A Study in Greedy Oracle Improvement of Translation Hypotheses

Benjamin Marie^{1,2}*, Aurélien Max*^{1,3}

(1) LIMSI-CNRS, Orsay, France
(2) Lingua et Machina, Le Chesnay, France
(3) Univ. Paris Sud, Orsay, France
{firstname.lastname}@limsi.fr

Abstract

This paper describes a study of translation hypotheses that can be obtained by iterative, greedy oracle improvement from the best hypothesis of a state-of-the-art phrase-based Statistical Machine Translation system. The factors that we consider include the influence of the rewriting operations, target languages, and training data sizes. Analysis of our results provide new insights into some previously unanswered questions, which include the reachability of previously unreachable hypotheses *via* indirect translation (thanks to the introduction of a `rewrite` operation on the source text), and the potential translation performance of systems relying on pruned phrase tables.

1. Introduction

There are two opposing ways in which one may look at the current level of performance reached by Statistical Machine Translation (SMT) systems. One is that the results of SMT systems are still quite unreliable and not appropriate for dissemination or even post-editing by human translators, in particular for low-resourced and/or difficult language pairs, and for situations where domain adaptation is difficult. The other, opposing view is that some contexts allow SMT systems to reach very high performance, including when large enough quantities of adapted data are available, e.g. by using SMT systems in conjunction with translation memories, which has yielded much interest into the study and use of human post-editing and tools for supporting this activity.

Such performance levels typically correspond to the utilization of the *best* translation hypothesis produced by a given system, which is a reflection of the system's relative evaluation of the translations in its search space. Previous oracle studies have shown that the best attainable performance of such systems was in fact much higher than their best output [1]. This is achieved by relaxing pruning and reordering constraints imposed on decoders, and maximizing some evaluation metrics score rather than the system's own scoring function. Such studies are useful, in particular, to make explicit the potential of a given system configuration (training data, extraction procedures, etc.) and to possibly exhibit the difficult parts of a source text (e.g. [2])) as well as the possible defects of reference translations. A lesson that can be drawn from these results is the poor adequacy of the internal scores of translation quality used by current systems.

Another interesting potential use of oracle studies is that they can produce useful data under the form of individual post-editing steps that may be used to improve existing translation hypotheses. Initial attempts at *automatic post-editing* of SMT output approached the problem as one of second-pass translation between automatic predictions and correct translations [3]. Among the drawbacks of such approaches, large quantities of texts have to be translated to learn post-editing models, which are then furthermore specific to a given version of a given system and consequently not straightforwardly reusable. Some large collections of manually revised translations have been collected [4, 5], which can be used e.g. for sub-sentential confidence estimation. However, such data sets are costly to acquire, in particular for some language pairs, and may again be, on some aspects, too specific to a given version of the MT system used.

In this article, we describe an approach to build a related resource, but for a modest cost and with possibly wider applicability. We resort to greedy rewriting of translation hypotheses, in a similar spirit to Langlais et al. [6], to find the sequence of rewriting steps which maximizes the quality of translation hypotheses with respect to some evaluation metrics and reference translation(s). Individual rewritings are based on the repertoire of biphrases units of some phrase-based SMT systems, and thus do not have to correspond to plausible rewritings made by human translators.

While we aim to use such a resource to learn to identify improvable fragments (e.g. [4]) and learn discriminative rerankers (e.g. [7]), we will here focus on a systematic study of such an artificial resource. Our experiments will study the following factors:

- rewriting operations: we will use a revised and extended set of previously used operations [6], and introduce an original operation which allows source sentence rewriting (`rewrite`), as well as a target phrase deletion operation (`remove`);
- training data size: we will use 5 different sizes of training data, where training data are split independently from their relation to the test data;

- number of available reference translations: we will be able to verify whether phenomena observed when a single reference translation is available can also be observed when as many as 7 reference translations allow for a much more robust evaluation of translation quality;
- phrase table filtering: we will use unfiltered phrase tables and phrase tables filtered using a significance testing criterion [8];
- target language: we will use French as the source language, and 10 other European languages as target languages, with exactly the same training data;
- beam size: finally, we will also consider various beam sizes to get some account of the quantity of search errors made by our greedy decoder, although this aspect is not central to the present study.

The remainder of this article is organized as follows. Section 2 introduces greedy oracle decoding and describes the operations that we have used in this work. Section 3 presents our choice of data, systems, and search settings for this work. Our experiments are then detailed in Section 4. We finally summarize our main findings and present some of our future work in Section 5.

2. Greedy oracle decoding

Greedy decoding for Statistical Machine Translation was introduced in [9], as a fast solution to the NP-complete problem of finding the best translation hypothesis from a translation engine's search space.[1] Although such a technique was shown to produce more search errors than its dynamic programming-based counterpart for max-derivation approximation, Langlais et al. [6] described an implementation of greedy search decoding that could improve the best hypothesis from a then state-of-the-art DP-decoder. Subsequent work using a Gibbs sampler for approximating maximum translation decoding [10] showed, however, the adequacy of the approximations made by recent decoders for finding the best translation in their search space, leaving as the main source to account for current translation performance the scoring of translation hypotheses.

Our objective in the present work is not to improve the decoder score of the translation hypotheses that are found, but rather to obtain, by construction, iteratively better hypotheses by using a sentence-level measure of actual translation performance (hence, some approximation of an *oracle*). The sub-optimality of the search is not a problem for our purpose, so we resort to a straightforward greedy algorithm to build such sequences of iteratively improving translation hypotheses.

Algorithm 1 Greedy oracle search algorithm

```
Require: $source$ (input sentence), $beamSize$
  $nbest \leftarrow$ NBEST_LIST($source, beamSize$)
  $oneBest \leftarrow$ GET_ONE_BEST($nbest$)
  loop
    $newNbestList \leftarrow$ INITIALIZE_LIST()
    $sCurrent \leftarrow$ sBLEU($oneBest$)
    $s \leftarrow sCurrent$
    for all $h \in$ NEIGHBORHOOD_BEAM($nbest$) do
      $c \leftarrow$ sBLEU($h$)
      $newNbestList \leftarrow$ ADD($h, c, beamSize$)
      if $c > s$ then
        $s \leftarrow c$
      end if
    end for
    if $s = sCurrent$ then
      return $oneBest$
    else
      $nbest \leftarrow newNbestList$
      $oneBest \leftarrow$ GET_ONE_BEST($newNbestList$)
    end if
  end loop
```

Our greedy oracle decoding is illustrated as pseudo-code in Algorithm 1. We take as seeds the n-best, segmented translation hypotheses of a phrase-based SMT system. At each iteration, a number of best hypotheses relative to our evaluation metrics are kept in a beam until convergence is obtained. Each surviving hypothesis undergoes a number of modifications by means of a repertoire of rewriting operations on bi-phrases that define a neighborhood function. We used the following operations (N denotes the number of biphrases, T the maximum number of entries per source phrase in a translation table, R the maximum number of entries per source phrase in a source rewriting table, and S the average number of tokens per source phrase)[2]:

1. `replace` ($\mathcal{O}(N.T)$): replaces the translation of a source phrase with another translation from the phrase table;
2. `split` ($\mathcal{O}(N.S.T^2)$): splits a source phrase into all possible sets of two (contiguous) phrases, and uses `replace` on each of the resulting phrases;
3. `merge` ($\mathcal{O}(T.N)$): merges two contiguous source phrases and uses `replace` on the resulting new phrase;
4. `move` ($\mathcal{O}(N^2)$): moves the target phrase of a biphrase to all inter-phrase positions in the translation hypothesis;

[1] An optimal, but more costly solution, relying on integer programming, was also proposed in the same article.

[2] Complexity is expressed in terms of the maximum number of hypotheses that will be considered given a seed hypothesis. Note that some of our operations have a much higher complexity than those in [6], which is justified by the fact that we want to explore a larger search space.

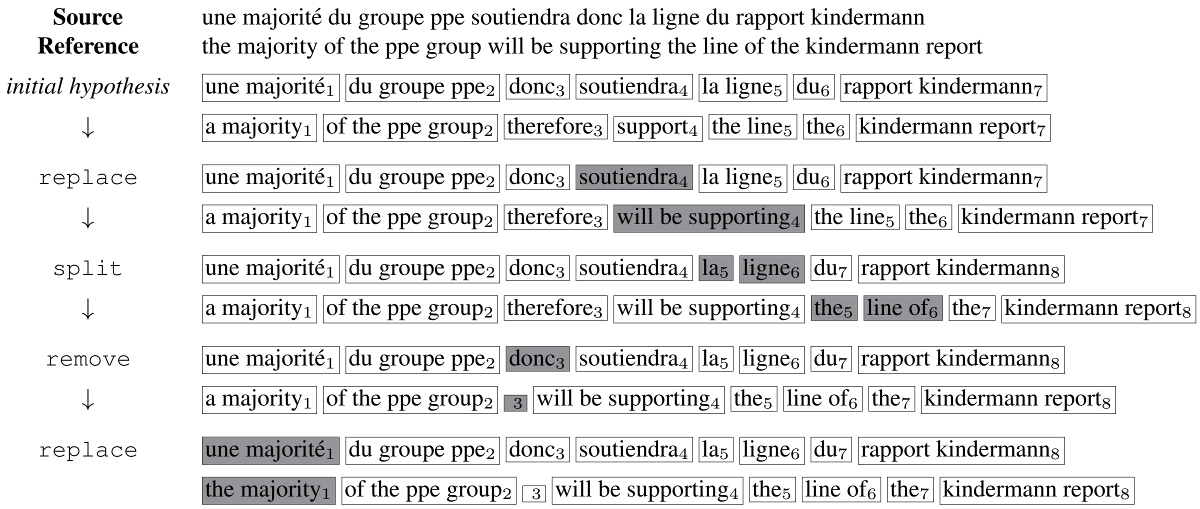
Source une majorité du groupe ppe soutiendra donc la ligne du rapport kindermann
Reference the majority of the ppe group will be supporting the line of the kindermann report
initial hypothesis
une majorité$_1$ du groupe ppe$_2$ donc$_3$ soutiendra$_4$ la ligne$_5$ du$_6$ rapport kindermann$_7$
↓
a majority$_1$ of the ppe group$_2$ therefore$_3$ support$_4$ the line$_5$ the$_6$ kindermann report$_7$
replace
une majorité$_1$ du groupe ppe$_2$ donc$_3$ soutiendra$_4$ la ligne$_5$ du$_6$ rapport kindermann$_7$
↓
a majority$_1$ of the ppe group$_2$ therefore$_3$ will be supporting$_4$ the line$_5$ the$_6$ kindermann report$_7$
split
une majorité$_1$ du groupe ppe$_2$ donc$_3$ soutiendra$_4$ la$_5$ ligne$_6$ du$_7$ rapport kindermann$_8$
↓
a majority$_1$ of the ppe group$_2$ therefore$_3$ will be supporting$_4$ the$_5$ line of$_6$ the$_7$ kindermann report$_8$
remove
une majorité$_1$ du groupe ppe$_2$ donc$_3$ soutiendra$_4$ la$_5$ ligne$_6$ du$_7$ rapport kindermann$_8$
↓
a majority$_1$ of the ppe group$_2$ $_3$ will be supporting$_4$ the$_5$ line of$_6$ the$_7$ kindermann report$_8$
replace
une majorité$_1$ du groupe ppe$_2$ donc$_3$ soutiendra$_4$ la$_5$ ligne$_6$ du$_7$ rapport kindermann$_8$
the majority$_1$ of the ppe group$_2$ $_3$ will be supporting$_4$ the$_5$ line of$_6$ the$_7$ kindermann report$_8$

Figure 1: Trace of an example greedy oracle decoding between French and English. The final state is reached after a sequence of 4 operations (replace, split, remove, replace). Indices in the frames around phrases indicate bilingual alignments originating from the seed hypothesis produced by the Moses decoder.

5. remove ($\mathcal{O}(N)$): deletes the translation of a given biphrase (which remains available as a placeholder for later rewritings);
6. rewrite ($\mathcal{O}(N.R)$): replaces the source phrase of a biphrase with some other source phrase, and replaces its translation with the translations of this new source phrase; note that, by construction, we only need to put in the source rewriting table biphrases that allow to reach n-grams that are not reachable using other operations.

Such a greedy oracle decoder has several limitations. As said previously, it cannot perform a full exploration of the search space and will consequently make search errors; we will report in Section 4 some effects of beam size. Furthermore, our operations are applied on some bilingual phrase segmentation of the source sentence and the translation hypothesis, and split and merge operations will only allow to visit a subset of all rewritings that would be licenced if considering word alignments only. However, this is acceptable for our purpose, as a subsequent objective will be to improve the output of a state-of-the-art phrase-based system using a repertoire of such phrase-based rewriting operations.

One may also keep in mind that some increases in translation scores will not always correspond to actual improvements as judged by human translators. Indeed, some attempts at maximizing a single metrics will result in inappropriate transformations, such as arbitrarily removing words or moving them to positions where e.g. they do not break any longer substrings from the reference translation. One solution may be to make use of a mixture of complementary translation metrics, which may however make computation much more expensive; we leave this to our future work, accepting for now the fact that important metrics score differences (e.g. up to 37 BLEU points and 31 TER points for French to English translation in this study) should always correspond to a majority of clear improvements.

Figure 1 shows an example of a trace by our system of iterative improvement of a translation from French into English, starting from a competitive initial hypothesis (see section 3). A local maximum is here reached after 4 rewriting operations. Examples for the other types of rewriting operations are shown on Figure 2.

3. Experimental settings

In order to experiment with several target languages under the same conditions, we used the Europarl corpus of parliamentary debates[3], and computed the intersection for 11 languages using English as pivot. From the collected data, we extracted held-out, later entries as tuning and test sets (see Table 1). We used French as our sole source language, and experimented with all other possible target languages. English was used as the main target language of the study, notably in settings where the training data was reduced to smaller fractions. Furthermore, in order to verify how our oracles would behave in situations where the evaluation metrics could make use of several possible reference translations, we also used the BTEC corpus of basic traveling expressions [11], allowing us to use 16 references for tuning our baseline systems and 7 references for evaluating them on the French to English language pair (see Table 1).

We built state-of-the-art phrase-based SMT systems using the open source Moses system[4], using standard settings and models and MERT [12] for optimizing the parameters on the tuning set. Trigrams target language models were es-

[3] http://statmt.org/europarl/

[4] http://www.statmt.org/moses

previous	... le projet qui ferait gagner le plus de temps sur un ferroviaire$_{15}$ trajet$_{16}$ très long$_{17}$
	... the project which would win the more time on a rail$_{15}$ route$_{16}$ very long$_{17}$
`move`	... le projet qui ferait gagner le plus de temps sur un ferroviaire trajet très long
	... the project which would win the more time on a very long rail route
previous	il est évident que parler d' intermodalité présuppose un profond changement de la culture d' entreprise .
	it is clear that speak intermodality presupposes a profound change in the business culture .
`rewrite`	il est évident que débat d' intermodalité présuppose un profond changement de la culture d' entreprise .
	it is clear that discussion on intermodality presupposes a profound change in the business culture .
previous	qu' il me soit$_3$ permis$_4$ dès lors de le placer dans une perspective plus historique .
	it would therefore be$_3$ allowed$_4$ to put it into a more a more historical perspective .
`merge`	qu' il me soit permis dès lors de le placer dans une perspective plus historique .
	it would therefore be permitted to put it into a more a more historical perspective .

Figure 2: Examples of applications of rewriting operations not already illustrated on the trace of Figure 1.

	train	**tune**	**test**		
	# M-tok.	# K-tok.	# K-tok.	BLEU	TER
			Europarl corpora		
fr	10.2	32.8	32.8	-	-
en	8.8	28.3	28.6	29.1	54.0
/2	4.4	|	|	28.6	54.4
/4	2.2	|	|	27.6	55.4
/8	1.1	|	|	26.1	56.8
/16	0.5	|	|	25.2	58.4
da	8.4	27.0	27.2	23.2	61.3
de	8.4	27.1	27.1	17.0	68.0
el	8.8	28.5	28.5	23.5	62.2
es	9.2	29.5	29.7	35.9	49.7
fi	6.4	20.6	20.5	11.2	79.7
it	10.2	28.9	29.0	31.6	55.3
nl	8.9	28.2	28.7	21.2	64.6
pt	9.1	29.4	29.3	33.4	52.8
sv	7.9	25.7	25.8	21.0	62.7
			BTEC corpus		
fr	0.2	0.5	0.5	-	-
en	0.2	0.5^*	0.5^{**}	59.6	24.6

Table 1: Top: Statistics for our Europarl training (up to 310K bi-sentences), tune (1K bi-sentences) and test (1K bi-sentences) corpora. Translation performance is given for all baseline systems using French as the source language. Bottom: Statistics for our BTEC training, tuning (16 references*) and test (7 references**) corpora.

timated from the bilingual training data only, using Kneser-Ney smoothing. Results for all baseline systems and all training conditions are reported in Table 1, using BLEU and TER as complementary indicators of translation performance.

We used the greedy search operations described in Section 2. We implemented various approximations to speed up decoding. In particular, we limited candidate replacements for `replace`, `split` and `merge` to phrases that contain at least one token in common with the reference translation, except for the 50 most frequent tokens.[5] We used sentence-level smoothed BLEU [13] as our objective function for greedy decoding (using a single (Europarl) or several reference translations (BTEC)), but will use corpus-level BLEU and individual n-gram precisions, as well as TER, to report translation performance.

4. Experiments and analysis

4.1. Rewriting operations

Using our main language pair, French to English, we experimented with each individual rewriting operation, as well as with the full set; see Table 2. The two operations that individually lead to the largest improvements are not surprisingly those that have access to replacement translations from the phrase table, `replace` and `split`. The larger improvements with the latter are due to the combination of sub-`replace` operations, which encompass translations attainable by composition as well as possibly more combinations not seen associated with the larger source phrases. Conversely, `merge` is of moderate use, but still manages to capture some cases where translations cannot be obtained by composition. As the sole operation, `remove` has almost no impact on translation, and may in fact only artificially inflate low-order n-gram precision values. `move` has a moderate impact, not too surprisingly more apparent on BLEU and higher-order n-gram precision than on TER, which may be attributed in part to the language pair (see Section 4.3). The impact of the `rewrite` operation will be specifically discussed in section 4.4.

[5] Although lowering this value led to fewer search errors, we deemed the chosen value a good compromise time-wise.

	Europarl fr→en (1 ref.)							**BTEC fr→en (7 refs.)**						
	BLEU					**TER**	**avg #**	**BLEU**					**TER**	**avg #**
	score	1g	2g	3g	4g	score	**iterations**	score	1g	2g	3g	4g	score	**iterations**
baseline	29.0	63.2	35.5	22.6	14.6	54.0	-	59.62	85.08	67.13	53.33	41.48	24.60	-
beam size = 1														
`merge`	31.8	65.3	38.3	25.2	16.9	51.7	0.75	60.43	85.43	67.84	54.32	42.35	24.32	0.07
`move`	32.0	63.2	39.1	25.8	17.3	53.3	1.01	61.70	85.08	69.52	55.84	43.87	24.60	0.16
`remove`	29.7	67.1	39.2	25.6	16.9	50.0	1.03	59.62	85.08	67.13	53.33	41.48	24.60	0.00
`replace`	42.1	73.9	48.8	34.8	25.1	42.5	4.40	66.50	88.72	73.40	60.90	49.33	23.67	0.91
`rewrite`	29.8	64.5	36.2	23.0	14.0	53.5	0.38	59.69	85.07	67.12	53.48	41.57	24.68	0.04
`split`	45.7	74.3	52.7	39.1	28.6	41.3	4.46	69.34	88.05	75.36	64.62	53.90	27.07	1.24
all	66.5	88.2	73.8	62.6	53.0	23.1	11.04	77.30	91.17	81.67	73.78	64.98	23.47	1.92
beam size = 2														
all	66.6	88.1	73.9	62.8	53.2	23.0	11.19	77.88	91.44	82.36	74.37	65.68	23.16	2.28
beam size = 5														
all	67.8	88.5	74.9	64.3	55.0	22.3	11.26	79.06	91.88	83.29	75.67	67.47	22.94	2.12

Table 2: Effects of individual operations and beam size (left: Europarl; right: BTEC).

Potential improvements to translation hypotheses using the original phrase table are very large. However, this may not reflect accurately *actual improvements*. One important reason for this is the fact that a single reference translation usually does not represent all the acceptable wordings of a translation. Looking at the BTEC condition, where the baseline evaluated on 7 reference translations is much stronger than in the Europarl condition, we still find significant increases in BLEU score with a relative contribution of operations that is well correlated to that obtained on the more difficult, single-reference Europarl condition. The main source of improvement for translation hypotheses thus resides in translating using generally smaller phrases (`split`) and choosing more appropriate translations for phrases (`replace`).

Next, we look at when each operation is used when they are all activated. The distribution of operations on Europarl is given on Figure 3 by looking at operations from each quarter of complete sequences (thus each corresponding to an average of $11.04/4 = 2.76$ operations). The first quarter of operations, yielding almost half of the full improvement, mostly consists of alternative translations (`split` and `replace`). The `move` operation contributes more after the initial burst of operations, while `remove` progressively acts on phrases for which `split` cannot propose any further improvement from the reached hypotheses.

All subsequent experiments will be conducted with a beam size of 1 to limit computation time.[6] Table 2 additionally provides results for larger beams, which gives some account of the reduction in search errors corresponding to a larger number of iterations per sentence (on average, there is 0.22 more iteration per sentence using a beam of 5, but at the cost of a running time multiplied by a factor of more than 3).

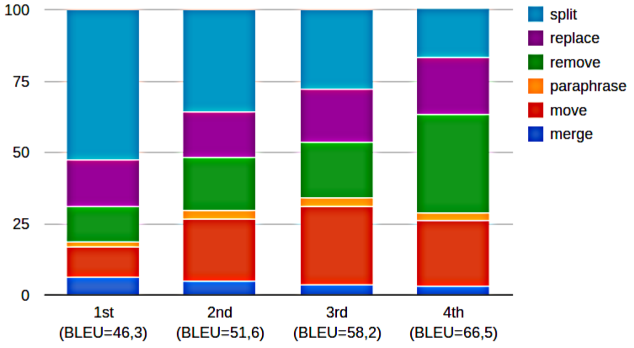

Figure 3: Distribution of types of operations per quarters of operations during greedy oracle search. Corresponding BLEU scores obtained after each quarter of iterations are indicated on the legend.

4.2. Training data size and phrase table filtering

Predicting translation performance given the available amount of training data is a useful problem [14]. Here, we look at how much training data size impacts the performance attainable by our oracle decoder. We reduce training data size up to 16 times on the Europarl condition, without selecting data in any way relative to the dev and test set. Results are given in Table 3. Whereas reducing by half the quantity of training data roughly corresponds to the loss of 1 BLEU point or less, we find that loss in oracle performance, although also regular for each training data size reduction, is close to 5 BLEU points. This fact may be often overlooked in the SMT research community, where it is commonly known that doubling the size of the training data typically has only a small impact on translation performance. Our results show that this is mostly a result of the limitations of the scoring function used by decoders, and that attainable improvements benefit much more from the added training data.

[6] On a single core of a 2.2Ghz machine with 64Gb memory, decoding our whole test sets took roughly 6 hours for a beam size of 1, 8 hours for a beam size of 2, and more than 20 hours for a beam size of 5.

A related question is whether pruned phrase tables, which can yield competitive translation performance while retaining only small fractions of the original phrase table entries, would be significantly different in terms of attainable translations. We used the widely used significance pruning of Johnson et al. [8], and selected a configuration where phrase pairs occurring once in the bilingual corpus and composed from phrases also occurring once on their respective side of the corpus (so-called 1-1-1 configurations) are pruned. Looking at the results on Table 3, we find that keeping only 27% of the original phrase table entries indeed yielded no loss in translation performance at rank 1 for the decoder. Although the intuitions for filtering such phrase pairs include the fact that they may correspond to noise or offer too little reusability, the important drop in oracle performance (-11.2 BLEU points and +8.8 TER points) clearly indicates that a significant part of the filtered entries, although apparently poorly scored by the translation system, would have in fact largely benefited the system.[7]

4.3. Target languages

Classes of language pairs correspond to very different challenges for SMT systems, as exemplified by the large-scale study reported in [15]. In this set of experiments, we wanted to assess oracle performance for a number of target languages with various types of relationship to the source language (e.g. closely related (Spanish), completely unrelated (Finnish), different sentence structure (German), etc.) Results are shown in Table 4 for the 10 target languages of our study in the Europarl condition. Relative improvements in BLEU scores range from roughly +100% (for Spanish and Portuguese) to more than +300% (Finnish). This latter case seems particularly instructive: although not directly comparable to the absolute values reached for other target languages, phrase tables do contain entries that can significantly improve automatic translation into such a complex language as Finnish.[8] We observe, in particular, a very large increase in *n*-gram precision at all sizes.

Another interesting result concerns romance target languages, which obtain both the smallest relative increase in BLEU (around +100%) and the largest relative reduction in TER (up to -63%). Our hypothesis to account for this fact is that the improvements on *n*-gram precisions do not result in the strongest increases overall in BLEU, but that given that many such improvements for long target phrases are indeed possible, this globally results in sentence orderings that are more symmetric between oracle outputs and reference translations.

		BLEU		**TER**		**#. iterations**
		score	`+rew`	score	`+rew`	**avg. per sent.**
da	*baseline*	23.2		61.3		-
	oracle	58.4	+0.9	29.5	-0.8	10.7
de	*baseline*	17.0		68.0		-
	oracle	55.1	+1.4	32.0	-1.2	13.3
el	*baseline*	23.5		62.2		-
	oracle	62.8	+1.0	26.5	-0.6	11.5
en	*baseline*	29.0		54.0		-
	oracle	66.5	+0.6	23.1	-0.4	11.0
es	*baseline*	35.9		49.7		-
	oracle	74.0	+0.5	18.2	-0.5	10.7
fi	*baseline*	11.2		79.7		-
	oracle	46.1	+1.2	38.1	-1.2	11.3
it	*baseline*	31.6		55.2		-
	oracle	71.2	+1.1	20.4	-1.7	11.3
nl	*baseline*	21.2		64.6		-
	oracle	56.3	+1.6	32.4	-0.7	12.9
pt	*baseline*	33.4		52.8		-
	oracle	69.8	+0.7	21.5	-0.5	10.2
sv	*baseline*	21.0		62.7		-
	oracle	59.9	+1.0	27.8	-1.1	11.2

Table 4: Effects of target language (Europarl). '`+rew(rite)`' indicates the specific contribution of the corresponding improvement (in BLEU or TER) of the oracle score.

We further look at the distributions of rewriting operations per target language, given on Figure 4. `replace` operations appear uniformely useful for all languages, illustrating the relative inadequacy of the translation models used by the decoders across languages. `split` operations are more numerous for target languages with good baseline performance (e.g. English and Portuguese). This can be attributed to some over-confidence in long bi-phrases that can be extracted from the training data, which not always permit to attain the expected reference translation. Conversely, we note slightly more `merge` operations for romance languages and Greek, a fact that should be investigated further. While phrases used by the decoder used should be generally shorter, a significant number of source fragments are nonetheless inaccurately translated compositionally when their correct translation is available.[9]

Not surprisingly, we also note a larger use of `move` operations for translating into German (and, to a lesser extent, Dutch and Scandinavian languages). Likewise, we find, at no surprise, that Finnish required a more important number of deletions of target words associated to source phrases, a reflection of the much compositional morphology of the language, which makes capturing appropriate biphrases difficult when such a language is involved.

[7] We note, however, that using a filtered phrase table already yields an interesting level of oracle translation improvement, with a very modest running time (less than half an hour on a single core for decoding the 1000 sentences of our test set).

[8] We must, however, acknowledge the fact that the target language model used for baseline decoding could not be very competitive here, which is particularly true for this target language.

[9] Among other possibilities, a stronger language model may help correct this to some extent. In this study, priority was put on ensuring that all systems were built from the same data.

	BLEU						**TER**		**#. biphrases**
	baseline	oracle	1g	2g	3g	4g	baseline	oracle	**in phrase table**
`full`	29.1	65.9	87.9	73.3	61.9	52.4	54.0	23.5	735,273
/2	28.6	60.8	85.5	68.8	56.5	46.4	54.4	27.5	419,716
/4	27.6	55.6	82.8	64.3	51.0	40.7	55.4	31.1	239,647
/8	26.1	51.1	79.8	60.0	46.3	36.0	56.8	35.1	137,719
/16	25.2	46.0	76.9	55.2	41.1	30.7	58.4	39.0	79,837
`sigtest`	29.1	54.7	81.4	63.0	50.3	40.4	54.1	32.3	203,672

Table 3: Effects of training data size and phrase table filtering (all operations but `rewrite`) (Europarl).

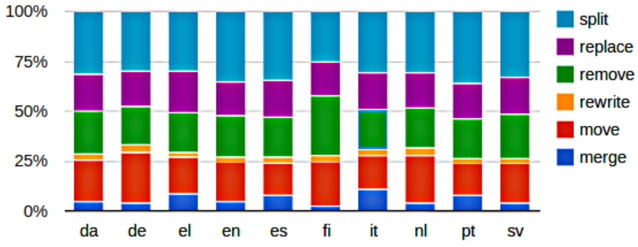

Figure 4: Distribution of operations per target language.

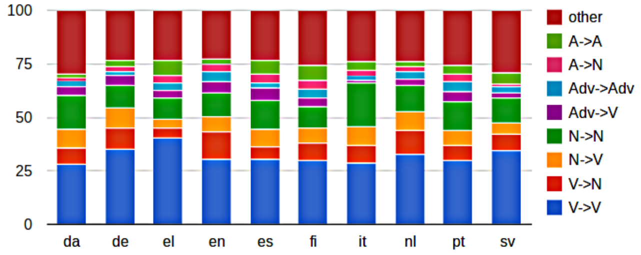

Figure 5: Distribution of main part-of-speech patterns of source `rewrite` for translation from French.

4.4. Reachability of new reference fragments

Our `rewrite` operation allows to reach fragments from the reference translations that are not directly reachable using `replace` only. Using this operation alone for French to English translation on the Europarl condition (Table 2) led to an improvement of +0.8 BLEU and -0.5 TER, for an average number of 0.38 applications per sentence decoding. Results across target languages (Table 4) show that languages that benefit the most from this increased reachability (more than +1 BLEU and -1 TER) mostly corresponds to languages with lower baseline scores, indicating that alignment difficulty (considering that the exact same training data were used for all language pairs) is responsible to some extent.

Positive applications of such an operation, as previously proposed by [16, 17] using source paraphrase lattices, include a large typology of configurations largely not limited to strict paraphrase phenomena, as illustrated on Figure 5. For instance, using English as the source language for illustration purposes, correctly translating the English word *buying* (in *not by buying other countries' quotas*) by *rachat* (in the expected translation *non par le rachat du "droit à polluer" d'un autre pays*) can only be done by translating the noun *purchase* instead. Studying source rewriting patterns on part-of-speeches (see Table 5) shows that French, with a rich verbal inflection system, mostly requires rewriting of verbs into verbs, with significantly fewer cases for nouns into nouns, and fewer yet for adjectives into adjectives. The most represented types with a change of category are verbs into nouns, nouns into verbs, and adverbs into verbs.

source	**reference**	**`rewrite` phrases**
abused	dénaturé	different
buying	rachat	purchase
complex	multitude	number | series | wealth
damaging	désastreuse	disastrous
drivers	des personnels	people
excuse	argument	argument | grounds | reason

Table 5: Examples of English source rewritings (note that English was used as source language here for illustration purposes) and their new reachable French reference translation fragment.

5. Conclusion

This article has presented a study of iteratively improved translation hypotheses, starting from competitive baseline hypotheses up to translation hypotheses of very high quality, even for comparatively difficult language pairs. Although we implemented a non-optimal solution to finding the hypotheses that maximize a single automatic metrics score, several useful facts were empirically demonstrated. Our study first confirmed the important potential for improvement of current phrase-based SMT systems, both in situations where a single or several reference translations are available, and the difficulty of the translation scoring problem. Such conclusions naturally pave the way for further research in discriminatively training systems, more particularly based on dynamic reranking using so-called pseudo-references [7], by focusing more particularly on the rewriting of possibly ill-translated phrases [2, 4].

We have also made explicit the relative contribution of a number of rewriting operations, including an original one, `rewrite`, which allows us to turn around the common acceptance that unique reference translations are poor representations of acceptable translations, and to claim that the specificities of a unique source text sometimes are responsible for (automatic) translation difficulty. Previously, Schroeder et al. [16] had shown the potential of using many human rewritings of input texts, and Khalilov and Sima'an [18] had shown the potential of using reorderings of input texts, but to our knowledge this work is the first to focus on the contribution of local indirect translation.[10] Paraphrasing the training data [19, 20] in a carefull manner is one way to provide access to such knowledge during translation.

Other salient results of our study include the empirical demonstration that pruned phrase tables significantly limit the potential of SMT systems, and that current SMT systems have the potential to already produce very good translation hypotheses even for difficult language pairs, however difficult this may be to achieve in practice. Part of our intended future work will focus on identifying high-quality greedy sequences of rewriting operations, and to compare them to edit sequences made by human post-editors, for whom finding a close-to-shortest route to translation improvement can be difficult.

6. Acknowledgements

This work was partially supported by ANR project Trace (ANR-09-CORD-023) and by a grant from LIMSI.

7. References

[1] G. Wisniewski, A. Allauzen, and F. Yvon, "Assessing Phrase-Based Translation Models with Oracle Decoding," in *EMNLP*, Cambridge, USA, 2010.

[2] B. Mohit and R. Hwa, "Localization of Difficult-to-Translate Phrases," in *WMT*, Prague, Czech Republic, 2007.

[3] M. Simard, C. Goutte, and P. Isabelle, "Statistical Phrase-based Post-editing," in *NAACL*, Rochester, USA, 2007.

[4] N. Bach, F. Huang, and Y. Al-Onaizan, "Goodness: A Method for Measuring Machine Translation Confidence," in *ACL*, Portland, USA, 2011.

[5] M. Potet, E. Esperança-Rodier, L. Besacier, and H. Blanchon, "Collection of a Large Database of French-English SMT Output Corrections," in *LREC*, Istanbul, Turkey, 2012.

[6] P. Langlais, A. Patry, and F. Gotti, "A Greedy Decoder for Phrase-Based Statistical Machine Translation," in *TMI*, Skovde, Sweden, 2007.

[7] P. Liang, A. Bouchard-Côté, D. Klein, and B. Taskar, "An End-to-End Discriminative Approach to Machine Translation," in *ACL*, Sydney, Australia, 2006.

[8] H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving Translation Quality by Discarding Most of the Phrasetable," in *EMNLP*, Prague, Czech Republic, 2007.

[9] U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada, "Fast decoding and optimal decoding for machine translation," in *ACL*, Toulouse, France, 2001.

[10] A. Arun, P. Blunsom, C. Dyer, A. Lopez, B. Haddow, and P. Koehn, "Monte Carlo inference and maximization for phrase-based translation," in *CoNLL*, Boulder, USA, 2010.

[11] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto, "Toward a Broad-coverage Bilingual Corpus for Speech Translation of Travel Conversations in the Real World," in *LREC*, Las Palmas, Spain, 2002.

[12] F. J. Och, "Minimum Error Rate Training in Statistical Machine Translation," in *ACL*, Sapporo, Japan, 2003.

[13] C.-Y. Lin and F. J. Och, "ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation," in *COLING*, Geneva, Switzerland, 2004.

[14] P. Kolachina, N. Cancedda, M. Dymetman, and S. Venkatapathy, "Prediction of Learning Curves in Machine Translation," in *ACL*, Jeju, Korea, 2012.

[15] P. Koehn, A. Birch, and R. Steinberger, "462 machine translation systems for Europe," in *MT Summit*, Ottawa, Canada, 2009.

[16] J. Schroeder, T. Cohn, and P. Koehn, "Word Lattices for Multi-Source Translation," in *EACL*, Athens, Greece, 2009.

[17] T. Onishi, M. Utiyama, and E. Sumita, "Paraphrase Lattice for Statistical Machine Translation," in *ACL, short papers*, Upsala, Sweden, 2010.

[18] M. Khalilov and K. Sima'an, "Statistical Translation After Source Reordering: Oracles, Context-Aware Models, and Empirical Analysis," *Natural Language Engineering*, vol. 18, no. 4, pp. 491–519, 2012.

[19] A. Max, "Example-based Paraphrasing for Improved Phrase-Based Statistical Machine Translation," in *EMNLP*, Cambridge, USA, 2010.

[20] W. He, S. Zhao, H. Wang, and T. Liu, "Enriching SMT Training Data via Paraphrasing," in *IJCNLP*, Chiang Mai, Thailand, 2011.

[10] We should, of course, repeat such experiments using several reference translations and larger training data sets.