Interactive machine translation based on partial statistical phrase-based alignments

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

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

Francisco Casacuberta
Dpto. de Sist. Inf. y Comp.
Univ. Politécnica 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. We present a new technique for IMT which is based on the generation of partial alignments at phrase-level. The proposed technique partially aligns the source sentence with the user prefix and then translates the unaligned portion of the source sentence. The generation of such partial alignments is driven by statistical phrase-based models. Our technique relies on the application of smoothing techniques over the phrase models to appropriately assign probabilities to unseen events. We report experiments investigating the impact of the different smoothing techniques in the accuracy of our system. In addition, we compare the results obtained by our system with those obtained by other well-known IMT systems.

Keywords

Statistical machine translation, interactive machine translation, phrase-based translation, phrase-based alignments, smoothing.

1 Introduction

Information technology advances in modern society have led to the need of more efficient methods of translation. It is worth mentioning that current MT systems are not able to produce ready-to-use texts. Indeed, MT systems usually require human post-editing in order to achieve high-quality translations.

One way of taking advantage of MT systems is to combine them with the knowledge of a human translator, constituting the Interactive Machine Translation (IMT) paradigm. This IMT paradigm can be considered a special type of the so-called Computer-Assisted Translation (CAT) paradigm.

An important contribution to IMT technology was carried out within the TransType (TT) project [11, 7, 5]. This project entailed a focus shift in which interaction directly aimed at the production of the target text, rather than at the disambiguation of the source text, as in former interactive systems. The idea proposed in that work was to embed data driven MT techniques within the interactive translation environment.

Following these TT ideas, [1] proposed a new approach to IMT. In this approach, 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 further, hopefully improved suggestions. Figure 1 illustrates a typical IMT session.

In this paper, we also focus on the IMT approach to CAT. Specifically, we propose a new IMT engine based on the generation of partial alignments at phrase-level. The proposed technique partially aligns the source sentence with the user prefix and then translates the unaligned portion of the source sentence. The partial alignments are generated using the statistical knowledge provided by a phrase-based model. As it will be shown, the techniques proposed here require the application of smoothing techniques over the phrase-based models to correctly assign probabilities to unseen events.

2 Statistical interactive MT

IMT can be seen as an evolution of the SMT framework. The fundamental equation of the statistical approach to MT is:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \left\{ Pr(\mathbf{f} \,|\, \mathbf{e}) \cdot Pr(\mathbf{e}) \right\} \qquad (1)$$

where $Pr(\mathbf{f} \,|\, \mathbf{e})$ is approached by a *translation model* that tries to represent the correlation between source and target sentence and $Pr(\mathbf{e})$ is approached by *language model* representing the well-formedness of the candidate translation $\mathbf{e}$.

Current MT systems are based on the use of phrase-based models [19, 10] 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 arrive to the following expression:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{K, \tilde{a}_1^K} Pr(\tilde{f}_1^K, \tilde{a}_1^K \,|\, \tilde{e}_1^K) \qquad (2)$$

330

source	Para ver la lista de recursos
interaction-0	To view the resources list
interaction-1	To view $\boxed{a}$ list of resources
interaction-2	To view a list $\boxed{i}$ ng resources
interaction-3	To view a listing $\boxed{o}$ f resources
acceptance	To view a listing of resources

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

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 Eq. (2), and following a maximum approximation, the problem stated in Eq. (1) can be reframed as:

$$\hat{\mathbf{e}} \approx \arg \max_{\mathbf{e},\mathbf{a}} \left\{ p(\mathbf{e}) \cdot p(\mathbf{f},\mathbf{a} \,|\, \mathbf{e}) \right\} \qquad (3)$$

State-of-the-art statistical machine translation systems model $p(\mathbf{f},\mathbf{a}|\mathbf{e})$ following a loglinear approach [14], that is:

$$p(\mathbf{f},\mathbf{a}|\mathbf{e}) \propto \exp \left[\sum_i \lambda_i f_i(\mathbf{f},\mathbf{e},\mathbf{a}) \right] \qquad (4)$$

In the IMT scenario we have to find an extension $\mathbf{e}_s$ for a given prefix $\mathbf{e}_p$. For this purpose we reformulate Eq. (3) as follows:

$$\hat{\mathbf{e}}_s \approx \arg \max_{\mathbf{e}_s,\mathbf{a}} \left\{ p(\mathbf{e}_s \,|\, \mathbf{e}_p) \cdot p(\mathbf{f},\mathbf{a} \,|\, \mathbf{e}_p,\mathbf{e}_s) \right\} \qquad (5)$$

where the term $p(\mathbf{e}_p)$ has been dropped since it does not depend on $\mathbf{e}_s$ and $\mathbf{a}$.

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 Eq. (5) and Eq. (3) (note that $\mathbf{e}_p\mathbf{e}_s \equiv \mathbf{e}$) allow us to use the same models if the search procedures are adequately modified [2, 1].

3   Related work

Several IMT systems have been proposed in the literature. For example, in [7] a maximum entropy version of IBM 2 model is used as word-based translation model. In [15] the Alignment Template approach to IMT is proposed. In that work a pre-computed word translation graph is used in order to achieve fast response times. This approach is compared with the use of a direct translation modeling [2]. In [4] an IMT approach based on stochastic finite-state transducers is presented. In that work, also word translation graphs are used to resolve real-time constraints. In [18] a phrase-based approach is presented.

Recently, in [1] the IMT approach to CAT is proposed, establishing the state-of-the-art in this discipline. In this work the last three approaches mentioned above are compared.

In the following sections, we present a new IMT technique which is based on the generation of partial alignments at phrase-level. The proposed technique partially aligns the source sentence with the user prefix and then translates the unaligned portion of the source sentence. The generation of such partial alignments is driven by statistical phrase-based models. Our technique relies on the application of smoothing techniques over the phrase models to appropriately assign probabilities to unseen events.

The IMT system we propose is similar to those presented in [2] and [18]. The so-called *interactive* generation strategy presented in [2] does not use word graphs as well as our proposal. The key difference between their system and the system we propose is that they use error-correcting techniques instead of smoothing techniques to assign probabilities to unseen events. Specifically, the error correcting costs are introduced as an additional weight in their log-linear model. We think that our approach is better motivated from a theoretical point of view, as it has been deeply studied and demonstrated in the field of language modelling. In addition, our system needs much less time per iteration (hundredths of seconds vs. seconds, as will be shown in section 7) than the system presented in [2].

The work presented in [18] is based on filtering the phrase table to obtain translations that are compatible with the user prefix. Since this approach seems too restrictive (phrase models always present coverage problems in complex tasks, as is discussed in section 4), we guess that also any sort of smoothing is taken into account, but as far as we know the exact technique that is used is not explained. Because of this, we think that the work presented in [18] can benefit from the study on smoothing techniques presented here.

4   Phrase-based alignments

The problem of finding the best alignment at phrase level has been studied in [16, 8, 13]. The concept of phrase-based alignment can be formalized as follows:

Let $\mathbf{f} \equiv f_1, f_2, \ldots, f_J$ be a source sentence and $\mathbf{e} \equiv e_1, e_2, \ldots, e_I$ the corresponding target sentence in a bilingual corpus. A phrase-alignment between $\mathbf{f}$ and $\mathbf{e}$ is defined as a set $\mathcal{S}$ of ordered pairs included in $\mathcal{P}(\mathbf{f}) \times \mathcal{P}(\mathbf{e})$, where $\mathcal{P}(\mathbf{f})$ and $\mathcal{P}(\mathbf{e})$ are the set of all subsets of consecutive sequences of words, of $\mathbf{f}$ and $\mathbf{e}$, respectively. In addition, the ordered pairs contained in $\mathcal{S}$ have to include all the words of both the source and target sentences.

A phrase-based alignment of length K ($\tilde{A}_K$) of a sentence pair $(\mathbf{f},\mathbf{e})$ is defined as a triple $\tilde{A}_K \equiv$

$(\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$, where $\tilde{a}_1^K$ is a specific one-to-one mapping between the K segments/phrases of both sentences $(1 \leq K \leq \min(J, I))$.

Then, given a pair of sentences $(\mathbf{f}, \mathbf{e})$ and a phrase alignment model, we have to obtain the best phrase-alignment $\tilde{A}_K$ (or Viterbi phrase-alignment $V(\tilde{A}_K)$) between them. Assuming a phrase-alignment of length K, $V(\tilde{A}_K)$ can be computed as:

$$V(\tilde{A}_K) = \arg\max_{\tilde{A}_K} \left\{ p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K) \right\} \qquad (6)$$

where, following the assumptions of [19], $Pr(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K)$ can be efficiently computed as:

$$p(\tilde{f}_1^K, \tilde{a}_1^K | \tilde{e}_1^K) = \prod_{k=1}^{K} p(\tilde{f}_k | \tilde{e}_{\tilde{a}_k}) \qquad (7)$$

The model parameters $(\{p(\tilde{f}|\tilde{e})\})$ are typically estimated via relative frequencies as $p(\tilde{f}|\tilde{e}) = N(\tilde{f}, \tilde{e})/N(\tilde{e})$, where $N(\tilde{f}|\tilde{e})$ is the number of times that $\tilde{f}$ has been seen as a translation of $\tilde{e}$ within the training corpus.

On the basis of Eq. (7), a very straightforward technique can be proposed for finding the best phrase-alignment of a sentence pair $(\mathbf{f}, \mathbf{e})$. This can be conceived as a sort of *constrained* translation. In this way, the search process only requires the use of a regular SMT system which filters its phrase-table in order to obtain those translations of $\mathbf{f}$ that are compatible with $\mathbf{e}$.

As noted in [16], this technique has no practical interest when applied on regular tasks. Specifically, the technique is not applicable when the alignments cannot be generated due to coverage problems of the phrase-based model (i.e. one or more phrase pairs required to compose a given alignment have not been seen during the training process). Coverage problems are very frequent in complex translation tasks as will be shown in section 7. In order to solve this problem, an alternative technique is proposed. The alternative technique is able to consider every source phrase of $\mathbf{f}$ as a possible translation of every target phrase of $\mathbf{e}$. For this purpose, it uses a general mechanism for assigning probabilities to phrase pairs based on the application of smoothing techniques over the phrase-table. In addition, the search algorithm that is used no longer filters its phrase-table to generate the sentence $\mathbf{e}$, but instead it can efficiently explore the set of possible alignments between $\mathbf{f}$ and $\mathbf{e}$ (see [16] for more details).

4.1 A log-linear approach to phrase-to-phrase alignments

The score for a given alignment can be calculated according Eq (7). This scoring function does not allow control of basic aspects of the phrase alignment, such as the lengths of the source and target phrases, and the reorderings of phrase alignments. This problem can be alleviated following the approach stated in Eq. (4), thus introducing different feature functions as scoring components in a log-linear fashion.

We use the same set of feature functions proposed in [16]:

- $f_1(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{e}_{\tilde{a}_k} | \tilde{f}_k))$: direct phrase model log-probability

- $f_2(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{f}_k | \tilde{e}_{\tilde{a}_k}))$: inverse phrase model log-probability

- $f_3(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(|\tilde{e}_k|))$: target phrase length model. This component can be modeled by means of a uniform distribution (penalizes the length of the segmentation) or a geometric distribution (penalizes the length of the target phrases)

- $f_4(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(\tilde{a}_k | \tilde{a}_{k-1}))$: distortion model. This component is typically modeled by means of a geometric distribution (penalizes the reorderings)

- $f_5(\mathbf{f}, \mathbf{e}, \mathbf{a}) = \log(\prod_{k=1}^{K} p(|\tilde{f}_k| \mid |\tilde{e}_{\tilde{a}_k}|))$: source phrase length model given the length of the target phrase. This component can be modeled by means of different distributions: uniform (does not take into account the relationship between the length of source and target phrase), Poisson or geometric

The corresponding weights $\lambda_i, i \in \{1, 2, \ldots, 5\}$ can be computed by means of MERT training.

5 Smoothing

As was mentioned in section 4, the application of smoothing techniques is crucial in the generation of phrase-alignments. Most of the well-known language model smoothing techniques (see for example [12]) can be imported to the SMT field and specifically to the PBT framework, as it is shown in [6]. However, PBT and the generation of phrase-alignments differ in a key aspect. While in PBT the probabilities of unseen events are not important (since the decoder only proposes phrase translations contained in the model, see [6]), in the generation of phrase alignments, assigning probabilities to unseen events is one of the most important problems that has to be solved (see [16]).

In the rest of this section, we describe the smoothing techniques that has been used in our work.

5.1 Statistical estimators

Training data can be exploited in different ways to estimate statistical models. Regarding the phrase-based models, the standard estimation technique is based on the relative frequencies of the phrase pairs. Taking this standard estimation technique as a starting point, a number of alternative estimation techniques can be derived.

We have implemented the following estimation techniques for phrase-based models: Maximum-likelihood (ML), Good-Turing (GT), Absolute-discount (AD), Kneser-Ney smoothing (KN), and Simple discount (SD). The SD estimation technique works in a similar way to AD estimation but it subtracts a fixed probability mass instead of a fixed count.

A good way to tackle the problem of unseen events is the use of probability distributions that decompose phrases into words. In our work we have used the

IBM 1 model as defined in [3] to assign probabilities to phrase pairs instead of sentence pairs (this distribution will be referred to as LEX).

5.2 Combining estimators

The statistical estimators described above can be combined in the hope of producing better models. We have chosen three different techniques for combining estimators: Linear interpolation, Backing-off, and Log-linear interpolation. Specifically, we have implemented combinations of two estimators, a phrase-based model estimator (ML, GT, AD, KN or SD estimator) and the LEX estimator.

The key difference between interpolation and backing off is that the latter only uses information from the smoothing distribution (the LEX distribution) for low frequency or unseen events. Since for phrase alignment generation, better prediction of unseen events has a great impact, backing-off seems a specially suitable approach.

Finally, the main difference between linear and log-linear combination is that the former moderates extreme probability values and preserves intermediate values, whereas the latter preserves extreme values and makes intermediate values more extreme. When assigning probabilities to unseen events, the phrase-based model statistical estimators will produce very low or zero probabilities that will be moderated by linear combination (using the LEX distribution), and preserved by log-linear combination. Because of this, we expect linear combination to work better than log-linear combination.

6 IMT based on partial phrase-based alignments

In this section we propose a new IMT technique based on the generation of *partial phrase-alignments* between the source sentence $\mathbf{f}$ and the user prefix $\mathbf{e}_p$. The concept of partial phrase-alignment is similar to the concept of complete phrase alignment described in section 4. Specifically, we define a partial alignment between $\mathbf{f}$ and $\mathbf{e}_p$ as the set $\mathcal{S}'$ of ordered pairs that contains all the words of $\mathbf{e}_p$ and only a subset of the words of $\mathbf{f}$.

The generation of the suffix in IMT can be seen as a two-stage process. First we partially align the prefix $\mathbf{e}_p$ with a part of $\mathbf{f}$, and second, we translate the unaligned portion of $\mathbf{f}$ (if any) giving the suffix $\mathbf{e}_s$. For this purpose, we propose the use of a *stack-decoding* algorithm [9]. The stack-decoding algorithm attempts to iteratively *expand* partial solutions, called hypotheses, until a complete translation is found. The expanded hypotheses are stored into a stack data structure which allows the efficient exploration of the search space.

The expansion process consists of appending target phrases as translation of previously uncovered source phrases of a given hypothesis. Let us suppose that we are translating the sentence $\mathbf{f} \equiv$ "Para ver la lista de recursos", and that the user has validated the prefix $\mathbf{e}_p \equiv$ "To view a" (interaction 1 of the IMT session given in Figure 1). Figure 2 shows an example of the

results obtained by the expansion algorithm that we propose for two hypotheses h_1 and h_2.

Hypothesis h_1 has covered the source phrase "Para ver la" (covered phrases are noted with underlined words in Figure 2), appending the target phrase "To view a". Since for h_1, the user prefix $\mathbf{e}_p$ has already been generated, the expansion process works in the same way as the one executed in a regular translator. Let us suppose that we are covering the source phrase $\tilde{f} \equiv$ "lista de recursos" given by the source positions $u = \{4,5,6\}$. We generate the new hypotheses h_3 and h_4 by appending target phrases $\tilde{e}$ from the set $\mathcal{T}_{\tilde{f}}$ of translations for $\tilde{f}$ contained in the phrase table.

Regarding the hypothesis h_2, it has covered the source phrase "Para" appending the target phrase "To". In this case, the prefix $\mathbf{e}_p$ has not been completely generated. Let $\mathbf{e}_r \equiv$ "view a" be the remaining words that are to be appended to h_2 to complete the user prefix. In this case, we have to take into account whether we are covering the last source phrase positions or not. For example, let us suppose that we cover the phrase positions $u = \{2,3,4,5,6\}$ ($\tilde{f} \equiv$ "ver la lista de recursos"). Since those are the last positions to be covered, we have to ensure that the whole prefix $\mathbf{e}_p$ is generated. For this purpose, we append $\mathbf{e}_r$ to h_2, resulting in the hypothesis h_5. In addition, we can append phrases $\tilde{e}$ contained in the set $\mathcal{T}_{\tilde{f}}$ having $\mathbf{e}_r$ as sub-prefix (if any). This allows the generation of hypotheses like h_6 that takes advantage of the information contained in the phrase table.

In contrast, if we are not covering the last phrase positions of h_2, we can also append strings from the set $\mathcal{S}_{\mathbf{e}_r}$ of sub-prefixes of $\mathbf{e}_r$ to the newly generated hypotheses, allowing the translation system to complete the whole prefix $\mathbf{e}_p$ in subsequent expansion processes. For example, let us suppose that we cover the phrase positions $u = \{2\}$ ($\tilde{f} \equiv$ "ver"). In this case we can append the phrase "view" which is a subprefix of $\mathbf{e}_r$, resulting in the hypothesis h_7. In addition, we can also append $\mathbf{e}_r$ itself, resulting in the hypothesis h_8. Finally, appending phrases from $\mathcal{T}_{\tilde{f}}$ having $\mathbf{e}_r$ as sub-prefix (if any) can also be considered, although this situation has not been depicted in Figure 2.

Algorithm 1 shows the expansion algorithm that we propose for its application in IMT. The algorithm is a formalization of the ideas depicted in Figure 2.

The time cost of the IMT expansion algorithm can be reduced by the introduction of pruning techniques. Such pruning techniques include hypotheses recombination, stack length limitation and restrictions on the maximum number of target phrases that can be linked to an unaligned source phrase during the expansion process. Specifically, in those cases where $\mathbf{e}_p$ has not already been generated, only a subset of the strings contained in the set $\mathcal{S}_{\mathbf{e}_r}$ are considered as candidates for the expansion process. One possible criterion to choose the substrings is based on the length of the phrase $\tilde{f}$ to be translated determined by u. Only those substrings with lengths similar to the length of $\tilde{f}$ are considered. In addition, the set of expanded hypotheses that is returned by the algorithm can be sorted by score, keeping only the best ones.
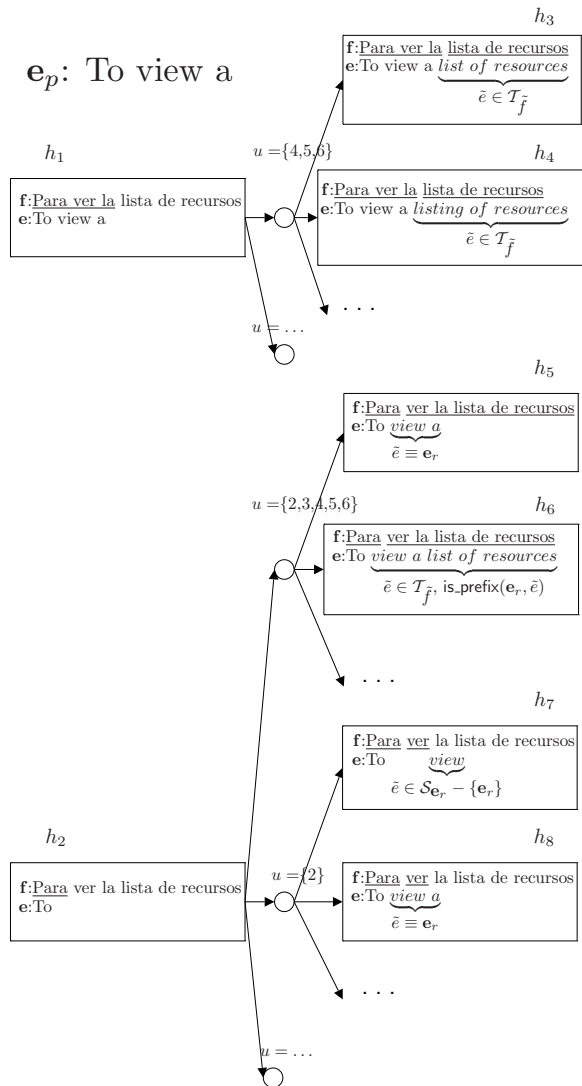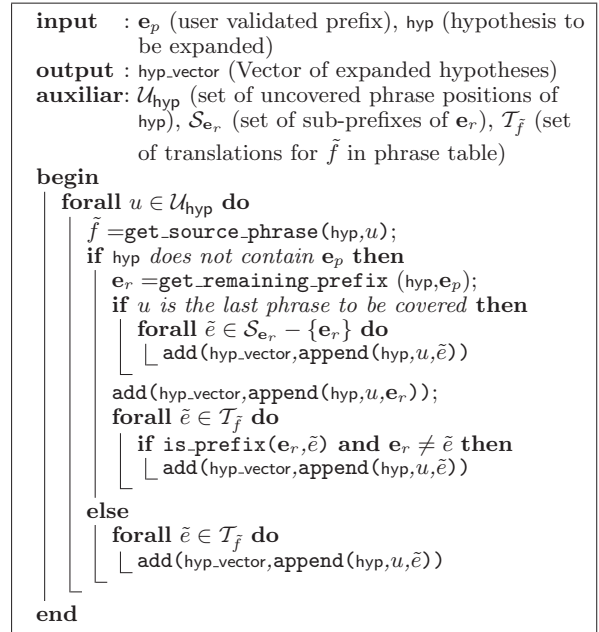
Fig. 2: *Example of the expansion of two hypotheses h_1 and h_2 given $\mathbf{f} \equiv$ "Para ver la lista de recursos" and the user prefix $\mathbf{e}_p \equiv$ "To view a"*

```
input   : e_p (user validated prefix), hyp (hypothesis to
          be expanded)
output  : hyp_vector (Vector of expanded hypotheses)
auxiliar: U_hyp (set of uncovered phrase positions of
          hyp), S_e_r (set of sub-prefixes of e_r), T_f̃ (set
          of translations for f̃ in phrase table)
begin
   forall u ∈ U_hyp do
      f̃ =get_source_phrase(hyp,u);
      if hyp does not contain e_p then
         e_r =get_remaining_prefix (hyp,e_p);
         if u is the last phrase to be covered then
            forall ẽ ∈ S_e_r − {e_r} do
               add(hyp_vector,append(hyp,u,ẽ))
         add(hyp_vector,append(hyp,u,e_r));
         forall ẽ ∈ T_f̃ do
            if is_prefix(e_r,ẽ) and e_r ≠ ẽ then
               add(hyp_vector,append(hyp,u,ẽ))
      else
         forall ẽ ∈ T_f̃ do
            add(hyp_vector,append(hyp,u,ẽ))
end
```

Algorithm 1: Pseudocode for the IMT hypothesis expansion algorithm

---

of the complexity of these corpora, the BLEU score and the number of sentences with coverage problems of the test partition (for both translation directions) are also reported. As can be seen, there is a great number of sentences that present coverage problems for the different corpora.

It is worth noting that the manuals were not the same in each pair of languages, therefore the figures for the different English counterparts are shown.

IMT experiments were carried out for both directions of the three different corpora.

7.2 Assessment criteria

The evaluation of the techniques presented in this papers were carried out using the *Key-stroke and mouse-action ratio* (KSMR) measure [1], which 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), divided by the total number of reference characters.

In the experiments we carried out only one reference translation was considered.

7.3 IMT results

In Table 2 the IMT results using different phrase-to-phrase alignment smoothing techniques are presented, for three different language pairs and translation directions, Geometric distributions were selected to implement both the f_3 and f_5 feature functions. The first row of the table shows the baseline, which consists of the results obtained using a maximum likelihood estimation (ML) without smoothing. The rows labelled with (GT, AD, KN, and SD) show the results for the phrase-based model estimators presented

7 Experimental results

In this section we describe the experiments we carried out to test the IMT techniques that we have presented in previous sections.

7.1 Experimental setup

The experiments were performed using the XEROX XRCE corpus [17], which consist of translation of Xerox printer manual 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 by randomly selecting (without replacement) a specific number of development and test sentences and leaving the remaining ones for training. In order to get a first impression

		Spa	**Eng**	**Fre**	**Eng**	**Ger**	**Eng**
Train	Sent. pairs	55761		52844		49376	
	Running words	657172	571960	573170	542762	440682	506877
	Vocabulary	29565	25627	27399	24958	37338	24899
Dev	Sent. pairs	1012		994		964	
	Running words	13808	12111	9801	9480	8283	9162
	Perplexity (3-grams)	34.0	46.2	74.1	96.2	124.3	68.4
Test	Sent. pairs	1125		984		996	
	Running words	9358	7634	9805	9572	9823	10792
	Running characters	57536	45770	62885	54757	70963	61327
	Perplexity (3-grams)	59.6	107.0	135.4	192.6	169.2	92.8
	BLEU score	58.7	51.1	26.7	24.9	24.8	16.4
	Sents. with coverage problems	617	633	615	653	724	722

Table 1: XEROX *corpus statistics for three differente language pairs*

in section 5.1. The rest of the rows corresponds to different estimation techniques combined with linear interpolation (LI), backing-off (BO), and log-linear interpolation (LL). As was expected (see section 5.2) linear interpolation and backing-off obtains better results than log-linear interpolation.

The baseline system obtained by far the worst results. In contrast, all those experiments that included the LEX distribution outperformed the others due to improved assignment of probabilities to unseen events.

Smooth.	**Spa-Eng**	**Fre-Eng**	**Ger-Eng**
ML	36.7/32.5	59.4/53.2	63.6/57.2
GT	28.6/29.4	51.9/49.4	57.7/53.0
AD	30.3/28.1	50.4/46.7	58.4/52.5
KN	30.3/28.1	50.4/46.7	58.4/52.4
SD	28.5/29.4	51.6/49.2	57.1/52.5
ML+LEX$_{LI}$	21.2/21.3	39.9/39.2	43.9/42.4
GT+LEX$_{LI}$	21.1/21.3	39.9/39.2	44.2/42.2
AD+LEX$_{LI}$	21.4/22.2	40.2/40.5	45.1/42.2
KN+LEX$_{LI}$	21.5/22.2	40.1/40.5	45.0/42.2
SD+LEX$_{LI}$	21.2/21.2	39.9/39.0	44.0/41.8
GT+LEX$_{BO}$	21.1/21.0	39.8/39.0	45.3/42.3
SD+LEX$_{BO}$	21.2/21.0	39.8/39.2	45.1/42.3
ML+LEX$_{LL}$	37.5/35.5	59.5/53.7	64.3/58.0
GT+LEX$_{LL}$	24.0/25.8	43.2/43.3	50.9/46.9
AD+LEX$_{LL}$	30.8/29.2	51.3/46.9	59.7/52.1
KN+LEX$_{LL}$	30.9/29.1	51.4/46.9	59.7/52.0
SD+LEX$_{LL}$	23.6/27.7	43.2/42.7	50.7/45.9

Table 2: *KSMR results for the three* XEROX *corpora (for both direct and inverse translation directions separated by the symbol "/") for different smoothing techniques. Geometric distributions were selected to implement the f_3 and f_5 feature functions*

In order to study the effect of the different probability distributions used for the feature functions f_3 (target phrase length model) and f_5 (source phrase length model) an exhaustive experimentation was carried for all smoothing techniques, and their respective combinations with the Lexical distribution. Table 3 reports the KSMR results for all possible combinations of the probability distributions used for f_3 (Uniform (U) and Geometric (G)) and for f_5 (Uniform (U), Geometric (G), and Poisson (P)). As can be seen in this table slight KSMR differences are obtained. In Table 3 only the results obtained for the best smoothing technique (Good-Turing) are reported. The best results were obtained when U+G distribution were used for

the GT smoothing estimation, and G+G for the BO combination. As was mentioned in section 4.1, the use of a uniform distribution for f_3 penalizes the length of the segmentation and the use of a geometric distribution penalized the length of the source phrases. Correspondingly, the use of a geometric distribution for f_5 makes it possible to establish a relationship between the length of source and target phrases (the use of a Poisson distribution also worked well).

Smooth.	f_3, f_5	**Spa-Eng**	**Fre-Eng**	**Ger-Eng**
GT	U,U	30.1/29.0	53.8/50.7	58.0/53.9
	U,P	29.5/28.6	52.9/49.7	57.6/53.4
	U,G	28.7/28.0	51.7/48.7	57.3/52.7
	G,U	30.5/29.7	54.6/51.5	58.5/54.4
	G,P	29.7/29.4	53.3/50.5	58.2/53.7
	G,G	28.6/29.4	51.9/49.4	57.7/53.0
GT+ **LEX$_{BO}$**	U,U	21.8/21.6	40.4/39.1	44.8/42.2
	U,P	21.5/21.4	40.2/39.0	44.3/42.0
	U,G	21.3/21.4	40.1/38.8	44.0/41.8
	G,U	21.6/21.5	40.3/39.1	44.6/42.1
	G,P	21.4/21.3	40.0/39.0	44.2/41.9
	G,G	21.1/21.0	39.8/39.0	45.3/42.3

Table 3: *KSMR results for the three* XEROX *corpora (for both direct and inverse translation directions separated by the symbol "/") for all possible combinations of the probability distributions for the f_3 and f_5 feature functions when using two different smoothing techniques*

In Table 4 the IMT results for the three considered corpora (for both translation directions) are shown. MERT training for the development corpus was performed to adjust the weights of the log-linear model. In this case, only the Good-Turing (GT) and Simple Discount (SD) results are reported, showing that both techniques yielded similar results. The last column of Table 4 shows the average time in seconds per iteration needed to complete a new translation given a user validated prefix. Clearly, these times allow the system to work on a real time scenario.

Finally, in Table 5 a comparison of the best results obtained in this work (Partial Statistical Phrase-based Alignments (PSPBA)) with state-of-the-art IMT systems is reported (95% confidence intervals are shown). We compared our system with those presented in [1]: the alignment templates (AT), the stochastic finite-state transducer (SFST), and the phrase-based (PB)

Corpus	Smooth.	KSMR	secs./iter.
Spa–Eng	GT+LEX$_{BO}$	19.6	0.086
	SD+LEX$_{BO}$	19.6	0.090
Eng–Spa	GT+LEX$_{BO}$	17.5	0.093
	SD+LEX$_{BO}$	17.6	0.094
Fre–Eng	GT+LEX$_{BO}$	36.9	0.204
	SD+LEX$_{BO}$	37.0	0.205
Eng–Fre	GT+LEX$_{BO}$	34.4	0.148
	SD+LEX$_{BO}$	34.4	0.147
Ger–Eng	GT+LEX$_{BO}$	39.5	0.170
	SD+LEX$_{BO}$	39.5	0.184
Eng–Ger	GT+LEX$_{BO}$	39.1	0.152
	SD+LEX$_{BO}$	39.2	0.154

Table 4: *KSMR results for the three* XEROX *corpora, using geometric distributions for f_3 and f_5 feature functions. MERT training was performed. The average time (in secs.) per iteration is also reported*

Corpus	AT	PB	SFST	PSPBA
Spa–Eng	24.0±1.3	18.1±1.2	26.9±1.3	19.6±1.1
Eng–Spa	23.2±1.3	16.7±1.2	21.8±1.4	17.6±1.1
Fre–Eng	40.5±1.4	37.2±1.3	45.5±1.3	37.0±1.4
Eng–Fre	40.4±1.4	35.8±1.3	43.8±1.6	34.4±1.2
Ger–Eng	45.9±1.2	36.7±1.2	46.6±1.4	39.5±1.1
Eng–Ger	44.7±1.2	40.1±1.2	45.7±1.4	39.2±1.1

Table 5: *KSMR results comparison of our system and three different state-of-the art IMT systems. 95% confidence intervals are shown*

approaches to IMT. As can be seen, our system obtains similar results and in some cases clearly outperforms the results obtained by these IMT systems. Specifically, our results were better than those obtained by the SFST and the AT systems. In contrast, the KSMR results with respect to the PB approach were similar.

8   Conclusions

We have presented a new technique for IMT which is based on the generation of partial alignments at phrase-level. The generation of such partial alignments is driven by statistical phrase-based models and relies on the application of smoothing techniques to assign probabilities to unseen events.

The experiments we carried out show the great impact of the smoothing techniques in the accuracy of our system. The combination of a phrase-based model estimator with a lexical distribution yielded the best results. Three different combination techniques were tested: backing-off, linear interpolation and log-linear interpolation. As we expected, backing-off and linear interpolation worked better than log-linear

Finally, we have compared the results obtained by our system with those obtained by state-of-the-art IMT systems. Our system obtained similar results and in some cases clearly outperformed the results obtained by the state-of-the-art systems.

Acknowledgments

References

[1] S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. H. Ney, J. Tomás, and E. Vidal. Statistical approaches to computer-assisted translation. *Computational Linguistics*, page In press, 2008. doi: 10.1162/coli.2008.07-055-R2-06-29.

[2] O. Bender, S. Hasan, D. Vilar, R. Zens, and H. Ney. Comparison of generation strategies for interactive machine translation. In *Conference of the European Association for Machine Translation*, pages 33–40, Budapest, Hungary, May 2005.

[3] P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, 1993.

[4] J. Civera, J. M. Vilar, E. Cubel, A. L. Lagarda, S. Barrachina, E. Vidal, F. Casacuberta, D. Picó, and J. González. From machine translation to computer assisted translation using finite-state models. In *Proc. of EMNLP*, Barcelona, 2004.

[5] G. Foster. *Text Prediction for Translators*. PhD thesis, Université de Montréal, 2002.

[6] G. Foster, R. Kuhn, and H. Johnson. Phrasetable smoothing for statistical machine translation. In *Proc. of the EMNLP*, pages 53–61, Sydney, Australia, July 2006. ACL.

[7] G. Foster, P. Langlais, and G. Lapalme. User-friendly text prediction for translators. In *Proc. of EMNLP'02*, pages 148–155, 2002.

[8] I. García-Varea, D. Ortiz, F. Nevado, P. A. Gómez, and F. Casacuberta. Automatic segmentation of bilingual corpora: A comparison of different techniques. In *Proc. of the 2nd IbPRIA*, volume 3523 of *LNCS*, pages 614–621. Estoril (Portugal), June 2005.

[9] F. Jelinek. A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, 13:675–685, 1969.

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

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

[12] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts 02142, 2001.

[13] D. Marcu and W. Wong. A phrase-based, joint probability model for statistical machine translation. In *emnlp*, pages 1408–1414, Philadelphia, USA, July 2002.

[14] F. J. Och and H. Ney. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the 40th ACL*, pages 295–302, Philadelphia, PA, July 2002.

[15] F. J. Och, R. Zens, and H. Ney. Efficient search for interactive statistical machine translation. In *EACL '03: Tenth Conf. of the Europ. Chapter of the Association for Computational Linguistics*, pages 387–393, Budapest, Hungary, Apr. 2003.

[16] D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model. In *Proc. of EAMT'08*, Hamburg, Germany, September 2008.

[17] SchlumbergerSema S.A., I. T. de Informática, R. W. T. H. A. L. für Informatik VI, R. A. en Linguistique Informatique Laboratory University of Montreal, C. Soluciones, S. Gamma, and X. R. C. Europe. TT2. TransType2 - computer assisted translation. Project technical annex., 2001.

[18] J. Tomás and F. Casacuberta. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the Coling/ACL2006*, pages 835–841, Sydney, Australia, 17th-21th July 2006. http://acl.ldc.upenn.edu/P/P06/P06-2107.pdf.

[19] R. Zens, F. J. Och, and H. Ney. Phrase-based statistical machine translation. In *Advances in artificial intelligence. 25. Annual German Conference on AI*, volume 2479 of *LNCS*, pages 18–32. Springer Verlag, Sept. 2002.