Improving On-line Handwritten Recognition using Translation Models in Multimodal Interactive Machine Translation

Vicent Alabau, Alberto Sanchis, Francisco Casacuberta
Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera, s/n, Valencia, Spain
{valabau,asanchis,fcn}@iti.upv.es

Abstract

In interactive machine translation (IMT), a human expert is integrated into the core of a machine translation (MT) system. The human expert interacts with the IMT system by partially correcting the errors of the system's output. Then, the system proposes a new solution. This process is repeated until the output meets the desired quality. In this scenario, the interaction is typically performed using the keyboard and the mouse. In this work, we present an alternative modality to interact within IMT systems by writing on a tactile display or using an electronic pen. An on-line handwritten text recognition (HTR) system has been specifically designed to operate with IMT systems. Our HTR system improves previous approaches in two main aspects. First, HTR decoding is tightly coupled with the IMT system. Second, the language models proposed are context aware, in the sense that they take into account the partial corrections and the source sentence by using a combination of n-grams and word-based IBM models. The proposed system achieves an important boost in performance with respect to previous work.

1 Introduction

Although current state-of-the-art machine translation (MT) systems have improved greatly in the last ten years, they are not able to provide the high quality results that are needed for industrial and business purposes. For that reason, a new interactive paradigm has emerged recently. In interactive machine translation (IMT) (Foster et al., 1998; Barrachina et al., 2009; Koehn and Haddow, 2009) the system goal is not to produce "perfect" translations in a completely automatic way, but to help the user build the translation with the least effort possible.

A typical approach to IMT is shown in Fig. 1. A source sentence $\boldsymbol{f}$ is given to the IMT system. First, the system outputs a translation hypothesis $\hat{\boldsymbol{e}}_s$ in the target language, which would correspond to the output of fully automated MT system. Next, the user analyses the source sentence and the decoded hypothesis, and validates the longest error-free prefix $\boldsymbol{e}_p$ finding the first error. The user, then, corrects the erroneous word by typing some keystrokes κ, and sends them along with $\boldsymbol{e}_p$ to the system, as a new validated prefix $\boldsymbol{e}_p, \kappa$. With that information, the system is able to produce a new, hopefully improved, suffix $\hat{\boldsymbol{e}}_s$ that continues the previous validated prefix. This process is repeated until the user agrees with the quality of the resulting translation.

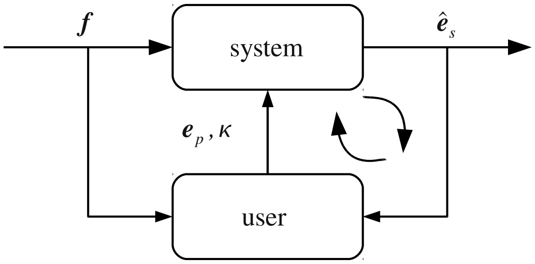

Figure 1: Diagram of a typical approach to IMT

The usual way in which the user introduces the corrections κ is by means of the keyboard. However, other interaction modalities are also possible. For example, the use of speech interaction was studied in (Vidal et al., 2006). In that work, several sce-

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 389–394,
Portland, Oregon, June 19-24, 2011. ©2011 Association for Computational Linguistics

narios were proposed, where the user was expected to speak aloud parts of the current hypothesis and possibly one or more corrections. On-line HTR for interactive systems was first explored for interactive transcription of text images (Toselli et al., 2010). Later, we proposed an adaptation to IMT in (Alabau et al., 2010). For both cases, the decoding of the on-line handwritten text is performed independently as a previous step of the suffix $\boldsymbol{e}_s$ decoding. To our knowledge, (Alabau et al., 2010) has been the first and sole approach to the use of on-line handwriting in IMT so far. However, that work did not exploit the specific particularities of the MT scenario.

The novelties of this paper with respect to previous work are summarised in the following items:

- in previous formalisations of the problem, the HTR decoding and the IMT decoding were performed in two steps. Here, a sound statistical formalisation is presented where both systems are tightly coupled.
- the use of specific language modelling for on-line HTR decoding that take into account the previous validated prefix $\boldsymbol{e}_p, \kappa$, and the source sentence $\boldsymbol{f}$. A decreasing in error of 2% absolute has been achieved with respect to previous work.
- additionally, a thorough study of the errors committed by the HTR subsystem is presented.

The remainder of this paper is organised as follows: The statistical framework for multimodal IMT and their alternatives will be studied in Sec. 2. Section 3 is devoted to the evaluation of the proposed models. Here, the results will be analysed and compared to previous approaches. Finally, conclusions and future work will be discussed in Sec. 4.

2 Multimodal IMT

In the traditional IMT scenario, the user interacts with the system through a series of corrections introduced with the keyboard. This iterative nature of the process is emphasised by the loop in Fig. 1, which indicates that, for a source sentence to be translated, several interactions between the user and the system should be performed. In each interaction, the system produces the most probable suffix $\hat{\boldsymbol{e}}_s$ that completes the prefix formed by concatenating the longest correct prefix from the previous hypothesis $\boldsymbol{e}_p$ and the keyboard correction κ. In addition, the concatenation of them, $(\boldsymbol{e}_p, \kappa, \hat{\boldsymbol{e}}_s)$, must be a translation of $\boldsymbol{f}$. Statistically, this problem can be formulated as

$$\hat{\boldsymbol{e}}_s = \operatorname*{argmax}_{\boldsymbol{e}_s} Pr(\boldsymbol{e}_s | \boldsymbol{e}_p, \kappa, \boldsymbol{f}) \qquad (1)$$

The multimodal IMT approach differs from Eq. 1 in that the user introduces the correction using a touch-screen or an electronic pen, $\boldsymbol{t}$. Then, Eq. 1 can be rewritten as

$$\hat{\boldsymbol{e}}_s = \operatorname*{argmax}_{\boldsymbol{e}_s} Pr(\boldsymbol{e}_s | \boldsymbol{e}_p, \boldsymbol{t}, \boldsymbol{f}) \qquad (2)$$

As $\boldsymbol{t}$ is a non-deterministic input (contrarily to κ), $\boldsymbol{t}$ needs to be decoded in a word d of the vocabulary. Thus, we must marginalise for every possible decoding:

$$\hat{\boldsymbol{e}}_s = \operatorname*{argmax}_{\boldsymbol{e}_s} \sum_{d} Pr(\boldsymbol{e}_s, d | \boldsymbol{e}_p, \boldsymbol{t}, \boldsymbol{f}) \qquad (3)$$

Furthermore, by applying simple Bayes transformations and making reasonable assumptions,

$$\begin{aligned} \hat{\boldsymbol{e}}_s \;\approx\; & \operatorname*{argmax}_{\boldsymbol{e}_s} \max_{d} Pr(\boldsymbol{t}|d)\; Pr(d|\boldsymbol{e}_p, \boldsymbol{f}) \\ & \qquad\qquad Pr(\boldsymbol{e}_s | \boldsymbol{e}_p, d, \boldsymbol{f}) \qquad (4) \end{aligned}$$

The first term in Eq. 4 is a morphological model and it can be approximated with hidden Markov models (HMM). The last term is an IMT model as described in (Barrachina et al., 2009). Finally, $Pr(d|\boldsymbol{e}_p, \boldsymbol{f})$ is a constrained language model. Note that the language model is conditioned to the longest correct prefix, just as a regular language model. Besides, it is also conditioned to the source sentence, since d should result of the translation of it.

A typical session of the multimodal IMT is exemplified in Fig. 2. First, the system starts with an empty prefix, so it proposes a full hypothesis. The output would be the same of a fully automated system. Then, the user corrects the first error, *not*, by writing on a touch-screen. The HTR subsystem mistakenly recognises *in*. Consequently, the user falls back to the keyboard and types `is`. Next, the system proposes a new suffix, in which the first word, *not*, has been automatically corrected. The user amends *at* by writing the word , which is correctly recognised by the HTR subsystem. Finally, as the new proposed suffix is correct, the process ends.

SOURCE ($\boldsymbol{f}$): si alguna función no se encuentra disponible en su red
TARGET ($\boldsymbol{e}$): if any feature is not available in your network

ITER-0	$(\boldsymbol{e}_p)$	
ITER-1	$(\hat{\boldsymbol{e}}_s)$	if any feature not is available on your network
	$(\boldsymbol{e}_p)$	*if any feature*
	$(\boldsymbol{t})$	is
	$(\hat{d})$	~~in~~
	(κ)	`is`
ITER-2	$(\hat{\boldsymbol{e}}_s)$	not available at your network
	$(\boldsymbol{e}_p)$	*not available*
	$(\boldsymbol{t})$	in
	$(\hat{d})$	**in**
FINAL	$(\hat{\boldsymbol{e}}_s)$	your network
	$(\boldsymbol{e}_p \equiv \boldsymbol{e})$	if any feature is not available in your network

Figure 2: Example of a multimodal IMT session for translating a Spanish sentence $\boldsymbol{f}$ from the Xerox corpus to an English sentence $\boldsymbol{e}$. If the decoding of the pen strokes $\hat{d}$ is correct, it is displayed in **boldface**. On the contrary, if $\hat{d}$ is incorrect, it is shown ~~crossed out~~. In this case, the user amends the error with the keyboard κ (in `typewriter`).

2.1 Decoupled Approach

In (Alabau et al., 2010) we proposed a decoupled approach to Eq. 4, where the on-line HTR decoding was a separate problem from the IMT problem. From Eq. 4 a two step process can be performed. First, $\hat{d}$ is obtained,

$$\hat{d} \approx \underset{d}{\operatorname{argmax}} Pr(\boldsymbol{t}|d)\ Pr(d|\boldsymbol{e}_p, \boldsymbol{f}) \tag{5}$$

Then, the most likely suffix is obtained as in Eq 1, but taking $\hat{d}$ as the corrected word instead of κ,

$$\hat{\boldsymbol{e}}_s = \underset{\boldsymbol{e}_s}{\operatorname{argmax}} Pr(\boldsymbol{e}_s|\boldsymbol{e}_p, \hat{d}, \boldsymbol{f}) \tag{6}$$

Finally, in that work, the terms of Eq. 5 were interpolated with a unigram in a log-linear model.

2.2 Coupled Approach

The formulation presented in Eq. 4 can be tackled directly to perform a coupled decoding. The problem resides in how to model the constrained language model. A first approach is to drop either the $\boldsymbol{e}_p$ or $\boldsymbol{f}$ terms from the probability. If $\boldsymbol{f}$ is dropped, then $Pr(d|\boldsymbol{e}_p)$ can be modelled as a regular n-gram model. On the other hand, if $\boldsymbol{e}_p$ is dropped, but the position of d in the target sentence $i = |\boldsymbol{e}_p| + 1$ is kept, $Pr(d|\boldsymbol{f}, i)$ can be modelled as a word-based translation model. Let us introduce a hidden variable j that accounts for a position of a word in $\boldsymbol{f}$ which is a candidate translation of d. Then,

$$\begin{aligned} Pr(d|\boldsymbol{f}, i) &= \sum_{j=1}^{|\boldsymbol{f}|} Pr(d, j|\boldsymbol{f}, i) && (7) \\ &\approx \sum_{j=1}^{|\boldsymbol{f}|} Pr(j|\boldsymbol{f}, i) Pr(d|f_j) && (8) \end{aligned}$$

Both probabilities, $Pr(j|\boldsymbol{f}, i)$ and $Pr(d|f_j)$, can be estimated using IBM models (Brown et al., 1993). The first term is an alignment probability while the second is a word dictionary. Word dictionary probabilities can be directly estimated by IBM1 models. However, word dictionaries are not symmetric. Alternatively, this probability can be estimated using the inverse dictionary to provide a smoothed dictionary,

$$Pr(d|f_j) = \frac{Pr(d)\ Pr(f_j|d)}{\sum_{d'} Pr(d')\ Pr(f_j|d')} \tag{9}$$

Thus, four word-based translation models have been considered: direct IBM1 and IBM2 models, and inverse IBM1-inv and IBM2-inv models with the inverse dictionary from Eq. 9.

However, a more interesting set up than using language models or translation models alone is to combine both models. Two schemes have been studied.

The most formal under a probabilistic point of view is a linear interpolation of the models,

$$Pr(d|\boldsymbol{e}_p, \boldsymbol{f}) = \alpha Pr(d|\boldsymbol{e}_p) + (1-\alpha) Pr(d|\boldsymbol{f}, i) \tag{10}$$

However, a common approach to combine models nowadays is log-linear interpolation (Berger et al., 1996; Papineni et al., 1998; Och and Ney, 2002),

$$Pr(d|\boldsymbol{e}_p, \boldsymbol{f}) = \frac{\exp\left(\sum_m \lambda_m h_m(d, \boldsymbol{f}, \boldsymbol{e}_p)\right)}{Z} \tag{11}$$

λ_m being a scaling factor for model m, h_m the log-probability of each model considered in the log-lineal interpolation and Z a normalisation factor.

Finally, to balance the absolute values of the morphological model, the constrained language model and the IMT model, these probabilities are combined in a log-linear manner regardless of the language modelling approach.

3 Experiments

The Xerox corpus, created on the TT2 project (SchulmbergerSema S.A. et al., 2001), was used for these experiments, since it has been extensively used in the literature to obtain IMT results. The simplified English and Spanish versions were used to estimate the IMT, IBM and language models. The corpus consists of $56k$ sentences of training and a development and test sets of $1.1k$ sentences. Test perplexities for Spanish and English are 33 and 48, respectively.

For on-line HTR, the on-line handwritten UNIPEN corpus (Guyon et al., 1994) was used. The morphological models were represented by continuous density left-to-right character HMMs with Gaussian mixtures, as in speech recognition (Rabiner, 1989), but with variable number of states per character. Feature extraction consisted on speed and size normalisation of pen positions and velocities, resulting in a sequence of vectors of six features (Toselli et al., 2007).

The simulation of user interaction was performed in the following way. First, the publicly available IMT decoder Thot (Ortiz-Martínez et al., 2005) [1] was used to run an off-line simulation for keyboard-based IMT. As a result, a list of words the system failed to predict was obtained. Supposedly, this is the list of words that the user would like to correct with handwriting. Then, from UNIPEN corpus, three users (separated from the training) were selected to simulate user interaction. For each user, the handwritten words were generated by concatenating random character instances from the user's data to form a single stroke. Finally, the generated handwritten words of the three users were decoded using the corresponding constrained language model with a state-of-the-art HMM decoder, *iAtros* (Luján-Mares et al., 2008).

System	Spanish		English	
	dev	test	dev	test
independent HTR (†)	9.6	10.9	7.7	9.6
decoupled (⋆)	9.5	10.8	7.2	9.6
best coupled	**6.7**	**8.9**	**5.5**	**7.2**

Table 1: Comparison of the CER with previous systems. In **boldface** the best system. (†) is an independent, context unaware system used as baseline. (⋆) is a model equivalent to (Alabau et al., 2010).

3.1 Results

Results are presented in *classification error rate* (CER), i.e. the ratio between the errors committed by the on-line HTR decoder and the number of handwritten words introduced by the user. All the results have been calculated as the average CER of the three users.

Table 1 shows a comparison between the best results in this work and the approaches in previous work. The log-linear and linear weights were obtained with the simplex algorithm (Nelder and Mead, 1965) to optimise the development set. Then, those weights were used for the test set.

Two baseline models have been established for comparison purposes. On the one hand, (†) is a completely independent and context unaware system. That would be the equivalent to decode the handwritten text in a separate on-line HTR decoder. This system obtains the worst results of all. On the other hand, (⋆) is the most similar model to the best system in (Alabau et al., 2010). This system is clearly outperformed by the proposed coupled approach.

A summary of the alternatives to language mod-

[1] http://sourceforge.net/projects/thot/

System	Spanish		English	
	dev	test	dev	test
4gr	7.8	10.0	6.3	8.9
IBM1	7.9	9.6	7.0	8.2
IBM2	7.1	**8.6**	6.1	7.9
IBM1-inv	8.4	9.5	7.5	9.2
IBM2-inv	7.9	9.1	7.1	9.1
4gr+IBM2 (L-Linear)	7.0	9.1	6.0	7.9
4gr+IBM2 (Linear)	**6.7**	8.9	**5.5**	**7.2**

Table 2: Summary of the CER results for various language modelling approaches. In **boldface** the best system.

elling is shown in Tab. 2. Up to 5-grams were used in the experiments. However, the results did not show significant differences between them, except for the 1-gram. Thus, context does not seem to improve much the performance. This may be due to the fact that the IMT and the on-line HTR systems use the same language models (5-gram in the case of the IMT system). Hence, if the IMT has failed to predict the correct word because of poor language modelling that will affect on-line HTR decoding as well. In fact, although language perplexities for the test sets are quite low (33 for Spanish and 48 for English), perplexities accounting only erroneous words increase until 305 and 420, respectively.

On the contrary, using IBM models provides a significant boost in performance. Although inverse dictionaries have a better vocabulary coverage (4.7% vs 8.9% in English, 7.4% vs 10.4% in Spanish), they tend to perform worse than their direct dictionary counterparts. Still, inverse IBM models perform better than the n-grams alone. Log-linear models show a bit of improvement with respect to IBM models. However, linear interpolated models perform the best. In the Spanish test set the result is not better that the IBM2 since the linear parameters are clearly over-fitted. Other model combinations (including a combination of all models) were tested. Nevertheless, none of them outperformed the best system in Table 2.

3.2 Error Analysis

An analysis of the results showed that 52.2% to 61.7% of the recognition errors were produced by punctuation and other symbols. To circumvent this problem, we proposed a contextual menu in (Alabau et al., 2010). With such menu, errors would have been reduced (best test result) to 4.1% in Spanish and 2.8% in English. Out-of-vocabulary (OOV) words also summed up a big percentage of the error (29.1% and 20.4%, respectively). This difference is due to the fact that Spanish is a more inflected language. To solve this problem on-line learning algorithms or methods for dealing with OOV words should be used. Errors in gender, number and verb tenses, which rose up to 7.7% and 5.3% of the errors, could be tackled using linguistic information from both source and target sentences. Finally, the rest of the errors were mostly due to one-to-three letter words, which is basically a problem of handwriting morphological modelling.

4 Conclusions

In this paper we have described a specific on-line HTR system that can serve as an alternative interaction modality to IMT. We have shown that a tight integration of the HTR and IMT decoding process and the use of the available information can produce significant HTR error reductions. Finally, a study of the system's errors has revealed the system weaknesses, and how they could be addressed in the future.

5 Acknowledgments

Work supported by the EC (FEDER/FSE) and the Spanish MEC/MICINN under the MIPRCV ”Consolider Ingenio 2010” program (CSD2007-00018), iTrans2 (TIN2009-14511). Also supported by the Spanish MITyC under the erudito.com (TSI-020110-2009-439) project and by the Generalitat Valenciana under grant Prometeo/2009/014 and GV/2010/067, and by the ”Vicerrectorado de Investigación de la UPV” under grant UPV/2009/2851.

References

[Alabau et al.2010] V. Alabau, D. Ortiz-Martínez, A. Sanchis, and F. Casacuberta. 2010. Multimodal interactive machine translation. In *Proceedings of the 2010 International Conference on Multimodal Interfaces (ICMI-MLMI'10)*, pages 46:1–4, Beijing, China, Nov.

[Barrachina et al.2009] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L.

Lagarda, H. Ney, J. Tomás, E. Vidal, and J. M. Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

[Berger et al.1996] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71.

[Brown et al.1993] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The mathematics of machine translation. 19(2):263–311.

[Foster et al.1998] G. Foster, P. Isabelle, and P. Plamondon. 1998. Target-text mediated interactive machine translation. *Machine Translation*, 12:175–194.

[Guyon et al.1994] Isabelle Guyon, Lambert Schomaker, Réjean Plamondon, Mark Liberman, and Stan Janet. 1994. Unipen project of on-line data exchange and recognizer benchmarks. In *Proceedings of International Conference on Pattern Recognition*, pages 29–33.

[Koehn and Haddow2009] P. Koehn and B. Haddow. 2009. Interactive assistance to human translators using statistical machine translation methods. In *Proceedings of MT Summit XII*, pages 73–80, Ottawa, Canada.

[Luján-Mares et al.2008] Míriam Luján-Mares, Vicent Tamarit, Vicent Alabau, Carlos D. Martínez-Hinarejos, Moisés Pastor i Gadea, Alberto Sanchis, and Alejandro H. Toselli. 2008. iATROS: A speech and handwritting recognition system. In *V Jornadas en Tecnologías del Habla (VJTH'2008)*, pages 75–78, Bilbao (Spain), Nov.

[Nelder and Mead1965] J. A. Nelder and R. Mead. 1965. A simplex method for function minimization. *Computer Journal*, 7:308–313.

[Och and Ney2002] F. J. Och and H. Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th ACL*, pages 295–302, Philadelphia, PA, July.

[Ortiz-Martínez et al.2005] D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Proceedings of the MT Summit X*, pages 141–148.

[Papineni et al.1998] K. A. Papineni, S. Roukos, and R. T. Ward. 1998. Maximum likelihood and discriminative training of direct translation models. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'98)*, pages 189–192, Seattle, Washington, USA, May.

[Rabiner1989] L. Rabiner. 1989. A Tutorial of Hidden Markov Models and Selected Application in Speech Recognition. *Proceedings IEEE*, 77:257–286.

[SchulmbergerSema S.A. et al.2001] SchulmbergerSema S.A., Celer Soluciones, Instituto Técnico de Informática, R.W.T.H. Aachen - Lehrstuhl für Informatik VI, R.A.L.I. Laboratory - University of Montreal, Société Gamma, and Xerox Research Centre Europe. 2001. X.R.C.: TT2. TransType2 - Computer assisted translation. Project technical annex.

[Toselli et al.2007] Alejandro H. Toselli, Moisés Pastor i Gadea, and Enrique Vidal. 2007. On-line handwriting recognition system for tamil handwritten characters. In *3rd Iberian Conference on Pattern Recognition and Image Analysis*, pages 370–377. Girona (Spain), June.

[Toselli et al.2010] A. H. Toselli, V. Romero, M. Pastor, and E. Vidal. 2010. Multimodal interactive transcription of text images. *Pattern Recognition*, 43(5):1814–1825.

[Vidal et al.2006] E. Vidal, F. Casacuberta, L. Rodríguez, J. Civera, and C. Martínez. 2006. Computer-assisted translation using speech recognition. *IEEE Transaction on Audio, Speech and Language Processing*, 14(3):941–951.