CHAPTER 8

Incremental and Adaptive Learning for Interactive Machine Translation

WITH CONTRIBUTION OF:
Daniel Ortiz-Martínez and Ismael García-Varea

High-quality translation between any pair of languages can be achieved by a human post-editing of the outputs of a MT system or, as mentioned in Chp. 6, by following the interactive machine translation (IMT) approach. In the interactive pattern recognition framework, IMT can predict the translation of the next words in the output, and suggest them to the human translator who, iteratively, can accept or correct the suggested translations. The consolidated translations obtained through the successive steps of the interaction process can be considered as "perfect translations" due to the fact that they have been validated by a human expert. Therefore, this consolidated translations can easily be converted into new, fresh, training data, useful for dynamically adapting the system to the changing environment. Taking that into account, on the one hand, the IMT paradigm offers an appropriate framework for incremental and adaptive learning in SMT. On the other hand, incremental and adaptive learning offers the possibility to substantially save human effort by simply avoiding the user to perform the same corrections again and again.

Chapter Outline

8.1 Introduction

IMT can be seen as an evolution of the SMT framework (see Sec. 6.2), where the same statistical models and search algorithms can be used with little modification. Because of these similarities, a whole set of proposals for domain adaptation [3, 11, 8, 2, 19] that were initially introduced in the SMT framework can also be applied to IMT. However, human interaction offers another unique opportunity to improve the performance of the IMT systems by tuning the statistical models involved in the translation process. In particular, at each interaction, the text segments validated by the user together with the corresponding aligned source segments can generally be converted into new, fresh training data, useful for adapting the system to a changing environment. An early example of this can be found in the TransType framework, where a cache-based technique both for target language models and translation models is used [14]. Another example can be found in [5], where the output of a post-editing system is used for adaptation. More recently, in [17], a purely statistical IMT system with online learning capabilities is introduced, where the system can incrementally update the parameters of all of the different models that are involved in the translation process, and therefore automatically adapted to new users, new translation styles, or even to new target languages.

In the following sections, we describe different proposals that try to take advantage of user feedback to extend the statistical models of the IMT system. Specifically, in Sec. 8.2, an IMT system with online learning capabilities is described. In addition to this, other topics related with adaptation in IMT are presented in Sec. 8.3.

8.2 On-line Learning

8.2.1 Concept of On-line Learning

In the IMT framework, the target sentences validated by the user along with the corresponding source sentences constitute new training samples that can be used to extend the statistical models of the IMT system. Unfortunately, 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. To solve this problem, the *on-line learning* paradigm to the IMT framework can be applied. In the on-line learning paradigm, the training and prediction stages are no longer separated.

The on-line learning paradigm has been previously applied to train discriminative models in SMT [12, 1, 21, 7]. The application of online learning techniques in IMT has not been extensively studied, one previous work is [6], where a very constrained version of on-line learning is presented. This constrained version of on-line learning is not able to extend the translation models due to technical problems with the efficiency of the learning process.

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. Our proposal uses a conventional IMT system which is based on a log-linear model. Such log-linear model is composed of a set of feature functions governing different aspects of the translation process. We will briefly describe the required techniques to incrementally update the statistical models used by the system. To do this, a set of sufficient statistics is incrementally updated for each feature function.

8.2.2 Basic IMT System

As was mentioned above, the basic IMT system that we propose uses a log-linear model to generate its translations. According to Eq. (6.3), we introduce a set of seven feature functions (from f_1 to f_7): a language model (f_1), an inverse sentence-length model (f_2), inverse and direct translation models (f_3 and f_4 respectively), a target phrase-length model (f_5), a source phrase-length model (f_6), and a distortion model (f_7).

The f_1 feature function is implemented by means of smoothed n-gram language models. In particular, we adopt an interpolated n-gram model with Kneser-Ney smoothing.

The f_2 feature function constitutes an inverse sentence length model implemented by means of a set of gaussian distributions whose parameters are estimated for each source sentence length.

The inverse translation model (feature function f_3) is implemented with an inverse phrase-based model. This phrase-based model is smoothed with an HMM-based alignment model [20] by means of linear interpolation. It is illustrative to consider the exact formulation of f_3 so as to later explain how it can be incrementally updated in the following section. Specifically, f_3 is defined as follows:

$$f_3(x,h,a) = \log(\prod_{k=1}^{K} P(\tilde{x}_k|\tilde{h}_{a_k})) \tag{8.1}$$

where x and h are the source and the target sentences, respectively, and the hidden alignment variable a determines a phrase alignment of length K between x and h. The probability of translation between phrase pairs, $P(\tilde{x}_k|\tilde{h}_{a_k})$, is defined as follows:

$$P(\tilde{x}_k|\tilde{h}_{a_k}) = \beta \cdot P_{phr}(\tilde{x}_k|\tilde{h}_{a_k}) + (1-\beta) \cdot P_{hmm}(\tilde{x}_k|\tilde{h}_{a_k}) \tag{8.2}$$

In Eq. (8.2), β is the linear interpolation parameter, $P_{phr}(\tilde{x}_k|\tilde{h}_{a_k})$ denotes the probability given by a statistical phrase-based dictionary used in regular phrase-based models (see [10] for more details) and $P_{hmm}(\tilde{x}_k|\tilde{h}_{a_k})$ is the probability given by an HMM-based alignment model defined at phrase level:

$$P_{hmm}(\tilde{x}|\tilde{h}) = \epsilon \sum_{b_1^{|\tilde{x}|}} \prod_{j=1}^{|\tilde{x}|} P(\tilde{x}_j|\tilde{h}_{b_j}) \cdot P(b_j|b_{j-1},|\tilde{h}|) \tag{8.3}$$

where ϵ is a small, fixed number which accounts for the probability of the length of the source sentence (see [4] for more details), $b_1^{|\tilde{x}|}$ represents a word-level hidden alignment variable between $\tilde{x}$ and $\tilde{h}$, $P(\tilde{x}_j|\tilde{h}_{b_j})$ is the lexical probability and $P(b_j|b_{j-1},|\tilde{h}|)$ is the alignment probability.

Analogously, feature function f_4 is implemented by means of a direct, smoothed phrase based model.

Regarding the target phrase-length model (feature function f_5), it is implemented by means of a geometric distribution. The use of a geometric distribution penalises the length of the target phrases. A geometric distribution can also be used to model f_6, such distribution penalises the difference between the source and target phrase lengths. Finally, we use again a geometric distribution to implement a distortion model for the phrase translations (feature function f_7). Such distribution penalises the reorderings.

The log-linear model, which includes the above described feature functions, is used to generate the suffix s given the user-validated prefix p. Specifically, the IMT system generates a partial phrase-based alignment between the user prefix p and a portion of the source sentence x, and returns the suffix s as the translation of the remaining portion of x (see [16]).

8.2.3 Online IMT System

After translating a source sentence x, a new sentence pair (x, h) is available to feed the IMT system. In this section we describe how the log-linear model described in the previous section 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 $f_i(\cdot)$. A sufficient statistic for a statistical model is a statistic that captures all the information that is relevant to estimate this model. If the estimation of the statistical model does not require the use of the EM algorithm (e.g. n-gram language models), then it is generally easy to incrementally extend the model given a new training sample. By contrast, if the EM algorithm is required (e.g. word alignment models), the estimation procedure has to be modified, since the conventional EM algorithm is designed for its use in batch learning scenarios. To solve this problem, we apply the incremental version of the EM algorithm [13].

The parameters of the n-gram language model with Kneser-Ney smoothing that implements the f_1 feature function can be incrementally adjusted with an appropriate algorithm which is shown in [17]. Since the estimation does not involve the EM algorithm, the algorithm is relatively simple.

Regarding the f_2 feature function, its incremental estimation requires updating the parameters of a set of gaussian distributions. This problem has been extensively studied in the literature, specifically, we apply the incremental update rules given in [9].

Feature functions f_3 and f_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 [15]). It is interesting to see how this inverse phrase-based model can be incrementally updated, since as it will be explained below, such incremental update involves the application of the incremental version of the EM algorithm.

Given a new sentence pair (x, h), the standard phrase-based model estimation method uses a word alignment matrix between x and h to extract the set of phrase pairs that are *consistent* with the word alignment matrix (see [10] for more details). Once the consistent phrase pairs have been extracted, the phrase counts are updated following the equation:

$$P(\tilde{x}|\tilde{h}) = \frac{c(\tilde{x}, \tilde{h})}{\sum_{\tilde{x}'} c(\tilde{x}', \tilde{h})} \tag{8.4}$$

where $c(\tilde{x}, \tilde{h})$ represents the count of the phrase pair $(\tilde{x}, \tilde{h})$ in the set of consistent phrase pairs extracted from the training corpus.

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 f_3 and f_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. 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. However, the conventional EM algorithm is not appropriate for its use in our online learning scenario. To solve this problem, the incremental version of the EM algorithm that was mentioned above is applied. As was mentioned above, HMM-based alignment models are composed of lexical and alignment probabilities. Lexical probabilities are obtained following the equation:

$$P(u|v) = \frac{c(u|v)}{\sum_{u'} c(u'|v)} \tag{8.5}$$

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

$$P(b_j|b_{j-1}, l) = \frac{c(b_j|b_{j-1}, l)}{\sum_{b'_j} c(b'_j|b_{j-1}, l)} \tag{8.6}$$

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

Finally, the parameters of the geometric distributions associated to the feature functions f_5, f_6 and f_7 are left fixed. Because of this, there are no sufficient statistics to store for these feature functions.

8.3 Related Topics

8.3.1 Active Learning on IMT via Confidence Measures

As outlined in Sec. 1.3, the interactive pattern recognition scenario allows to use the valuable user feedback signals produced in the interaction in an adaptive training process which progressively tunes the models to the specific task and/or to the way the user makes use of the systems in this task. Final outputs contain lot of information provided by the user to help the system refine or improve its hypothesis. For the IMT scenario this information consists in correct translations of the input sentences that can be used to improve the translation models as shown in Sec. 8.2.

Regarding the alternative IMT scenario presented in Sec. 6.2.1, confidence measures are used to reduce the effort required for the user, so final translations are composed of segments validated by the user and segments automatically translated by the system. This information can be used

to perform a mixed active and semi-supervised learning to exploit both what the system think it knows about the unlabelled data and the new information provided by the user for the data the system is not confident enough.

8.3.2 Bayesian Adaptation

In IMT, it is quite common to have a model trained on a given domain where lots of data is available, but the purpose is to translate another domain, where only small data is available. Such problem has not been confronted as of yet in the context of IMT, although some work exists in the context of traditional SMT. This is the case of Bayesian adaptation [18], which is an instantiation of the Bayesian learning paradigm to the case of adaptation in SMT.

The weights of the log-linear combination in Eq. (6.3), can be adapted to new situations by the application of the MERT algorithm of Chp. 6, estimating a new set of weights on the adaptation data and substituting the old ones. However, if the adaptation set made available to the system is not big enough, MERT will most likely become unstable and fail in obtaining an appropriate weight vector. This situation is common in IMT.

The main idea behind Bayesian learning is that parameters are viewed as random variables which have some kind of a priori distribution. Given a training set T to build the initial models and an adaptation set A, the problem can be stated as:

$$\begin{aligned} \hat{h} &= \arg\max_{h} P(h|x;T,A) \\ &= \arg\max_{h} \int P(\lambda|T,A)P(h|x;\lambda)\, d\lambda \end{aligned} \tag{8.7}$$

In last equation, the integral over the complete parametric space forces the model to take into account all possible values of the model parameters (in this case, only the λ in Eq. (6.3)), although the prior over the parameters implies that our model will prefer parameter values which are closer to our prior knowledge. Two assumptions are adopted: first, the output sentence h only depends on the model parameters (and not on the complete training and adaptation data), and second, that the model parameters do not depend on the actual input sentence x. Such simplifications lead to a decomposition of the integral into two parts: the first one, $P(\lambda|T,A)$ will assess how good the current model parameters are, and the second one, $P(h|x;\lambda)$, will account for the quality of the translation h given the current model parameters.

8.4 Results

We carried out experiments to test the IMT system with online learning techniques proposed in Sec. 8.2. The proposals that were presented in Sec. 8.3 have not yet been adequately tested in the IMT framework.

All the experiments were executed using the XEROX task described in Sec. 6.4. The XEROX task was used to perform IMT experiments in two different scenarios. In the first one, the first

10 000 sentences extracted from the English–Spanish 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. The results obtained in this first experimentation scenario are shown in Fig. 8.1. In the figure, the evolution of the WSR measure as a function of the number of interactively translated sentences is represented. As can be seen, the results clearly demonstrate that the IMT system is able to learn from scratch. The results obtained for the translation from English to Spanish were similar for the rest of language pairs of the XEROX corpora, as it is shown in [17].

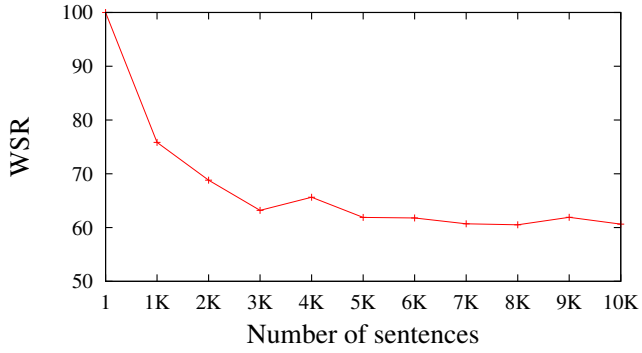

Figure 8.1: WSR evolution when translating a portion of the XEROX English–Spanish training corpora by means of an online IMT system

Alternatively, we carried out experiments in a different learning scenario. Specifically, we compared the performance of a batch IMT system with that of an online IMT system when translating the XEROX test corpora. In this experiment, the statistical models used by both the batch and the online IMT systems were initialised by means of the XEROX training corpora. The obtained WSR results for the English–Spanish language pair are shown in Tab. 8.1.

Table 8.1: On-line learning results for the English–Spanish XEROX task

Task	System	WSR
XEROX (Eng–Spa)	batch	32.0
	on-line	26.6

As can be seen in the table, the proposed on-line learning techniques allow the IMT system to learn from previously estimated models (further results in this experimentation scenario can be found in [17]).

8.5 Conclusions

Human interaction offers a unique opportunity to improve the performance of the IMT systems by tuning the translation models. In particular, at each interaction of the IMT process, the text validated by the user along with the source sentence constitute new training data that can be used to extend the statistical models used by the system. The IMT techniques proposed in this chapter allow us to take advantage of such user feedback by means of different techniques, including online learning, active learning and bayesian adaptation.

Bibliography

[1] Arun, A. and Koehn, P. (2007). Online learning methods for discriminative training of phrase based statistical machine translation. In *Proceedings of the Machine Translation Summit XII (MT Summit 07)*, pages 15–20, Copenhagen, Denmark.

[2] Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the EACL 09 Fourth Workshop on Statistical Machine Translation (WSMT 09)*, pages 182–189, Athens, Greece.

[3] Bing Zhao, M. E. and Vogel, S. (2004). Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the International Conference on Computational Linguistics (COLING 04)*, pages 411–417, Geneva, Switzerland.

[4] Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–310.

[5] Callison-burch, C., Bannard, C., and Schroeder, J. (2004). Improving statistical translation through editing. In *Proceedings of the 9th EAMT Workshop Broadening horizons of machine translation and its applications*, Malta.

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

[7] Chiang, D., Marton, Y., and Resnik, P. (2008). Online large-margin training of syntactic and structural translation features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 08)*, pages 224–233, Honolulu, Hawaii.

[8] Civera, J. et al. (2004). A syntactic pattern recognition approach to computer assisted translation. In Fred, A., Caelli, T., Campilho, A., Duin, R. P., and de Ridder, D., editors, *Advances in Statistical, Structural and Syntactical Pattern Recognition*, Lecture Notes in Computer Science, pages 207–215. Springer-Verlag.

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

[10] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 03)*, pages 48–54, Edmonton, Canada.

[11] Koehn, P. and Schroeder, J. (2007). Experiments in domain adaptation for statistical machine translation. In *Proceedings of the ACL 2007 Second Workshop on Statistical Machine Translation (WSMT 07)*, pages 224–227, Prague, Czech Republic.

[12] Liang, P., Bouchard-Côté, A., Klein, D., and Taskar, B. (2006). An end-to-end discriminative approach to machine translation. In *In Proceedings of the Joint International Conference on Computational Linguistics and Association of Computational Linguistics (COLING/ACL 06)*, pages 761–768, Sydney, Australia.

[13] Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers.

[14] Nepveu, L., Lapalme, G., Langlais, P., and Foster, G. F. (2004). Adaptive language and translation models for interactive machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 04)*, pages 190–197, Barcelona, Spain.

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

[16] Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2009). Interactive machine translation based on partial statistical phrase-based alignments. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 09)*, pages 330–336, Borovets, Bulgaria.

[17] Ortiz-Martínez, D., García-Varea, I., and Casacuberta, F. (2010). Online learning for interactive statistical machine translation. In *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 10)*, pages 546–554, Los Angeles, USA.

[18] Sanchis, G. and Casacuberta, F. (2010). Bayesian adaptation for statistical machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 10)*, pages 1077–1085, Beijing, China.

[19] Sanchis-Trilles, G. and Cettolo, M. (2010). Online language model adaptation via n-gram mixtures for statistical machine translation. In *Proceedings of the Conference of the European Association for Machine Translation (EAMT 10)*, Saint-Raphaël, France.

[20] Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 96)*, pages 836–841, Copenhagen, Denmark.

[21] Watanabe, T., Suzuki, J., Tsukada, H., and Isozaki, H. (2007). Online large-margin training for statistical machine translation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing (EMNLP 07) and Computational Natural Language Learning (CoNLL 07)*, pages 764–773, Prage, Czeck Republic.