translations. According to Equation (7), we introduce a set of seven feature functions (from h_1 to h_7):

- **n-gram language model (h_1)**
  $h_1(\mathbf{e}) = \log(\prod_{i=1}^{|\mathbf{e}|+1} p(e_i|e_{i-n+1}^{i-1}))$, [1] where $p(e_i|e_{i-n+1}^{i-1})$ is defined as follows:

$$p(e_i|e_{i-n+1}^{i-1}) = \frac{\max\{c_X(e_{i-n+1}^{i}) - D_n, 0\}}{c_X(e_{i-n+1}^{i-1})} + \frac{D_n}{c_X(e_{i-n+1}^{i-1})} N_{1+}(e_{i-n+1}^{i-1}\bullet) \cdot p(e_i|e_{i-n+2}^{i-1}) \quad (8)$$

  where $D_n = \frac{c_{n,1}}{c_{n,1}+2c_{n,2}}$ is a fixed discount ($c_{n,1}$ and $c_{n,2}$ are the number of n-grams with one and two counts respectively), $N_{1+}(e_{i-n+1}^{i-1}\bullet)$ is the number of unique words that follows the history e_{i-n+1}^{i-1} and $c_X(e_{i-n+1}^{i})$ is the count of the n-gram e_{i-n+1}^{i}, where $c_X(\cdot)$ can represent true counts $c_T(\cdot)$ or modified counts $c_M(\cdot)$. True counts are used for the higher order n-grams and modified counts for the lower order n-grams. Given a certain n-gram, its modified count consists in the number of different words that precede this n-gram in the training corpus.

  Equation (8) corresponds to the probability given by an n-gram language model with an interpolated version of the Kneser-Ney smoothing (Chen and Goodman, 1996).

[1] $|\mathbf{e}|$ is the length of $\mathbf{e}$, e_0 denotes the *begin-of-sentence* symbol, $e_{|e|+1}$ denotes the *end-of-sentence* symbol, $e_i^j \equiv e_i ... e_j$

- **target sentence-length model (h_2)**
  $h_2(\mathbf{e}, \mathbf{f}) = \log(p(|\mathbf{f}| \,|\, |\mathbf{e}|)) = \log(\phi_{|\mathbf{e}|}(|\mathbf{f}|+0.5) - \phi_{|\mathbf{e}|}(|\mathbf{f}|-0.5))$, where $\phi_{|\mathbf{e}|}(\cdot)$ denotes the cumulative distribution function (cdf) for the normal distribution (the cdf is used here to integrate the normal density function over an interval of length 1). We use a specific normal distribution with mean $\mu_{|\mathbf{e}|}$ and standard deviation $\sigma_{|\mathbf{e}|}$ for each possible target sentence length $|\mathbf{e}|$.

- **inverse and direct phrase-based models (h_3, h_4)**
  $h_3(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log(\prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}))$, where $p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$ is defined as follows:

$$p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) = \beta \cdot p_{phr}(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) + (1-\beta).p_{hmm}(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) \quad (9)$$

  In Equation (9), $p_{phr}(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$ denotes the probability given by a statistical phrase-based dictionary used in regular phrase-based models (see (Koehn et al., 2003) for more details). $p_{hmm}(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$ is the probability given by an HMM-based (intra-phrase) alignment model (see (Vogel et al., 1996)):

$$p_{hmm}(\tilde{f}|\tilde{e}) = \epsilon \sum_{a_1^{|\tilde{f}|}} \prod_{j=1}^{|\tilde{f}|} p(\tilde{f}_j|\tilde{e}_{a_j}) \cdot p(a_j|a_{j-1}, |\tilde{e}|) \quad (10)$$

  The HMM-based alignment model probability is used here for smoothing purposes as described in (Ortiz-Martínez et al., 2009).

  Analogously h_4 is defined as:
  $h_4(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log(\prod_{k=1}^{K} p(\tilde{e}_{\tilde{a}_k}|\tilde{f}_k))$

- **target phrase-length model (h_5)**
  $h_5(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log(\prod_{k=1}^{K} p(|\tilde{e}_k|))$, where $p(|\tilde{e}_k|) = \delta(1-\delta)^{|\tilde{e}_k|}$. h_5 implements a target phrase-length model by means of a geometric distribution with probability of success on each trial δ. The use of a geometric distribution penalizes the length of target phrases.

- **source phrase-length model (h_6)**
  $h_6(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log(\prod_{k=1}^{K} p(|\tilde{f}_k| \,|\, |\tilde{e}_{\tilde{a}_k}|))$,
  where $p(|\tilde{f}_k| \,|\, |\tilde{e}_{\tilde{a}_k}|) = \delta(1-\delta)^{abs(|\tilde{f}_k| - |\tilde{e}_{\tilde{a}_k}|)}$ and $abs(\cdot)$ is the absolute value function. A geometric distribution is used to model this feature (it penalizes the difference between the source and target phrase lengths).

- **distortion model (h_7)**
  $h_7(\mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{a}_k|\tilde{a}_{k-1}))$, where $p(\tilde{a}_k|\tilde{a}_{k-1}) = \delta(1-\delta)^{abs(b_{\tilde{a}_k} - l_{\tilde{a}_{k-1}})}$, $b_{\tilde{a}_k}$ denotes the beginning position of the source phrase covered by $\tilde{a}_k$ and $l_{\tilde{a}_{k-1}}$ denotes the last position of the source phrase covered by $\tilde{a}_{k-1}$. A geometric distribution is used to model this feature (it penalizes the reorderings).

The log-linear model, which includes the above described feature functions, is used to generate the suffix $\mathbf{e}_s$ given the user-validated prefix $\mathbf{e}_p$. Specifically, the IMT system generates a partial phrase-based alignment between the user prefix $\mathbf{e}_p$ and a portion of the source sentence $\mathbf{f}$, and returns the suffix $\mathbf{e}_s$ as the translation of the remaining portion of $\mathbf{f}$ (see (Ortiz-Martínez et al., 2009)).

4.2 Extending the IMT system from user feedback

After translating a source sentence $\mathbf{f}$, a new sentence pair $(\mathbf{f}, \mathbf{e})$ is available to feed the IMT system (see Figure 1). In this section we describe how the log-linear model described in section 4.1 is updated given the new sentence pair. To do this, a set of *sufficient statistics* that can be incrementally updated is maintained for each feature function $h_i(\cdot)$. A sufficient statistic for a statistical model is a statistic that captures all the information that is relevant to estimate this model.

Regarding feature function h_1 and according to equation (8), we need to maintain the following data: $c_{k,1}$ and $c_{k,2}$ given any order k, $N_{1+}(\cdot)$, and $c_X(\cdot)$ (see section 4.1 for the meaning of each symbol). Given a new sentence $\mathbf{e}$, and for each k-gram e_{i-k+1}^{i} of $\mathbf{e}$ where $1 \leq k \leq n$ and $1 \leq i \leq |\mathbf{e}|+1$, we modify the set of sufficient statistics as it is shown in Algorithm 1. The algorithm checks the changes in the counts of the k-grams to update the set of sufficient statistics. Sufficient statistics for D_k are updated following the auxiliar procedure shown in Algorithm 2.

Feature function h_2 requires the incremental calculation of the mean $\mu_{|\mathbf{e}|}$ and the standard deviation $\sigma_{|\mathbf{e}|}$ of the normal distribution associated to a target sentence length $|\mathbf{e}|$. For this purpose the procedure described in (Knuth, 1981) can be used. In this procedure, two quantities are maintained for each normal distribution: $\mu_{|\mathbf{e}|}$ and $S_{|\mathbf{e}|}$. Given a new sentence

input : n (higher order), e_{i-k+1}^{i} (k-gram), $\mathcal{S} = \{\forall j(c_{j,1}, c_{j,2}), N_{1+}(\cdot), c_X(\cdot)\}$ (current set of sufficient statistics)
output : $\mathcal{S}$ (updated set of sufficient statistics)
begin
  if $c_T(e_{i-k+1}^{i}) = 0$ **then**
    if $k - 1 \geq 1$ **then**
      updD ($\mathcal{S}$,k-1,$c_M(e_{i-k+2}^{i-1})$,$c_M(e_{i-k+2}^{i-1})+1$)
      if $c_M(e_{i-k+2}^{i-1}) = 0$ **then**
        $N_{1+}(e_{i-k+2}^{i-1}) = N_{1+}(e_{i-k+2}^{i-1}) + 1$
      $c_M(e_{i-k+2}^{i-1}) = c_M(e_{i-k+2}^{i-1}) + 1$
      $c_M(e_{i-k+2}^{i}) = c_M(e_{i-k+2}^{i}) + 1$
    if $k = n$ **then**
      $N_{1+}(e_{i-k+1}^{i-1}) = N_{1+}(e_{i-k+1}^{i-1}) + 1$
  if $k = n$ **then**
    updD ($\mathcal{S}$,k,$c_T(e_{i-k+1}^{i})$,$c_T(e_{i-k+1}^{i}) + 1$)
  $c_T(e_{i-k+1}^{i-1})$=$c_T(e_{i-k+1}^{i-1}) + 1$
  $c_T(e_{i-k+1}^{i})$=$c_T(e_{i-k+1}^{i}) + 1$
end

Algorithm 1: Pseudocode for updating the sufficient statistics of a given k-gram

input : $\mathcal{S}$ (current set of sufficient statistics),k (order), c (current count), c' (new count)
output : $(c_{k,1}, c_{k,2})$ (updated sufficient statistics)
begin
  if $c = 0$ **then**
    if $c' = 1$ **then** $c_{k,1} = c_{k,1} + 1$
    if $c' = 2$ **then** $c_{k,2} = c_{k,2} + 1$
  if $c = 1$ **then**
    $c_{k,1} = c_{k,1} - 1$
    if $c' = 2$ **then** $c_{k,2} = c_{k,2} + 1$
  if $c = 2$ **then** $c_{k,2} = c_{k,2} - 1$
end

Algorithm 2: Pseudocode for the updD procedure

pair $(\mathbf{f}, \mathbf{e})$, the two quantities are updated using a recurrence relation:

$$\mu_{|\mathbf{e}|} = \mu'_{|\mathbf{e}|} + (|\mathbf{f}| - \mu'_{|\mathbf{e}|})/c(|\mathbf{e}|) \quad (11)$$

$$S_{|\mathbf{e}|} = S'_{|\mathbf{e}|} + (|\mathbf{f}| - \mu'_{|\mathbf{e}|})(|\mathbf{f}| - \mu_{|\mathbf{e}|}) \quad (12)$$

where $c(|\mathbf{e}|)$ is the count of the number of sentences of length $|\mathbf{e}|$ that have been seen so far, and $\mu'_{|\mathbf{e}|}$ and $S'_{|\mathbf{e}|}$ are the quantities previously stored ($\mu_{|\mathbf{e}|}$ is initialized to the source sentence length of the first sample and $S_{|\mathbf{e}|}$ is initialized to zero). Finally, the standard deviation can be obtained from S as follows: $\sigma_{|\mathbf{e}|} = \sqrt{S_{|\mathbf{e}|}/(c(|\mathbf{e}|) - 1)}$.

Feature functions h_3 and h_4 implement inverse and direct smoothed phrase-based models respectively. Since phrase-based models are symmetric models, only an inverse phrase-based model is maintained (direct probabilities can be efficiently obtained using appropriate data structures, see (Ortiz-Martínez et al., 2008)). The inverse phrase model probabilities are estimated from the phrase counts:

$$p(\tilde{f}|\tilde{e}) = \frac{c(\tilde{f}, \tilde{e})}{\sum_{\tilde{f}'} c(\tilde{f}', \tilde{e})} \quad (13)$$

According to Equation (13), the set of sufficient statistics to be stored for the inverse phrase model consists of a set of phrase counts ($c(\tilde{f}, \tilde{e})$ and $\sum_{\tilde{f}'} c(\tilde{f}', \tilde{e})$ must be stored separately). Given a new sentence pair $(\mathbf{f}, \mathbf{e})$, the standard phrase-based model estimation method uses a word alignment matrix between $\mathbf{f}$ and $\mathbf{e}$ to extract the set of phrase pairs that are *consistent* with the word alignment matrix (see (Koehn et al., 2003) for more details). Once the consistent phrase pairs have been extracted, the phrase counts are updated. The word alignment matrices required for the extraction of phrase pairs are generated by means of the HMM-based models used in the feature functions h_3 and h_4.

Inverse and direct HMM-based models are used here for two purposes: to smooth the phrase-based models via linear interpolation and to generate word alignment matrices. The weights of the interpolation can be estimated from a development corpus. Equation (10) shows the expression of the probability given by an inverse HMM-based model. The probability includes lexical probabilities $p(f_j|e_i)$ and alignment probabilities $p(a_j|a_{j-1}, l)$. Since the alignment in the HMM-based model is determined by a hidden variable, the EM algorithm is required to estimate the parameters of the model (see (Och and Ney, 2003)). However, the standard EM algorithm is not appropriate to incrementally extend our HMM-based models because it is designed to work in batch training scenarios. To solve this problem, we apply the incremental view of the EM algorithm described in (Neal and Hinton, 1998). According to (Och and Ney, 2003), the lexical probability for a

pair of words is given by the expression:

$$p(f|e) = \frac{c(f|e)}{\sum_{f'} c(f'|e)} \quad (14)$$

where $c(f|e)$ is the *expected* number of times that the word e is aligned to the word f. The alignment probability is defined in a similar way:

$$p(a_j|a_{j-1}, l) = \frac{c(a_j|a_{j-1}, l)}{\sum_{a'_j} c(a'_j|a_{j-1}, l)} \quad (15)$$

where $c(a_j|a_{j-1}, l)$ denotes the expected number of times that the alignment a_j has been seen after the previous alignment a_{j-1} given a source sentence composed of l words.

Given the equations (14) and (15), the set of sufficient statistics for the inverse HMM-based model consists of a set of expected counts (numerator and denominator values are stored separately). Given a new sentence pair $(\mathbf{f}, \mathbf{e})$, we execute a new iteration of the incremental EM algorithm on the new sample and collect the contributions to the expected counts.

The parameters of the direct HMM-based model are estimated analogously to those of the inverse HMM-based model. Once the direct and the inverse HMM-based model parameters have been modified due to the presentation of a new sentence pair to the IMT system, both models are used to obtain word alignments for the new sentence pair. The resulting direct and inverse word alignment matrices are combined by means of the *symmetrization* alignment operation (Och and Ney, 2003) before extracting the set of consistent phrase pairs.

HMM-based alignment models are used here because, according to (Och and Ney, 2003) and (Toutanova et al., 2002), they outperform IBM 1 to IBM 4 alignment models while still allowing the exact calculation of the likelihood for a given sentence pair.

The δ parameters of the geometric distributions associated to the feature functions h_5, h_6 and h_7 are left fixed. Because of this, there are no sufficient statistics to store for these feature functions.

Finally, the weights of the log-linear combination are not modified due to the presentation of a new sentence pair to the system. These weights can be adjusted off-line by means of a development corpus and well-known optimization techniques.

5 Experiments

This section describes the experiments that we carried out to test our online IMT system.

5.1 Experimental setup

The experiments were performed using the XEROX XRCE corpus (SchlumbergerSema S.A. et al., 2001), which consists of translations of Xerox printer manuals involving three different pairs of languages: French-English, Spanish-English, and German-English. The main features of these corpora are shown in Table 1. Partitions into training, development and test were performed. This corpus is used here because it has been extensively used in the literature on IMT to report results.

IMT experiments were carried out from English to the other three languages.

5.2 Assessment criteria

The evaluation of the techniques presented in this paper were carried out using the *Key-stroke and mouse-action ratio* (KSMR) measure (Barrachina et al., 2009). This is calculated as the number of keystrokes plus the number of *mouse movements* plus one more count per sentence (aimed at simulating the user action needed to accept the final translation), the sum of which is divided by the total number of reference characters. In addition to this, we also used the well-known BLEU score (Papineni et al., 2001) to measure the translation quality of the first translation hypothesis produced by the IMT system for each source sentence (which is automatically generated without user intervention).

5.3 Online IMT results

To test the techniques proposed in this work, we carried out experiments in two different scenarios. In the first one, the first 10 000 sentences extracted from the training corpora were interactively translated by means of an IMT system without any pre-existent model stored in memory. Each time a new sentence pair was validated, it was used to incrementally train the system. Figures 3a, 3b and 3c show the evolution of the KSMR with respect to the number of sentence pairs processed by the IMT system; the results correspond to the translation from English to Spanish, French and German, respectively. In addi-

		En	Sp	En	Fr	En	Ge
Train	Sent. pairs	55761		52844		49376	
	Running words	571960	657172	542762	573170	506877	440682
	Vocabulary	25627	29565	24958	27399	24899	37338
Dev.	Sent. pairs	1012		994		964	
	Running words	12111	13808	9480	9801	9162	8283
	Perplexity (3-grams)	46.2	34.0	96.2	74.1	68.4	124.3
Test	Sent. pairs	1125		984		996	
	Running words	7634	9358	9572	9805	10792	9823
	Perplexity (3-grams)	107.0	59.6	192.6	135.4	92.8	169.2

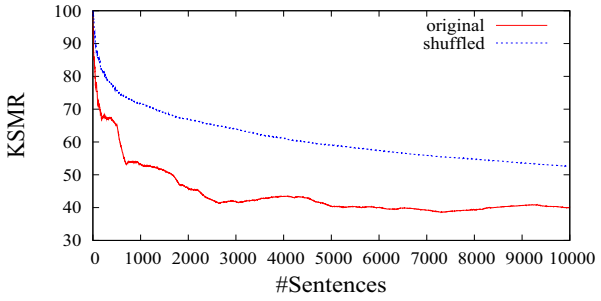
Table 1: XEROX corpus statistics for three different language pairs (from English (En) to Spanish (Sp), French (Fr) and German (Ge))

tion, for each language pair we interactively translated the original portion of the training corpus and the same portion of the original corpus after being randomly shuffled.

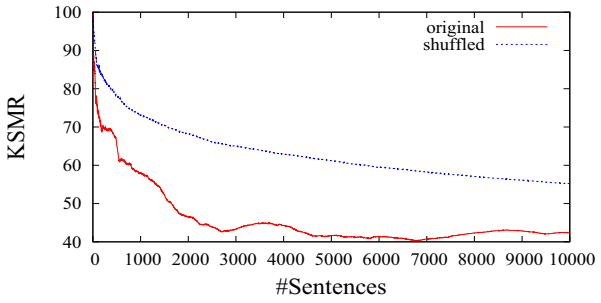
As these figures show, the results clearly demonstrate that the IMT system is able to learn from scratch. The results were similar for the three languages. It is also worthy of note that the obtained results were better in all cases for the original corpora than for the shuffled ones. This is because, in the original corpora, similar sentences appear more or less contiguously (due to the organization of the contents of the printer manuals). This circumstance increases the accuracy of the online learning, since with the original corpora the number of *lateral effects* ocurred between the translation of similar sentences is decreased. The online learning of a new sentence pair produces a lateral effect when the changes in the probability given by the models not only affect the newly trained sentence pair but also other sentence pairs. A lateral effect can cause that the system generates a wrong translation for a given source sentence due to undesired changes in the statistical models.

The accuracy were worse for shuffled corpora, since shuffling increases the number of lateral effects that may occur between the translation of similar sentences (because they no longer appear contiguously). A good way to compare the quality of different online IMT systems is to determine their robustness in relation to sentence ordering. However, it can generally be expected that the sentences to be translated in an interactive translation session will be in a non-random order.

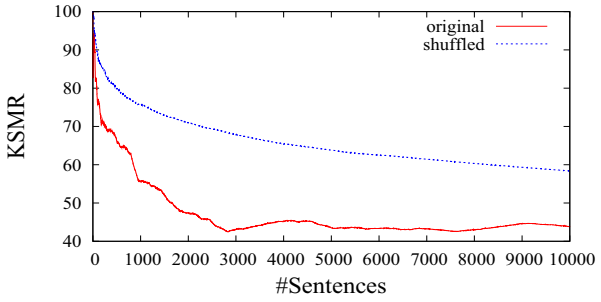
Alternatively, we carried out experiments in a different learning scenario. Specifically, the XEROX



(a) English-Spanish



(b) English-French



(c) English-German

Figure 3: KSMR evolution translating a portion of the training corpora

test corpora were interactively translated from the English language to the other three languages, comparing the performance of a batch IMT system with

that of an online IMT system. The batch IMT system is a conventional IMT system which is not able to take advantage of user feedback after each translation while the online IMT system uses the new sentence pairs provided by the user to revise the statistical models. Both systems were initialized with a log-linear model trained in batch mode by means of the XEROX training corpora. The weights of the log-linear combination were adjusted for the development corpora by means of the downhill-simplex algorithm. Table 2 shows the obtained results. The table 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. The table also shows the average online learning time (LT) for each new sample presented to the system[2]. All the improvements obtained with the online IMT system were statistically significant. Also, the average learning times clearly allow the system to be used in a real-time scenario.

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

Table 2: BLEU and KSMR results for the XEROX test corpora using the batch and the online IMT systems. The average online learning time (LT) in seconds is shown for the online system

Finally, in Table 3 a comparison of the KSMR results obtained by the online IMT system with state-of-the-art IMT systems is reported (95% confidence intervals are shown). We compared our system with those presented in (Barrachina et al., 2009): the alignment templates (AT), the stochastic finite-state transducer (SFST), and the phrase-based (PB) approaches to IMT. The results were obtained using the same Xerox training and test sets (see Table 1) for the four different IMT systems. Our system outperformed the results obtained by these systems.

[2]All the experiments were executed on a PC with a 2.40 Ghz Intel Xeon processor with 1GB of memory.

	AT	PB	SFST	Online
En-Sp	23.2±1.3	16.7±1.2	21.8±1.4	15.8± 1.0
En-Fr	40.4±1.4	35.8±1.3	43.8±1.6	27.9± 1.3
En-Ge	44.7±1.2	40.1±1.2	45.7±1.4	35.0± 1.3

Table 3: KSMR results comparison of our system and three different state-of-the-art batch systems

6 Conclusions

We have presented an online IMT system. The proposed system is able to incrementally extend the statistical models involved in the translation process, breaking technical limitations encountered in other works. Empirical results show that our techniques allow the IMT system to learn from scratch or from previously estimated models.

One key aspect of the proposed system is the use of HMM-based alignment models trained by means of the incremental EM algorithm.

The incremental adjustment of the weights of the log-linear models and other parameters have not been tackled here. For the future we plan to incorporate this functionality into our IMT system.

The incremental techniques proposed here can also be exploited to extend SMT systems (in fact, our proposed IMT system is based on an incrementally updateable SMT system). For the near future we plan to study possible aplications of our techniques in a fully automatic translation scenario.

Finally, it is worthy of note that the main ideas presented here can be used in other interactive applications such as Computer Assisted Speech Transcription, Interactive Image Retrieval, etc (see (Vidal et al., 2007) for more information). In conclusion, we think that the online learning techniques proposed here can be the starting point for a new generation of interactive pattern recognition systems that are able to take advantage of user feedback.

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), the Univ. Politécnica de Valencia (grant 20091027) and the Spanish JCCM (grant PBI08-0210-7127).

References

A. Arun and P. Koehn. 2007. Online learning methods for discriminative training of phrase based statistical machine translation. In *Proc. of the MT Summit XI*, pages 15–20, Copenhagen, Denmark, September.

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.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

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, Mar.

S.F. Chen and J. Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proc. of the ACL*, pages 310–318, San Francisco.

D. Chiang, Y. Marton, and P. Resnik. 2008. Online large-margin training of syntactic and structural translation features. In *Proc. of EMNLP*.

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

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

D.E. Knuth. 1981. *Seminumerical Algorithms*, volume 2 of *The Art of Computer Programming*. Addison-Wesley, Massachusetts, 2nd edition.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the HLT/NAACL*, pages 48–54, Edmonton, Canada, May.

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

P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proc. of the 44th ACL*, pages 761–768, Morristown, NJ, USA.

R.M. Neal and G.E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Proceedings 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. of EMNLP*, pages 190–197, Barcelona, Spain, July.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th ACL*, pages 295–302, Philadelphia, PA, July.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

D. Ortiz-Martínez, I. García-Varea, and Casacuberta F. 2008. The scaling problem in the pattern recognition approach to machine translation. *Pattern Recognition Letters*, 29:1145–1153.

Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. 2009. Interactive machine translation based on partial statistical phrase-based alignments. In *Proc. of RANLP*, Borovets, Bulgaria, sep.

Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September.

SchlumbergerSema S.A., ITI Valencia, RWTH Aachen, RALI Montreal, Celer Soluciones, Société Gamma, and XRCE. 2001. TT2. TransType2 - computer assisted translation. Project Tech. Rep.

Kristina Toutanova, H. Tolga Ilhan, and Christopher Manning. 2002. Extensions to hmm-based statistical word alignment models. In *Proc. of EMNLP*.

E. Vidal, L. Rodríguez, F. Casacuberta, and I. García-Varea. 2007. Interactive pattern recognition. In *Proc. of the 4th MLMI*, pages 60–71. Brno, Czech Republic, 28-30 June.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proc. of COLING*, pages 836–841, Copenhagen, Denmark, August.

T. Watanabe, J. Suzuki, H. Tsukada, and H. Isozaki. 2007. Online large-margin training for statistical machine translation. In *Proc. of EMNLP and CoNLL*, pages 764–733, Prage, Czeck Republic.