

Assessing Phrase-Based Translation Models with Oracle Decoding

Guillaume Wisniewski and Alexandre Allauzen and François Yvon

Univ. Paris Sud ; LIMSI—CNRS

91403 ORSAY CEDEX

France

{wisniews, allauzen, yvon}@limsi.fr

Abstract

Extant Statistical Machine Translation (SMT) systems are very complex softwares, which embed multiple layers of heuristics and embark very large numbers of numerical parameters. As a result, it is difficult to analyze output translations and there is a real need for tools that could help developers to better understand the various causes of errors.

In this study, we make a step in that direction and present an attempt to evaluate the quality of the phrase-based translation model. In order to identify those translation errors that stem from deficiencies in the phrase table (PT), we propose to compute the *oracle BLEU-4 score*, that is the best score that a system based on this PT can achieve on a reference corpus. By casting the computation of the oracle BLEU-1 as an Integer Linear Programming (ILP) problem, we show that it is possible to efficiently compute accurate lower-bounds of this score, and report measures performed on several standard benchmarks. Various other applications of these oracle decoding techniques are also reported and discussed.

1 Phrase-Based Machine Translation

1.1 Principle

A Phrase-Based Translation System (PBTS) consists of a ruleset and a scoring function (Lopez, 2009). The ruleset, represented in the phrase table, is a set of phrase¹ pairs $\{(f, e)\}$, each pair expressing that the source phrase f can be rewritten (translated) into a target phrase e . *Translation hypotheses* are generated by iteratively rewriting portions of the source sentence as prescribed by the ruleset, until each source word has been consumed by exactly one rule. The order of target words in an hypothesis is uniquely determined by the order in which the rewrite operation are performed. The *search space* of the translation model corresponds to the set of all possible sequences of

rules applications. The scoring function aims to rank all possible translation hypotheses in such a way that the best one has the highest score.

A PBTS is learned from a parallel corpus in two independent steps. In a first step, the corpus is aligned at the word level, by using alignment tools such as Giza++ (Och and Ney, 2003) and some symmetrisation heuristics; phrases are then extracted by other heuristics (Koehn et al., 2003) and assigned numerical weights. In the second step, the parameters of the scoring function are estimated, typically through Minimum Error Rate training (Och, 2003).

Translating a sentence amounts to finding the best scoring translation hypothesis in the search space. Because of the combinatorial nature of this problem, translation has to rely on heuristic search techniques such as greedy hill-climbing (Germann, 2003) or variants of best-first search like multi-stack decoding (Koehn, 2004). Moreover, to reduce the overall complexity of decoding, the search space is typically pruned using simple heuristics. For instance, the state-of-the-art phrase-based decoder Moses (Koehn et al., 2007) considers only a restricted number of translations for each source sequence² and enforces a distortion limit³ over which phrases can be re-ordered. As a consequence, the best translation hypothesis returned by the decoder is not always the one with the highest score.

1.2 Typology of PBTS Errors

Analyzing the errors of a SMT system is not an easy task, because of the number of models that are combined, the size of these models, and the high complexity of the various decision making processes. For a SMT system, three different kinds of errors can be distinguished (Germann et al., 2004; Auli et al., 2009): *search errors*, *induction errors* and *model errors*. The former corresponds to cases where the hypothesis with the best score is missed by the search procedure, either because of the use of an ap-

¹Following the usage in statistical machine translation literature, we use “phrase” to denote a subsequence of consecutive words.

²the `ttl` option of Moses, defaulting to 20.

³the `d1` option of Moses, whose default value is 7.

proximate search method or because of the restrictions of the search space. *Induction errors* correspond to cases where, given the model, the search space does not contain the reference. Finally, *model errors* correspond to cases where the hypothesis with the highest score is not the best translation according to the evaluation metric.

Model errors encompass several types of errors that occur during learning (Bottou and Bousquet, 2008)⁴. *Approximation errors* are errors caused by the use of a restricted and oversimplistic class of functions (here, finite-state transducers to model the generation of hypotheses and a linear scoring function to discriminate them) to model the translation process. *Estimation errors* correspond to the use of sub-optimal values for both the phrase pairs weights and the parameters of the scoring function. The reasons behind these errors are twofold: first, training only considers a finite sample of data; second, it relies on error prone alignments. As a result, some “good” phrases are extracted with a small weight, or, in the limit, are not extracted at all; and conversely that some “poor” phrases are inserted into the phrase table, sometimes with a really optimistic score.

Sorting out and assessing the impact of these various causes of errors is of primary interest for SMT system developers: for lack of such diagnoses, it is difficult to figure out which components of the system require the most urgent attention. Diagnoses are however, given the tight intertwining among the various component of a system, very difficult to obtain: most evaluations are limited to the computation of global scores and usually do not imply any kind of failure analysis.

1.3 Contribution and organization

To systematically assess the impact of the multiple heuristic decisions made during training and decoding, we propose, following (Dreyer et al., 2007; Auli et al., 2009), to work out *oracle scores*, that is to evaluate the best achievable performances of a PBTS. We aim at both studying the expressive power of PBTS and at providing tools for identifying and quantifying causes of failure.

Under standard metrics such as BLEU (Papineni et al., 2002), oracle scores are difficult (if not impossible) to compute, but, by casting the computation of the oracle unigram recall and precision as an Integer Linear Programming (ILP) problem, we show that it is possible to efficiently compute accurate lower-bounds of the oracle BLEU-4 scores and report measurements performed on several standard benchmarks.

The main contributions of this paper are twofold. We first introduce an ILP program able to efficiently find the best hypothesis a PBTS can achieve. This program can be easily extended to test various improvements to

⁴We omit here *optimization errors*.

phrase-base systems or to evaluate the impact of different parameter settings. Second, we present a number of complementary results illustrating the usage of our oracle decoder for identifying and analyzing PBTS errors. Our experimental results confirm the main conclusions of (Turchi et al., 2008), showing that extant PBTs have the potential to generate hypotheses having very high BLEU-4 score and that their main bottleneck is their scoring function.

The rest of this paper is organized as follows: in Section 2, we introduce and formalize the oracle decoding problem, and present a series of ILP problems of increasing complexity designed so as to deliver accurate lower-bounds of oracle score. This section closes with various extensions allowing to model supplementary constraints, most notably reordering constraints (Section 2.5). Our experiments are reported in Section 3, where we first introduce the training and test corpora, along with a description of our system building pipeline (Section 3.1). We then discuss the baseline oracle BLEU scores (Section 3.2), analyze the non-reachable parts of the reference translations, and comment several complementary results which allow to identify causes of failures. Section 4 discuss our approach and findings with respect to the existing literature on error analysis and oracle decoding. We conclude and discuss further prospects in Section 5.

2 Oracle Decoder

2.1 The Oracle Decoding Problem

Definition To get some insights on the errors of phrase-based systems and better understand their limits, we propose to consider the *oracle decoding problem* defined as follows: given a source sentence, its reference translation⁵ and a phrase table, what is the “best” translation hypothesis a system can generate? As usual, the quality of an hypothesis is evaluated by the similarity between the reference and the hypothesis. Note that in the oracle decoding problem, we are only assessing the ability of PBT systems to generate good candidate translations, irrespective of their ability to score them properly.

We believe that studying this problem is interesting for various reasons. First, as described in Section 3.4, comparing the best hypothesis a system could have generated and the hypothesis it actually generates allows us to carry on both quantitative and qualitative failure analysis. The oracle decoding problem can also be used to assess the *expressive power* of phrase-based systems (Auli et al., 2009). Other applications include computing acceptable pseudo-references for discriminative training (Tillmann and Zhang, 2006; Liang et al., 2006; Arun and

⁵The oracle decoding problem can be extended to the case of multiple references. For the sake of simplicity, we only describe the case of a single reference.

Koehn, 2007) or combining machine translation systems in a multi-source setting (Li and Khudanpur, 2009). We have also used oracle decoding to identify erroneous or difficult to translate references (Section 3.3).

Evaluation Measure To fully define the oracle decoding problem, a measure of the similarity between a translation hypothesis and its reference translation has to be chosen. The most obvious choice is the BLEU-4 score (Papineni et al., 2002) used in most machine translation evaluations.

However, using this metric in the oracle decoding problem raises several issues. First, BLEU-4 is a metric defined at the corpus level and is hard to interpret at the sentence level. More importantly, BLEU-4 is not decomposable⁶: as it relies on 4-grams statistics, the contribution of each phrase pair to the global score depends on the translation of the previous and following phrases and can not be evaluated in isolation. Because of its non-decomposability, maximizing BLEU-4 is hard; in particular, the phrase-level decomposability of the evaluation metric is necessary in our approach.

To circumvent this difficulty, we propose to evaluate the similarity between a translation hypothesis and a reference by the number of their common words. This amounts to evaluating translation quality in terms of unigram precision and recall, which are highly correlated with human judgements (Lavie et al.,). This measure is closely related to the BLEU-1 evaluation metric and the Meteor (Banerjee and Lavie, 2005) metric (when it is evaluated without considering near-matches and the distortion penalty). We also believe that hypotheses that maximize the unigram precision and recall at the sentence level yield corpus level BLEU-4 scores close the maximal achievable. Indeed, in the setting we will introduce in the next section, BLEU-1 and BLEU-4 are highly correlated: as all correct words of the hypothesis will be compelled to be at their correct position, any hypothesis with a high 1-gram precision is also bound to have a high 2-gram precision, etc.

2.2 Formalizing the Oracle Decoding Problem

The oracle decoding problem has already been considered in the case of word-based models, in which all translation units are bound to contain only one word. The problem can then be solved by a *bipartite graph matching* algorithm (Leusch et al., 2008): given a $n \times m$ binary matrix describing possible translation links between source words and target words⁷, this algorithm finds the subset of links maximizing the number of words of the reference that have been translated, while ensuring that each word

is translated only once.

Generalizing this approach to phrase-based systems amounts to solving the following problem: given a set of possible translation links between potential phrases of the source and of the target, find the subset of links so that the unigram precision and recall are the highest possible. The corresponding oracle hypothesis can then be easily generated by selecting the target phrases that are aligned with one source phrase, disregarding the others. In addition, to mimic the way OOVs are usually handled, we match identical OOV tokens appearing both in the source and target sentences. In this approach, the unigram precision is always one (every word generated in the oracle hypothesis matches exactly one word in the reference). As a consequence, to find the oracle hypothesis, we just have to maximize the recall, that is the number of words appearing both in the hypothesis and in the reference.

Considering phrases instead of isolated words has a major impact on the computational complexity: in this new setting, the optimal segmentations in phrases of both the source and of the target have to be worked out in addition to links selection. Moreover, constraints have to be taken into account so as to enforce a proper segmentation of the source and target sentences. These constraints make it impossible to use the approach of (Leusch et al., 2008) and concur in making the oracle decoding problem for phrase-based models more complex than it is for word-based models: it can be proven, using arguments borrowed from (De Nero and Klein, 2008), that this problem is NP-hard even for the simple unigram precision measure.

2.3 An Integer Program for Oracle Decoding

To solve the combinatorial problem introduced in the previous section, we propose to cast it into an Integer Linear Programming (ILP) problem, for which many generic solvers exist. ILP has already been used in SMT to find the optimal translation for word-based (Germann et al., 2001) and to study the complexity of learning phrase alignments (De Nero and Klein, 2008) models. Following the latter reference, we introduce the following variables: $f_{i,j}$ (resp. $e_{k,l}$) is a binary indicator variable that is true when the phrase contains all spans from between-word position i to j (resp. k to l) of the source (resp. target) sentence. We also introduce a binary variable, denoted $a_{i,j,k,l}$, to describe a possible link between source phrase $f_{i,j}$ and target phrase $e_{k,l}$. These variables are built from the entries of the phrase table according to *selection strategies* introduced in Section 2.4. In the following, index variables are so that:

$$0 \leq i < j \leq n, \text{ in the source sentence and} \\ 0 \leq k < l \leq m, \text{ in the target sentence,}$$

⁶Neither at the sentence (Chiang et al., 2008), nor at the phrase level.

⁷The (i, j) entry of the matrix is 1 if the i^{th} word of the source can be translated by the j^{th} word of the reference, 0 otherwise.

where n (resp. m) is the length of the source (resp. target) sentence.

Solving the oracle decoding problem then amounts to optimizing the following objective function:

$$\max_{i,j,k,l} \sum_{i,j,k,l} a_{i,j,k,l} \cdot (l - k), \quad (1)$$

under the constraints:

$$\forall x \in \llbracket 1, m \rrbracket : \sum_{k,l \text{ s.t. } k \leq x \leq l} e_{k,l} \leq 1 \quad (2)$$

$$\forall y \in \llbracket 1, n \rrbracket : \sum_{i,j \text{ s.t. } i \leq y \leq j} f_{i,j} = 1 \quad (3)$$

$$\forall k, l : \sum_{i,j} a_{i,j,k,l} = f_{k,l} \quad (4)$$

$$\forall i, j : \sum_{k,l} a_{i,j,k,l} = e_{i,j} \quad (5)$$

The objective function (1) corresponds to the number of target words that are generated. The first set of constraints (2) ensures that each word in the reference \mathbf{e} appears in no more than one phrase. Maximizing the objective under these constraints amounts to maximizing the unigram recall. The second set of constraints (3) ensures that each word in the source \mathbf{f} is translated exactly once, which guarantees that the search space of the ILP problem is the same as the search space of a phrase-based system. Constraints (4) bind the $f_{k,l}$ and $a_{i,j,k,l}$ variables, ensuring that whenever a link $a_{i,j,k,l}$ is active, the corresponding phrase $f_{k,l}$ is also active. Constraints (5) play a similar role for the reference.

The Relaxed Problem Even though it accurately models the search space of a phrase-based decoder, this program is not really useful as is: due to out-of-vocabulary words or missing entries in the phrase table, the constraint that all source words should be translated yields infeasible problems⁸. We propose to relax this problem and allow some source words to remain untranslated. This is done by replacing constraints (3) by:

$$\forall y \in \llbracket 1, n \rrbracket : \sum_{i,j \text{ s.t. } i \leq y \leq j} f_{i,j} \leq 1$$

To better reflect the behavior of phrase-based decoders, which attempt to translate all source words, we also need to modify the objective function as follows:

$$\sum_{i,j,k,l} a_{i,j,k,l} \cdot (l - k) + \sum_{i,j} f_{i,j} \cdot (j - i) \quad (6)$$

The second term in this new objective ensures that optimal solutions translate as many source words as possible.

⁸An ILP problem is said to be *infeasible* when every possible solution violates at least one constraint.

The Relaxed-Distortion Problem A last caveat with the `Relaxed` optimization program is caused by frequently occurring source tokens, such as function words or punctuation signs, which can often align with more than one target word. For lack of taking distortion information into account in our objective function, all these alignments are deemed equivalent, even if some of them are clearly more satisfactory than others. This situation is illustrated on Figure 1.

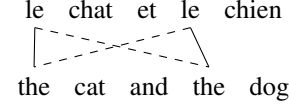


Figure 1: Equivalent alignments between “le” and “the”. The dashed lines corresponds to a less interpretable solution.

To overcome this difficulty, we propose a last change to the objective function:

$$\sum_{i,j,k,l} a_{i,j,k,l} \cdot (l - k) + \sum_{i,j} f_{i,j} \cdot (j - i) - \alpha \sum_{i,j,k,l} a_{i,j,k,l} |k - i| \quad (7)$$

Compared to the objective function of the relaxed problem (6), we introduce here a supplementary penalty factor which favors monotonous alignments. For each phrase pair, the higher the difference between source and target positions, the higher this penalty. If α is small enough, this extra term allows us to select, among all the optimal alignments of the `relaxed` problem, the one with the lowest distortion. In our experiments, we set α to $\min\{n, m\}$ to ensure that the penalty factor is always smaller than the reward for aligning two single words.

2.4 Selecting Indicator Variables

In the approach introduced in the previous sections, the oracle decoding problem is solved by selecting, among a set of possible translation links, the ones that yield the solution with the highest unigram recall.

We propose two strategies to build this set of possible translation links. In the first one, denoted *exact match*, an indicator $a_{i,j,k,l}$ is created if there is an entry (f, e) so that f spans from word position i to j in the source and e from word position k to l in the target. In this strategy, the ILP program considers exactly the same ruleset as conventional phrase-based decoders.

We also consider an alternative strategy, which could help us to identify errors made during the phrase extraction process. In this strategy, denoted *inside match*, an indicator $a_{i,j,k,l}$ is created when the following three criteria are met: *i*) f spans from position i to j of the source; *ii*) a *substring* of e , denoted \bar{e} , spans from position k to l

of the reference; *iii*) (f, \bar{e}) is not an entry of the phrase table. The resulting set of indicator variables thus contains, at least, all the variables used in the exact match strategy. In addition, we license here the use of phrases containing words that do not occur in the reference. In fact, using such solutions can yield higher BLEU scores when the reward for additional correct matches exceeds the cost incurred by wrong predictions. These cases are symptoms of situations where the extraction heuristic failed to extract potentially useful subphrases.

2.5 Oracle Decoding with Reordering Constraints

The ILP problem introduced in the previous section can be extended in several ways to describe and test various improvements to phrase-based systems or to evaluate the impact of different parameter settings. This flexibility mainly stems from the possibility offered by our framework to express arbitrary constraints over variables. In this section, we illustrate these possibilities by describing how reordering constraints can easily be considered.

As a first example, the Moses decoder uses a distortion limit to constrain the set of possible reorderings. This constraint “enforces (...) that the last word of a phrase chosen for translation cannot be more than d^9 words from the leftmost untranslated word in the source” (Lopez, 2009) and is expressed as:

$$\forall a_{ijkl}, a_{i'j'k'l'} \text{ s.t. } k > k', \\ a_{ijkl} \cdot a_{i'j'k'l'} \cdot |j - i' + 1| \leq d,$$

The maximum distortion limit strategy (Lopez, 2009) is also easily expressed and take the following form (assuming this constraint is parameterized by d):

$$\forall l < m - 1, \\ a_{i,j,k,l} \cdot a_{i',j',l+1,l'} \cdot |i' - j - 1| < d$$

Implementing the “local” or MJ-d (Kumar and Byrne, 2005) reordering strategy is also straightforward, and implies using the following constraints:

$$\forall i, k, \left| \sum_{i' \leq i} a_{i',j',k',l'} - \sum_{k' \leq k} a_{i',j',k',l'} \right| \leq d$$

Similarly, It is possible to simulate decoding under the so-called IBM(d) reordering constraints¹⁰ by considering the following constraints:

$$\forall \mu \leq m, \max_{\substack{i,k,l \\ j \leq \mu}} a_{i,j,k,l} \cdot j - \sum_{i,j,k,l} a_{i,j,k,l} \cdot (j - i) \leq d$$

⁹This corresponds to the dl parameter of Moses

¹⁰Under IBM(d) constraints, the translation is done, phrase by phrase, from the beginning of the sentence until the end and only one of the first d untranslated phrase can be selected for translation.

In these constraints, the first factor corresponds to the rightmost translated word of the source and the second one to the number of translated source words. The constraints simply enforce that, at each step of the decoding, there are no more than d source words that were skipped.

Note that the constraints introduced above are not all linear in the problem variables; however they can easily be linearized using standard ILP techniques (Roth and Yih, 2005).

3 Oracle Decoding for Failure Analysis

3.1 Experimental Setting

We propose to use our oracle decoder to study the ability of a PBTS to translate from English to French and from German to English. These two languages pairs present different challenges: English to French translation is considered a relatively easy pair, notwithstanding the difficulties of generating the right inflection marks in French. Translating from German into English is more difficult, notably due to the productivity of inflectional and compounding processes in German, and also to significant differences in word ordering between these languages.

Our experiments are based on the corpora distributed for the WMT'09 constrained tasks (Callison-Burch et al., 2009). All data are tokenized, cleaned and converted to lowercase letters using the tools provided by the organizers. We then used a standard training pipeline to construct the translation model: the bitexts were aligned using Giza++¹¹, symmetrized using the grow-diag-final-and heuristic; the phrase table was extracted and scored using the tools distributed with Moses.¹² Finally, baseline systems were optimized using WMT'08 test set as development using MERT. Note that, for all these steps, we used the default value of the various parameters. The extracted phrase table is then used to find the oracle alignment on the task test set. Recall that oracle decoding do not use the scores estimated by Moses in any way.

In the experiments reported below, two settings are considered. In the first one, denoted NEWS-CO, Moses was trained only on a small data set taken from the News Commentary corpus. Using a small sized corpus reduces both training time and decoding time, which allows us to quickly test different configurations of the decoder. In a second setting, denoted EURO-PARL, Moses was trained on a larger corpora containing the entirety of the Europarl Corpus, but no in-domain data, to provide results on more realistic conditions. Statistics regarding the different corpora used are reported in Table 1. These statistics are computed on the lowercase cleaned corpora.

¹¹<http://www.fjoch.com/GIZA++.html>

¹²<http://statmt.org/moses>

	en → fr		de → en	
	NEWSCO	EUROPART	NEWSCO	EUROPART
#words	1,023,401	21,616,114	1,530,693	22,898,644
#sentences	51,375	1,050,398	71,691	1,118,399
#vocabulary	31,416	78,071	78,140	242,219
#phrase table	3,061,701	46,003,525	4,133,190	44,402,367
% OOV	5.3%	3.1%	8.0%	5.2%

Table 1: Statistics regarding the training corpora: number of words, number of sentences, vocabulary and phrase table size and percentage of test words not appearing in the train set (OOV).

Finding the oracle alignment amounts to solving the ILP problems introduced above. Even though ILP problems are NP-hard in general, there exist several off-the-shelf ILP solvers able to efficiently find an optimal solution or decide that the problem is infeasible. In our experiments, we used the free solver SCIP (Achterberg, 2007). An optimal solution was found for all problems we considered. Decoding the 3,027 sentences of WMT’09 test set takes about 10 minutes (wall time) for the NEWSCO setting, and several hours for the EUROPART setting¹³.

3.2 Oracle BLEU Score

Table 2 reports, for all considered settings, the BLEU-4 scores¹⁴ achieved by our oracle decoder, as well as the number of source words used to generate the oracle hypothesis and the number of target words that are reachable. In these experiments, two objective functions were considered: first, we only consider the objective function corresponding to the relaxed problem defined by Eq. (6); second, we introduced an extra term in the objective to penalize distortion, as described by Eq. (7). Unless explicitly stated otherwise, we always used the exact match strategy.

The main result in 2 is that, for the two language pairs considered, the expressive power of PBTS is not the limiting factor to achieve high translation performance. In fact, for most sentences in the test set, excellent oracle hypotheses, which contain a very high proportion of reference words, are found. This remains true even when the phrase table is extracted from a small corpus. Given that the best BLEU-4 scores achieved during the WMT’09 evaluation are about 28 for the English to French task and 24 for the German to English task ((Callison-Burch et al., 2009), Tables 26 and 25), these results strongly suggest that the main bottleneck of current phrase-based translation systems is their scoring function rather than their expressive power. As we will discuss in Section 4, similar conclusions were drawn by (Auli et al., 2009) and (Turchi et al., 2008).

Several additional comments on these numbers are in

¹³All our experiments are run on a 8 cores computer, each core being a 2.2GHz Intel Processor; the decoder is multi-threaded.

¹⁴These are computed on lowercase with the default tokenization.

order. Despite these very high BLEU scores, in most cases, the reference is only partly translated. In the most favorable case, for the English to French EUROPART setting, only 26% of the references could be fully generated¹⁵. These numbers are consistent with the results reported in (Auli et al., 2009). Similarly, only about 31% of the source sentences are completely translated by the oracle decoder, which supports our choice to consider a relaxed version of the ILP problem. Finally, Table 2 also shows that introducing the distortion penalty does not affect the oracle performance of the decoder.

Considering the inside match strategy improves the performance of the oracle decoder: for instance, for the English to French NEWSCO setting, oracle decoder with the inside match strategy achieves a BLEU-4 score of 70.15 (a 2.5 points improvement over the baseline). To achieve this score, 21.45% of the phrases used during decoding were phrases that are not considered by the exact match strategy. Similar results can be observed for other settings, which highlights the significance of one kind of failure of the extraction heuristic: useful “subphrases” of actual phrase pairs are not always extracted.

The numbers in Table 2, no matter how good they may look, should be considered with caution: they only imply that, for most test sentences, all the information necessary to produce a good translation is available in the phrase table. However, the alignment decisions underlying these oracle hypotheses are sometimes hard to justify, and one has to accept that part of these good hypotheses translations are due to a series of lucky alignment errors. This is illustrated on Figure 2, which displays one such lucky oracle alignment based on the misalignment, during training, of the French preposition “des” (*of the*) with the English noun “stock”. Such lucky errors are naturally also observed in the outputs of conventional decoders, even though phrase table filtering heuristics probably makes them somewhat more rare.

3.3 Analyzing Non-Reachable Parts of a Reference

Table 3 contains typical examples of sentence pairs that could not be fully generated by our oracle decoder. They

¹⁵Similar numbers were obtained, albeit much more slowly, with the `--constraint` option of Moses.

	training set	objective function	% source translated	% target generated	4-BLEU
en → fr	NEWSCO	RELAXED	86.04%	84.74%	67.65
		RELAXED-DISTORTION	85.99%	84.77%	67.77
	EUROPARL	RELAXED	93.66%	93.06%	85.05
		RELAXED-DISTORTION	93.65%	93.06%	85.08
de → en	NEWSCO	RELAXED	82.57%	82.33%	64.60
		RELAXED-DISTORTION	82.59%	82.30%	64.65
	EUROPARL	RELAXED	90.34%	91.16%	81.77
		RELAXED-DISTORTION	90.36%	91.12%	81.77

Table 2: Translation score of the ILP oracle decoder for the various settings described in Section 3.1

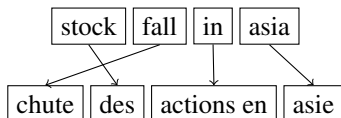


Figure 2: Example of alignment obtained by our oracle decoder

illustrate the three main reasons which cause some parts of the reference to remain *unreachable*:

- phrases are missing from the phrase table, either because they do not occur in the training corpus (OOVs) or because they failed to be extracted. In Table 3, OOV errors are mainly due to past tense forms translated into verbs conjugated in *passé simple* (“rejeta”, “rencontrèrent”, “renoua”) a French literary tense, mostly used in formal writings.
- obvious errors (misspelled words, misinterpretation or mistranslation, ...) in the reference. The reference of the fifth example contains one such error: the state name “Nevada” is translated to “n’évadiez” (literally “have not escaped”), yielding a very poor reference sentence.
- parts of the reference have no translation equivalence in the source. This can be either because references are produced in “context” and some pieces of information are moved across sentence boundaries or because these references are non-literal. The fourth example, which seems to be the translation of a title, falls into this category: the French part contains a reference to the context (“les SA” is referring to the bacteria the text is talking about) which is not in the source text. Non-literal translation are illustrated by the third example, where English “Monday” is translated into French “la veille” (*the day before*).

While the first kind of errors is inherent to the use of a statistical approach, the last two kinds result from the quality of the data used in the evaluation and directly impact both training and evaluation of automatic translation

systems: if they should not distort too much comparisons of MT systems, these errors prevent us from assessing the “global” quality of automatic translation and, if similar errors are found in the train set, they make learning harder as some probability mass is wasted to model them.

To provide a more quantitative analysis, we manually looked at all the non-aligned parts of some WMT’09 references and found that out of 800 references, more than 133 contain either an obvious translation error or can not be achieved by a PBTS¹⁶. Note that, while identifying these errors could be done in many ways, our oracle decoder makes it far easier.

3.4 Identifying Causes of Failure

By comparing the hypotheses found by the oracle decoder and the ones found by the phrase-based decoder, causes of failure can be easily identified. In this section, we will present several measures that allow us to identify and quantify several causes of failure.

Errors Caused by Search Space Pruning Recall from Section 1.1 that Moses uses several heuristics to prune the search space. In particular, there is a distortion limit and a limit on the number of target phrases considered for one source phrase. In this paragraph, we evaluate the impact of these two heuristics on translation quality.

Table 4 presents the average distortion computed on the oracle hypotheses, as well as the percentage of phrases used that have a distortion strictly greater than 6 (the default distortion limit of Moses). All these numbers are obtained by solving the RELAXED-DISTORTION problem. Surprisingly enough, the average distortion of oracle hypotheses is quite small, even for the German to English task, and the distortion constraint seems to be violated only in a few cases. It also appears that the distortion of the hypotheses generated in the NEWSCO setting is significantly larger than in the EUROPARL setting. This can be explained by the extra degrees of freedom in the

¹⁶Annotation at a finer level is an on-going effort; the annotated corpus is available from http://www.limsi.fr/Individu/wisnews/oracle_decoding.

①	<ul style="list-style-type: none"> – On Monday the American House of Representatives rejected the plan to support the financial system, into which up to 700 billion dollars (nearly 12 billion Czech crowns) was to be invested. – Lundi, la chambre des représentants américaine <i>rejeta</i> le projet de soutien du système financier, auquel elle aurait dû consacrer jusqu’à 700 milliards de dollars (près de 12 <i>bilions</i> de Kč).
②	<ul style="list-style-type: none"> – Representatives of the legislators met with American Finance Minister Henry Paulson Saturday night in order to give the government fund a final form. – Dans la nuit de samedi à dimanche, des représentants des législateurs <i>rencontrèrent</i> le ministre des finances américain Henry Paulson, afin de donner au fond du gouvernement une forme finale.
③	<ul style="list-style-type: none"> – The Prague Stock Market immediately continued its fall from Monday at the beginning of Tuesday’s trading , when it dropped by nearly six percent. – Mardi, dès le début des échanges, la bourse de prague <i>renoua</i> avec sa chute de la veille, lorsqu’elle perdait presque six pour cent.
④	<ul style="list-style-type: none"> – Antibiotic Resistance – Les SA résistent aux antibiotiques.
⑤	<ul style="list-style-type: none"> – According to Nevada Democratic senator Harry Reid, that is how that legislators are trying to have Congress to reach a definitive agreement as early as on Sunday. – D’après le sénateur démocrate n’<i>évadiez</i> Harry Reid, les législateurs font de sorte que le Congrès aboutisse à un accord définitif dès dimanche.

Table 3: Output examples of our oracle decoder on the English to French task. Words in bold are non-aligned words and words in italic are non-aligned out-of-vocabulary words. For clarity the examples have been detokenized and recased.

	training set	avg. distortion	%phrases with a dist. > 6
en → fr	NEWSCO	4.57	22.02%
	EUROPAL	3.21	13.32%
de → en	NEWSCO	5.16	25.37%
	EUROPAL	3.81	17.21%

Table 4: Average distortion and percentage of phrases with a distortion greater than Moses default distortion limit.

alignment decisions enabled by the use of larger training corpora and phrase table.

To evaluate the impact of the second heuristic, we computed the number of phrases discarded by Moses (because of the default *ttl* limit) but used in the oracle hypotheses. In the English to French NEWSCO setting, they account for 34.11% of the total number of phrases used in the oracle hypotheses. When the oracle decoder is constrained to use the same phrase table as Moses, its BLEU-4 score drops to 42.78. This shows that filtering the phrase table prior to decoding discards many useful phrase pairs and is seriously limiting the best achievable performance, a conclusion shared with (Auli et al., 2009).

Search Errors Search errors can be identified by comparing the score of the best hypothesis found by Moses and the score of the oracle hypothesis. If the score of the oracle hypothesis is higher, then there has been a search error; on the contrary, there has been an estimation error when the score of the oracle hypothesis is lower than the score of the best hypothesis found by Moses.

Based on the comparison of the score of Moses hypotheses and of oracle hypotheses for the English to French NEWSCO setting, our preliminary conclusion is that the number of search errors is quite limited: only about 5% of the hypotheses of our oracle decoder are actually getting a better score than Moses solutions. Again, this shows that the scoring function (model error) is one of the main bottleneck of current PBTS. Comparing these hypotheses is nonetheless quite revealing: while Moses mostly selects phrase pairs with high translation scores and generates monotonous alignments, our ILP decoder uses larger reorderings and less probable phrases to achieve better solutions: on average, the reordering score of oracle solutions is -5.74 , compared to -76.78 for Moses outputs. Given the weight assigned through MERT training to the distortion score, no wonder that these hypotheses are severely penalized.

The Impact of Phrase Length The observed outputs do not only depend on decisions made during the search, but also on decisions made during training. One such decision is the specification of maximal length for the source and target phrases. In our framework, evaluating the impact of this decision is simple: it suffices to change the definition of indicator variables so as to consider only alignments between phrases of a given length.

In the English-French NEWSCO setting, the most restrictive choice, when only alignments between single words are authorized, yields an oracle BLEU-4 of 48.68; however, authorizing phrases up to length 2 allows to achieve an oracle value of 66.57, very close to the score achieved when considering all extracted phrases (67.77).

This is corroborated with a further analysis of our oracle alignments, which use phrases whose average source length is 1.21 words (respectively 1.31 for target words). If many studies have already acknowledged the predominance of “small” phrases in actual translations, our oracle scores suggest that, for this language pair, increasing the phrase length limit beyond 2 or 3 might be a waste of computational resources.

4 Related Work

To the best of our knowledge, there are only a few works that try to study the expressive power of phrase-based machine translation systems or to provide tools for analyzing potential causes of failure.

The approach described in (Auli et al., 2009) is very similar to ours: in this study, the authors propose to find and analyze the limits of machine translation systems by studying the *reference reachability*. A reference is reachable for a given system if it can be exactly generated by this system. Reference reachability is assessed using Moses in forced decoding mode: during search, all hypotheses that deviate from the reference are simply discarded. Even though the main goal of this study was to compare the search space of phrase-based and hierarchical systems, it also provides some insights on the impact of various search parameters in Moses, delivering conclusions that are consistent with our main results. As described in Section 1.2, these authors also propose a typology of the errors of a statistical translation systems, but do not attempt to provide methods for identifying them.

The authors of (Turchi et al., 2008) study the learning capabilities of Moses by extensively analyzing learning curves representing the translation performances as a function of the number of examples, and by corrupting the model parameters. Even though their focus is more on assessing the scoring function, they reach conclusions similar to ours: the current bottleneck of translation performances is not the representation power of the PBTS but rather in their scoring functions.

Oracle decoding is useful to compute reachable pseudo-references in the context of discriminative training. This is the main motivation of (Tillmann and Zhang, 2006), where the authors compute high BLEU hypotheses by running a conventional decoder so as to maximize a per-sentence approximation of BLEU-4, under a simple (local) reordering model.

Oracle decoding has also been used to assess the limitations induced by various reordering constraints in (Dreyer et al., 2007). To this end, the authors propose to use a beam-search based oracle decoder, which computes lower bounds of the best achievable BLEU-4 using dynamic programming techniques over finite-state (for so-called local and IBM constraints) or hierarchically structured (for ITG constraints) sets of hypotheses. Even

though the numbers reported in this study are not directly comparable with ours¹⁷, it seems that our decoder is not only conceptually much simpler, but also achieves much more optimistic lower-bounds of the oracle BLEU score. The approach described in (Li and Khudanpur, 2009) employs a similar technique, which is to guide a heuristic search in an hypergraph representing possible translation hypotheses with n-gram counts matches, which amounts to decoding with a n-gram model trained on the sole reference translation. Additional tricks are presented in this article to speed-up decoding.

Computing oracle BLEU scores is also the subject of (Zens and Ney, 2005; Leusch et al., 2008), yet with a different emphasis. These studies are concerned with finding the best hypotheses in a word graph or in a consensus network, a problem that has various implications for multi-pass decoding and/or system combination techniques. The former reference describes an exponential approximate algorithm, while the latter proves the NP-completeness of this problem and discuss various heuristic approaches. Our problem is somewhat more complex and using their techniques would require us to built word graphs containing all the translations induced by arbitrary segmentations and permutations of the source sentence.

5 Conclusions

In this paper, we have presented a methodology for analyzing the errors of PBTS, based on the computation of an approximation of the BLEU-4 oracle score. We have shown that this approximation could be computed fairly accurately and efficiently using Integer Linear Programming techniques. Our main result is a confirmation of the fact that extant PBTS systems are expressive enough to achieve very high translation performance with respect to conventional quality measurements. The main efforts should therefore strive to improve on the way phrases and hypotheses are scored during training. This gives further support to attempts aimed at designing context-dependent scoring functions as in (Stroppa et al., 2007; Gimpel and Smith, 2008), or at attempts to perform discriminative training of feature-rich models. (Bangalore et al., 2007).

We have shown that the examination of difficult-to-translate sentences was an effective way to detect errors or inconsistencies in the reference translations, making our approach a potential aid for controlling the quality or assessing the difficulty of test data. Our experiments have also highlighted the impact of various parameters.

Various extensions of the baseline ILP program have been suggested and/or evaluated. In particular, the ILP formalism lends itself well to expressing various constraints that are typically used in conventional PBTS. In

¹⁷The best BLEU-4 oracle they achieve on Europarl German to English is approximately 48; but they considered a smaller version of the training corpus and the WMT’06 test set.

our future work, we aim at using this ILP framework to systematically assess various search configurations. We plan to explore how replacing non-reachable references with high-score pseudo-references can improve discriminative training of PBTS. We are also concerned by determining how tight is our approximation of the BLEU-4 score is: to this end, we intend to compute the best BLEU-4 score within the n -best solutions of the oracle decoding problem.

Acknowledgments

Warm thanks to Houda Bouamor for helping us with the annotation tool. This work has been partly financed by OSEO, the French State Agency for Innovation, under the Quaero program.

References

- Tobias Achterberg. 2007. *Constraint Integer Programming*. Ph.D. thesis, Technische Universität Berlin. <http://opus.kobv.de/tuberlin/volltexte/2007/1611/>.
- Abhishek Arun and Philipp Koehn. 2007. Online learning methods for discriminative training of phrase based statistical machine translation. In *Proc. of MT Summit XI*, Copenhagen, Denmark.
- Michael Auli, Adam Lopez, Hieu Hoang, and Philipp Koehn. 2009. A systematic analysis of translation model search spaces. In *Proc. of WMT*, pages 224–232, Athens, Greece.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proc. of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan.
- Srinivas Bangalore, Patrick Haffner, and Stephan Kanthak. 2007. Statistical machine translation through global lexical selection and sentence reconstruction. In *Proc. of ACL*, pages 152–159, Prague, Czech Republic.
- Léon Bottou and Olivier Bousquet. 2008. The tradeoffs of large scale learning. In *Proc. of NIPS*, pages 161–168, Vancouver, B.C., Canada.
- Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. of WMT*, pages 1–28, Athens, Greece.
- David Chiang, Steve DeNeefe, Yee Seng Chan, and Hwee Tou Ng. 2008. Decomposability of translation metrics for improved evaluation and efficient algorithms. In *Proc. of ECML*, pages 610–619, Honolulu, Hawaii.
- John De Nero and Dan Klein. 2008. The complexity of phrase alignment problems. In *Proc. of ACL: HLT, Short Papers*, pages 25–28, Columbus, Ohio.
- Markus Dreyer, Keith B. Hall, and Sanjeev P. Khudanpur. 2007. Comparing reordering constraints for smt using efficient bleu oracle computation. In *NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 103–110, Rochester, New York.
- Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proc. of ACL*, pages 228–235, Toulouse, France.
- Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2004. Fast and optimal decoding for machine translation. *Artificial Intelligence*, 154(1-2):127–143.
- Ulrich Germann. 2003. Greedy decoding for statistical machine translation in almost linear time. In *Proc. of NAACL*, pages 1–8, Edmonton, Canada.
- Kevin Gimpel and Noah A. Smith. 2008. Rich source-side context for statistical machine translation. In *Proc. of WMT*, pages 9–17, Columbus, Ohio.
- Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL*, pages 48–54, Edmonton, Canada.
- Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL, demonstration session*.
- Philipp Koehn. 2004. Pharaoh: A beam search decoder for phrase-based statistical machine translation models. In *Proc. of AMTA*, pages 115–124, Washington DC.
- Shankar Kumar and William Byrne. 2005. Local phrase reordering models for statistical machine translation. In *Proc. of HLT*, pages 161–168, Vancouver, Canada.
- Alon Lavie, Kenji Sagae, and Shyamsundar Jayaraman. The significance of recall in automatic metrics for MT evaluation. In *In Proc. of AMTA*, pages 134–143, Washington DC.
- Gregor Leusch, Evgeny Matusov, and Hermann Ney. 2008. Complexity of finding the BLEU-optimal hypothesis in a confusion network. In *Proc. of EMNLP*, pages 839–847, Honolulu, Hawaii.
- Zhifei Li and Sanjeev Khudanpur. 2009. Efficient extraction of oracle-best translations from hypergraphs. In *Proc. of NAACL*, pages 9–12, Boulder, Colorado.
- Percy Liang, Alexandre Bouchard-Côté, Dan Klein, and Ben Taskar. 2006. An end-to-end discriminative approach to machine translation. In *Proc. of ACL*, pages 761–768, Sydney, Australia.
- Adam Lopez. 2009. Translation as weighted deduction. In *Proc. of EACL*, pages 532–540, Athens, Greece.
- Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51.
- Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proc. of ACL*, pages 160–167, Sapporo, Japan.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. Technical report, Philadelphia, Pennsylvania.
- D. Roth and W. Yih. 2005. Integer linear programming inference for conditional random fields. In *Proc. of ICML*, pages 737–744, Bonn, Germany.
- Nicolas Stroppa, Antal van den Bosch, and Andy Way. 2007. Exploiting source similarity for smt using context-informed

- features. In Andy Way and Barbara Gawronska, editors, *Proc. of TMI*, pages 231–240, Skövde, Sweden.
- Christoph Tillmann and Tong Zhang. 2006. A discriminative global training algorithm for statistical mt. In *Proc. of ACL*, pages 721–728, Sydney, Australia.
- Marco Turchi, Tijl De Bie, and Nello Cristianini. 2008. Learning performance of a machine translation system: a statistical and computational analysis. In *Proc. of WMT*, pages 35–43, Columbus, Ohio.
- Richard Zens and Hermann Ney. 2005. Word graphs for statistical machine translation. In *Proc. of the ACL Workshop on Building and Using Parallel Texts*, pages 191–198, Ann Arbor, Michigan.