Manuscript Number: PR-D-11-00647R1

Title: Explicit Length Modelling for Statistical Machine Translation

Article Type: Special Issue: IbPRIA'2011

Keywords: Length modelling; log-linear models; phrase-based models; statistical machine translation.

Corresponding Author: Dr. Jorge Civera,

Corresponding Author's Institution:

First Author: Joan Albert Silvestre-Cerdà

Order of Authors: Joan Albert Silvestre-Cerdà; Jesús Andrés-Ferrer; Jorge Civera

Abstract: Explicit length modelling has been previously explored in statistical pattern recognition with successful results. In this paper, two length models along with two parameter estimation methods and two alternative parametrisation for statistical machine translation (SMT) are presented. More precisely, we incorporate explicit bilingual length modelling in a state-of-the-art log-linear SMT system as an additional feature function in order to prove the contribution of length information. Finally, a systematic evaluation on reference SMT tasks considering different language pairs prove the benefits of explicit length modelling.

Highlights

- Development of novel phrase-length models in statistical machine translation (SMT).

- Proposal of parameter estimation methods and parametrisations for these models.

- Analysis and discussion of the performance of phrase-length models.

- Comparison between estimation methods and parametrisations.

- Automatic evaluation, in terms of BLEU score, on a reference task in SMT.

- Experimental results proved the benefits of length modelling in SMT.

Explicit Length Modelling for Statistical Machine Translation

Submission PR-D-11-00647

Cover Letter

December 20, 2011

Reviewer 1

The paper presents four new approaches to model phrase length into a SMT framework, achieving limited (and not statistically significant) improvement. The paper is well organized and can be easily readable also for not-MT people.

1. *Unfortunately, the presented idea is not novel, as authors sincerely claim. They simply apply the idea to a different underlying model.*

   Novel contributions of the current work are now clearly stated in the Introduction.

2. *The evaluation section is not very strong. It would be appreciated, results on different tasks, in which the difference of source and target sentence lengths is higher (German, Finnish, or other more agglutinative languages)*

   Experimental evaluation has been significantly extended reporting results on German-English and English-German (Europarl), as well as Chinese-English (BTEC) tasks.

3. *Which is the size of the specific model? Is it comparable with that of the translation model? For the Europarl domain, this is not an issue, but it could for much larger tasks. Please, say something about that. Is there any increment of the computational cost of the training or translation phases (in time or memory) dur to the new models?*

   A paragraph discussing the spatial complexity of standard and specific models, and the temporal complexity of the training phase have been now included.

4. *You do not report what kind of significance test you performed. This could be useful for the reader.*

   The significance test was cited: “P. Koehn. Statistical significance tests for machine translation evaluation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004, pp. 388-395.”

   In addition, we have incorporated a pairwise significance test described in “M. Bisani, H. Ney, Bootstrap estimates for confidence intervals in ASR performance evaluation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, pp. 409–412."

5. *It is not explicit which test set you experiment on (I suppose test2006)*

   Training, development and test sets are now clearly defined in the description of the experimental setting.

6. *It is not clear whether your results refers to case-sensitive output, or not. From your figures it seems you compare case-insensitive outputs; if so, the baseline performance are a bit lower the actual state-of-the-art systems (built on Moses); please, refer to the latest performance reported in the Euro-Matrix (maintained by U.Edinburgh, http://matrix.statmt.org/matrix).*

   Case-insensitive outputs were reported, now it is clearly stated. Our baseline results are comparable to those reported by the University of Edinburgh system (Moses) on the WMT07 shared task (Please refer to P. Koehn and J. Schroeder. Experiments in Domain Adaptation for Statistical Machine Translation, in: Proceedings of the ACL 2007 Second Workshop on Statistical Machine Translation, pp. 224-227). Minor differences may be due to the incorporation of the News Commentary corpora in the training and development sets, different Moses versions or different language model training parameters.

7. *In figure 1, the different y-axis scale is a bit annoying because the reader can not directly compare the two plots (left and right)*

   Plots for the same language pairs are now in the same scale to ease comparison.

8. *In the bib entry, publication years are mostly not reported; there is at least one wrong conference acronym (IWSL - IWSLT); I would suggest to use the extended version of the conference names.*

   Publication years were omitted due to a bug in the bibliography stylesheet. Extended version of conference names are now available.

9. *Minor corrections:*

   - *pg. 8: estiation - estimation*
   - *pg. 14:please force the segmentation of the term "parametrization"*

   Minor corrections fixed.

Reviewer 3

1. *First of all, hasn't this work already been published here:*

   http://www.springerlink.com/content/y1w58633845rj07k/?

   If there is no significant difference in technical contributions between the two versions (for e.g., same methods, experiments, etc.), perhaps this submission should be withdrawn.

   The current version of the manuscript is a significantly extended version of that published in the Proceedings of the Iberian Conference on Pattern Recognition and Image

Analysis (IbPRIA) 2011. Indeed, this manuscript was submitted to the "Special Issue: IbPRIA'2011". The main extensions of the current version over the conference version of the manuscript are:

- Poisson parametrisation of the standard and specific models.
- Additional experiments on English-German, German-English and Chinese-English.
- Detailed analysis of the contribution of phrase length models.

2. *The authors propose to explicitly model phrase-length information for SMT and study its effect on MT quality (BLEU scores). They propose several different methods to model the phrase-length information. The models are interesting and they do extensive evaluations comparing the different variants on a Spanish-English dataset.*

   It is interesting to note that adding these features to MOSES produced some BLEU improvements (though not statistically significant). But perhaps a more interesting study would be to test this phenomenon on some other (unrelated) language pair (e.g., Arabic/English or Chinese/English) where the effects (and maybe even the BLEU gains) could be prominent.

   As mentioned above, additional experiments on Chinese-English have been carried out to endeavour the effects of phrase length modelling.

3. *you should do a spell-check, I noticed several mistakes, including grammatical ones (pg.3: "However, any of the previous..." =¿ "However, none of the previous...", "modelisation" =)*

   The manuscript has been spell-checked and carefully reviewed to fix grammatical mistakes.

4. *you used the terms "source" and "target" inconsistently in some places (e.g., on pg 5: when referring to l, m). This should be corrected*

   This has been corrected now.

5. *"year" information missing in References, make sure the references are complete.*

   Publication years were omitted due to a bug in the bibliography stylesheet.

6. *if you already have experiment results on German/English, I would suggest adding the tables/figures to the paper. Did you do any analysis comparing the effect of length modeling with language-relatedness?*

   Additional experiments on English-German, German-English and Chinese-English have been included in the current version of the manuscript to study the contribution of phrase length modelling as a function of language-relatedness.

7. *it was interesting that the Poisson model has fewer features (since you condition on length instead of actual phrases), but still performs slightly better. Is this due to the sparsity issue?*

   Although the Poisson parametrisation has fewer parameters, it introduces the same number of features into the log-linear model, one for the direct and one for the inverse length models. We think this confusion might have been introduced by the notation employed in Section 4 for the Poisson parameters We have changed the notation in this version to avoid this confusion. It is also worth noting that the two proposed models (standard and specific) are parametrised in two ways: using a contingency table or assuming a Poisson distribution. Regarding parameter sparseness and its relation with respect to system performance, we have extended the experimental section with a discussion on this matter.

8. *what was the additional feature count for the different length models? how was the training time affected when adding the length features?*

   A paragraph discussing the spatial complexity of standard and specific models, and the temporal complexity of the training phase have been now included.

9. *there are several other optimization methods you could try instead of MERT (for e.g., MIRA can scale to many more features)*

   Due to time limitations, we have preferred to explore additional language pairs and perform a detailed analysis before exploring alternative optimization methods. This is an interesting idea that we leave out for future work, since phrase length models only contribute with 2 additional features to the conventional Moses features.

Reviewer 5

The manuscript is about explicit length modeling for statistical machine translation. The authors propose and investigate two length models and integrate them as additional feature functions into a state-of-the-art SMT system (Moses). All experiments use BLEU score as evaluation metric.

I have two major concerns with the current version of the manuscript (where the second one weighs higher than the first one):

1. *Some parts in the description are sometimes misleading or imprecise.*

2. *The experimental results do not support the authors' points nor do they give more inside wrt the effect of the new length models.*

Concerning 1

1. *The authors say in the Introduction that "Length modelling is a well-known problem in pattern recognition which is often disregarded." This is a misleading statement and should be rephrased. "Disregarded" in this context means that length modeling is actually ignored. However, as the authors point out later, the difference is between implicit length modeling vs explicit length modeling, and most state-of-the-art systems use implicit length modeling.*

   We agree with reviewer's comments. We meant “Explicit length modelling is a well-known ...”. The introduction has been rewritten accordingly, and a discussion on implicit length model in phrase-based translation system has been included.

2. *(In language modeling, the sentence-end-symbol is used to achieve some crude length model, in machine translation, the number_of_words or number_of_phrases feature function is used to implicitly model length.) The description of the baseline features used leaves it unclear whether an implicit length model is applied or not. As far as I know, Moses uses implicit length modeling as part of its baseline features, and it should be pointed out by the authors whether they use this or not.*

Experimental results were obtained using the conventional features provided with Moses. This includes word and phrase penalty features. We have clearly stated this fact in the description of the baseline system. While these features may be understood to be modelling "similar" length information, the proposed length models and the problem tackled by these standard features are essentially different. This has been made explicit in the current version of the manuscript.

Concerning 2

1. *My major concern is with the experimental section. Part of the problem is that the evaluation is entirely done in terms of BLEU scores, and the way this is done does not give much insight into the power of the new length models.*

   This paper is about explicit phrase length modelling of the bilingual segmentation process induced by a phrase-based model. When this process is properly analysed, it yields conditional phrase length probabilities. These phrase length probabilities better model the segmentation process of bilingual sentences into phrases, and, hence, any improvement at predicting translation sentence lengths will be a side effect and not a direct consequence of phrase length modelling. TER results are now reported together with pairwise significance tests. As expected, the prediction performance of translation sentence lengths is in essence the same. However, n-gram overlapping precision of translations is sometimes improved because of a wiser selection of translated phrases produced by the explicit conditional phrase length models.

2. *An important part of the BLEU metric is its brevity penalty which severely penalizes translations that are too short. However, the authors only present final BLEU score results; the BLEU brevity penalties are not given, and this is at least something I would expect if explicit length modeling is the main subject of this investigation.*

   We have added a discussion on BLEU brevity penalties in "Experimental Results". However, brevity penalties turned out to be very similar between the baseline system and the systems incorporating phrase length models.

3. *With your new length models, I think you should evaluate whether you do a much better job on predicting the actual length of the translation compared to the reference translation than what the baseline system is able to do. The plots in the result section do not show this, and it is important to clearly analyze this effect.*

   We have added TER and pairwise significance results to clearly gauge the contribution of phrase length models.

4. *The explicit length models are introduced as feature functions whose weights are then trained via MERT. Although this is okay, readjusting new MERT weights also introduces quite a lot of variability/noise which makes it hard to analyze the outcome. Therefore, it is even more important that the predictive power of the new length models is analyzed independently of the BLEU metric. The BLEU score results must be presented with confidence intervals. On page 12 of the manuscript, the authors say that "... when comparing them to the baseline. Although these differences are not statistically significant, ..." "The most significant improvement over the baseline was 0.4 BLEU ..." The second statement seems to be in contradiction to the first. The authors should specify:*

   - *the significance test used for the experiments*

- *provide the confidence intervals*

Otherwise this is confusing.

Statements about system performance have been rewritten using a precise language to avoid confusion about the statistically significance of the results. In the current version of the manuscript, we have used significance tests described in:

- M. Bisani and H. Ney. Bootstrap estimates for confidence intervals in ASR performance evaluation, in: Procceedings of the International Conference in Audio, Speech and Signal Processing, 2004, pp. 409-412.

and

- P. Koehn. Statistical significance tests for machine translation evaluation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004, pp. 388-395.

Confidence intervals are now reported. Furthermore, a pairwise significance test between the baseline systems and ours based on (Bisani and Ney, 2004) is also presented.

Minor comments

Most of the references are incomplete. Almost all of them lack the year of publication. The authors should double-check that they conform to the journal's reference and citation guidelines.

page 2, last paragraph "Next section describes ..." - "The next section describes ..."

page 4, 1st paragraph "However, state-of-the-art ..." - Remove "However"

Eq(9) either use m_t on the LHS or m on the RHS

page 8, last paragraph typo "estiation" - "estimation"

page 9, 1st paragraph "This approach is referred as to ..." - "This approach is referred to as ..."

page 9, 2nd paragraph "This approach only considers ..." - "This approach considers only ..."

page 10, last paragraph "... since the rest of features ..." - "since the rest of the features ..."

page 14 "... similar or sligh better ..." - "... similar or slightly better ..."

page 2, last paragraph "Next section describes ..." - "The next section describes ..."

Detailed comments are appreciated and have taken into account when rewriting the current version.

Explicit Length Modelling for Statistical Machine Translation

Joan Albert Silvestre-Cerdà, Jesús Andrés-Ferrer and Jorge Civera

Universitat Politècnica de València
Departament de Sistemes Informàtics i Computació
Camí de Vera s/n
46022 València (Spain)

Abstract

Explicit length modelling has been previously explored in statistical pattern recognition with successful results. In this paper, two length models along with two parameter estimation methods and two alternative parametrisation for statistical machine translation (SMT) are presented. More precisely, we incorporate explicit bilingual length modelling in a state-of-the-art log-linear SMT system as an additional feature function in order to prove the contribution of length information. Finally, a systematic evaluation on reference SMT tasks considering different language pairs prove the benefits of explicit length modelling.

Keywords:
Length modelling, log-linear models, phrase-based models, statistical machine translation.

1. Introduction

Explicit length modelling is a well-known problem in pattern recognition which is often disregarded. However, it has provided positive results in applications such as author recognition [25], handwritten text and speech recognition [29], and text classification [10], whenever it is taken into consideration.

Length modelling may be considered under two points of view. On the one hand, the so-called implicit modelling in which the information about the length of the sequence is indirectly captured by the model structure. This is often the case of handwritten text and speech recognition [11], language modelling [5] and machine translation [18], which often include additional states to convey length information. On the other hand, we may perform an explicit modelling by incorporating a probability distribution in the model to represent length variability in our data sample [23]. Explicit modelling can be found in language modelling [12, 19], and bilingual sentence alignment and segmentation [3, 9], among others.

This work focuses on explicit length modelling for statistical machine translation (SMT). The aim of SMT is to provide automatic translations between languages, based on statistical models inferred from translation examples. State-of-the-art translation systems grounded on phrase-based models implicitly model sentence length information through features, such as word and phrase penalty, that controls the number of words and phrases in the resulting translation. As discussed in more detail later, the word penalty compensates for the bias towards short sentences [4] or prevents the generation of spurious words [17], while the phrase penalty avoids the bias towards long phrases. However, in this work, we address the problem of explicit conditional length modelling at the phrase level.

State-of-the-art phrase-based systems are basically based on a large bilingual phrase dictionary, known as phrase table. Phrase tables do not model conditional phrase length correlation between corresponding phrase translations, that is, the probability of translating a source phrase made up of l words by a target phrase of m words. However, conditional phrase length models seamlessly emerge in the generative process of a bilingual phrase-based segmentation [1].

The main contribution of the current work is a systematic and extensive evaluation of explicit conditional phrase length modelling in a state-of-the-art phrase-based SMT system. To this purpose, two conditional phrase length models are proposed along with two alternative parametrisations and two different parameter estimation methods. Furthermore, strong experimental results are reported on language pairs with different degree of relatedness.

The rest of the paper is structured as follows. The next section describes related work in SMT regarding explicit length modelling. Section 3 introduces the log-linear framework in the context of SMT and Section 4 explains the proposed conditional phrase length models. Experimental results are reported in Section 5. Finally, conclusions and future work are discussed in Section 6.

2. Related work

Explicit length modelling in SMT has received little attention since Brown's seminal paper [4] until recently. Nowadays state-of-the-art SMT systems are grounded on the paradigm of phrase-based translation [18], in which sentences are translated as segments of consecutive words. Thereby, most recent work related to explicit length modelling has been performed at the phrase level with a notable exception [27]. Explicit phrase length modelling was initially presented in [26] where the difference ratio between source and target phrase length is employed to phrase extraction and scoring with promising results. Zhao and Vogel [28] discussed the estimation of a phrase length model from a word fertility model [4], using this model as an additional score in their SMT system. In [8], a word-to-phrase model is proposed which includes a word-to-phrase length model. Finally, [1] describes the derivation and estimation of a phrase-to-phrase model including a model for the source and target phrase lengths.

However, none of the previous works report results on how explicit phrase length modelling contributes to the performance of a state-of-the-art phrase-based SMT system. Furthermore, phrase-length models proposed so far depend on their underlying model or phrase extraction algorithm, which differ from those employed in state-of-the-art SMT systems. The current work is inspired on the explicit phrase length model proposed in [1], but applied to a state-of-the-art phrase-based SMT system [17] and assessed on diverse language pairs in order to systematically evaluate the contribution of explicit phrase length modelling in SMT.

3. Log-linear modelling

In SMT, we formulate the problem of translating a sentence as the search of the most probable target sentence $\hat{y}$ given the source sentence x

$$\hat{y} = \operatorname*{argmax}_{y} Pr(y \mid x) \,. \tag{1}$$

State-of-the-art SMT systems are based on log-linear models that combine a set of feature functions to directly model this posterior probability

$$Pr(y \mid x) = \tfrac{1}{Z(x)} \exp\left(\textstyle\sum_i \lambda_i \, f_i(x, y) \right), \tag{2}$$

being λ_i, the weight for the i-th feature function $f_i(x, y)$ and $Z(x)$, a normalisation term so that the posterior probability sums up to 1. Feature weights are usually optimised according to minimum error rate training (MERT) on a development set [20].

Conventional feature functions in phrase-based SMT systems range from those depending on word-based and phrase-based translation models [18], over that directly derived from an n-gram language model [5], to those inspired on word and phrase reordering models, and word and phrase penalties. In fact, n-gram language models and word and phrase penalties capture to some extend length information at the sentence level.

In general, n-gram language models incorporate the special end-of-sentence symbol that implicitly models sentence length information, even though it is not able to incorporate long-term constraints. This limitation produces that ill-formed sentences receive an exponentially growing probability mass depending on their length [4]. Hence, the probability of well-formed sentences exponentially decays with their length. In order to alleviate this bias towards short sentences, the word penalty feature introduces a constant bonus for each new word added to the translation. However, in phrase-based SMT systems, the word penalty avoids the generation of spurious words [17]. In any case, the word penalty feature aims at implicitly modelling sentence length information, not phrase length information, as the models proposed in this work do.

On the other hand, phrase tables suffer from a bias towards long phrases due to a similar modelling deficiency. Indeed, the phrase penalty adds a constant bonus for each additional phrase incorporated into the translation. In fact, as shown in Section 4, the phrase penalty is complementary to the proposed conditional phrase length models.

In this work, in addition to the conventional features mentioned above, additional features derived from conditional phrase length models [1] are introduced. These additional features are presented in the next section.

4. Explicit length modelling

In the phrase-based approach to SMT, the translation model considers that the source sentence x is generated by segments of consecutive words defined over the target sentence y. As in [1], in order to define these segments we introduce two hidden segmentation variables

$$p(x \mid y) = \textstyle\sum_T \sum_{l_1^T} \sum_{m_1^T} p(x, l_1^T, m_1^T \mid y)\,, \tag{3}$$

being T the number of phrases into which both sentences are to be segmented, and being l_1^T and m_1^T the source and target segmentation variables, respectively. Thus, we can factor Eq. (3) as follows

$$p(x, l_1^T, m_1^T \mid y) = p(m_1^T \mid y)\, p(l_1^T \mid m_1^T, y)\, p(x \mid l_1^T, m_1^T, y)\,, \tag{4}$$

where $p(m_1^T \mid y)$ and $p(l_1^T \mid m_1^T, y)$ are phrase length models, whilst $p(x \mid l_1^T, m_1^T, y)$ constitutes the phrase-based translation model. We can independently factorise terms in Eq. (4) from left to right,

$$p(m_1^T \mid y) = \textstyle\prod_t \; p(m_t \mid m_1^{t-1}, y) \,, \tag{5}$$

$$p(l_1^T \mid m_1^T, y) = \textstyle\prod_t \; p(l_t \mid l_1^{t-1}, m_1^T, y) \,, \tag{6}$$

$$p(x \mid l_1^T, m_1^T, y) = \textstyle\prod_t \; p(x(t) \mid x(1), \ldots, x(t-1), l_1^T, m_1^T, y) \,, \tag{7}$$

where t ranges over the possible segmentation positions of the target sentence, l_t and m_t are the length of the t-th source and target phrase, respectively, and $x(t)$ is the t-th source phrase.

In state-of-the-art systems, the model in Eq. (5) is approximated by the phrase penalty, which is intended to control the number of phrases involved in the construction of a translation, as previously discussed. Eq. (7) is simplified by conditioning only on the t-th target phrase to obtain the conventional phrase table, which is used as another feature,

$$p(x(t) \mid x(1), \ldots, x(t-1), l_1^T, m_1^T, y) := p(x(t) \mid y(t)) \,, \tag{8}$$

with parameter set, $\theta = \{p(u \mid v)\}$, for each source, u, and target, v, phrase. Finally, Eq. (6) is used to derive conditional phrase length models that become new feature functions of our log-linear model, and the corresponding phrase-based SMT system.

Next sections present two conditional phrase length models, namely, *standard* and *specific*, as a result of different assumptions on Eq. (6). In addition, two alternative parametrisations will be considered for each of these models, referred to as *parametric* and *non-parametric*.

4.1. Standard length models

The standard length model is derived from Eq. (6) by taking the assumption that the source length l_t only depends on the corresponding target phrase length m_t as follows

$$p(l_t \mid l_1^{t-1}, m_1^T, y) \approx p(l_t \mid m_t)\,. \tag{9}$$

The parametric model further assumes that the rightmost probability in Eq. (9) follows a Poisson distribution

$$p_{\gamma_{m_t}}(l_t \mid m_t) \propto \gamma_{m_t}^{l_t} \exp(-\gamma_{m_t}) \tag{10}$$

where the mass probability function is renormalised to sum 1, if a maximum phrase length is specified. Therefore, the parameter set is $\gamma = \{\gamma_m\}$ for each target phrase length m.

On the contrary, in the non-parametric model, each $p(l_t \mid m_t)$ term in Eq. (9) plays the role of a parameter, and, consequently, the parameter set is given by $\gamma = \{p(l \mid m)\}$ for each source, l, and target, m, lengths. This model is more sparse than the parametric model and it is smoothed to alleviate this problem as follows

$$\tilde{p}(l \mid m) := (1 - \varepsilon) \cdot p(l \mid m) + \varepsilon \cdot \frac{1}{M}\,, \tag{11}$$

where M stands for the maximum phrase length.

For a given maximum phrase length M, say 7, the parametric standard model requires M parameters, i.e. $\{\gamma_1, \gamma_2, \ldots, \gamma_M\}$, while the non-parametric model needs M^2 parameters, i. e. $\{p(1 \mid 1), p(2 \mid 1), \ldots,\ p(M \mid 1), p(1 \mid 2), \ldots, p(M \mid M)\}$.

4.2. Specific length models

In the specific model, we take a more *specific* assumption for Eq. (6) than that of Eq. (9) by considering the dependency on the actual phrase $y(t)$, instead of its length,

$$p(l_t \mid l_1^{t-1}, m_1^T, y) \approx p(l_t \mid y(t)) \,, \tag{12}$$

being $p(l_t \mid y(t))$, a source phrase-length model conditioned on the t-th target phrase. This latter probability $p(l_t \mid y(t))$ in Eq. (12) can be regarded as a parameter itself, yielding the non-parametric model. In this case, the parameter set is defined by $\gamma = \{p(l \mid v)\}$, for any target phrase v. In practice, v is any target phrase observed in the training set.

Similarly to the standard length model, the parametric model will assume that the probability in Eq. (12) follows a Poisson distribution

$$p_{\gamma_{y(t)}}(l_t \mid y(t)) \propto \gamma_{y(t)}^{l_t} \exp(-\gamma_{y(t)}) \,, \tag{13}$$

where the probability mass function is renormalised so that it sums up to 1 if a maximum phrase length is specified. Hence, the parameter set is $\gamma = \{\gamma_v\}$ for each target phrase v. It is worth noting the difference between Eq. (10) and Eq. (13). In the former, a Poisson distribution is considered for each *target phrase length*, while in the latter a Poisson distribution is assumed for each *target phrase*.

Specific length models, both parametric and non-parametric, are considerably more sparse than those of the standard model. In order to alleviate overfitting problems, the specific parameters are smoothed with the standard parameters as follows

$$\tilde{p}(l \mid v) := (1-\varepsilon) \cdot p(l \mid v) + \varepsilon \cdot \tilde{p}(l \mid |v|) \,, \tag{14}$$

denoting by $| \cdot |$ the length of the corresponding phrase. The interpolation parameter ε is adjusted on a validation set in order to maximise BLEU.

Given a maximum phrase length M, and the set of all unique target phrases that have been extracted from the training data $\mathcal{V} = \{v_1, \ldots, v_n\}$, the parametric specific model requires one Poisson parameter for each phrase, i.e., $\{\gamma_{v_1}, \gamma_{v_2}, \ldots, \gamma_{v_n}\}$. On the other hand, the non-parametric specific model requires M parameters for each target phrase, $\{p(1 \mid v_1), \ldots, p(M \mid v_1), p(1 \mid v_2), \ldots, p(M \mid v_2), \ldots, p(1 \mid v_n), \ldots, p(M \mid v_n)\}$.

As usual in statistical modelling, there is a trade-off between the complexity of the model, basically the number of parameters to be learnt, and the number of data samples available. The more parameters to be learnt, the more data samples are needed to properly train the corresponding model. If we compare the specific model with the standard model, the former possesses a significantly greater number of parameters with respect to the latter, and hence, smoothing becomes critical. For instance, most of the target phrases occur only once, and then, the ratio of samples to parameter for the non-parametric specific model is less than 1, which requires a strong smoothing. However, the standard model possesses fewer parameters, and hence, it does not suffer from severe overfitting problems, i.e., the ratio will always be much larger than 1.

Another problem that arises in the context of data sparsity is the excessive generalisation of parametric models. Our Poisson model makes a stronger assumption regarding the probability length distribution than that of the non-parametric model. However, it could be the case that a target phrase occurs only once with its corresponding source phrase translation. If we assume that source and target phrase share the same length, say $m = l$, then the maximum likelihood estimation of a non-parametric model gives $(1 - \varepsilon)$

probability mass to the single observed hypothesis. In contrast, the parametric model, which follows a Poisson distribution, would smoothly decrease this probability according to $(1-\varepsilon)\, l^m \exp(-l)$ for source lengths m different from l. Depending on the language pairs involved, either the parametric or the non-parametric model would be a better hypothesis. These trade-offs are experimentally analysed in Section 5.

4.3. Estimation of phrase-length models

The parameters of the models introduced in the previous section could be estimated by maximum likelihood criterion using the EM algorithm [7]. As shown in [1], the phrase-based translation model is estimated as

$$p(u \mid v) = \frac{N(u,v)}{\sum_{u'} N(u',v)}\,, \tag{15}$$

being $N(u,v)$, the expected counts for the bilingual phrase (u,v). The estimation of $p(l \mid m)$ is computed as

$$p(l \mid m) = \frac{N(l,m)}{\sum_{l'} N(l',m)}\,, \tag{16}$$

where

$$N(l,m) = \sum_{u,v} \delta(l,|u|)\, \delta(m,|v|)\, N(u,v)\,, \tag{17}$$

being δ, the Kronecker delta. Conversely, for the Poisson model, the estimation of the parameter γ_m is similar to that of Eq. (16), and is given by

$$\gamma_m = \frac{\sum_l l \cdot N(l,m)}{\sum_l N(l,m)}\,. \tag{18}$$

The parameters for the specific models, $p(l \,|\, v)$, are estimated analogously.

Although expected counts $N(u,v)$ were exactly computed in [1], this was done at the expense of limitating the expressiveness of the model to only consider monotonic bilingual segmentations, otherwise the computational cost of the expected counts would require exponential time [13].

However, the parameter estimation of conventional log-linear phrase-based system approximate expected counts $N(u,v)$ with approximated counts $N^*(u,v)$, derived from a heuristic phrase-extraction algorithm [16]. Similarly, our first approach is also to approximate $N(l,m)$ in Eq. (16) as follows

$$N^*(l,m) = \sum_{u,v} \delta(l, |u|)\, \delta(m, |v|)\, N^*(u,v)\,. \tag{19}$$

This approach is referred to as *phrase-extract* estimation, or simply *extract* estimation. The estimation of the proposed models with the phrase-extract estimation can be implemented adding a constant time to the the phrase-extract algorithm for each phrase extracted. But for simplicity reasons, it has been implemented as an additional pass over the extracted phrases.

A second approach to the estimation of phrase-length parameters is based on the idea of a Viterbi approximation to Eq. (3). This approach considers only the source and target segmentation that maximises Eq. (3)

$$\hat{l_1^T}, \hat{m_1^T} = \operatorname*{argmax}_{l_1^T, m_1^T} \left\{ Pr(x, l_1^T, m_1^T \mid y) \right\}. \tag{20}$$

So, the hidden segmentation variables are uncovered and the counts in Eq. (15) are not expected fractional counts, but integer counts approximated by the Viterbi segmentation.

The search denoted by Eq. (20) is performed using a conventional log-linear phrase-based system which is based on a A^* search algorithm. It must be noted that the source and target sentences are available during the training phase, so this search becomes a guided search in which the target sentence is known.

In terms of computational complexity, the Viterbi-based estimation introduces a large additional computational cost to the standard training phase. First, the Viterbi segmentation of each training sample needs to be computed, which is an NP-hard problem approximated by a A^* search algorithm. Then, counts are collected from Viterbi segmentations in order to estimate phrase length parameters.

Regarding the estimation method, it is not clear whether Viterbi or phrase-extract counts better approximate actual expected counts, or even more important, which counts yield a better estimation. On the one hand, Viterbi counts are more sparse than extract counts since they are obtained only from a single segmentation, that is, the most probable segmentation. On the other hand, phrase-extract counts are extracted from several "heuristic segmentations". For example, let $y = (y_1, y_2)$ a target sentence, and $x = (x_1, x_2)$ its source counterpart. Despite its simplicity, this example allows us to illustrate the sparseness of the different estimation methods. The heuristic extraction is based on word alignments between source and target words. For this example, we further assume that x_1 is aligned with y_1, and x_2 with y_2. Provided this example, the Viterbi approximation would probably consider the full sentence as a phrase, (x_1x_2, y_1y_2), counting it once. In contrast, the phrase-extraction heuristic, would produce 3 phrases: $(x_1x_2, y_1y_2), (x_1, y_1)$, and (x_2, y_2), and each of them would be counted once.

5. Experimental results

In this section, an systematic evaluation is performed to elucidate the benefits of explicit phrase length modelling in phrase-based SMT. To this purpose, three language pairs were involved in the experiments: English-Spanish (En-Es), Spanish-English (Es-En), English-German (En-De), German-English (De-En) and Chinese-English (Zh-En). Experiments on Spanish and German were carried out using the Europarl-v3 parallel corpora [15], which is a reference task in the SMT field, while the Chinese-English experiments were performed using the BTEC parallel corpora provided in the evaluation campaign for the IWSLT09 [22]. Basic statistics for both corpora, Europarl and BTEC, are shown in Tables 1 and 2, respectively.

The experimental setup for the Europarl-v3 corpora provides three separate sets for the purpose of evaluation campaigns: training, development, and test. The training set consists of two datasets. The first of them is a monolingual dataset, that is devoted to train language models, while the second dataset is a parallel corpus to train translation models. Also, two development sets, known as *dev2006* and *devtest2006*, are provided. On the one hand, the dataset *dev2006* is used to perform Minimum Error Rate Training (MERT) of the weights involved in the log-linear SMT model [20]. On the other hand, the development set *devtest2006* is dedicated in this work to adjust the interpolation smoothing parameter ε. Finally, final performance results are reported on the *test2007* set.

Similarly, the BTEC corpora was also divided in three sets. The training set consists in an unique parallel dataset which is used to train both language and translation models. The development set *devset6* was divided into two

Table 1: Basic statistics for Europarl-v3.

Training sets	Monolingual			Bilingual			
Language pairs	En	Es	De	En	Es	En	De
Bilingual sentences	1.4M			965K		995K	
Vocabulary size	115.7K	167.6K	327.2K	81.8K	113.0K	74.6K	226.9K
Running words	38.3M	40.3M	36.7M	20.3M	20.9M	21.5M	20.4M

	Development						Test		
	dev2006			devtest2006			test2007		
Language	En	Es	De	En	Es	De	En	Es	De
Sentences	2K			2K			2K		
Vocab. size	6.1K	7.7K	8.8K	6.1K	7.8K	8.7K	6.0K	7.8K	8.8K
Run. words	58.8K	60.5K	55.1K	58.1K	60.2K	54.2K	59.2K	61.3K	55.6K
Perplexity	74	75	119	73	76	118	71	76	121

datasets. The first dataset, referred to as *dev-mert*, is devoted to perform Minimum Error Rate Training, and the second dataset, *dev-smooth*, is used to optimize the interpolation smoothing parameter, ε. Finally, *devset7* is the test set on which final results are reported.

The performance of phrase length models was assessed on the freely available Moses toolkit [17]. Basically, we compare the performance of the Moses baseline system (including word and phrase penalties) to that of an augmented version of the Moses system incorporating the phrase length models as additional features. More precisely, the phrase-length augmented system includes two additional features, a source-conditioned and a target-conditioned phrase length models.

Table 2: Basic statistics for BTEC (IWSLT09).

	Training		Development (devset6)				Test (devset7)	
			dev-mert		dev-smooth			
Language pairs	Zh	En	Zh	En	Zh	En	Zh	En
Bilingual sentences	20K		389		100		511	
Number of references	1	1	1	6	1	6	1	10
Vocabulary size	8.4K	7.1K	726	724	307	321	888	872
Running words	171.5K	188.9K	2.5K	3K	682	871	3.3K	4.2K
Perplexity	-	-	47	26	60	31	51	33

A selection of the most relevant and representative experiments are shown in this section. In order to gauge the translation quality of the different systems, the well-known BLEU [21] and TER [24], were used. The BLEU score is composed by the product of two terms: an accuracy measure of the degree of n-gram overlapping between the system and the reference translation, and a brevity penalty (BP) which exponentially penalises short references. The translation edit rate (TER) is defined as the minimum number of edits, which include the insertion, deletion, and substitution of single words as well as shifts of word sequences, needed to convert a hypothesis translation into one of the reference translations. In all cases, we reported the performance of the system on case-insensitive translations.

First, BLEU scores as a function of maximum phrase length are plotted for each language direction to provide an initial performance analysis of phrase length models (standard vs. specific), parametrisations (Poisson vs. contingency table) and estimation methods (extract vs. Viterbi). In these plots confidence intervals are not reflected for the sake of clarity. Afterwards, we report BLEU and TER for maximum phrase length (maxPL=7) including confidence intervals and pairwise statistical significance tests.

To be more precise, two bootstrapping methods, referred to as standard [14] and pairwise [2], were applied to all experiments in order to verify the statistically significance of our results. The standard bootstrapping method computes absolute confidence intervals for BLEU or TER for each system [14], while the second performs pairwise system comparison [2]. Although the standard bootstrapping method yields trustful confidence intervals, it ignores the current bootstrapping sample complexity and thereby, typically produces large confidence intervals. In other words, it does not consider that there are sentences which are more difficult to translate than others, such as long sentences. In contrast, the pairwise bootstrapping method provides smaller variances, since it takes into account the sample variance by computing the difference with respect to a baseline system. For this reason, we also report the so-called *probability of improvement* (PI) [2] that aims at minimising the variety of bootstrapping sample complexity by simply counting the number of times a system is better than other without taking into account the absolute improvement. PI figures for BLEU and TER evaluations are reported when comparing the performance of our proposed models to that of the conventional Moses baseline system.

Figures 1, 2 and 3 show the evolution of the BLEU score (y-axis) as a function of the maximum phrase length (x-axis) for experiments involving Spanish, German and Chinese, respectively. In the case of Spanish and German experiments, the left-most column of plots presents BLEU trends with English as the source language, while the right-most column does the same with English as the target language. Reading Figures 1 and 2 by rows from top to bottom, first the standard (std) or specific (spc) model is set, depending which the best performing system is, leaving the other two experimental parameter (estimation method and parametrisation) free. The second row sets the estimation method (Viterbi or extract) and the rest of experimental parameters are left free. Finally, the third row leaves the parametrisation constant, Poisson (param) or contingency table (non-param), and explores the other two experimental parameters.

In Figure 1, the experiments regarding Spanish are presented. In both directions, the best performing system involved the standard model with Poisson parametrisation and Viterbi parameter estimation. As shown later, the PI for BLEU proved that the improvement over the baseline is statistically significant, and not only for the best performing system. However, given the standard model, there seem not to be a clearly better estimation method or parametrisation.

Figure 2 shows BLEU trends involving German. Generally speaking, the best results are obtained again with the standard model, but the specific model using the extract estimation model obtains similar performance in the English-German pair. In all cases, the extract estimation methods seems to perform the best on average. These results can be explained in the light of the trade-off between the number of parameters and data samples available, mentioned in Section 4.2. As shown in Table 1, the German language possesses a higher perplexity compared to Spanish, so this fact reduces the parameter to sample ratio. To compensate this effect more samples are required to train the same model (standard or specific). For this reason, the

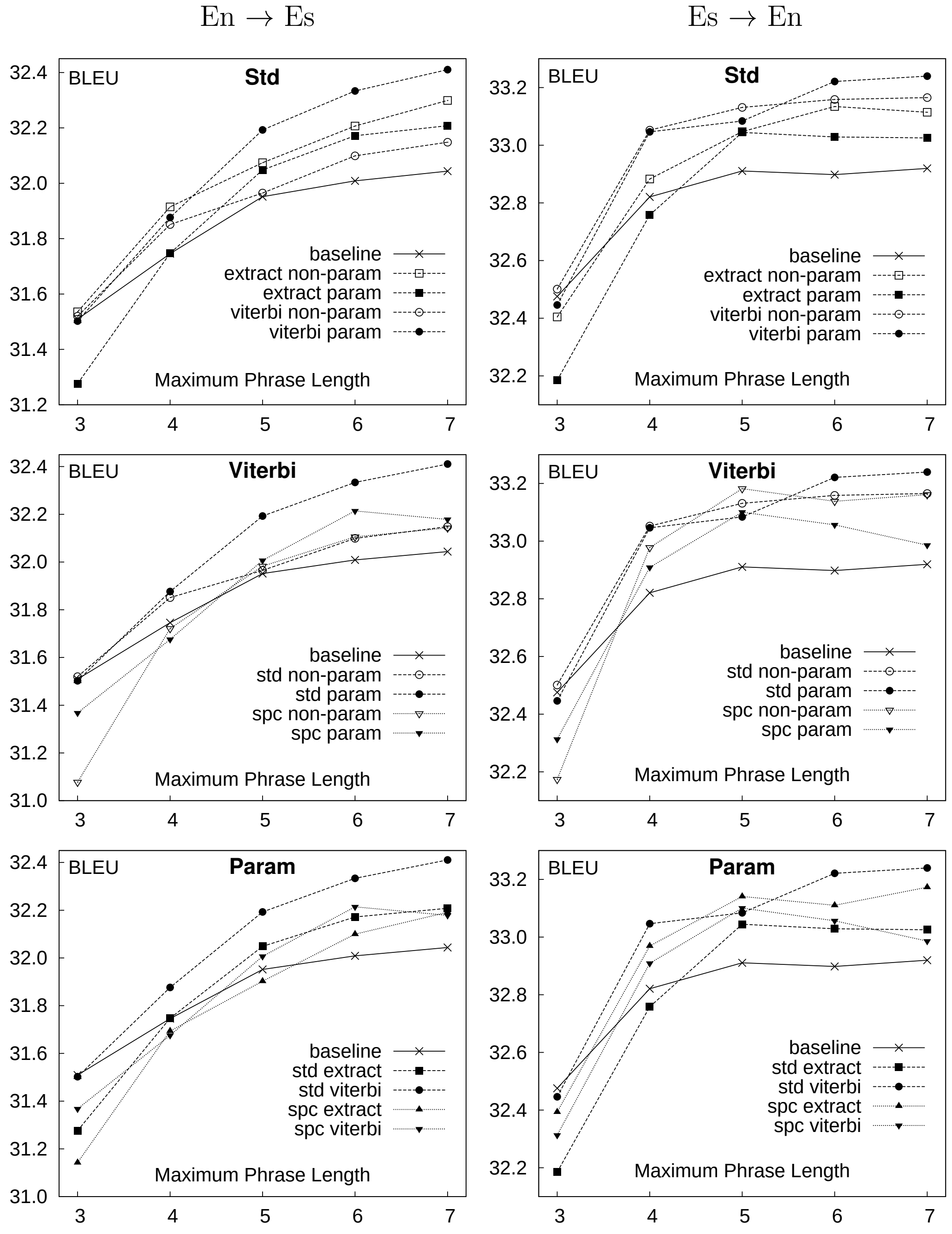

Figure 1: BLEU scores as a function of the maximum phrase length in English-Spanish and Spanish-English.

extract estimation method is preferred over the Viterbi method.

Interestingly enough, the non-parametric parametrisation based on a contingency table outperforms the Poisson parametrisation in all cases. This phenomenon is related to the stronger assumption on the probability length distribution of the Poisson compared to that of the non-parametric approach explained in Section 4.2.

Figure 3 presents BLEU score trends in the Chinese-English BTEC task. The leftmost plot sets the specific model (best performing model for maximum phrase length equal to 7) to analyse the influence of the estimation method and the parametrisation, while the rightmost plot sets the non-parametric approach and compares the performance of the standard and specific models, and estimation methods. As shown, the non-parametric approach supersedes the Poisson parametrisation given the specific model. This phenomenon is the same than that observed in the experimental results with German. However, the specific model outperforms the standard model in all cases, although an estimation method is not clearly preferred over the other.

Tables 3, 4 and 5 show comparative performance results achieved by the baseline system and the different phrase length models proposed in this work for maximum phrase length equal to 7. Furthermore, confidence intervals at 95% computed according to the bootstrapping method proposed in [14] are reported just below the corresponding measure, while PI in percentage for BLEU and TER were calculated according to [2].

For all the experiments, we analyse the behaviour of the BP, and observed that it does not varies significantly. For instance, for the German to English task it is 0.995 ± 0.006 for all models including the baseline. The

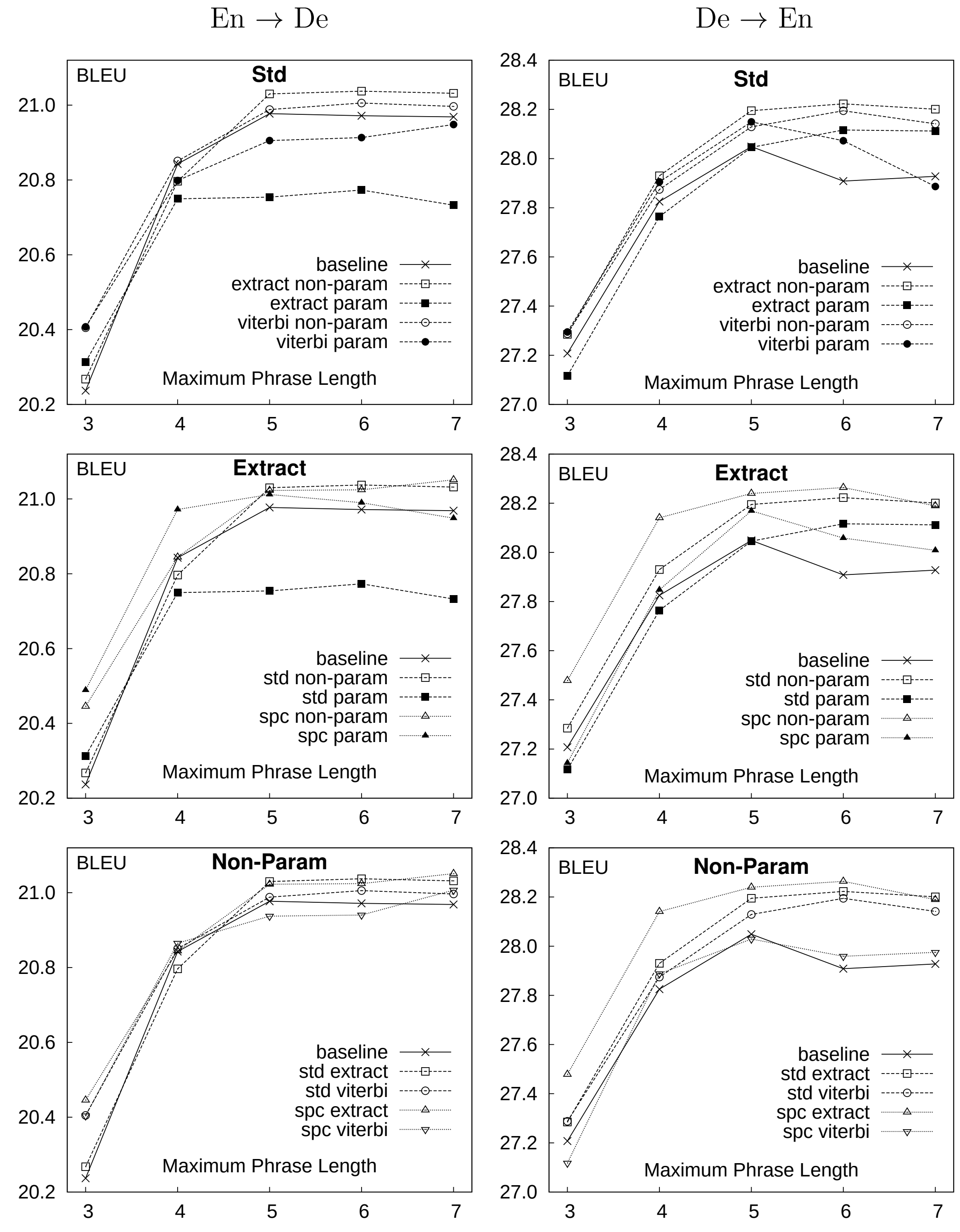

Figure 2: BLEU scores as a function of the maximum phrase length in English-German and German-English.

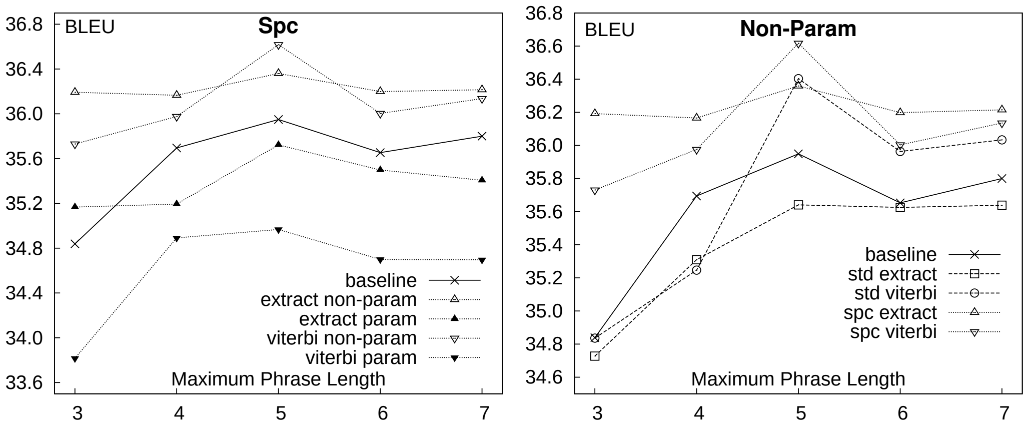

Figure 3: BLEU scores as a function of the maximum phrase length for Chinese-English BTEC task setting the specific model (left) and the non-parametric approach (right), while the other two experimental parameters are left free.

only exception in which the BP could have some effect in BLEU scores is the English-Spanish pair. In this pair, it is observed the greatest variability in BP 0.987 ± 0.009, but the BP of the baseline is similar to that of some of our proposed models. For this reason, we do not report a systematic evaluation in terms of BP values. Note that this result is in accordance with the discussion in Section 4, in which we concluded that the sentence length prediction is not expected to improve as a direct consequence of applying phrase length models.

Table 3 presents the results for the English-Spanish (En-Es) and Spanish-English (Es-En) pairs. As observed, confidence intervals overlap in all cases for TER and BLEU measures. However, PI for BLEU figures reflect that most of the systems proposed supersedes the baseline system in more than

Table 3: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on English-Spanish (En-Es) and Spanish-English (Es-En) pairs.

	En-Es				Es-En			
System	BLEU	BLEU	TER	TER	BLEU	BLEU	TER	TER
	± 0.9	PI	± 1.0	PI	± 1.0	PI	± 1.1	PI
baseline	32.0	-	54.2	-	32.9	-	52.6	-
std extr non-par	32.3	99.4	54.2	62.1	33.1	97.9	52.3	100.0
std extr param	32.2	90.8	54.3	34.3	33.0	78.1	52.6	55.0
std vite non-par	32.2	87.5	53.9	100.0	33.2	97.7	52.4	99.0
std vite param	32.4	100.0	54.2	57.9	33.2	98.7	52.3	99.6
spc extr non-par	32.2	93.9	54.1	76.8	33.2	98.6	52.5	94.6
spc extr param	32.2	91.7	54.3	26.7	33.2	98.4	52.4	96.1
spc vite non-par	32.1	84.9	54.3	15.1	33.2	97.3	52.4	98.9
spc vite param	32.2	92.1	54.1	87.8	33.0	76.4	52.4	98.9

Table 4: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on English-German (En-De) and German-English (De-En).

	En-De				De-En			
System	BLEU	BLEU	TER	TER	BLEU	BLEU	TER	TER
	± 0.8	PI	± 0.9	PI	± 1.0	PI	± 1.0	PI
baseline	21.0	-	65.8	-	27.9	-	58.6	-
std extr non-par	21.0	72.4	65.8	40.0	28.2	99.7	58.4	93.1
std extr param	20.7	4.7	65.8	46.7	28.1	92.9	58.2	100.0
std vite non-par	21.0	67.5	65.5	98.5	28.1	99.1	58.4	96.2
std vite param	20.9	47.7	65.5	98.0	27.9	53.5	58.7	19.2
spc extr non-par	21.1	73.2	65.6	95.5	28.2	99.5	58.3	99.4
spc extr param	21.0	52.2	65.5	99.5	28.0	89.4	58.3	98.8
spc vite non-par	21.0	55.9	65.3	100.0	28.0	74.7	58.5	74.8
spc vite param	20.9	21.8	65.8	61.2	27.9	71.7	58.5	77.8

90% of the bootstrapping rounds. This is not so clear in PI for TER on the English-Spanish pair, but again on the Spanish-English we observe the superiority of the phrase length models proposed.

Table 4 provides experimental results on English-German and German-English pairs in a similar fashion to Table 3. Again, we observed that the confidence intervals for BLEU and TER between the baseline system and the proposed systems overlap. Indeed, PI for BLEU on English-German do not reflect a notable superiority of phrase length systems over the baseline, but PI for TER clearly does for at least four of our models. Nevertheless, the analysis of PI on the German-English for BLEU and TER provides statistically significance evidence of the superiority of some of our proposed models.

Experimental results on the BTEC Chinese-English task are displayed

Table 5: Evaluation results in terms of BLEU, Probability of Improvement (PI) for BLEU, TER and PI for TER on Chinese-English (Zh-En).

System	Zh-En			
	BLEU	BLEU PI	TER	TER PI
baseline	35.8 ± 2.8	-	46.2 ± 2.3	-
std extract non-param	35.6 ± 2.9	42.3	44.0 ± 2.2	100.0
std extract param	35.7 ± 2.8	43.8	46.9 ± 2.4	11.3
std viterbi non-param	36.0 ± 2.9	64.7	45.4 ± 2.3	92.7
std viterbi param	35.1 ± 2.8	9.0	47.9 ± 2.3	0.0
spc extract non-param	36.2 ± 3.0	73.2	44.0 ± 2.2	100.0
spc extract param	35.4 ± 2.8	31.8	44.6 ± 2.2	99.5
spc viterbi non-param	36.1 ± 2.9	66.7	43.6 ± 2.1	100.0
spc viterbi param	34.7 ± 3.0	11.8	43.4 ± 2.0	100.0

on Table 5. As happened in the other language pairs, confidence intervals overlap, though some of the proposed models obtain a higher (not statistically significant) average performance than the baseline system. Nonetheless, four of the proposed models outperform in terms of TER the baseline system in all bootstrapping rounds. However, the same cannot be claimed when analysing the PI for BLEU.

As mentioned above, there are language pairs for which the Poisson distribution (parametric) is a better parametrisation than a contingency table (non-parametric). For instance, the Poisson distribution obtains surprisingly good results in the English-Spanish pair compared with its non-parametric counterpart. However, in the German-English pair, we observed that the non-parametric models performs better. A possible explanation is a mismatch between the underlying probability distribution and the Poisson distribu-

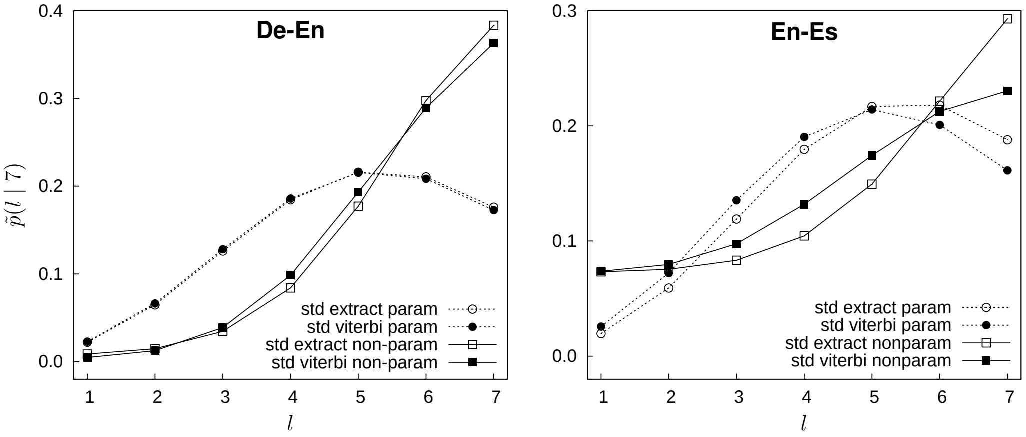

Figure 4: Probabilities learnt with standard model for both parametrisation and estimation algorithms. Vertical axis plots the learnt probability for a target phrase length of 7 as a function of the source phrase length in the horizontal axis.

tion. In order to verify this hypothesis, Figure 4 plots the standard model probabilities learnt with the parametric and non-parametric parametrisation for a fixed target length of 7 on the aforementioned language pairs. In the English-Spanish pair, the non-parametric models seem to approximate the learnt Poisson distribution. However, this is not the case in the German-English pair, in which the non-parametric model learns a completely different probability distribution. Note that we are already comparing smoothed models in order to avoid overfitting on the training data. Furthermore, we have also analysed this behaviour in the case of specific models, in which the disagreement is more accentuated.

In Table 6, positive and negative translation examples were selected to illustrate the behaviour of phrase length models on the Spanish-English task.

Each example shows the source sentence, the reference translation and the translation provided by the baseline system followed by translations generated by a system augmented with phrase length features. Those phrase length systems providing the same translation are referred to as *others* and common suffixes are replaced by "...". In the first example, as a side effect of a better phrase length model, the standard parametric model improves both, the quality of the translation and the sentence length as a byproduct. In Table 3, the standard model approximated by a Poisson distribution obtains the best results when trained using Viterbi counts. Indeed, it is the only system able to balance the implicit length modelling features and the conditional phrase length in this sentence. A similar result is observed in the second example, where both parametric models trained with Viterbi counts produce a better translation. Finally, the last example is a difficult sentence for which no system obtains a good translation. The proposed length models in this case decrease the performance of the system in terms of TER. However, the translation is similar for many of the length models. It is interesting to analyse in this case how the Spanish word "desde" which means "from" is not translated by the baseline system, but some of the length models introduce it, even at the expense of other word. In general, the appropriate length model yields similar or better translations than the baseline system both in terms of TER and BLEU, as shown in Table 3.

6. Conclusions and Future Work

In contrast to the conventional implicit length modelling features present in state-of-the-art SMT systems, we propose two novel explicit conditional

Table 6: Translation examples on the Spanish-English pair. Phrase length systems providing the same translation are referred to as *others* and common suffixes are replaced by "…".

	Length models improve evaluation measures
source	nosotros hemos votado en contra .
reference	we voted against it .
baseline	we voted against .
std vite param	we voted against *it* .
others	we voted against .
source	estos documentos sumamente secretos nos proporcionan una extraña mirada entre bastidores de …
reference	these extremely secret documents give us a rare look behind the scenes of …
baseline	these documents secret extremely strange, give us a hidden behind the scenes of …
std vite param	these *highly secret documents* give us a *strange* look behind the scenes of the policy of …
spc vite param	these *highly secret documents* give us a *strange* look behind the scenes of the policy of …
others	these documents extremely secrets provide us with a strange look behind the scenes of …
	Length models degrade evaluation measures
source	desde el grupo socialista estimamos que el actual funcionamiento de la administracin pública comunitaria es …
reference	the socialist group considers that the current functioning of the community's public administration is …
baseline	we in the socialist group believe that the current functioning of the european public service is …
std vite param	from the group of the party of european socialists, we believe that the current functioning of
spc extr non-par	from the socialist group , we believe that the current functioning of the european public service is …
spc vite non-par	from the socialist group we believe that the current functioning of the european public service is …
others	we in the socialist group believe that the current functioning of the european public service is …

phrase length models: the standard length model and the specific length model. These two models can be seamlessly derived from a generative bilingual segmentation process as shown in Section 4. Although previous work [1] addresses explicit phrase length modelling for a hidden semi-Markov model, no systematic evaluation in a state-of-the-art system has been performed to the best of our knowledge.

The proposed models have been parametrised in two different ways, using a contingency table or assuming a Poisson distribution. In addition, two alternative parameter estimation methods have been also presented: a heuristic algorithm based on the well-known phrase-extract algorithm, and a maximum likelihood estimation method based on the Viterbi segmentation.

These phrase-length models have been integrated in a state-of-the-art log-linear SMT system as additional feature functions, providing in most cases a systematic boost of translation quality on unrelated language pairs. This improvement, albeit not being large, has been proved to be statistically significant for several language pairs: English to/from Spanish, English to/from German and Chinese to English.

From the comparison of phrase-length models and parameter estimation approaches it has been observed that, as theoretically expected, there is a trade-off between model complexity and data sparseness. While for the Spanish pairs, a simple but properly estimated model (standard model) suffices, other languages require a more complex and flexible model (specific model). Regarding the estimation procedures a similar behavior is observed. On the one hand, whenever a simple model is enough to model bilingual phrase length correlations, the Viterbi approach obtains a reliable and accurate estimation making the most out of the model. On the other hand, for complex models, the phrase-extract produce a better estimation since more approximated counts are generated to better estimate the parameters.

In the light of the results, as future work, we plan to perform a full Viterbi-like iterative training algorithm that may improve the quality obtained by the proposed Viterbi-based estimation method, for example using n-best segmentation lists instead of one simple segmentation. Moreover, we would also study as a smoothing technique, the combination of Viterbi extracted counts with those heuristically extracted. Finally, alternative optimization methods to MERT, such as MIRA [6], will also be explored.

References

[1] J. Andrés-Ferrer, A. Juan, A phrase-based hidden semi-markov approach to machine translation, in: Procedings of European Association for Machine Translation, 2009, pp. 168–175.

[2] M. Bisani, H. Ney, Bootstrap estimates for confidence intervals in asr performance evaluation, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004, pp. 409–412.

[3] P.F. Brown, J.C. Lai, R.L. Mercer, Aligning sentences in parallel corpora, in: Proceedings of the 29th annual meeting on Association for Computational Linguistics, 1991, pp. 169–176.

[4] P.F. Brown, et al., The mathematics of statistical machine translation: Parameter estimation, Computational Linguistics 19 (1993) 263–311.

[5] S.F. Chen, J. Goodman, An empirical study of smoothing techniques for language modeling, in: Proceedings of the 34th annual meeting on Association for Computational Linguistics, 1996, pp. 310–318.

[6] D. Chiang, Y. Marton, P. Resnik, Online large-margin training of syntactic and structural translation features, in: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Honolulu, Hawaii, 2008, pp. 224–233.

[7] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. Roy. Statistical Society. Series B 39 (1977) 1–38.

[8] Y. Deng, W. Byrne, HMM word and phrase alignment for statistical machine translation, IEEE Trans. Audio, Speech, and Language Processing 16 (2008) 494–507.

[9] W.A. Gale, K.W. Church, A program for aligning sentences in bilingual corpora, in: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, 1996, pp. 177–184.

[10] A. Giménez, et al., Modelizado de la longitud para la clasificación de textos, in: Actas del I Workshop de Reconocimiento de Formas y Análisis de Imágenes, 2005, pp. 21–28.

[11] S. Günter, H. Bunke, Hmm-based handwritten word recognition: on the optimization of the number of states, training iterations and gaussian components, Pattern Recognition 37 (2004) 2069–2079.

[12] R. Kneser, Statistical language modeling using a variable context length, in: Proceedings of the 4th International Conference on Spoken Language, 1996, pp. 494–497.

[13] K. Knight, Decoding complexity in word-replacement translation models, Computational Linguistics 25 (1999) 607–615.

[14] P. Koehn, Statistical significance tests for machine translation evaluation, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2004, pp. 388–395.

[15] P. Koehn, Europarl: A parallel corpus for statistical machine translation, in: Proceedings of the 10th Machine Translation Summit, 2005, pp. 79–86.

[16] P. Koehn, Statistical Machine Translation, Cambridge University Press, 2010.

[17] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, Moses: Open source toolkit for statistical machine translation., in: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, 2007.

[18] P. Koehn, F.J. Och, D. Marcu, Statistical phrase-based translation, in: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, 2003, pp. 48–54.

[19] E. Matusov, A. Mauser, H. Ney, Automatic sentence segmentation and punctuation prediction for spoken language translation, in: Proceedings of the 3rd International Workshop on Spoken Language Translation, 2006, Kyoto, Japan, pp. 158–165.

[20] F.J. Och, Minimum error rate training in statistical machine translation, in: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, 2003, pp. 160–167.

[21] K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 2002, pp. 311–318.

[22] M. Paul, Overview of the IWSLT 2009 Evaluation Campaign, in: Proc. of the International Workshop on Spoken Language Translation, 2009, Tokyo, Japan, pp. 1–18.

[23] H.S. Sichel, On a distribution representing sentence-length in written prose, J. Roy. Statistical Society. Series A 137 (1974) 25–34.

[24] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, J. Makhoul, A study of translation edit rate with targeted human annotation, in: Proceedings of Association for Machine Translation in the Americas, 2006, pp. 223–231.

[25] O. Uzuner, B. Katz, A comparative study of language models for book and author recognition, Proceedings of the 2nd International Joint Conference on Natural Language Processing (2005) 969–980.

[26] A. Venugopal, et al., Effective phrase translation extraction from alignment models, in: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 2003, pp. 319–326.

[27] R. Zens, H. Ney, N-gram posterior probabilities for statistical machine translation, in: Human Language Technology Conference, North American Chapter of the Association for Computational Linguistics Annual Meeting, Workshop on Statistical Machine Translation, 2006, New York City, pp. 72–77.

[28] B. Zhao, S. Vogel, A generalized alignment-free phrase extraction, in: Proceedings of ACL Workshop on Building and Using Parallel Texts, 1995, pp. 141–144.

[29] M. Zimmermann, H. Bunke, Hidden markov model length optimization for handwriting recognition systems, in: Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition, 2002, pp. 369–374.

*Author Biography

Joan Albert Silvestre-Cerdà (jsilvestre@dsic.upv.es) received his Computer Science degree (2010) from the Universitat Politècnica de València (UPV), Spain, and nowadays is ending his Ms.C. Degree in Artificial Intelligence, Pattern Recognition and Digital Imaging from the UPV. He is also member of the Pattern Recognition and Human Language Technology research group and of the Spanish AERFAI Society. His main research interests are in the areas of statistical machine translation and speech recognition.

Dr. Jesús Andrés-Ferrer received his Computer Science Degree (2004), his Ms.C. Degree (2008) in Artificial Intelligence, Pattern Recognition and Digital Imaging and his European phD degree in Computer Science (2010) from the Universidad Politécnica de Valencia (UPV). His thesis was awarded by the "Asociación Española de Reconocimiento de Formas y Análisis de Imágenes" (AERFAI) in 2011. He is co-author of 6 articles in international journals and more than 15 articles in international conferences. One of his articles was awarded with the “IEEE Spoken Language Processing Student Travel Grant”. He is an active member of the Pattern Recognition and Human Language Technology research group and the Spanish AERFAI Society. He has been involved in several national and international projects, and he is currently working in one of them (erudito.com) as a post-Doctoral researcher. His research interests include pattern recognition and its application to statistical machine translation, speech recognition, handwritten recognition and language modelling.

Dr. Jorge Civera is an assistant professor of computer science in the Universitat Politècnica de València (UPV). He received his undergraduate degree from UPV, in 2003 he completed his Master's degree at Georgia Institute of Technology, and in 2008 he received his Ph.D. from the UPV. His Ph.D. thesis was awarded as the best thesis on Computer Science at the UPV 2007-2008. He is co-author of 8 articles in international journals and more than 20 articles in international conferences. He has been involved in several national and international projects, actively participating in the European project TransType2. He is currently leading a Spanish research project on interactive machine translation and speech transcription. He is a member of the Pattern Recognition and Human Language Technology research group and the Spanish AERFAI Society. His research interests include pattern recognition and its application to statistical machine translation, speech recognition, and handwriting recognition.