Identifying Infrequent Translations By Aligning Non Parallel Sentences

Julien Bourdaillet
Xerox Research Center Webster
800 Philipps Road
Webster, NY 14580
julien.bourdaillet@xerox.com

Philippe Langlais
DIRO/RALI
Université de Montréal
Montréal, Canada H3C 3J7
felipe@iro.umontreal.ca

Abstract

Aligning a sequence of words to one of its infrequent translations is a difficult task. We propose a simple and original solution to this problem that yields to significant gains over a state-of-the-art transpotting task. Our approach consists in aligning non parallel sentences from the training data in order to reinforce online the alignment models. We show that using only a few pairs of non parallel sentences allows to improve significantly the alignment of infrequent translations.

1 Introduction

The task of *transpotting* consists in identifying the translation of a given sequence of words, hereafter called the query, in a pair of parallel sentences (Simard, 2003). While transpotting may be seen as a special case of word aligning pairs of parallel sentences, which is the bread and butter of Machine Translation (MT), this task deserves to be evaluated as such. Indeed, transpotting is at the heart of bilingual concordancers (Wu et al., 2003; Callison-Burch et al., 2005; Bourdaillet et al., 2010; Désilets et al., 2010), and professional translators in the translation industry rely heavily on such Computer Assisted Translation (CAT) tools (Bowker and Barlow, 2008; Macklovitch et al., 2008; Koehn, 2009; Paulsen Christensen and Schjoldager, 2010; Karamanis et al., 2011).

Figure 1 illustrates the answer provided by the bilingual concordancer Tradooit[1] for the English query *meanwhile*. Typically, a concordancer returns pairs of sentences where the query and one of its translations are identified using word alignment. Based on these alignments, a distribution of translations, hereafter called *transpots*, is returned as well for the query. This gives a user information of how likely the transpots are. By clicking on a given transpot a user can consult the sentence pairs where it has been identified.

After inspecting a few bilingual concordancers, like Tradooit, TransSearch[2] or Linguee,[3] we observed that often, transpots are partial or even wrong. This is a precision problem which is due to word alignment errors. In Figure 1, the French transpot *part* is an example of a partial transpot; the translation being *pour sa part* (literally, *for his part*) which is not present in the transpot distribution.

In an ethnographic field study, Désilets *et al.* (2009) analyze that professional translators deal easily with errors present in lists of translations since they are bilingual; this means that they can handle precision problems. On the other hand, they are more concerned with missing translations, that is, a recall problem. They suggest to put efforts into improving the recall of CAT tools such as bilingual concordancers. Indeed, being able to provide a diversified set of infrequent and idiomatic translations to professional translators would be invaluable.

Rare translations are quite likely to be missing in the transpot distribution, due to word alignment errors. Indeed, since they co-occur only a few times with the query in the training data, their lexical

[1] http://www.tradooit.com

[2] http://tsrali3.com

[3] http://www.linguee.com

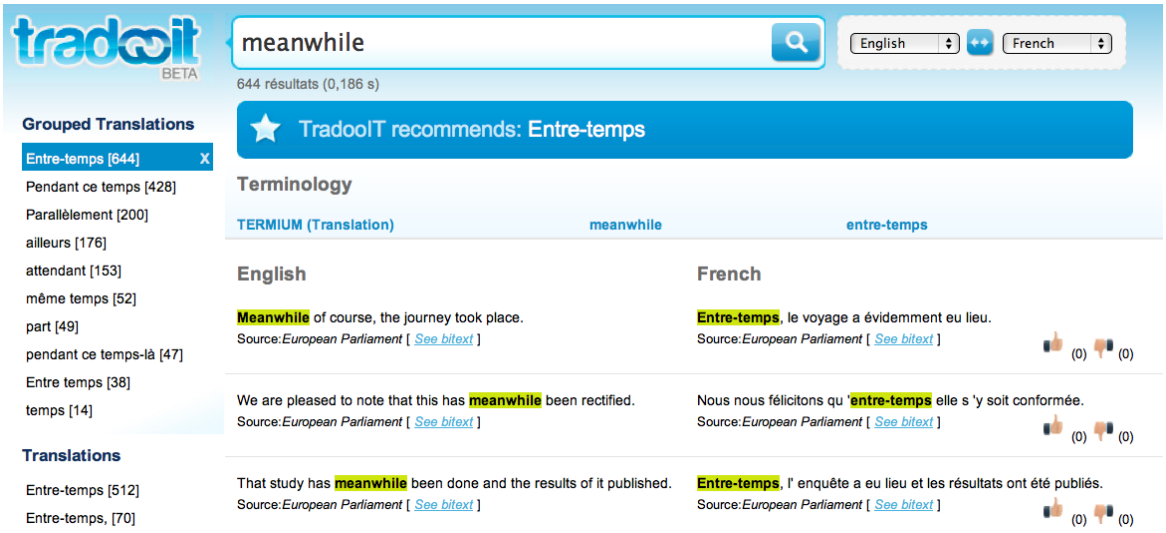

Figure 1: Screenshot of the web-based bilingual concordancer `Tradooit` for the query *meanwhile*. The left part displays a distribution of the transpots identified. By selecting one of them, here *entre-temps*, the user can go through the examples in which it occurs along with the query (main part of the display).

associations tend to be poorly estimated by statistical translation models. For instance, over more than 8 million sentence pairs we use in this work, the term *consistently* appears to be translated only once by *avec logique* (literally, *with logic*) and once by *de façon répétée* (literally, *of way repeated*), two valuable translations that none of the standard word alignment techniques we tried (see Section 3) aligned properly. Also, infrequent translations are often idiomatic expressions. This is the case of the idiomatic translation *sur ces entrefaits* (literally, *on these intermediate facts*), for the query *meanwhile*, which is not proposed by the tools we tried.

In the end, the low word-alignment precision leads to a recall problem in the transpot distribution. This is unfortunate since many translations are available in the bitext, but are simply not mined. In this work, we address this problem by proposing a two-stage transpotting method that aims at improving the transpotting of infrequent translations. After the first transpotting stage that retrieves the transpot distribution (as in Figure 1), a second stage focuses on the lower tail of this distribution whose transpots are suspected to be misaligned.

This second stage intends to realign the sentence pairs where these low frequency transpots occur, by taking advantage of a new word alignment model adapted online for each query/transpot pair. Due to the low cooccurrence of the query/transpot pairs in the training data, there is no additional data available for the adaptation. To overcome this, we propose to sentence-align non parallel sentence pairs sampled from the training material, whose source sentences contain (only) the query and target sentences contain the suspicious transpot. After transpotting this artificial bitext, we extract a lexical distribution local to the sentence pair containing the infrequent translation. Finally, this sentence pair is retranspotted after adapting online the lexical distribution, leading to significant improvements in terms of alignment.

On top of improving the state-of-the art of transpotting, a task of practical importance for the translation industry, there are several contributions we would like to underline:

- We show that a smart processing of non parallel sentence pairs from the training data can help statistical word alignment, a somehow surprising result. Our two-stage approach significantly outperforms the standard bidirectional IBM word-alignment combined with the mainstream `grow-dial-final` heuristic (Koehn et al., 2007) when aligning rare translations.

- We demonstrate that our approach works while

using only a few non parallel sentence pairs, making the approach very tractable.

- We show that it is possible to enhance word alignment for infrequent events in the setting of a bilingual concordancer, which is an important concern for professional translators.

The reminder of this paper is organized as follows. In Section 2, we describe our approach to enhance the alignment of infrequent translations. In Section 3, we present several state-of-the-art transpotting algorithms we compared. In Section 4, we present our experimental setup, and analyze our results in Section 5. We discuss related works in Section 6 and conclude in Section 7.

2 Approach

Figure 2 presents a transpot distribution produced by one of the transpotting algorithms we tested (see Section 3) for the query *wait and see*. Among those transpots, some are very frequent, while many appear rarely, following a typical zipfian law. In this work we concentrate on *frequency-1* transpots, that is, transpots that appear only once in the transpot distribution of a given query. Frequency-1 transpots might be correct rare translations, but are more likely the result of word alignment errors. Among erroneous transpots, some are entirely wrong, such as *manger* (*to eat*), while others are partially good, such as *verra*, which is part of the idiomatic (rare) translation *qui vivra verra* (literally, *who will live will see*).

After the first transpotting stage that generated the transpot distribution, we apply a second transpotting stage for each sentence pair containing a frequency-1 transpot. These pairs are called *seed sentence* pairs hereafter. During this second stage, each seed sentence pair is realigned thanks to a word-alignment model adapted online for this sentence pair, with the hope that it will improve the transpotting.

For each seed sentence, we create a small non parallel bitext by sampling non parallel sentence pairs from the training data used for training the alignment models in the first place. This bitext is used to estimate online a lexical model local to the seed sentence pair. It is important to note that we do not exploit extra data for adapting the alignment model, but instead make better use of available data.

Transpot	*Freq.*	*Freq.-1*	*Wrong*	*Partial*	*Good*
attendre pour voir	66	-	-	-	✓
voir	33	-	-	-	✓
attendre	32	-	-	-	✓
attentiste	14	-	-	-	✓
regarder	4	-	-	-	✓
...			...		
manger	1	✓	✓	-	-
il faudra	1	✓	✓	-	-
verra	1	✓	-	✓	-
inertie	1	✓	-	-	✓
patiente	1	✓	-	-	✓

Figure 2: Transpot distribution for the query *wait and see*. For each transpot, we mention: its frequency; whether it is a frequency-1 transpot; and whether it is a wrong, partial or good transpot.

For our approach to work, we need to specify two elements: how to gather the material used to adapt the lexical distributions of the alignment model, and how to adapt the model.

2.1 Gathering Non Parallel Sentence Pairs

Adapting lexical models usually involves more or dedicated parallel data (Foster and Kuhn, 2007; Koehn and Schroeder, 2007; Bertoldi and Federico, 2009). In this paper, we exploit an already large training set and we do not seek for external data. Indeed, since we are dealing with infrequent translations, there is no extra parallel data on top of the seed sentences that we can exploit.

Rather, we propose an empirical way to better exploit the training data. For each seed sentence pair, we randomly sample from the training data source sentences where the query occurs. We artificially associate them to the target seed sentence which contains the frequency-1 suspicious transpot. This gives a set of non parallel sentence pairs to which we add the seed sentence pair. Figure 3 illustrates such a "bitext" for the frequency-1 transpot *verra* transpotted instead of the (rare) translation *qui vivra verra*. As can be observed, to the exception of the seed sentence pair (the first one in Figure 3), all the other pairs of sentences are not translations, but contain by construction the query and the frequency-1 transpot.

I hope that the parliamentary secretary is here to tell us that is not true, but we shall **wait and see**.	J'espère que le secrétaire parlementaire est ici pour démentir ces rumeurs, mais qui vivra **verra.**
He continues to dodge and weave, **wait and see**, hide and seek.	J'espère que le secrétaire parlementaire est ici pour démentir ces rumeurs, mais qui vivra **verra**.
⋮	⋮
Instead it chose the **wait and see** approach, and what have we seen?	J'espère que le secrétaire parlementaire est ici pour démentir ces rumeurs, mais qui vivra **verra**.

Figure 3: Non parallel bitext gathered for the seed query *wait and see* and the frequency-1 transpot *verra* which was transpotted instead of the correct rare translation *qui vivra verra*. The first sentence pair is the seed pair. All other pairs are non parallel sentence pairs, where the source sentence contains the seed query and the target sentence is the seed target sentence where the frequency-1 transpot was identified.

2.2 Adapting Alignment Models

The artificial bitext is then transpotted with the same transpotting algorithm used during the first stage. The list of transpots $\mathcal{C}$ is extracted and each transpot is considered to be a translation of the query. A typical transpot list is provided for our running example in Figure 4. From this list, it is straightforward to compute an a posteriori lexical model of the query, by counting the frequency of each word and by normalizing.[4]

Although the transpot list $\mathcal{C}$ is typically noisy, it is interesting to note that (at least a part of) the reference translation occurs most of the time in the candidate transpots. Therefore, one might hope that the a posteriori lexical distribution contains useful lexical associations.

This a posteriori distribution (p_l), *local* to each seed sentence pair, can then combined with the *global* lexical distribution (p_g) trained once for all on the entire training data. This is done by linearly combining both distributions, where λ controls the combination:

$$p(t|s) = \begin{cases} \lambda p_g(t|s) + (1-\lambda) p_l(t|s) & \text{if } s \in query \\ p_g(t|s) & \text{otherwise} \end{cases} \quad (1)$$

[4]We tried an alternative and smarter way to estimate the a posteriori lexical model. We used GIZA++ to train an HMM alignment model on the artificial bitext formed by the query sentence-aligned with each transpot of the transpot list $\mathcal{C}$. Then the resulting lexical distribution is used as the local a posteriori distribution, the remainder of the 2-stage method being the same. In the end, we did not obtain any gain in experiments similar to those described in Section 5.

vivra
mais *qui vivra verra*
J'espère que le
rumeurs, mais *qui vivra verra*
qui vivra
est ici pour démentir ces rumeurs, mais *qui*
vivra verra
mais *qui vivra*
ces rumeurs, mais *qui vivra*
que le secrétaire parlementaire est ici pour
qui
vivra verra

Figure 4: Transpot list $\mathcal{C}$ obtained after aligning the artificial bitext of Figure 3, for the seed query *wait and see*. For computing the a posteriori lexical model, each word of these transpots is considered aligned to each word of the seed query. It is then straightforward to count word alignments and normalize their count in order to obtain a probability distribution. (The reference translation words are emphasized for the sake of the presentation.)

Finally, the seed sentence pair is transpotted again using Eq. (1). Since there is no reason why the local distribution would improve the alignment of words outside the query, the combination in Eq. (1) is only applied when transporting words of the query. For the other words, the global distribution is used, as in the first transpotting stage.

In the end, the result of this second transpotting stage is returned to the user with the hope that it is more accurate than the transpot identified in the first place for the query in the seed sentence pair.

3 Transpotting Algorithms

IBM models are natural contenders for tackling the transpotting task (Brown et al., 1993). Formally, given a source language sentence $S = s_1...s_n$ and its target language translation $T = t_1...t_m$, an IBM-style alignment $a = a_1...a_m$ links each word of T to a word of S ($a_j \in \{1, ..., n\}$) or to the empty word ($a_j = 0$) which is arbitrarily associated to untranslated words. For the IBM model 2, the joint probability of a target sentence and its alignment given the source sentence is expressed by:

$$p(t_1^m, a_1^m|s_1^n) = p(m|n) \prod_{j=1}^{m} p(t_j|s_{a_j}) \times p(a_j|j, m, n) \quad (2)$$

where $p(m|n)$ is the sentence length probability, the first term of the product is the lexical probability, and the second term is the alignment probability. According to this formulation, the most probable alignment of two sentences, $\text{argmax}_{a_1^m} p(a_1^m|t_1^m, s_1^n)$, can be efficiently computed in $O(mn)$ time; we call it (by abuse of language) the Viterbi alignment.

The hidden Markov alignment model (HMM) is a generalization of the IBM model 2 (Vogel et al., 1996). In this case, the alignment probability in Equation (2) is expressed by $p(a_j|a_{j-1}, n)$, where the alignment probability is designed by a first-order dependency: the alignment of a target word depends of the alignment of the preceding one. The Viterbi alignment is obtained by dynamic programming in $O(mn^2)$ time.

A simple transpotting algorithm consists in computing the Viterbi alignment of a sentence pair using either an IBM model 2 or an HMM, then the target words which are aligned to the (source) query words correspond to the transpot. We call those algorithms `IBM2` and `HMM` respectively. Unfortunately, this approach tends to produce discontinuous transpots, most of the time erroneously. To overcome this, we adapt an idea initially described in (Simard, 2003), where for each pair $\langle j_1, j_2 \rangle \in [1, m] \times [1, m], j_1 < j_2$, two Viterbi alignments are computed: one between the target word sequence $t_{j_1}^{j_2}$ and the source query $s_{i_1}^{i_2}$, and the other between the remaining of the two sentences $\bar{s}_{i_1}^{i_2} \equiv s_1^{i_1-1} s_{i_2+1}^n$ and $\bar{t}_{j_1}^{j_2} \equiv t_1^{j_1-1} t_{j_2+1}^m$. The transpot $\hat{t}_{\hat{j}_1}^{\hat{j}_2}$ is then obtained by maximizing:

$$\underset{j_1, j_2}{\text{argmax}} \left\{ \max_{a_{j_1}^{j_2}} p(a_{j_1}^{j_2}|s_{i_1}^{i_2}, t_{j_1}^{j_2}) \times \max_{\bar{a}_{j_1}^{j_2}} p(\bar{a}_{j_1}^{j_2}|\bar{s}_{i_1}^{i_2}, \bar{t}_{j_1}^{j_2}) \right\} \quad (3)$$

Thanks to dynamic programming, this maximization can be computed in $O(mn)$ with IBM model 2 and $O(mn^2)$ with HMM. We call these algorithms `C-IBM2` and `C-HMM` respectively.

These transpotting algorithms can be enhanced by using lexical distributions in both translation directions (IBM models are not symmetrical). For this, the lexical probability of a target word t given a source word s is reformulated as:

$$p_{bi}(t|s) = \phi(p_{S \to T}(t|s), p_{T \to S}(s|t)) \quad (4)$$

where $\phi(\cdot)$ is a function to define that combines lexical probabilities from both translation directions. This enhancement does not the alter the time complexity of the aforementioned algorithms. For HMM, we call this algorithm `C-HMM-bi`.

Finally, Callison-Burch et al. (2005) proposed to use phrase-based translation models for transpotting. MOSES offers a state-of-the-art toolkit for extracting a phrase-based model from a given bitext (Koehn et al., 2007). Using such a model, transpotting is done by searching in the phrase table the set of candidate segment pairs whose source segments equals the source query and the target segment occurs in the target sentence. The scores associated with each segment pair in the phrase table enables to keep the best transpot among candidates. We call this method `PBM`.

4 Experimental setup

4.1 Corpus

We used the Canadian Hansards bilingual corpus made of the proceedings of the Canadian parliament from 1986 to 2007. It is composed of more than 8.3 million sentence pairs. In order to measure how transpotting algorithms are impacted by the frequency of a translation, we designed two test corpora. TESTFREQ enables the evaluation of the transpotting algorithms when aligning frequent translations, and TESTRARE when aligning against rare translations.

Both corpora are built semi-automatically by extracting sentence pairs containing query/translation pairs and then by validating manually the correctness of those pairs. For this, we rely on real user queries we obtained from the logs of the commercial bilingual concordancer TransSearch.

We crossed the most frequent queries from these logs with an in-house bilingual dictionary, and for each query/translation pair obtained, we counted its number of occurrences in the Hansards. This allows to discriminate frequent from rare translations. This gave two sets of sentence pairs: one in which query/translation pairs are frequent, occurring in 594 sentence pairs on average in the Hansards, the other in which the query/translation pairs are rare, occurring in only one sentence pair in the Hansards.

It is interesting to remark that these rare query/translation pairs cooccur only once in the Hansards, but that the queries and the translations taken separately do occur frequently: the queries occur in 4959 sentence pairs on average, and the translations in 1977 on average. This means that the queries and the translations are not rare, but their cooccurrences are.

Last, since the translations were automatically identified using a dictionary, we had to manually check the correctness of the translations to ensure the quality of the reference corpora. We manually examined each sentence pair to validate the query/translation pair.

We ended up with 1516 sentence pairs in TESTFREQ, and 706 sentence pairs containing a query and a rare translation. From these, we kept 115 sentence pairs for a development corpus called DEVRARE; the remaining 591 pairs formed the TESTRARE corpus.

4.2 Metrics

We considered two accuracy metrics for evaluating the transpotting algorithms. The first one is the percentage of times the reference translation is correctly identified by an algorithm. For example, when the reference translation is *qui vivra verra*, the transpot has to match it exactly to be credited a point.

Often, transpotting algorithms are short of one word when identifying the translation of a query. Therefore, we computed a second metric where we give a point to a transpot with at least one word correct. For example, when the translation is *qui vivra verra*, the transpot *verra* is considered correct with this metric.

Since the transpotting algorithms always return a transpot, the number of returned transpots is the same as the number of reference translations; so for these tasks precision equals recall.

4.3 Training and Tuning

The distributions required by the word-based transpotting algorithms described in Section 3 were obtained by running GIZA++ on the Hansards (Och and Ney, 2003). IBM model 4 lexical distributions were used in this work, while the alignment distributions were coming from IBM2 or HMM models. The phrase table required by the phrase-based transpotting algorithm has been computed by MOSES in its default configuration.

Equation (4) which combines lexical distributions obtained by training models in both translation directions requires a combination operator. We obtained the best results on DEVRARE using the geometric mean. Equation (1) which combines global and local lexical distributions requires the optimization of the λ parameter that controls the linear combination. We optimized it on DEVRARE for each transpotting algorithm.

5 Experiments

5.1 Frequent Translation Spotting

The first experiment compares the transpotting algorithms described in Section 3 for transpotting the queries of the TESTFREQ corpus. Results are presented in Table 1.

Algorithm	% ref. found	% 1 word found
IBM2	61.8	84.4
HMM	61.0	83.1
C-IBM2	68.3	94.2
C-HMM	69.3	94.8
C-HMM-bi	72.3	**96.3**
PBM	**77.2**	93.5

Table 1: Scores of the transpotting algorithms on the TESTFREQ corpus.

As we could expect, PBM is the best algorithm for identifying the reference translations. Then, follow

the 3 constrained models that overpass the 2 unconstrained ones. For constrained models, the HMM-based model logically outperforms the IBM-based model, whereas surprisingly, the opposite result is observed for unconstrained models.

According to the percentage of times at least one word of the reference translation is correctly identified, the constrained models overpass PBM. This shows that constrained models are more recall oriented, while PBM is more balanced between recall and precision.

We also observe that the percentage of times where at least one word of the translation is found are around 20 points higher than those where the whole reference translation is found. This indicates that the algorithms are good at spotting the region of the translation in the target sentence, but often lack precision at identifying their exact boundaries. Often, this is due to the insertion of a grammatical word at the frontier of a reference translation. Those words tend to be highly ranked in lexical distributions, a problem analyzed in (Moore, 2004). We illustrate in Section 5.4 that missing word boundaries can lead to very misleading translations.

5.2 Rare Translation Spotting

The second experiment compares the algorithms when transpotting the queries of the TESTRARE corpus. Results are presented in Table 2.

Algorithm	% ref. found	% 1 word found
PBM	32.5	47.9
IBM2	43.8	58.5
HMM	47.4	63.3
C-IBM2	51.3	70.4
C-HMM	54.3	73.6
C-HMM-bi	**65.6**	**83.9**

Table 2: Scores of the transpotting algorithms on the TESTRARE corpus.

For both metrics, we observe that the scores of all algorithms drop significantly. This shows that identifying rare translations of a query is harder than identifying frequent ones, and this, even if rare translations in our test set are composed of frequent words (see Section 4.1).

The large drop of the PBM algorithm might seem surprising. It is explained by the fact that often, a query/rare translation pair is not kept by the grow-diag-final heuristic of MOSES, due to a failure of the word alignment models being used. This highlights a shortcoming of phrase-based models when dealing with rare events, as described in (Foster et al., 2006). Note that some kind of SMT decoder could be used to enhance the PBM algorithm. Nevertheless, it seemed more rational to us to focus on pure word-alignment based techniques.

5.3 2-stage Translation Spotting

The third experiment evaluates the performance of the 2-stage method described in Section 2. We considered the three best transpotting algorithms for identifying rare translations, as identified in the previous section. Results are presented in Table 3 when using 200 non parallel sentence pairs for building the artificial bitext of each sentence pair. In Figure 5, the results are plotted as a function of the number of non parallel sentence pairs used.

2-stage w.	% ref. found	% 1-w found	λ
C-IBM2	58.4 *+7.1*	79.0 *+8.6*	0.99
C-HMM	60.9 *+6.6*	83.2 *+9.6*	0.98
C-HMM-bi	**69.7** ***+3.9***	**86.5** ***+2.6***	0.50

Table 3: Scores of the 2-stage transpotting approach on the TESTRARE corpus, as well as its absolute gains, as a function of the transpotting algorithm used. λ stands for the best value given to the global lexical model in Eq. (1).

For the three transpotting algorithms, significant gains are observed when our method is applied with regards to the scores of Table 2. In terms of the percentage of reference translations correctly identified, we observe a gain of nearly 7 absolute points for C-IBM2 and C-HMM, and nearly 4 points for C-HMM-bi. A gain of 4 points is similar to the gain observed in Table 2 when going from IBM2 to HMM or from C-IBM2 to C-HMM.

As can be observed in Figure 5, most of the gains are obtained using only 10 non parallel sentence pairs. For instance, C-HMM-bi shows a gain of 3.1 absolute points. The fact that only a small number of non parallel sentence pairs is required to adapt the lexical model indicates that the cost of the method remains moderate and that it can be deployed in industrial applications.

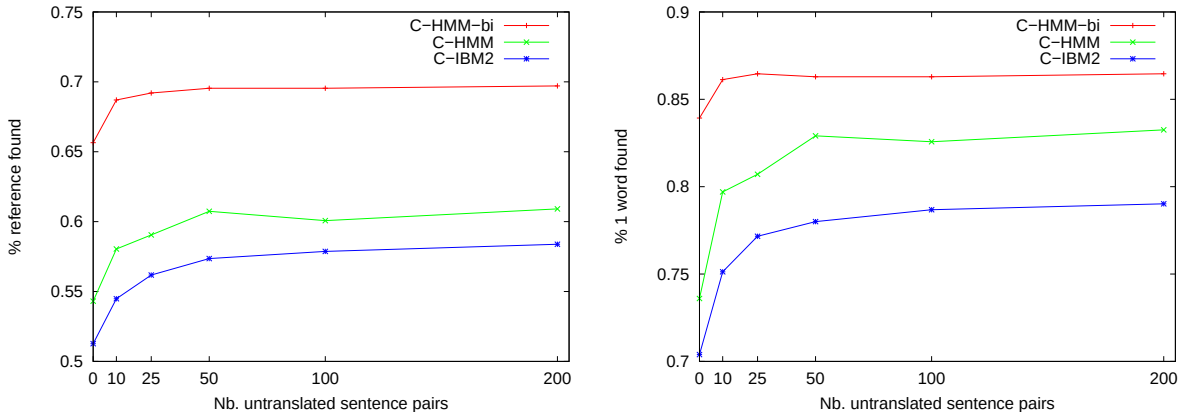

Figure 5: % of times where the reference translation is found (left), and where at least one word of the reference translation is found (right) as a function of the number of non parallel sentence pairs used.

The last column of Table 3 indicates the optimal value of λ used for Eq. (1). For the two unidirectional models C−IBM2 and C−HMM, the confidence given to the local model is very low. Indeed, we observed that when the value of λ is below 0.95, the gain of the 2-stage approach cancels out, and a value below 0.7 even degrades performances. On the contrary, for the bidirectional model C−HMM−bi, the confidence in the local model is higher. For all the values of λ we tested, our method yields significant gains.

These results show that even if the local distribution for a given query is estimated from non parallel sentence pairs, it embeds valuable lexical links between the words of the query and those of the rare translation. These links permit to reinforce the second alignment pass despite the noise contained in the local distribution, which means that it contains more valuable links than noise.

5.4 Qualitative Results

When comparing the results of C−HMM−bi alone and our two-stage approach using C−HMM−bi, we found that the transpot produced for the 591 sentence pairs of TESTRARE differed 82 times: 35 errors were corrected, 11 errors were introduced, and 36 wrong or partial transpots were modified, but not fully corrected. Table 4 shows some examples among those corrected transpots which are rare translations of their query in the training material. Note that more than often, the transpots identified during the first stage are very partial, if not misleading, while the ones identified after the second stage are typically more faithful to the reference translation. This is the case even if the transpots proposed by the two methods often differ in their boundaries only.

Query	Transpot with	
	C−HMM−bi	2-stage C−HMM−bi
input	entrée	entrée de données
lead the way	tête au	être en tête
liability	mort	poids mort
with all due respect	contredire	sans vouloir vous contredire
out of date	des	surannée
corporation	compagnie	compagnie commerciale
take charge	assumer	assumer la responsibilité
in some ways	à certains points	à certains points de vue
hard at work	plein travail	en plein travail
vantage point	avantageuse	position avantageuse
payroll	salaires	traitements et salaires

Table 4: Examples of rare translations from TEST-RARE which are wrong or partial when aligned with C−HMM−bi only, but correct when aligned with our 2-stage approach using the same alignment algorithm.

6 Related Works

Improving word-alignment by exploiting more data has been the focus of some studies. In particular, it was shown that acquiring comparable data from external resources is a fruitful strategy (Munteanu and Marcu, 2006; Abdul-Rauf and Schwenk, 2009; Cettolo et al., 2010; Gahbiche-Braham et al., 2011).

Exploiting external data is mainly a batch procedure, while our approach better exploits the available training data, and this is done online.

The works of (Simard, 2003; Vogel, 2005) are closely related to the constrained alignment approaches described in Section 3. Both allow to extract phrase pairs from bilingual data. Our work attempts to extend this family of methods with the adaptive online method described above.

Other works proposed discriminative approaches for word alignment (Moore et al., 2006; Blunsom and Cohn, 2006; Liu et al., 2010; Setiawan et al., 2010). They rely on manually word-aligned training data which render them hard to generalize and questionable for industrial applications. Dyer *et al.* (2011) proposed a discriminative framework that does not need such manual training data. Comparing our approach to this one would be interesting. In the eventuality that it outperforms our approach, embedding it into our 2-stage framework could be attempted, as planned in future work.

Some works have been proposed to smooth translation models. Toutanova et al. (2002) used POS tags for smoothing translation distributions in the HMM alignment model of Vogel et al. (1996). Moore (2004) proposed to smooth IBM model 1 translation model, especially the count of rare events. Foster et al. (2006) proposed to smooth a phrase-based translation model. These works correct the estimated probabilities of rare events by smoothing lexical distributions. While they attempt to smooth the whole lexical model once for all, we propose a smoothing local to each rare event and dependent of the translation process.

In paraphrase extraction, approaches have been proposed to extract paraphrases from bilingual data by relying on phrase-based alignment models (Bannard and Callison-Burch, 2005; Callison-Burch, 2008; Max, 2010). Although our approach is different, identifying a query's rare translation can been seen as recognizing that this translation is a paraphrase of a query's frequent translation.

7 Conclusion and Future Work

We proposed an original method that improves the transpotting of infrequent translations. To overcome the lack of additional data, this method uses non parallel sentences sampled from the training material in order to adapt the lexical distribution used to transpot a given query. The experiments we conducted exhibit significant gains in identifying rare translations by making use of only a small number of non parallel sentence pairs for each query. This suggests that the method could be implemented at a moderate additional cost in industrial bilingual concordancers.

In this study, we considered only frequency-1 transpots. It would be natural to extend this work to (slightly) more frequent translations and to investigate the frequency threshold over which the approach do not improve the alignment process.

In this work, the source sentences gathered for building the non parallel local bitexts are sampled randomly. This simple approach lead to significant gains. Still, seeking for source sentences which differ the most from the seed target sentence available seems to be intuitively attractive, possibly leading to further improvements.

Finally, our work shows that paying attention to rare events is fruitful in a transpotting task. We plan to investigate whether it can pay off as well in MT. The approach of Lopez (2008) computes online lexical and alignment models during decoding. Our method could easily be integrated in this approach.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve SMT performance. EACL'09, pages 16–23.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. ACL'05, pages 597–604.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proc. of the Fourth Workshop on SMT*, pages 182–189.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. ACL'06, pages 65–72.

Julien Bourdaillet, Stéphane Huet, Philippe Langlais, and Guy Lapalme. 2010. TransSearch: From a bilingual concordancer to a translation finder. *Machine Translation*, 24(3-4):241–271.

Lynne Bowker and Michael Barlow, 2008. *Topics in Language Resources for Translation and Localisation*, chapter A Comparative Evaluation of Bilingual Con-

cordancers and Translation Memories, pages 1–22. John Benjamins Publisher.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Chris Callison-Burch, Colin Bannard, and Josh Schroeder. 2005. A compact data structure for searchable translation memories. EAMT'05, pages 59–65.

Chris Callison-Burch. 2008. Syntactic constraints on paraphrases extracted from parallel corpora. EMNLP'08, pages 196–205.

Mauro Cettolo, Marcelo Federico, and Nicola Bertoldi. 2010. Mining parallel fragments from comparable texts. In *Proc. of the 7th International Workshop on Spoken Language Translation*, pages 227–234.

Alain Désilets, Christiane Melançon, Geneviève Patenaude, and Louise Brunette. 2009. How translators use tools and resources to resolve translation problems: an ethnographic study. In *Proc. of the Workshop Beyond Translation Memories: New Tools for Translators*, MT-Summit XII, pages 26–30.

Alain Désilets, Benoit Farley, Marta Stojanovic, and Frances Urdininea. 2010. Using webitext to search multilingual web sites. AMTA'10, pages 26–30.

Chris Dyer, Jonathan Clark, Alon Lavie, and Noah A. Smith. 2011. Unsupervised word alignment with arbitrary features. ACL'11, pages 409–419.

George Foster and Roland Kuhn. 2007. Mixture-model adaptation for SMT. In *Proc. of the Second Workshop on Statistical Machine Translation*, pages 128–135.

George Foster, Roland Kuhn, and Howard Johnson. 2006. Phrasetable smoothing for statistical machine translation. EMNLP'06, pages 53–61.

Souhir Gahbiche-Braham, Hélène Bonneau-Maynard, and François Yvon. 2011. Two ways to use a noisy parallel news corpus for improving statistical machine translations. In *Proc. of the 4th Workshop on Building and Using Comparable Corpora*, ACL'11, pages 44–51.

Nikiforos Karamanis, Saturnino Luz, and Gavin Doherty. 2011. Translation practice in the workplace: contextual analysis and implications for machine translation. *Machine Translation*, 25(1):35–52.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proc. of the Second Workshop on SMT*, pages 224–227.

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. ACL'07, pages 177–180.

Philipp Koehn. 2009. A process study of computer-aided translation. *Machine Translation*, 23(4):241–263.

Yang Liu, Qun Liu, and Shouxun Lin. 2010. Discriminative word alignment by linear modeling. *Computational Linguistics*, 36(3):303–339.

Adam Lopez. 2008. Tera-scale translation models via pattern matching. COLING'08, pages 505–512.

Elliott Macklovitch, Guy Lapalme, and Fabrizio Gotti. 2008. TransSearch: What are translators looking for? In *AMTA'08*, pages 412–419.

Aurélien Max. 2010. Example-based paraphrasing for improved phrase-based statistical machine translation. EMNLP'10, pages 656–666.

Robert C. Moore, Wen-Tau Yih, and Andreas Bode. 2006. Improved discriminative bilingual word alignment. ACL'06, pages 513–520.

Robert C. Moore. 2004. Improving IBM word alignment model 1. ACL'04, pages 518–525.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. COLING-ACL'08, pages 81–88.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Tina Paulsen Christensen and Anne Schjoldager. 2010. Translation-memory (TM) research: What do we know and how do we know it? *Journal of Language and Communication Studies*, 44:89–101.

Hendra Setiawan, Chris Dyer, and Philip Resnik. 2010. Discriminative word alignment with a function word reordering model. EMNLP'10, pages 534–544.

Michel Simard. 2003. Translation spotting for translation memories. In *Proc. of the Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and beyond*, NAACL'03, pages 65–72.

Kristina Toutanova, H. Tolga Ilhan, and Christopher D. Manning. 2002. Extensions to HMM-based statistical word alignment models. EMNLP'02, pages 87–94.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. COLING'96, pages 836–841.

Stephan Vogel. 2005. PESA: Phrase pair extraction as sentence splitting. MT-Summit X, pages 251–258.

Jian-Cheng Wu, Kevin C. Yeh, Thomas C. Chuang, Chung-Li Tao-Yuan, Wen-Chi Shei, and Jason S. Chang. 2003. TotalRecall: A bilingual concordance for computer assisted translation and language learning. ACL'03, pages 201–204.