Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 484580, 15 pages
doi:10.1155/2012/484580

Research Article

Learning to Translate: A Statistical and Computational Analysis

Marco Turchi,[1,2] Tijl De Bie,[2] Cyril Goutte,[3] and Nello Cristianini[2]

[1] *European Commission-Joint Research Centre (JRC), IPSC, GlobeSec, Via Fermi 2749, 21020 Ispra, Italy*
[2] *Intelligent Systems Laboratory, University of Bristol, MVB, Woodland Road, Bristol BS8 1UB, UK*
[3] *Interactive Language Technologies, National Research Council Canada, 283 Boulevard Alexandre-Taché, Gatineau, QC, Canada J8X 3X7*

Correspondence should be addressed to Marco Turchi, marco.turchi@gmail.com

Received 15 July 2011; Accepted 30 January 2012

Academic Editor: Peter Tino

We present an extensive experimental study of Phrase-based Statistical Machine Translation, from the point of view of its learning capabilities. Very accurate Learning Curves are obtained, using high-performance computing, and extrapolations of the projected performance of the system under different conditions are provided. Our experiments confirm existing and mostly unpublished beliefs about the learning capabilities of statistical machine translation systems. We also provide insight into the way statistical machine translation learns from data, including the respective influence of translation and language models, the impact of phrase length on performance, and various unlearning and perturbation analyses. Our results support and illustrate the fact that performance improves by a constant amount for each doubling of the data, across different language pairs, and different systems. This fundamental limitation seems to be a direct consequence of Zipf law governing textual data. Although the rate of improvement may depend on both the data and the estimation method, it is unlikely that the general shape of the learning curve will change without major changes in the modeling and inference phases. Possible research directions that address this issue include the integration of linguistic rules or the development of active learning procedures.

1. Introduction

Traditional approaches to machine translation (MT) [1] relied to a large extent on linguistic analysis. The (relatively) recent development of statistical approaches [2] and especially phrase-based machine translation, or PBMT [3, 4], has put the focus on the intensive use of large parallel corpora. In that statistical framework, translation is essentially viewed as the process of associating an input, the source sentence, with an output, the target sentence. Estimating a machine translation system is therefore similar to learning the mapping between the source/input and the target/output, a problem which has been extensively studied in statistics and in machine learning. This justifies our view of a typical phrase-based machine translation model as a learning system and motivates our analysis of the performance on that system.

A learning system typically considers a class of models, or hypotheses, and tries to find the one element in that class that provides the best prediction of the output on future, unseen input examples. The performance of every learning system is the result of (at least) two combined effects: the representation power of the hypothesis class, determining how well the system can approximate the target behaviour; statistical effects, determining how well the system can approximate the best element of the hypothesis class, based on finite and noisy training information. The two effects interact with richer classes being better approximators of the target behaviour but requiring more training data to reliably identify the best hypothesis. The resulting trade-off, equally well known in statistics and in machine learning, can be expressed in terms of bias versus variance, capacity control, or model selection. Various theories on learning curves have been proposed to deal with it, where a learning curve is a plot describing performance of a learning system as a function of some parameters, typically training set size. In practice this trade-off is easily observed, by noticing how the training error can be driven to zero by using a rich hypothesis class, which typically results into overfitting and increased test error.

In the context of statistical machine translation (SMT), where large bilingual corpora are used to train adaptive software to translate text, this task is further complicated by the peculiar distribution underlying the data, where the probability of encountering new words or expressions never vanishes. If we want to understand the potential and limitations of the current technology, we need to understand this interplay between the two factors affecting performance. In an age where the creation of intelligent behaviour is increasingly data driven, this is a question of great importance to all of artificial intelligence. These observations lead us to an analysis of learning curves in machine translation, and to a number of related questions, including an analysis of the flexibility of the representation class used, an analysis of the stability of the models with respect to perturbations of the parameters, and an analysis of the computational resources needed to train these systems.

In phrase-based approaches to statistical machine translation, translations are generated in response to some input source text. The quality of the output translation depends on the correctness of the generated language (*fluency*) as well as on how faithful it is to the original meaning (*adequacy*). This is reflected in the two main components of the typical SMT model: the language model and the translation model. While these will be formally defined in Section 2, let us mention that the translation model relies on a bilingual table of corresponding "phrases" (sequences of words), with an associated probability which reflects how likely it is that the source phrase will be translated as the target phrase. The language model is typically built using a table of n-grams, with associated probabilities, which is sufficient to define a Markov chain. Note that the overall behaviour of the translation system is largely controlled by the content of those two tables. They are automatically filled during the training phase, when a bilingual corpus is used to identify both phrases and their probabilities. Since future translations are produced by maximizing a scoring function estimating translation quality, using the content of the two tables, we see that the contents of the translation and language models tables correspond to the tunable parameters of the learning system. The hypothesis space of SMT systems is then the class of all possible "translation functions" that can be implemented, for all possible choices of language and translation tables. As this is an enormous search space, it is no wonder that both algorithmic and statistical challenges are encountered when training these systems.

From a statistical learning point of view, this raises interesting questions: How much of the overall error of the translation system is due to representation limitations, and how much to the difficulty of extracting suitable Translation and Language model tables from a finite sample? And what quantities control the trade off between the approximation and estimation errors, essentially playing the role of model selection?

We have undertaken a large scale experimental and theoretical investigation of these questions. Using the open source packages Moses [5] and Portage [6], and three different corpora: the Spanish-English Europarl [7], the UN Chinese-English, and the Giga Corpus French-English [8], we have performed a detailed investigation of the influence of data sizes and other design choices in training various components of the system, both on the quality of translations and on the computational cost. We use this data to inform a discussion about learning curves. We have also investigated the model-selection properties of n-gram size, where the n-grams are the phrases used as building blocks in the translation process. Note that our experiments have been performed using English as target language. Although many language pairs would yield different translation performance, in this paper, we are not interested in the translation performance *per se*: we focus our attention on analyzing the SMT system as a learning system.

Since our goal was to obtain high-accuracy learning curves, that can be trusted both for comparing different system settings and to extrapolate performance under unseen conditions, we conducted a large-scale series of tests, to reduce uncertainty in the estimations and to obtain the strongest possible signals. This was only possible, to the degree of accuracy needed by our analysis, by the extensive use of a high-performance computer cluster over several weeks of computation.

One of our key findings is that the current performance of phrase-based statistical machine translation systems is not limited by the representation power of the hypothesis class, but rather by model estimation from data. In other words, we demonstrate that parameter choices exist that can deliver significantly higher performance, but that inferring them from finite samples is the problem. We also suggest that increasing dataset size is not likely to bridge that gap (at least not for realistic amounts in the i.i.d. setting), nor is the development of new parameter estimation principles. The main limitation seems to be a direct consequence of Zipf's law, and the introduction of constraints from linguistics seems to be an unavoidable step, to help the system in the identification of the optimal models without resorting to massive increases in training data, which would also result in unmanageable training times, and model sizes. This is because the rate of improvement of translation performance is at best logarithmic with the training set size. We estimate that bridging the gap between training and test error would require about 10^{15} paired bilingual sentences, which is larger than the current estimated size of the web.

Parts of the results reported in this paper were presented at the third workshop on statistical machine translation [9]. Work related to our learning curve experiments can also be found in [10].

It is important to remark that while there are many discussions about automatic evaluation of SMT systems, this work does not consider them. We work within the well-defined setting where a loss function has been agreed upon, that can measure the similarity between two sentences, and a paired training set has been provided. The setting prescribes that the learning system needs to choose its parameters so that it can identify high-quality (low expected loss) translations. We investigate the learning-theoretic implications of this setting, including the interplay between approximation error and estimation error, model selection, and accuracy in parameters estimation. We do not address more general

themes about the opportunity for SMT to be evaluated by automatic metrics.

2. Phrase-Based Machine Translation

What is the best function class to map Spanish documents into English documents? This is a question of linguistic nature and has been the subject of a long debate. With the growing availability of bilingual parallel corpora, the 1990 s saw the development of *statistical* machine translation (SMT) models. Given a source ("foreign") language sentence $\mathbf{f}$ and a target ("English") language translation $\mathbf{e}$, the relationship between $\mathbf{e}$ and $\mathbf{f}$ is modelled using a statistical or probabilistic model such as $p(\mathbf{e} \mid \mathbf{f})$.

The first statistical models were word based [2, 11], combining a Markovian language model with a generative word-to-word translation model, in a noisy channel model inspired by speech recognition research.

Current state-of-the-art SMT uses *phrase-based* models, which generalized and superseded word-based models. They rely on three key ideas:

(i) the use of *phrases* (sequences of consecutive words) as basic translation units instead of words;

(ii) the use of a log-linear model instead of a simple product of the language and translation models;

(iii) the use of minimum error-rate training in order to estimate the parameters of the log-linear model, instead of maximum likelihood.

Many additional ideas contribute to the efficiency of these models, such as efficient hypothesis search, rescoring, or improved feature functions. The underlying log-linear model may be interpreted as a maximum entropy model:

$$p(\mathbf{e} \mid \mathbf{f}) = \frac{\exp\left(\sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f})\right)}{Z(\mathbf{f})}, \tag{1}$$

where $Z(\mathbf{f})$ is the normalization factor, and h_i are the feature functions, which we will discuss further in a moment.

Note that finding the best target translation $\mathbf{e}$ given a source sentence $\mathbf{f}$ amounts to maximizing the conditional probability $p(\mathbf{e} \mid \mathbf{f})$ with respect to $\mathbf{e}$, which yields

$$\hat{\mathbf{e}} = \arg\max_{\mathbf{e}} \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{f}) \approx \arg\max_{\mathbf{e}} \sum_i \lambda_i h_i(\mathbf{e}, \mathbf{a}, \mathbf{f}). \tag{2}$$

The *feature functions* $h_i(\mathbf{e}, \mathbf{a}, \mathbf{f})$ involve both the source and target sentences, and the approximation gives these feature functions access to $\mathbf{a}$, the alignment connecting $\mathbf{e}$ and $\mathbf{f}$. This model for $\mathbf{e}$ given $\mathbf{f}$ is linear in the log domain, which motivates the description of this framework as "log-linear model" [3, 4, 12]. In addition, many feature functions are defined as logarithms of probabilities. The search for the optimal translation in (2) is also referred to as *decoding* as, in the original analogy of the noisy channel, it corresponds to retrieving the clean message $\mathbf{e}$ from a noisy or encrypted observation $\mathbf{f}$.

One key aspect of the log-linear model in (2) is that it can take into account almost arbitrary feature functions, as long as they can be defined in terms of $\mathbf{e}$, $\mathbf{f}$, and $\mathbf{a}$. However, in order to be able to efficiently search for the optimal translation $\hat{\mathbf{e}}$, we usually assume that the feature functions decompose linearly across basic constituents of the sentences. In phrase-based MT, the sentences are decomposed into basic translation units called *phrases*. Note that these are not phrases in the linguistic sense, but simply subsequences of words. For a sentence $\mathbf{e}$ composed of the sequence of words $w_1, \ldots w_{|\mathbf{e}|}$, phrases e_k can be any contiguous subsequence of words w_j (and similarly for $\mathbf{f}$). A feature function that linearly decomposes across phrases takes the form $h_i(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \sum_k h_i(e_k, a_k, f_k)$, where e_k and f_k are phrases from $\mathbf{e}$ and $\mathbf{f}$ (resp.), and a_k is the alignment that connects them.

Typical examples of feature function that compose a basic phrase-based MT system are:

(i) one or several phrase translation features: $h_T(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \sum_k \log p(f_k \mid e_k)$;

(ii) one or more language model features: $h_L(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \log p(\mathbf{e}) = \sum_j \log p(w_j \mid w_{j-1}, \ldots w_1)$;

(iii) distortion feature $h_D(\mathbf{e}, \mathbf{a}, \mathbf{f}) = \sum_k -\|\text{start}(f_k) - \text{end}(f_{k-1}) - 1\|$;

(iv) word penalty and/or phrase penalty features.

The phrase translation probabilities $p(f_k \mid e_k)$ are defined over a set of phrase pairs (e_k, f_k) referred to as a *phrase table*. Part of the overall MT training process is to estimate this table and the associated probabilities. This is typically done by first aligning each sentence pair at the word level, and then extracting all phrase pairs that are compatible with the word alignment. Statistics on the phrase pair are accumulated over the entire corpus. In our experiments below, we rely on word-to-word IBM models [2] for alignment. Although more elaborate techniques have appeared more recently [13, 14], their impact on the resulting machine translation quality is still unclear [15].

The standard language model feature such as used below relies on an n-gram language model, combining the probabilities of each word in the target hypothesis given the preceding n-gram language models, combined with appropriate smoothing, are very efficient and naturally decomposable across words (and phrases), making them particularly well suited in this framework. Again, more recent alternatives exist, but their actual impact on MT performance is not always obvious.

The distortion feature controls the reordering between phrases. Note that only very short-range reordering may be handled within phrases. Long-range reordering must be handled by target phrase permutations. This feature allows to regulate the amount of reordering depending on, for example, the language pair.

The word length (or word penalty) feature regulates the length of the target sentence. This is useful because the language model feature typically favours shorter sentences (because each additional trigram can only lower the language model probability). This is a simple, yet effective feature.

The process of training a machine translation system involves estimating the various parameters of the model: the

log-linear parameters λ_i as well as the parameters internal to the feature functions, such as the phrase translation probabilities and language model n-gram and backoff probabilities. In a typical phrase-based system such as used in our work, training is modularized by first estimating the later on various corpora, for example, a large bilingual corpus for translation probabilities and a possibly larger monolingual corpus for the language model parameters.

The log-linear parameters are then estimated by *minimum error rate training* (MERT). The weights λ_i are tuned to optimize an automatic MT metric such as BLEU [16] over a number of bilingual sentence pairs from a development corpus:

$$\hat{\lambda} = \arg\max_{\lambda} BLEU(\hat{\mathbf{e}}(\mathbf{f}), \mathbf{e};\ (\mathbf{e}, \mathbf{f}) \in \mathcal{D}), \tag{3}$$

where $\hat{\mathbf{e}}(\mathbf{f})$ is the translation produced by solving (2) for source sentence $\mathbf{f}$, and $\mathbf{e}$ is the corresponding reference translation. $\mathcal{D}$ is the set of sentence pairs over which MERT is performed. Solving (3) is difficult because the decoding necessary to produce the hypothesis translation is expensive. The standard solution [17] is to approximate the exhaustive search in (2) by a search over a smaller set of n-best candidate translations for $\mathbf{f}$. For each new value of λ tried by MERT, new hypothesis translations are added to the list. As the number of hypotheses produced by the decoder is finite, this is guaranteed to converge, and in practice, it does fairly quickly. An additional difficulty is that the landscape of the cost, for example, BLEU, is piecewise constant and highly irregular. In practice, the optimization in (3) is done using the Powell's method [18], a straightforward coordinate descent method where optimization is performed along each λ_i in turn. The key observation is that when optimization is approximated as described above by a search over a list of n translations, there are at most $n - 1$ points along the line search where BLEU can change. This yields an efficient algorithm for obtaining the exact solution of each line search in Powell's method and therefore provides a way to iteratively optimize the log-linear weights λ using MERT. A number of alternatives have been proposed, such as on-line discriminative training [19, 20]. However, the two systems we use here both use a fairly traditional implementation of MERT.

This *phrase-based machine translation* approach relies on a specific representation of the translation process, such as the choice of contiguous word sequences (phrases) as basic units in the language and translation models. How far can this representation take us towards the target of improving translation quality? Are the current limitations due to the approximation error of this representation, or to estimation errors originating from insufficient training data? How much space for improvement is there, given new data or new statistical estimation methods or given different models with different complexities?

Before we present experimental results that address these questions, we will describe the setup that was used to obtain these results.

3. Experimental Setup

3.1. Data. We used three different sentence-aligned corpora, covering different language pairs and sizes:

(1) Europarl Release v3 Spanish-English [7],

(2) UN Chinese-English corpus provided by the Linguistic Data Consortium,

(3) Giga corpus French-English [8].

The details of these three corpora are given in Table 1. The language pairs cover European as well as non-European languages, and the sizes range from 1.2 M to 22.5 M sentence pairs. We expect that translation between European languages will be easier than from Chinese to English; however, we are not so much interested in the actual translation performance as in the way this performance evolves with increasing data and under a number of conditions.

The Europarl corpus contains material extracted from the proceedings of the European parliament, and the UN data contains material from the United Nations. Both therefore cover a wide range of themes, but are fairly homogeneous in terms of style and genre. The Giga corpus, on the other hand, was obtained through a targeted web crawl of bilingual web sites. These sites come from the Canadian government, the European Union, the United Nations, and other international organizations. In addition to covering a wide range of themes, they also contain documents with different styles and genres. We estimate that the rate of misaligned sentence pairs was around 13%.

These corpora are each divided in three sets, each with a different role. The *training* part is used to obtain the language model and phrase tables. The *development* set is used to estimate the log-linear weights λ using MERT, and the *test* set is set aside during the estimation process in order to provide an unbiased estimate of the translation performance.

This work contains several experiments on different types and sizes of data set. To be consistent and to avoid anomalies due to overfitting or particular data combinations, each set of pairs of sentences has been randomly sampled. The number of pairs is fixed, and a program selects them randomly from the whole original training, development, or test set using a uniform distribution. This process is iterated a certain number of time and redundancy of pairs is allowed inside each subset (bootstrap, see [18, 21, 22] for an application to PBSMT).

3.2. Software. Several software packages are available for training PBSMT systems. In this work, we use both Moses [5] and Portage [6]. Moses is a complete open-source phrase-based translation toolkit for academic purposes, while Portage is a similar package available to partners of the National Research Council Canada. Both provide all the state-of-the-art components needed to create a phrase-based machine translation system from one language to another. They contain different modules to preprocess data and train the language models and the translation models. These models can be tuned using minimum error rate training [17]. Both use standard external tools for training

Table 1: Number of total and distinct words in training, development, and test sets.

			Training	Development	Test
Europarl	English	No. of Sentences	1,259,914	2,000	2,000
		Total words	35,284,052	58,762	59,147
		Distinct words	124,080	6,551	6,429
	Spanish	Total words	36,695,628	60,536	61,160
		Distinct words	164,920	8,182	8,239
UN corpus	English	No. of Sentences	4,968,857	1,000	10,000
		Total words	146,980,344	29,545	295,085
		Distinct words	485,494	5,210	17,105
	Chinese	Total words	138,045,740	27,764	278,4256
		Distinct words	530,295	4,353	13,193
Giga corpus	English	No. of Sentences	22,515,400	2,000	3,000
		Total words	636,113,866	51,549	90,474
		Distinct words	2,603,907	8,691	12,580
	French	Total words	772,104,558	62,682	109,197
		Distinct words	2,512,286	10,124	14,614

the language model, such as SRILM [23], and Moses also uses GIZA++ [24] for word alignments. Moses and Portage are very sophisticated systems, capable of learning translation tables, language models, and decoding parameters from data. We will later analyze the contribution of each component to the overall score.

The typical processing pipeline is as follows. Given a parallel training corpus, long sentences are filtered out, and the remaining material is lowercased and tokenized. These sentences are used to train the language and translation models. Training the translation models requires several steps such as aligning words, computing the lexical translation, extracting and scoring the phrases, and creating the reordering model. When the models have been created, the development set is used to run the minimum error rate training (MERT) algorithm [17] to optimize their weights. We refer to that step as the optimization step in the rest of the paper. The test set is used to evaluate the quality of models on the data.

All experiments using Moses have been run using the default parameter configuration. GIZA++ used IBM models 1, 2, 3, and 4 with number of iterations for model 1 equal to 5, model 2 equal to 0, and model 3 and 4 equal to 3; SRILM used n-gram order equal to 3 and the Kneser-Ney smoothing algorithm; MERT has been run fixing to 100 the number of n best target sentence for each developed sentence, and it stops when none of the weights changed more than 1e-05 or the n best list does not change.

The training, development, and test set sentences are tokenized and lowercased. The maximum number of tokens for each sentence in the training pair has been set to 50, whilst no limit is applied to the development or test set. TMs were limited to a phrase length of 7 words, and LMs were limited to 3.

3.3. Hardware. All the experiments have been run on high-performance clusters of machines.

The first cluster (at U. Bristol) includes 96 nodes each with two dual-core Opteron processors, 8 GB of RAM per node (2 GB per core) and 4 thick nodes each with four dual-core Opteron processors, 32 GB of RAM per node (4 GB per core), for a total of 416 CPUs. Additional information: ClearSpeed accelerator boards on the thick nodes; SilverStorm Infiniband high-speed connectivity throughout for parallel code message passing; General Parallel File System (GPFS) providing data access from all the nodes with a total of 11 terabytes of storage. Each experiment has been run using one core and allocating 4 Gb of RAM.

The second cluster (at NRC) includes 29 nodes each with two dual-core processors and 16 GB RAM per node (4 GB per core) and 8 "fat" nodes with 4 quad-core processors each and 128 GB RAM per node (8 GB per core), for a total of 244 CPUs. The file system is Ibrix and provides data access from all nodes, with a total of 17 TB of storage. Experiments using Portage are distributed over several CPUs, the total number of which depends on the various stages in the estimation process.

3.4. Evaluation Metrics. The evaluation of a machine translation system is a lively and hotly debated topic in this

Table 2: Correlation coefficient between evaluation scores.

	BLEU	NIST	METEOR	TER
BLEU	1	0.724	0.8633	−0.866
NIST	0.724	1	0.6171	−0.8702
METEOR	0.8633	0.6171	1	−0.8173
TER	−0.866	−0.8702	−0.8173	1

field. Ideally, human beings can evaluate the quality of a translated sentence. However, this is unfeasible for rapid development of automatically trained systems with multiple parameter tuning, as human evaluation is expensive, slow, and sometimes inconsistent and subjective.

Therefore, instead of reporting human judgement of translation quality, various automatic measures have been proposed. An automatic score measures the quality of machine-translated sentences by comparing them to a set of human translations, called reference sentences. The score needs to be able to discriminate good translations from bad ones, whilst considering aspects such as adequacy and fluency.

Several metrics have been introduced: BLEU [16], NIST [25], Meteor [26, 27], and TER [28] are among the most well known. BLEU and NIST are based on averaging n-gram precisions, combined with a length penalty which penalizes short translations containing only sure words. These metrics differ on the way the precisions are combined and on the length penalty.

Meteor evaluates a translation by computing a score based on the word alignment between the translation and a given reference translation. TER relies on the computation of an edit distance, and it is defined as the minimum number of edit operations (insertions, deletions, substitutions, *and shifts*) needed to change an automatically translated sentence into one of the references, normalized by the average length of the references.

Table 2 reports the correlation coefficients between the measures (details on how these values have been computed are in Section A.3). Clearly, all measures correlate strongly with each other, such that the choice of the performance measure is fairly arbitrary, as long as one is consistent. For this reason, we have chosen to use BLEU throughout this paper as it is the most widely used automatic score in machine translation.

4. Learning Curve Analysis

4.1. Role of Training Set Size on Performance on New Sentences. In this section, we analyze how training set size affects the performance by creating learning curves (BLEU score versus training set size).

The general framework for this set of experiments consists of creating subsets of the complete corpus by subsampling from a uniform distribution without replacement. We have created 10 random subsets for each of the 20 chosen sizes, where each size represents 5%, 10%, and so forth of the complete corpus. For each subset, a new instance of the PBSMT system has been created, for a total of 200 models. Two hundred experiments have then been run on an independent test set (of 2,000 sentences, also not included in any other phase of the experiment).

Table 3: Different settings used to create the Learning Curves. Due to the large dimension of the Giga corpus, only three random subsets have been built.

Language pairs	Data	SMT system	No. random subsets
Es − En	Europarl	Moses	10
Es − En	Europarl	Portage	10
Fr − En	Giga	Moses	3
Zh − En	UN	Portage	10

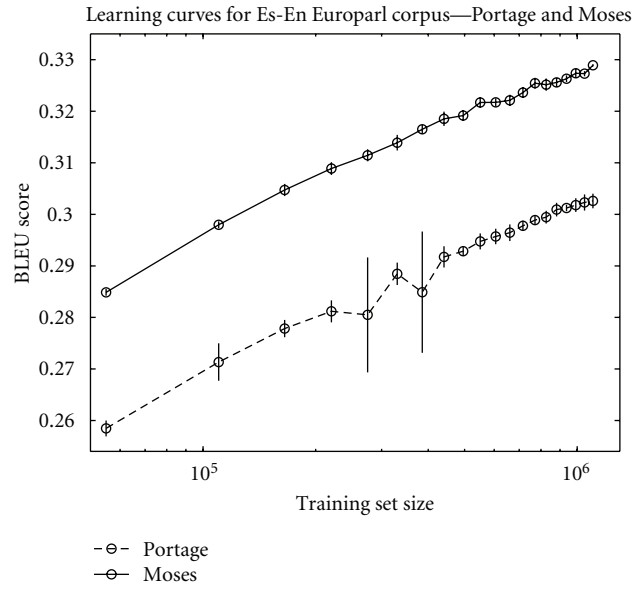

Figure 1: Spanish-English learning curve obtained using Europarl corpus, Portage and Moses.

In these experiments, we focus our attention on the growth rate of the learning curve, in particular we are interested to check if the learning curve has a logarithmic behaviour. In fact, a common belief in SMT is that learning curves follow logarithmic laws; to analyze this in our experiments, we show all the learning curves in the linear log scale, where we can study if the curve has a linear behaviour.

Note that sampling without replacement, error bar dimension reduces according to the increment of the training set size. We also believe that in this particular situation, the presence of the error bars may help to better understand the stability of the system.

Using the framework described above, four different settings have been set to produce learning curves, see Table 3. In Figures 1, 2, and 3, the learning curves are shown.

In all the figures, the curves are increasing linearly or slightly more slowly than that, suggesting a learning curve

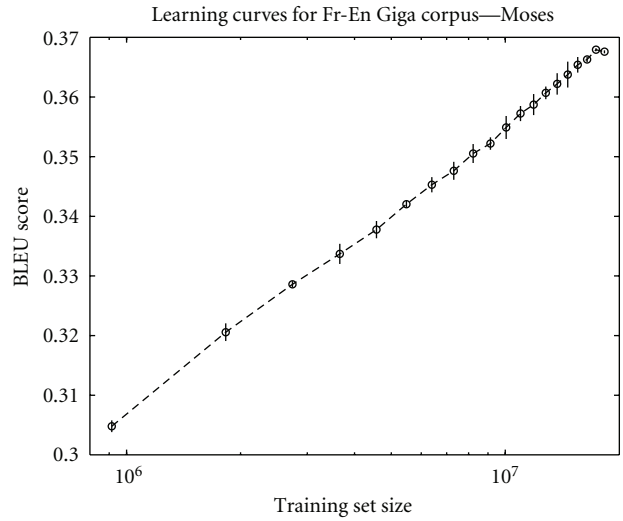

Figure 2: French-English learning curve obtained using Giga corpus and Moses.

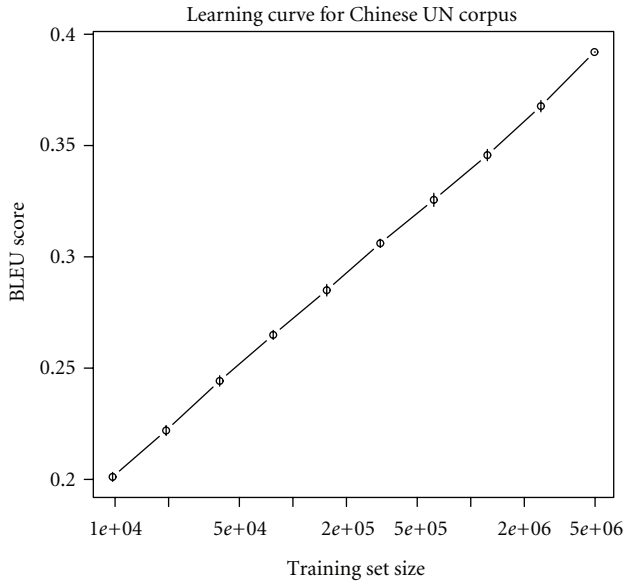

Figure 3: Chinese-English learning curve obtained using UN corpus and Portage.

that is "at best" logarithmically increasing with the training set size. Shape of the learning curves is comparable across different data size and language pairs. In any case, the addition of massive amounts of data from the same distribution will result in small improvements in the performance.

The small error bars that we have obtained also allow us to regard the stability of the SMT when trained on the same training set size.

Due to the large amount of data in the Giga corpus, Fr-En learning curve has been obtained running only three subsamples for training set size.

4.2. Relative Importance of TM and LM. In the previous section, experiments have been run using the same training set size for language and translation models. In general, there is a large difference in terms of cost of retrieving training data for language and translation models; the former can be trained using monolingual data, while the second needs bilingual texts. In recent years, several parallel corpus have been produced, for example, Europarl [7], JRC Acquis [29], and others, but they are not comparable to the amount of freely available monolingual datasets.

Google in [30] has shown that performance improves logarithmically in the linear scale with the number of tokens in the language model training set when this quantity is huge (from billions to trillions of tokens). In this section, we are interested to understand whether there is a trade-off between the training data size used to build language and translation models and how performances are affected by their differences. We propose a mathematical model to estimate the BLEU score's variations according to the language and translation training data sizes. Performance of an SMT system is a function of the dimension of the training data that can be logarithmic as seen in the previous section. We have modelled this relation in the following way:

$$\text{BLEU}(d) = \alpha_{\text{LM}} \times \log_2(d_{\text{LM}}) + \alpha_{\text{TM}} \times \log_2(d_{\text{TM}}) + \epsilon \,, \tag{4}$$

where d is the amount of training data, d_{LM} is the amount of training data used to build the language model, d_{TM} is the amount of training data used to build the translation model, α_{LM} and α_{TM} are weighting factors that identify the contribution of language and translation training data to the BLEU score, and ϵ is the residual. To evaluate the relation between the amount of training data used to build language and translation models, we estimated α_{LM} and α_{TM}.

Subsets of the training data have been selected using 10% of the data as increasing set size. For each training set size, one random set has been created without replacement. Development and test sets are fixed. One instance of the SMT system has been run for each of all possible combinations of the language and translation training data sizes. BLEU score value has been associated to each pair: language and translation set size. Multiple linear regression based on least squares [31] has been performed using the BLEU score values as response observations and the logarithmic values of the training set sizes as observations. This setting has been applied to Europarl and Giga corpus datasets using Moses as SMT system.

We performed two sets of experiments: in the first one, we estimated the weighting factors using all the data, see Table 4, and in the second, we tested the prediction capability of our model randomly splitting the data in training (80%) and test (20%) sets.

The results in Table 4 empirically confirm the common belief that adding data to the translation model is more important than to the language model ($\alpha_{\text{TM}} > \alpha_{\text{LM}}$). The values of α_{LM} and α_{TM} vary across the datasets and correspond to an increase of 1.3 to 1.5 BLEU point for the LM and 1.8 to 1.9 for the TM, for each doubling of the data. However, their ratio is rather stable.

For the second set of experiments, data has been randomly sampled in training and test sets one thousand

Table 4: Empirical estimation of the weighing factors α_{LM} and α_{TM}. Experiments have been performed independently on the Europarl and Giga corpus datasets.

Data	α_{LM}	α_{TM}	α_{TM}/α_{LM}	ϵ
Europarl	0.0147	0.0193	1.313	2×10^{-4}
Giga Corpus	0.0133	0.0182	1.368	4×10^{-6}

times. Training set has been used to estimate the alphas and the residual, and test set to predict the BLEU score values. At each iteration, estimation error was computed. Average α and error on the Europarl and Giga corpus datasets are shown in Table 5. The proposed model is able to approximate well enough the BLEU score using Moses as translation system and in-domain test sets. According to this setting and assuming that we are in the standard case where d_{LM} is equal to d_{TM}, it is possible to use our model to estimate the amount of training data needed to reach a certain Bleu score for example, more than 150 million sentences to obtain a Spanish-English Bleu score equal to 0.45.

4.3. Role of Phrase Length in the Translation Table (Model Selection). The richness of the hypothesis class controls the trade-off between training and test error. Richer hypothesis classes can fit the training data more accurately but generalize less well than poorer classes, a phenomenon known as overfitting. The choice of the appropriate expressive power, within a parametrized class of models, is called model selection and is one of the most crucial steps in the design of learning systems.

In phrase-based SMT, this is controlled by selecting the maximum length of phrases to be used as building blocks for translation. Using long phrases will help when the system has to translate sequences of words that match what was encountered in the training corpus, but this becomes increasingly unlikely as the phrases become longer. On the other hand, short sentences are more often reused, but may also be more ambiguous and lead to errors more often. This is where the trade-off between representation and estimation errors is controlled. When extracting longer phrases, we expect training set performance to be higher, but test performance to drop (overfitting). Optimizing test performance requires the right trade-off.

In this section, we analyze how the phrase length can affect the performance in terms of BLEU score. We also report the distribution of the phrase lengths in the translation table, as well as how the system uses the phrases of different length during the translation of both training and test material. We have created 10 random subsets of the complete Europarl corpus containing 629,957 pairs of sentences. For each subset, ten PBMT systems have been estimated. Each instance of Moses has been trained using a different maximum phrase length, from 1 to 10. Each model was then tested on the 2000 sentence test set, and on a random subset of 2000 training sentences. Translation of training sentences allows us to estimate the training error.

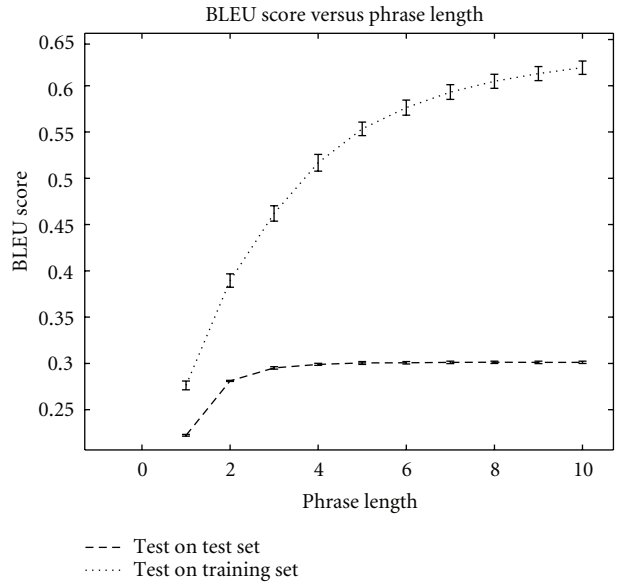

Figure 4: BLEU versus n-gram length. "Test on test set" has been obtained using a fixed test set and no optimization phase. "Test on training set" is a test set selected by the training set for each training set size and no optimization phase.

The learning curves in Figure 4 illustrate how the performance is affected by the phrase length. The "test on test set" curve is less influenced by the phrase length than the "test on training set" curve. The latter shows a large upward bias in translation performance, which is expected from a learning system when testing on material that has been used for training. Both learning curves show a big improvement when moving from the word-to-word translation (phrase length equal to one) to the phrase-based model (higher phrase lengths).

In the "test on test set" learning curve, there seems to be no significant advantage to using phrases longer than 4 words. By contrast, when testing on the subset of the training material, the performance continues to grow. This unrealistic case is not affected by the Zipf's law, because almost all the words necessary to translate the training material have, by definition, already been observed. The model is therefore able to match long phrases when producing the "test on training set" translations.

In order to explore this further, we also compute the distribution of the entries in the translation table as a function of the length of the source language phrases. Figure 5 shows that the number of phrases peaks around 4-grams and 5-grams, then steadily decreases. This means that the phrase extraction algorithm finds it more and more difficult to extract longer phrases.

We investigate this further by plotting the distribution of phrases actually used while translating. We randomly select 2 sets of 500 sentences, one from the training and one from the test material. In each case, we count the number of phrases of each length that were actually used to produce the translation. The right panel of Figure 5 reports these distribution. It shows that, while the models use a fair

Table 5: Performance obtained training the regressor on 80% of the data and testing on 20%. This process has been iterated 1,000 times. Experiments have been performed independently on the Europarl and Giga corpus dataset.

Data	α_{LM}	α_{TM}	*Estimation Error*
Europarl	$0.0102 \pm 2 \times 10^{-4}$	$0.0134 \pm 2 \times 10^{-4}$	$5.45 \times 10^{-5} \pm 14 \times 10^{-4}$
Giga Corpus	$0.0092 \pm 7 \times 10^{-5}$	$0.0126 \pm 7 \times 10^{-5}$	$-5.57 \times 10^{-5} \pm 2 \times 10^{-4}$

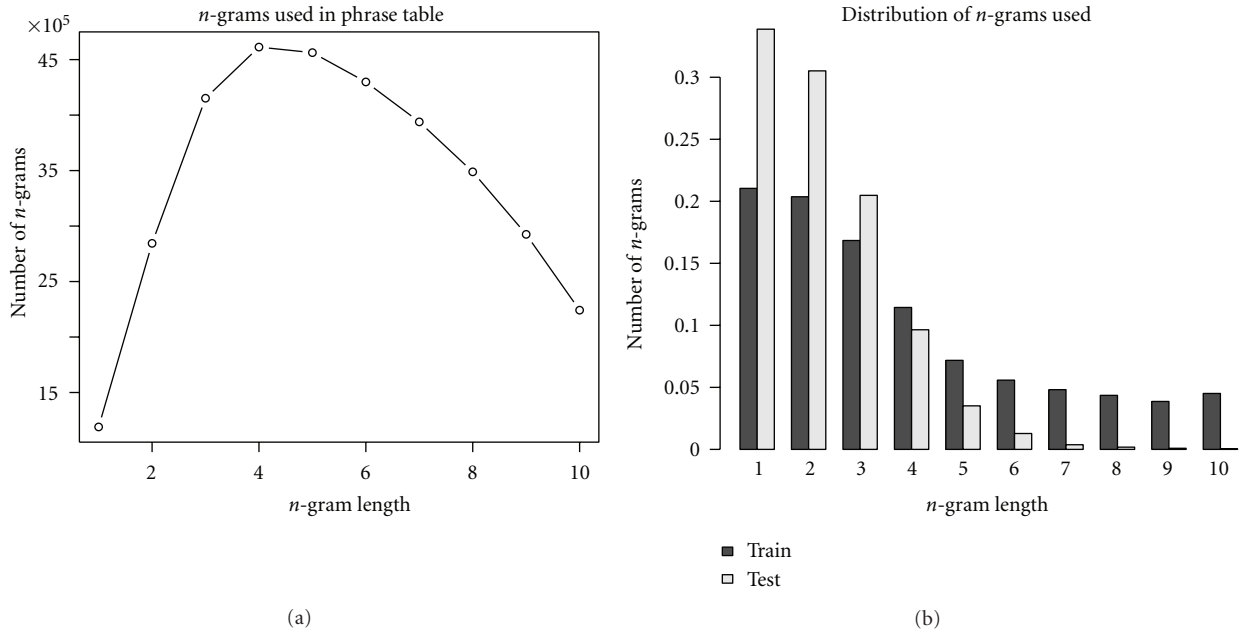

Figure 5: *N*-gram statistics: (left) number of *n*-grams of each size in the translation table; (right) distribution of phrase size used for translating training and test material, for $n = 1, \ldots, 10$.

amount of longer phrases to translate the training material, these longer phrases are essentially never used for translating the test set: 98% of the phrases are 5-grams or shorter. Again this is related to Zipf's law: while the model is able to match longer sequences from material it has already seen, test data is usually too different to reliably match beyond 4 or 5 consecutive words.

These experiments suggest that the phrase length has a limited impact on actual test performance. Going to larger *n*-grams seems to bring little benefit in terms of performance as the model continues to prefer short phrases during the decoding phase. This is due to both the diminishing number of longer phrases in the table and to the lower probability that these longer sequences match the test material.

5. Model Perturbation: Analysis and Unlearning Curves

Much research has focused on devising improved principles for the statistical estimation of the parameters in language and translation models. The introduction of discriminative graphical models has marked a departure from traditional maximum likelihood estimation principles, and various approaches have been proposed.

The question is how much information is contained in the probabilities estimated by the model? Does the performance improve with more data because certain parameters are estimated better, or just because the lists are growing? In the second case, it is likely that more sophisticated statistical algorithms to improve the estimation of probabilities will have limited impact.

In this section, we analyze what we call "unlearning curves." These are obtained by increasingly perturbing the parameters inferred by the system, in order to observe how performance deteriorates. This can either represent the effect of insufficient statistics in estimating them, or the use of imperfect parameter estimation biases. These parameters are probabilities, phrases, and associations between source/target phrases contained inside translation and language model tables.

We have performed two different types of perturbation.

(i) *Unlearning by Adding Noise.* A percentage of noise has been added to each probability, $\widetilde{p}$, in the Language model, including conditional probability, and translation model, bidirectional phrase translation probabilities and lexicalized weighting. The aim of this set of experiments is to test how robust the system is with respect to a reduced accuracy of its numeric parameters.

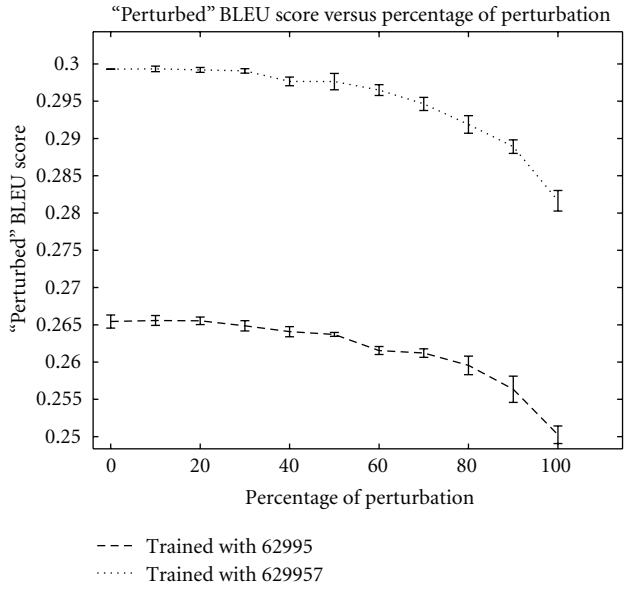

Figure 6: Each probability of the language and translation models has been perturbed adding a percentage of noise. This learning curve reports BLEU score versus the percentage of perturbation applied. These results have been obtained using a fixed training set size equal to 62,995 and 629,957 pairs of sentences and Moses as translation system.

The noised probability is obtained as $p' = \min(1, \widetilde{p} + \nu)$, where $\nu = \text{rand}(-\widetilde{p} \times k, +\widetilde{p} \times k)$ with percentage of noise $k \in [0, 1]$. Different noise levels k have been used. For each value of k, ten experiments have been run. In this case, we randomly select two fixed training set sizes equal to 62,995 and 629,957 pairs of sentences form the Europarl corpora and use Moses as translation system. The unlearning curves are shown in Figure 6.

(ii) *Unlearning by Randomization of Parameters.* The second kind of noise that we add to the model is based on a swap of a particular quantity inside two entries of language or translation model. This is meant to test how robust the system is to perturbations of the all-important associations between phrases/numbers and to the associations between source/target phrases.

We refer to "numerical swap" when, given two entries, probabilities are swapped. While we refer to "words swap" when, given two entries of the translation model, we swap the target language phrases.

Three different sets of experiments have been run applying "numerical swap" only to the language model, "numerical swap" only to the translation model and "words swap" only to the translation model. Different values of percentage of noise between 0 and 1 have been used. For each percentage value, ten experiments have been run. All the perturbations have been applied on a model trained with 629,957 pairs of sentences randomly selected form the Europal data using Moses as translation system. The unlearning curves are shown in Figure 7.

Various observations can be made based on these experiments. The first unlearning curve (Figure 6), obtained by adding to each parameter a random number (sampled from within a range) proportional to its size, is meant to test the role of detailed tuning of parameters. While the orders of magnitude are respected, the fine structure of the parameter set is randomized. The gentle decline in performance seems to suggest that fine tuning of parameters is not what controls the performance here, and that perhaps advanced statistical estimation or more observations of the same n-grams would not lead to much better performance. This is also compatible with what is seen in the learning curve.

It is important to notice, however, that introducing a more aggressive type of noise (Figure 7(b)) that essentially replaces entire parameters with random values does lead to a more significant decline in performance. This was obtained by swapping random entries, and so after 100 percent of swaps essentially every entry is a random number (because the locations to swap are chosen with replacement). It is interesting to see that the decline of the language model is much less pronounced than that of the translation model.

The set of experiments in Figure 7(a) is harder to explain without discussing the inner workings of the translation model and Moses. Here, we swapped n-grams in the translation table, essentially breaking the connection between words and their translation. A rapid decline should be expected. However, the mapping between words is stored in a very redundant way within the TM, and this depends on the way the translation table is created, based on sentence alignments. Once an alignment has been found between two sentences, essentially every n-gram (for every value of n) is a candidate for insertion in the translation table. So very often, longer n-grams are inserted, alongside shorter segments of the same n-grams, and are added to different entries of the table. So if we remove an n-gram, chances are that other similar (longer or shorter) n-grams are present and can take over. In this way, it is not possible to directly compare the unlearning curve for the n-grams part with that for the numeric part of the tables.

6. Discussion

The impressive capability of current machine translation systems is not only a testament to an incredibly productive and creative research community, but can also be seen as a paradigm for other artificial intelligence tasks. Data-driven approaches to all main areas of AI currently deliver the state-of-the-art performance, from summarization to speech recognition to machine vision to information retrieval. And statistical learning technology is central to all approaches to data-driven AI. Understanding how sophisticated behaviour can be learnt from data is hence not just a concern for machine learning, or to individual applied communities, such as statistical machine translation, but rather a general concern for modern artificial intelligence. The analysis of learning curves and the identification of the various

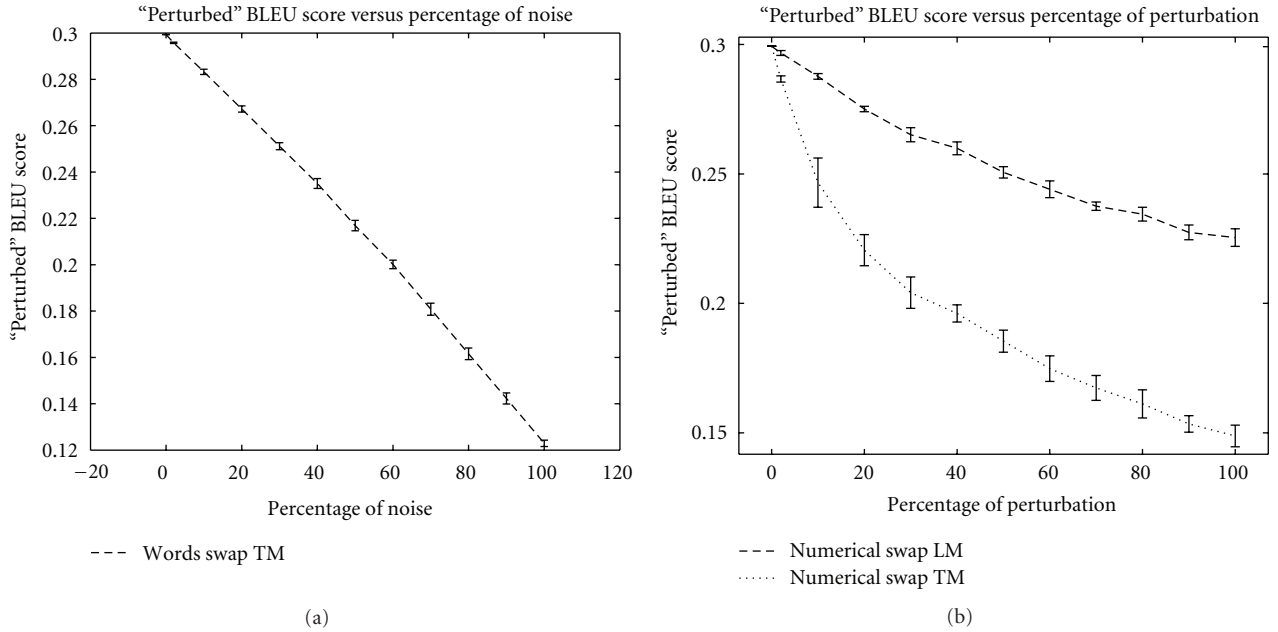

Figure 7: Unlearning curves. These results have been obtained using a fixed training set size equal to 629,957 pairs of sentences and Moses as translation system. In (a), "words swap TM" has been obtained by swapping the target phrases inside the TM. In (b), two unlearning curves have been compared. "Numerical swap LM" has been obtained applying numerical swaps only to the LM and "numerical swap TM" applying numerical swaps only to the LM.

limitations to performance are a crucial part of the machine learning method, and one where statistics and algorithms interact closely.

Using state-of-the-art phrase-based statistical machine translation packages and high-performance computing, we derived very accurate learning curves for a number of language pairs and domains. Our results suggest that performance, as measured by BLEU, increases by a constant factor for each doubling of the data. Although that factor varies depending on corpus and language pair, this result seems consistent over all experimental conditions we tried. Our findings confirm the results reported, for example, by [30, 32]. Authors in [30] reported "almost linear" improvements in BLEU score by doubling the training set size. In the presentation [32], the claim is that BLEU increases with each doubling of the training set size, by 0.5 and 2.5 BLEU points for the language and translation models, respectively, in the context of Arabic-English translation. Both claims seem to be qualitatively compatible with our observation of improvements that are at best logarithmic in the training set size, although our estimates are closer to 1.4 and 1.9 BLEU points for the LM and TM, respectively.

Our findings are also consistent with the curves presented by [33], although their results are limited to a much lower data set size (less than 10^5 sentences) and presented on a linear scale. Incidentally, that paper also presents a recent attempt into using active learning for improving MT and meets the challenge of "diminishing returns" identified in the learning curves: a constant performance improvement requires increasing amounts of data. Active learning is well-known in machine learning as an attempt to bypass the distribution-sampling limitations by actively seeking new examples. One way to achieve this could be to either introduce an oracle to which the system can ask for annotation when needed or a process that uses linguistic knowledge to create new table entries based on existing table entries and some grammatical rules. An oracle could be formed by a web agent capable of locating useful bilingual sentences, for any given task, or even a linguistic-based system that could turn SMT models into richer ones by essentially generating new entries and removing unreliable ones. Any way to enforce linguistic constraints might result in a reduced need for data, and ultimately in more complete models, given the same corpus [34]. Neither approach would change the statistical nature of the system, but they would help it bypass the phrase acquisition bottleneck. Note however that the development of active learning systems for statistical MT raises the issue of confidence estimation, which is typically at the heart of many classical active learning procedures. This is an active area of research in machine translation [35–37].

The results of the perturbation analysis in Section 5 suggest that the limiting factor in the translation tables is not in the numeric part of the model—the parameters being estimated—but in the phrases contained in it, the entries of the phrase table. Together these observations point to limitations to the phrase-acquisition process, that under i.i.d. conditions is controlled by a Zipf law and hence leads to very slow rate of discovery of new phrases. In other words, the essential limiting factor for phrase-based SMT systems seems to be the Zipf law found in natural language.

Our large-scale analysis also suggests that the current bottleneck of the phrase-based SMT approach is the lack of

sufficient data, not the function class used for the representation of translation systems. We reach this conclusion by noting that within a fixed class of parameters, an instance of the model has been trained on a set of sentences that includes also the test set (while limiting the phrase length to 7 words, to prevent full-sentence memorization), and this yields much improved performance. In addition, we also observed that the larger phrases (beyond 5 words) are in fact seldom used.

The high performance observed in the train-on-test conditions shows that there exists at least one choice of tunable parameters with which the phrase-based translation system can deliver much higher performance. This is useful to bound the space of "possible performances," although in ideal situations. The clear gap between performances on training and test set, together with the rate of the learning curves and our perturbation experiments, suggest that improvements in BLEU score are theoretically possible, if the right entries were present in the translation and language models. Unfortunately, current estimation procedures are unable to reach such high-performing regions of the parameter space. This was also noted by a recent paper by Wisniewski et al. who note that "*the current bottleneck of translation performances is not the representation power of the [phrase-based translation systems] but rather in their scoring functions*" [38].

Finally, let us note that we have not addressed the thorny issue of the reliability of automatic MT metrics. Some will be quick to point out that maximizing, for example, BLEU may neither be necessary for, nor guarantee good translation performance. Although we acknowledge that automatic MT metrics may not tell the whole story as far as translation quality is concerned, our systematic study aims at characterizing the behaviour of SMT systems that are built by maximizing such metrics. Deriving learning curves using human-generated translation quality score would definitely be interesting, but we are not aware of such effort, which would currently involve an overwhelming amount of human annotation.

7. Conclusion

Data-driven solutions to classic AI problems are now commonplace, ranging from computer vision to information retrieval tasks, and machine translation is one of the main successes of this approach. The idea of putting learning systems at the centre of all AI methodologies introduces however the need to understand the properties and limitations of these learning components. In this study, we have produced very accurate learning curves for the class of phrase-based SMT systems, using different implementations, different language pairs, and different datasets. In each case, we found the same overall behaviour, of a logarithmic growth in performance with training set size. The question becomes as follows: on which aspect of these systems should we act to achieve better performance? We have performed an extensive series of experiments to separately measure how different factors affect the performance of phrase-based SMT systems. Our first concerns were to distinguish between approximation and estimation error: the performance limitations due to the use of a limited language model versus those due to the need to estimate the parameters of that model from a finite sample. Our experiments show that the estimation part of the error is the dominant one, suggesting that performance can still improve if the appropriate entries were available in the language and translation models. The second concern was to distinguish between the role of the numerical and lexical parts in the language and translation models. Various perturbation experiments show that the accuracy in estimating the numerical parameters is not a crucial aspect of performance, while the estimation of the lexical parts of the tables is a major factor in determining performance. In a third set of experiments, we determined that the estimation of the translation model has a bigger effect than the estimation of the language model, on performance.

We therefore reach the conclusion that estimating entries in the phrase translation tables is the dominant factor in determining performance. What controls the creation of phrase-translation tables? This is mostly limited by Zipf's law, since the probability of encountering phrases that have not been seen in the training set does not vanish even after observing very large corpora. The question therefore becomes as follows: how can we fill translation tables with phrase pairs while Zipf's law seems to prevent us from generating them from the data alone? Among possible methods, two stand out as particularly promising. The first is to generate phrase pairs by using grammatical or various linguistic rules (e.g., turning existing entries into new entries, by applying various forms of inflection). The second is to allow the system to make queries, active learning style, in order to produce phrase-table entries without having to wait for them to appear by sampling additional textual data. Both of these ideas of course are being pursued at the moment. It is of course important to remark that these limitations only refer to the current systems, where language is modelled as a Markov chain, and by entirely changing language model, different limitations could be found.

Appendix

A. Supplementary Results

A.1. Effect of Data Size in Optimization Set. In this section, we study the role of the optimization/development set with regard to the quality of translation. In particular, we analyze how different sizes of the development set affect the performance and the computational cost of the optimization phase.

The whole Europarl training set has been randomly split into two parts without replacements. One, containing 1,159,914 pairs of sentences, has been used to train the model. This step has been done only once, and all the experiments use the same translation, language, and reordering models. The second set has 100,000 pairs of sentences, and it is used to randomly select the development sets. The test set contains 2,000 pairs of sentences and is the same for all the experiments.

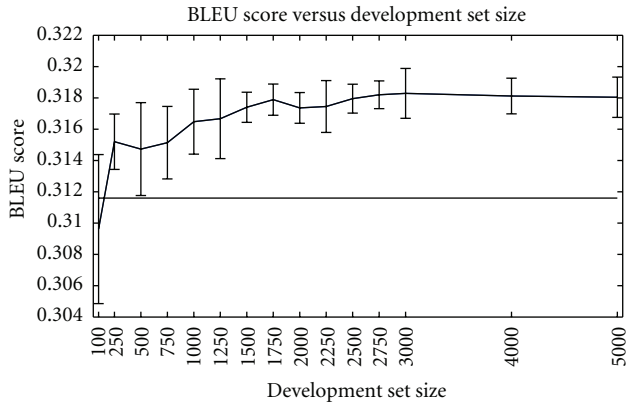

Figure 8: BLEU score versus development set size. The horizontal line is the BLEU score without any optimization. It is equal to 0.3116.

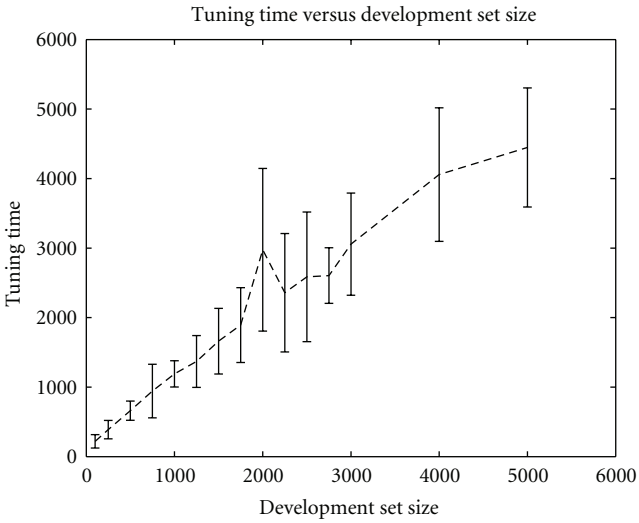

Figure 9: Tuning CPU times versus development set size.

Different sizes of the development set (100, 250, 500, 750, 1,000, 1,250, 1,500, 1,750, 2,000, 2,250, 2,500, 2,750, 3,000, 4,000, and 5,000 pairs of sentences) have been chosen, focusing our attention on small, rather than large dimensions. Replacements are not allowed. These choices also depend on the high computational cost of the tuning algorithm. For each size, ten random sets have been selected. For each set, an instance of the system has been run. The optimized model is used to test, and the results are evaluated.

In Figure 8, BLEU score as function of the development size is reported. The optimization procedure increases the quality of the translations. This improvement does not seem to be significant after a certain size of the development set. In fact, when we increase the development set size beyond 2,000 sentence pairs, BLEU does not change significantly. On the other hand, optimization is really expensive in terms of computational cost. In Figure 9, it increases roughly linearly with the development set size. It is nice to note how the computational time is strongly related to the number of optimization steps in Figure 10.

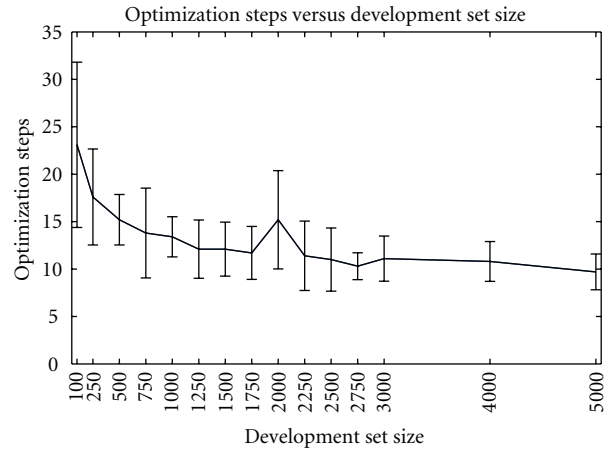

Figure 10: Optimization steps versus development set size.

A.2. Role of the Unknown Words. Understanding the most important reasons for failure of a PBSMT system is a fundamental task. In [39], a classification of different types of error has been proposed. In this section, we focus our attention on a particular type of error: unknown words. This type of error is considered a source error because it depends on the source sentence and not the translation process. It has been distinguished between truly unknown words (or stems) and unseen forms of known stems. The unknown words are the direct effect of Zipf's law in a language, as new words can come, but the training set is not flexible enough to cover them.

We have created 10 random subsets for each of the 10 chosen sizes, where each size represents 10%, 20%, and so forth of the complete corpus. For each subset, a new instance of the PBSMT system has been created, for a total of 100 models. Each model has been tested on the test set and on a subset of 2,000 pairs the training set. The optimization step has not been run. For each model, we count the unknown words.

Figure 11 shows unknown words as function of the training model. It is clear that small training sets are able to cover a small part of the word space. When increasing the dimension of the training set, the number of unknown words decreases. These curves reflect how machine translation is strongly affected by Zipf's law and confirm the results of the previous sections. A briefly discussion about the presence of unknown words when we test on a subset of the training set is given by Section 4.3.

A.3. Role of Test Set Size on Measuring Performance. BLEU score, the metric used in this work to evaluate the quality of the translation, is test set dependent. It means that different test sets regardless of the dimension can produce variation in the value of the BLEU score. In this section, we investigate how BLEU is affected by the test set size.

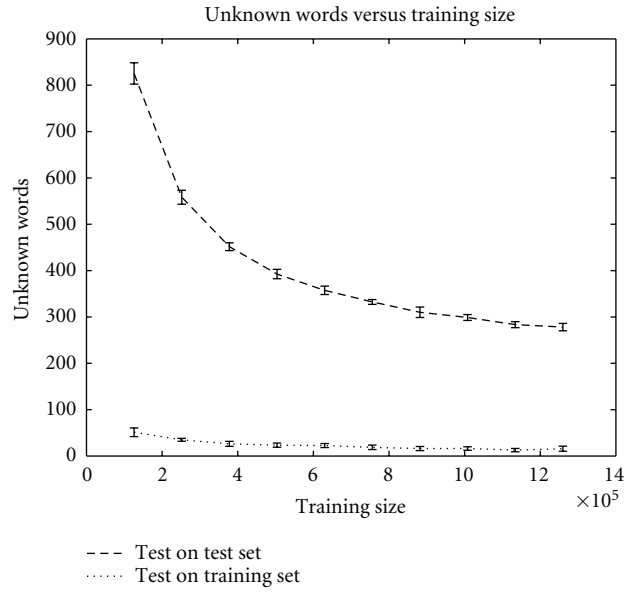

Figure 11: Number of unknown words translating training and test sets versus training set size.

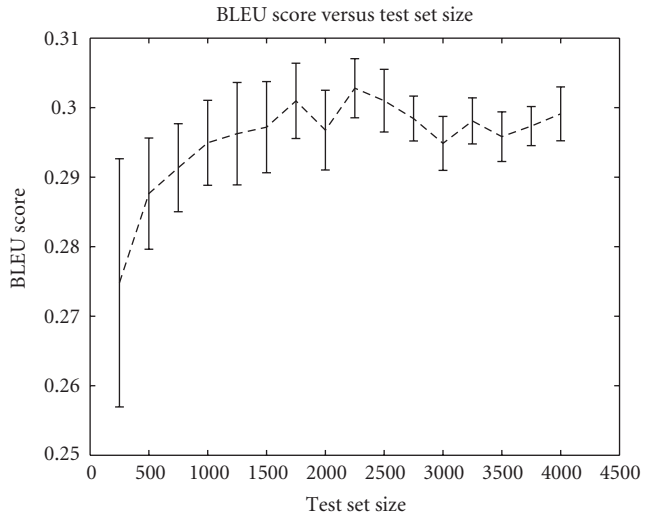

Figure 12: BLEU score versus test set size.

We have isolated 4,000 pairs of sentences from the Europarl training set, and we have selected from the remaining part 629,957 pairs. A Moses model is trained using this set. Using the 4,000 sentences pairs, we have created 10 random subsets for each of the 16 chosen sizes, where each size can contain a number of pairs from 250 to 4,000 by a step of 250 pairs. The model is tested over all these subsets, and the learning curve is reported in Figure 12. Small test set sizes produce a big variance in BLEU score. When increasing the test set size, the error bars tend to reduce.

In this work, a test set with 2,000 sentences pairs has been used. In the learning curve in Figure 12, the average value and standard deviation for this size are equal to 0.29677 ± 0.0057. This implies that differences in BLEU score smaller than 1.9% using two different test sets of the same size depend on the test set choice and not on different techniques. In recent years, this trouble has been partially solved using a standard test set obtained by the Europarl corpus.

In Section 3.4, we report the correlation coefficient between all the measures. Each correlation coefficient is computed using the results of the 160 experiments described above.

Acknowledgments

The authors thank Callum Wright, Bristol HPC Systems Administrator (http://www.acrc.bris.ac.uk/acrc/hpc.htm), Moses mailing list, Philip Koehn for various advices and for making available the Europarl corpus, Alessia Mammone, Omar Ali, and George Foster for many useful comments and for help with Portage. This work is supported by the EU IST Project SMART (FP6-033917).

References

[1] W. Locke and A. Booth, *Machine Translation of Languages*, MIT Press, Cambridge, Mass, USA, 1955.

[2] P. F. Brown, S. D. Pietra, V. J. D. Pietra, and R. L. Mercer, "The mathematic of statistical machine translation: parameter estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1994.

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pp. 48–54, Edmonton, Canada, 2003.

[4] R. Zens, F.-J. Och, and H. Ney, "Phrase-based statistical machine translation," in *Proceedings of the 25th Annual German Conference on AI (KI '02)*, pp. 18–32, Springer, London, UK, 2002.

[5] P. Koehn, H. Hoang, A. Birch et al., "Moses: open source toolkit for statistical machine translation," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics, Demonstration Session*, pp. 177–180, Columbus, Ohio, USA, 2007.

[6] N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson, "NRC's PORTAGE system for WMT," in *Proceedings of ACL 2nd Workshop on Statistical Machine Translation*, pp. 185–188, Prague, Czech Republic, 2007.

[7] P. Koehn, "Europarl: a parallel corpus for statistical machine translation," in *Proceedings of the 10 th Machine Translation Summit*, pp. 79–86, Phuket, Thailand, 2005.

[8] C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, "Findings of the 2009 Workshop on Statistical Machine Translation," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder, Eds., pp. 263–311, Association for Computational Linguistics, Athens, Greece, 2009.

[9] M. Turchi, T. DeBie, and N. Cristianini, "Learning performance of a machine translation system: a statistical and computational analysis," in *Proceedings of the 3rd Workshop on Statistical Machine Translation*, pp. 35–43, Columbus, Ohio, USA, 2008.

[10] Y. Al-Onaizan, J. Curin, M. Jahr et al., "Statistical machine translation: final report," Tech. Rep., Johns Hopkins University, Summer Workshop on Language Engineering, Center for Speech and Language Processing, Baltimore, Md, USA, 1999.

[11] F. J. Och and H. Weber, "Improving statistical natural language translation with categories and rules," in *Proceedings of the 17th International Conference on Computational Linguistics*, vol. 2, pp. 985–989, Stroudsburg, Pa, USA, 1998.

[12] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 295–302, Philadelphia, Pa, USA, 2001.

[13] Y. Liu, Q. Liu, and S. Lin, "Discriminativeword alignment by linear modeling," *Computational Linguistics*, vol. 36, no. 3, pp. 303–339, 2010.

[14] R. C. Moore, W.-T. Yih, and A. Bode, "Improved discriminative bilingual word alignment," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL '06)*, pp. 513–520, Sydney, Australia, 2006.

[15] A. Fraser and D. Marcu, "Measuring word alignment quality for statistical machine translation," *Computational Linguistics*, vol. 33, no. 3, pp. 293–303, 2007.

[16] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pa, USA, 2001.

[17] F. J. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pp. 160–167, Sapporo, Japan, 2003.

[18] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++*, Cambridge University Press, Cambridge, Mass, USA, 2002.

[19] A. Arun and P. Koehn, "Online learning methods for discriminative training of phrase based statistical machine translation," in *Proceedings of 11th the Machine Translation Summit*, Copenhagen, Denmark, 2007.

[20] D. Chiang, Y. Marton, P. Resnik, and P. Resnik, "Online large-margin training of syntactic and structural translation features," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*, pp. 224–233, Association for Computational Linguistics, Stroudsburg, Pa, USA, 2008.

[21] B. Efron and R. J. Tibshirani, *An Introuction to the Bootstrap*, CRC Press, Boca Raton, Fla, USA, 1994.

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

[23] A. Stolcke, "Srilm—an extensible language modeling toolkit," in *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colo, USA, 2002.

[24] F. J. Och and H. Ney, "A systematic comparison of various statistical alignment models," *Computational Linguistics*, vol. 29, no. 1, pp. 19–51, 2003.

[25] G. Doddington, "Automatic evaluation of machine translation quality using n-gram co-occurrence statistics," in *Proceedings of the 2nd International Conference on Human Language Technology Research*, pp. 138–145, Morgan Kaufmann Publishers, San Francisco, Calif, USA, 2002.

[26] S. Banerjee and A. Lavie, "Meteor: an automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Ann Arbor, Mich, USA, 2005.

[27] A. Lavie and A. Agarwal, "Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments," in *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*, pp. 228–231, Association for Computational Linguistics, Prague, Czech Republic, 2007.

[28] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A study of translation edit rate with targeted human annotation," in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pp. 223–231, Association for Machine Translation in the Americas, Los Angeles, Calif, USA, 2006.

[29] R. Steinberger, B. Pouliquen, A. Widiger et al., "The jrc-acquis: a multilingual aligned parallel corpus with 20+ languages," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 2142–2147, Genova, Italy, 2006.

[30] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pp. 858–867, Association for Computational Linguistics, Prague, Czech Republic, 2007.

[31] N. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, NY, USA, 1981.

[32] F. J. Och, "Statistical machine translation: foundations and recent advances," in *Proceedings of the Tutorial at MT Summit*, 2005.

[33] M. Bloodgood and C. Callison-Burch, "Bucking the trend: large-scale cost-focused active learning for statistical machine translation," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL '10)*, pp. 854–864, Association for Computational Linguistics, Stroudsburg, Pa, USA, 2010.

[34] P. Koehn and H. Hoang, "Factored translation models," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pp. 868–876, 2007.

[35] J. Blatz, E. Fitzgerald, G. Foster et al., "Confidence estimation for machine translation," in *Proceedings of the 20th international Conference on Computational Linguistics (COLING '04)*, vol. 315, Association for Computational Linguistics, Morristown, NJ, USA, 2004.

[36] L. Specia, M. Turchi, Z. Wang, J. Shawe-Taylor, and C. Saunders, "Improving the confidence of machine translation quality estimates," in *Proceedings of 12th the Machine Translation Summit*, 2009.

[37] N. Ueffing and H. Ney, "Word-level confidence estimation for machine translation using phrase-based translation models," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*, pp. 763–770, Association for Computational Linguistics, Morristown, NJ, USA, 2005.

[38] G. Wisniewski, A. Allauzen, and F. Yvon, "Assessing phrase-based translation models with oracle decoding," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pp. 933–943, Association for Computational Linguistics, Stroudsburg, Pa, USA, 2010.

[39] D. Vilar, J. Xu, L. F. D'Haro, and H. Ney, "Error analysis of statistical machine translation output," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, Genova, Italy, 2006.