Towards Contextual Adaptation for Any-text Translation

Li Gong, Aurélien Max, François Yvon

LIMSI-CNRS & Univ. Paris Sud
Orsay, France
{firstname.lastname}@limsi.fr

Abstract

Adaptation for Machine Translation has been studied in a variety of ways, using an ideal scenario where the training data can be split into "out-of-domain" and "in-domain" corpora, on which the adaptation is based. In this paper, we consider a more realistic setting which does not assume the availability of any kind of "in-domain" data, hence the name "any-text translation". In this context, we present a new approach to contextually adapt a translation model *on-the-fly*, and present several experimental results where this approach outperforms conventionaly trained baselines. We also present a document-level contrastive evaluation whose results can be easily interpreted, even by non-specialists.

1. Introduction

It is now a well-established fact in Statistical Machine Translation that systems must be adapted to each particular input text. Adaptation has been tackled in a variety of ways (see e.g. [1, 2, 3]), most notably by adapting the translation model, by adapting the target language model, and by adaptating the tuning set. In most of these works, it is assumed that the bilingual training corpus can be partitioned into "in-domain" and "out-of-domain" subsets relative to the input text, and that there exists some smaller "in-domain" held-out corpus to tune the system. In typical settings, large bilingual corpora are collected opportunistically; as a result, the amount of data that do not resemble closely the input text largely outweights the data that appear to be the most relevant.

Using as much data as is available for a given language pair is necessary to alleviate the data sparseness issue through better coverage: in particular, it seems to improve the alignment of some rare translation units, which would otherwise be misaligned, and yield inappropriate phrase pairs. On the other hand, adding more bilingual data increases the possibility of encountering new translations, and makes the translation of phrases more ambiguous, sometimes in a detrimental way, since not all corresponding translations (or senses) are appropriate for the input text. The data sparseness and the ambiguity problem thus entertain a repulsion relationship that is at the core of the adaptation problem (see e.g. [4]), even though the recent work of Haddow and Koehn [5] concludes that good coverage is more important than appropriate scoring: adding out-of-domain corpora containing examples of rare units benefits more to translation than the inclusion of inappropriate examples of frequent units harms it.

A practical solution is to use all the available training data, but to consider differently translation examples depending on their relevance to the input text, possibly at the corpus [1], sentence [6] or phrase [3] level. As noted e.g. by Haddow and Koehn [5], although the in-domain vs. out-of-domain distinction is frequently used, precise definitions are still lacking; in their words, "it is normally understood that data from the same domain is in some sense similar (for example in the words and grammatical constructions used)" and, in their experiments, they characterize domain differences in terms of word distributions and out-of-vocabulary rates. While some domain distinctions are clearly undebatable, such as when opposing e.g. News commentaries and parliamentary speeches, other distinctions may in fact be more difficult to draw when one considers arbitrary text inputs, as may be submitted to online translation services.

In this work, we consider a case that has been so far comparatively less studied, where the characteristics of the input text are completely unknown before translation. We thus make the following assumptions:

- The input text is short and corresponds to a coherent discourse (i.e. is not made by concatenating unrelated documents).
- The text can be from any arbitrary domain, which precludes any realistic off-line adaptation using any predefined specific bilingual corpora; therefore, the only "in-domain" corpus available is the input text itself.
- No adapted development corpus is available, which precludes the use of tuning techniques relying on a development bitext from the same data source or domain.
- Training data was collected opportunistically and no specific document metadata (e.g. genre, document boundaries) are available for the full data set.

Note that the issues of adapting alignments and target language models will not be considered in this work. As to the former, it has previously been shown that using all the available corpora during word alignment tends to improve

translation performance [7, 5], so our word alignment models will be built offline using all available parallel data. As to the latter, there is a large body of works addressing language model adaptation which all report improvements over non-adapted language models (e.g. [1]). We leave it to our future work to evaluate whether the effects of all types of adaptations can be compounded.

This paper is to our knowledge the first attempt at studying the scenario of what we call here "any-text translation", with the notable absence of some predefined identifiable in-domain training and tuning corpora. An important aspect of our scenario is that there is no guarantee that appropriate data will be available for the input text as regards e.g. genre, phraseology, theme vocabulary, or even effects of original language. Thus, adaptation will be performed with the objective of modeling some *a priori* confidence into the system's ability to translate short translation units.

Another consequence of our setting is that online adaptation is necessary and is in fact the only solution. We therefore propose an *on-the-fly* pipeline consisting of the following stages : sampling at the level of translation units is performed (similarly to [8, 9]) for selecting sentences from the training data, and instance weighting is applied for scoring phrase pairs (e.g. [6]). Based on these computations, two additional scores are then produced: the first estimates the *goodness* of each collected source phrase as a translation unit for the language pair at hand; the second estimates how much confidence should be put in the adaptated translation distribution for each source phrase[1]. An important result of the paper will be the description of a document-level contrastive evaluation scheme that enables a more interpretable analysis of the differences between two systems.

The rest of this article is organized as follows. We first describe our approach to on-the-fly instance weighting for adapting translation models (section 2). We then describe how to model the goodness of source phrases (section 3) and to compute confidence scores for (adapted) translation distributions (section 4). The experimental section (section 5) is decomposed into a description of data sets (section 5.1), systems (section 5.2), and evaluation settings (section 5.3). We next present the main experimental results (section 5.4) and discuss them in relation to previous works (section 6). We finally conclude and describe plans for future work.

2. Instance-weighting for contextual adaptation

Adaptation can be tackled as a data selection problem: given an in-domain training corpus and out-of-domain corpora, a fixed number of sentences are selected in the out-of-domain corpora on the basis of their similarity to the in-domain corpus. These sentences may be denoted as *pseudo in-domain data* [11], where it is hoped that, given the selected number of sentences to draw, performance will be improved. This approach is in fact flawed in a particular respect, as it does not provide any guarantee that instances of rare units will be selected, specifically if they do not occur in sentences resembling the in-domain data. This has been sometimes solved by ad-hoc strategies to recover infrequent units [12].

We would like instead to make use of all available training corpora. Sampling at the level of phrases is an efficient solution to achieve this goal [8, 9]. Indeed, suffix arrays [13] offer fast access to phrase instances in large corpora, and can be used to select a given number of instances of phrases, rather than sentences, thereby ensuring that all the phrases present in a corpus are appropriately covered.[2]

Previous approaches to sampling have resorted to *random deterministic sampling*, which picks a given number of examples by scanning the suffix array index at fixed intervals (hence the apparently random, and actually deterministic, behavior). This, of course, is sub-optimal as it does not attempt to select the most appropriate data for the input text. We may instead resort to criteria that are often used in data selection approaches: Information Retrieval similarity measures such as `tf.idf` and Information Theory measures such as perplexity.

Once a sample has been collected for every source phrase, (pre-computed) word alignments are retrieved to extract the corresponding translations. Assuming a set of retrieved sentences and their individual similarity score, denoted as w_i, the adapted translation model can be estimated by weighting each example with the corresponding sentence weight [6]:

$$p_{iw}(e|f) = \frac{\sum_{j \in T_f \cap T_e} w_j c_j(e,f)}{\sum_{j \in T_f} w_j c_j(f)}, \qquad (1)$$

where T_f (resp. T_e) is the set of source (resp. target) sentences containing f (resp. e), and $c_j()$ is the count function.

3. Estimating the goodness of translation units

Given that our sampling strategy ensures that all occurrences (up to a maximum sampling size) of each source phrase will be retrieved, all source phrases that are found in the training corpus will also be present in the phrase table. Although no definitive criterion as to what constitutes a good phrase translation unit has emerged[3], the two following criteria have been proposed:

[1] Note that in the present work, the effect of this score will only be to act as a *segmentation* model, so that some segmentation may be preferred over some other. Future work will include searching for more translation examples for those unreliable phrases, as hinted by [5], and having recourse to automatic paraphrasing (e.g. [10]) of those phrases.

[2] Callison-Burch *et al.* [8] found that a sample size of 100 was sufficient for German-to-English phrase-based SMT, while Lopez [9] determined that 300 was an appropriate value for Chinese-English hierarchical SMT. We will use a larger sample size of 1,000 in our experiments in an attempt to let instance weighting find the most appropriate examples from a larger sample.

[3] For instance, limiting phrases to constituents was found to be sub-optimal [14]. The very definition of what a *phrase* is with respect to the SMT problem poses many interesting research questions, see e.g. [15].

- Given some word alignment between a source and target parallel corpora, the absence of an aligned target phrase for a given source phrase may suggest that the corresponding failure of the extraction process should be accounted for in the translation model. Lopez [9] therefore proposes the following *coherent* estimate of the translation conditional probability:

$$p_{coherent}(e|f) = \frac{c(f,e)}{c(f)} \quad (2)$$

  where $c(f)$, the number of occurrences of the source phrase, corresponds to the total number of attempted extractions, in lieu of the traditional summation over all extracted translations for f, $\sum_{e'} c(e', f)$.

- It has been observed that the traditional heuristic approach to phrase pair extraction does not offer a consistent view over the training and the actual use of phrases by decoders. It is thus possible to have recourse to a forced alignment which results in the decoder producing what it believes is the best alignment for a given training sentence. Wuebker *et al.* [16] implement this idea using *leaving-one-out*, so that the phrase examples for each training bi-sentence are not used to decode it, and subsequently estimate their system's models on the resulting alignment. Even though this intuition does not guarantee that the retained phrases are *intrinsically* good translation units, they were selected as pertaining to best derivations allowing to reproduce the reference target sentence.

We exploit the two above ideas as follows. First, we use some pre-trained standard phrase-based system to translate its own training corpus. Instead of sticking strictly to leaving-one-out, we simply remove from the system's phrase table all source phrases occurring only once, corresponding mostly to long phrases. In addition, we consider all phrases coherent with the resulting alignment (i.e. coherent sub- or super-phrases) as candidates for extraction. Then, for all the selected occurrences of a given source phrase f, we count how many times f has both a coherent alignment in the original alignment (using GIZA++ in our case) and in the decoder alignment, and normalize by the number of occurrences of that source phrase[4]. The following calculation was used as a new feature in our experiments:

$$h_{goodness}(f) = \frac{c_{coherent}(f)}{c(f)}, \quad (3)$$

where $c_{coherent}(f)$ denotes the count of instances of phrase f being coherent with respect to both the training and decoding alignment.

[4]This can be done w.r.t. to the full corpus or a to particular sample, depending on the configuration studied.

4. Confidence estimation for adapted translations probabilities

Phrase scoring strategies used in conventional phrase-based SMT systems are based on simple count ratios and can thus be criticized on the following grounds :

1. A source phrase occuring rarely will result in its translations being over-estimated[5].
2. A majority of *inappropriate* examples for a given source phrase will result in incorrect translations being more likely for the translation model[6].

The instance-weighting scheme presented in section 2 allows us to assign an adapted weight to each individual example: in some sense, this weight should reflect the confidence that the associated translation is contextually appropriate. Intuitively, an example matching only loosely the context of the input sentence should not partipate much to the confidence that the final translation distribution is correctly estimated. The worst-case scenario would be if all available examples were poor matches (such as examples for an incorrect translation sense for a polysemous phrase). Conversely, a perfect match (such as finding in the training data the very input sentence or a very close match) would indicate that the translation distribution was derived from appropriate data, at least for this example.

In addition to the appropriateness of the examples used, their number should also participate in estimating the confidence in a translation distribution. Given a particular number of examples for a source phrase, the least informative, or least *committing*, situation would be one in which all translation examples are different, yielding the following conditional entropy:

$$H_{unif}(f) = -\sum_e p(e|f)\log(p(e|f)) = \log(\frac{1}{c(f)}) \quad (4)$$

Intuitively, the better the examples used for contextual estimation of a phrase's translations, and the better the instance-weighting scheme, the more the conditional entropy for that phrase should be reduced, as translation alternatives should be restricted to a few synonymous translations. The information gain measured as a difference of entropy values between the previous situation and the more informative situation of a given model provides some account of how much confidence should be put in the collective contribution of all weighted examples. We thus used the following as a new feature in our experiments involving adapted translation models:

$$\begin{aligned} h_{confidence}(f) &= H_{unif}(f) - H(f) \quad (5) \\ &= -\log(\frac{1}{c(f)}) + \sum_e p_{iw}(e|f)\log(p_{iw}(e|f)) \end{aligned}$$

[5]Inverse translation models and lexical weighting are in a way meant to compensate for this.

[6]Context-dependent phrase tables (e.g. [17]) is a way to address this.

Corpus		#lines	#tok.en	#tok.fr	ppl.en	ppl.fr	oov.en	oov.fr
tuning	newsco (in)	934	22.4K	25.3K	316.19	211.07	629	273
	ted (out)	934	19.6K	20.3K	265.63	164.57	238	273
test	newsco	1,859	44.2K	48.8K	307.14	222.79	1,700	1,558

Table 1: Tuning and test documents statistics

Corpus	domain w.r.t. test	# lines	# tokens.en	# tokens.fr
newsco	in	137K	3,381M	4,017M
epps	out	1,982M	54,170M	59,702M
newsco+epps	mixed	2,119M	57,551M	63,790M

Table 2: Training corpora statistics

This value increases when either the number of examples for f is high or when the entropy of the adapted translation distribution is low.

5. Experiments

We now describe experiments intended to show whether on-the-fly contextual adaptation can improve over standard estimation of translation models, as well as over a standard way of combining translation models estimated from different corpora. For this, we resort to data conditions that simulate short input documents and training corpora for which the in-domain part is either clearly identified or dissolved in a larger corpora, and use three scenarios where an out-of-domain, an in-domain and a perfect tuning set is available[7]. For each system configuration, we compute traditional evaluation metrics over the full document collection (as is typically done with corpus-based metrics such as BLEU). We also propose a new document-based evaluation method that is more appropriate for the problem at hand.

5.1. Data sets

Experiments were performed on the English-French language pair in both directions, using data released for the evaluation track of the Workshop on Statistical Machine Translation[8]. Our test document collection, described in Table 1, also stems from WMT data: it consists of a set of 76 News commentary documents (from newstest2009).

We use the tuning sets described in Table 1: one is "in-domain" (in its traditional sense in SMT) w.r.t. to our test corpus (newsco), and one is out-of-domain and is taken from presentations from TED talks[9] (ted). These conditions allow us to compare situations where tuning corpora of various degrees of appropriateness are available and can be identified as more appropriate; we will also simulate the availability of a "perfect" tuning set by performing self-tuning.

Lastly, our training corpus, described in Table 2, contains two sub-corpora of in-domain News commentaries (newsco) and out-of-domain parliamentary debates (epps). These sub-corpora will be either used separately or jointly.

[7]Performing tuning set adaptation at the document-level as in [18] will be part of our future work.

[8]http://www.statmt.org/wmt12

[9]Available from IWSLT'11: http://iwslt2011.org

5.2. Systems

5.2.1. *Off-line baseline systems*

We build standard phrase-based systems using moses[10], and use MERT for tuning parameters. We compare the following conditions: training on all available data (newsco+epps), as well as using two separate phrase tables built from newsco and epps (i.e. multiple alternative decoding paths) as is standard practice in domain adaptation where corpus boundaries are known [1].

5.2.2. *On-the-fly adapted systems*

We build various adapted systems on-the-fly. All use the word alignments produced by Giza++[11] on the full newsco+epps corpus, as out-of-domain data may improve alignment quality in our situation [7]. We test the three following sampling and instance-weighting strategies for estimating translation model: (a) random sampling and uniform weighting [8, 9] (RND), (b) using tf.idf values of training sentences [19] (IR), and (c) perplexity values of training sentences relative to each test document (PPL).[12]

An important difference with our baseline systems is that we do not estimate a back-translation model ($p(f|e)$) as this proves costly using sampling; [9] reported that this model does not have a significant impact on translation performance for large training corpora. Furthermore, we believe that such a model should in fact not be needed, were the translation model appropriately estimated (i.e. contextually appropriate), as there would be no need to compensate for the "ambiguity" in this model by considering the reverse direction.[13]

We build variants where we consider one translation model in isolation (RND, IR, PPL) as well as our source phrase goodness model (section 3) and our translation distribution confidence estimation (section 4). Parameter tuning is performed once for all with MERT, considering the tuning set as a single document. For testing, we build an adapted translation model for each document and use the previously tuned parameters to decode using the moses decoder. For self-tuning, which simulates the availability of a (smallish)

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

[11]http://code.google.com/p/giza-pp/

[12]Note that scoring test examples at the sentence level, as done e.g. by [6], might be sub-optimal: we would rather consider thematically-coherent units from the training corpus. We did not have this information at our disposal here, but plan to perform automatic thematic segmentation of the training corpora as part of our future work. Note also that the target side of our "in-domain" corpus (i.e. test documents) was not available for adaptation.

[13]The argument also holds for lexical weighting models, which are meant to model intra-biphrase cohesion.

perfect tuning set for each input document, each document is tuned independently using the reference corpus and the best optimization point is used for testing; this is obviously an oracle situation, and will be denoted as such in our results for `moses` and our adapted systems.

5.3. Evaluation setting and contrastive document-level evaluation

We will compare our various settings using the well-established BLEU [20] and TER [21] metrics, using initially the full test corpus made up of the full collection of documents. Absolute values being always difficult to interpret, we propose to resort to contrastive evaluation between two systems. Our contrastive document-level evaluation is performed as follows: given two systems we wish to compare, a single configuration, and a target evaluation metrics, we look on *a per document basis* which system outperformed the other for some interval (e.g. "1-2 BLEU increase", "0.5-0.75 TER decrease"). We then compute statistics over the entire document set. Considering a particular significance level for the selected metrics, we can then report the percentage of cases the first system outperformed the second system, the other way round, and when they leveled each other out, corresponding to figures that are easier to interpret.

5.4. Results

The results of all systems under our three tuning conditions are given in Table 3. It immediately stands out that we are looking at two very different situations: on the one hand, French to English translation shows a clear advantage of the adapted systems over both `moses` and the adapted baseline (`2-tables`) under all tuning conditions; on the other hand, English to French translation calls for a closer look at results as no immediate conclusion can be drawn.

Tuning condition Considering first the most likely scenario for any-text translation, we look at results obtained when using an out-of-domain tuning set for all systems. On English to French, we find that the `IR` and `PPL` system can achieve slightly better performance than `moses`, which in turn performs slightly better than `RND`. The `2-tables` adaptation system clearly failed to improve over any other system. On French to English, the situation is comparable with the exception of two differences: `IR` and `PPL` now achieve a significant improvement over `moses` (resp. +1.15 and +1.12 BLEU point), and `2-tables` now performs slightly better than `moses`.

The in-domain condition, where a tuning set from the same domain as the test set is used, exhibits a similar pattern: on English to French, `moses` and our best adapted systems are almost indistinguishable, and the `2-tables` system performs comparatively poorly. On French to English, `2-tables` now performs slightly better than `moses`, while our adapted systems outperform again the latter (+0.77 BLEU point for `IR` and +0.57 BLEU point for `PPL`).

	English → French						French → English					
	tuning condition											
	out		in		oracle		out		in		oracle	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
`moses`	28.15	55.27	28.32	**56.72**	30.07	56.61	28.36	55.66	29.46	52.07	32.11	52.69
`2-tables`	27.80	55.31	26.91	58.71	-	-	28.49	55.16	29.53	51.90	-	-
`RND`	28.01	55.15	28.17	57.12	-	-	28.24	56.44	29.99	51.83	-	-
`IR`	**28.36**	**54.83**	**28.42**	56.86	-	-	28.59	55.74	29.57	52.08	-	-
`+good`	28.07	55.34	28.13	57.13	-	-	29.11	54.69	30.01	**51.68**	-	-
`+conf`	27.74	55.27	28.25	57.13	-	-	**29.51**	**54.12**	29.66	52.05	-	-
`+all`	28.17	55.07	27.92	57.45	**30.12**	56.70	28.76	54.98	**30.23**	51.80	31.74	53.52
`PPL`	**28.32**	55.09	27.99	57.46	-	-	28.76	55.27	30.03	51.81	-	-
`+good`	**28.34**	55.15	28.21	57.39	-	-	28.86	54.75	29.54	52.33	-	-
`+conf`	28.22	55.42	28.12	57.60	-	-	29.36	**54.16**	29.51	52.16	-	-
`+all`	27.89	55.25	27.87	57.74	30.03	**56.33**	**29.48**	54.34	29.76	51.95	**32.78**	**51.70**

Table 3: BLEU and TER results. Highest values in a given column appear in bold.

Comparing results between the out-of-domain and in-domain conditions makes the English to French situation look even more complex: there seems to be no marked regular differences between systems tuned with these two conditions (e.g. only +0.17 BLEU point improvement for `moses`). The situation is much clearer on French to English, where all systems benefit from in-domain tuning (e.g. +1.1 BLEU point improvement for `moses`).

Lastly, oracle tuning conditions yield again two different results: `moses` and the two adapted systems are indistinguishable in English to French, while on French to English we find `PPL` to be superior to `moses` (+0.67 BLEU point), itself superior to `IR` (+0.37 BLEU point). In all conditions, we note a substantial improvement over out-of-domain and in-domain tuning (e.g. for `PPL` up to 2.16 BLEU point over in-domain tuning on English to French and 3.02 on French to English). This last result clearly emphasizes the need for performing document-level adaptation for tuning, something that will be addressed in our future work. It also shows that improvements through better tuning are possible even for the (apparently difficult) English to French language pair, where in-domain tuning did not achieve a superior result than out-of-domain tuning.

Adaptation scenarios No instance-weighting scheme (`IR` or `PPL`) appears to clearly outperform the other: they stand in close range in the out-of-domain tuning condition, while `IR` has a slight advantage in the in-domain condition and the `PPL` oracle performs better in French to English. Our two additional features (`+good` and `+conf`) both proved useful under different situations; we can only observe a small tendancy of `conf` to perform better in the out-of-domain condition in French to English. Furthermore, their combination never leads to improvements on English to French, adding to the previously mentioned complexity of this language pair in our experiments.

Contrastive document-level evaluation Pair-wise contrastive results for a set of selected systems and the full range of tuning conditions are given in Table 4, where we consider differences over 0.5 BLEU point. These results allow us to obtain a more interpretable analysis of the comparison between any two systems. For instance, `IR+all` obtained a small advantage of +0.40 BLEU point over `moses` in the

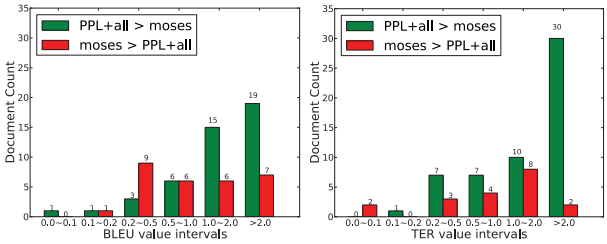

Figure 1: Document-level comparison for any-text translation: green bars (resp. red bars) show number of documents per BLEU (left side) or TER (right side) intervals for which `PPL+all` outperformed `moses` (resp. the other way round) in the out-of-domain tuning condition for French to English.

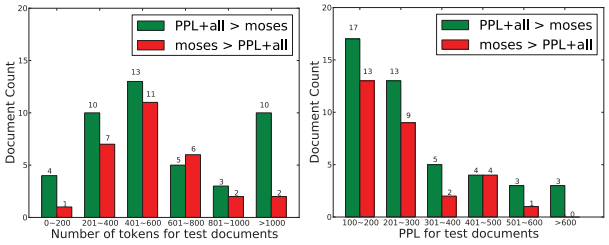

Figure 2: Document-level comparison depending on: (left) test document length (in tokens); (right): test document perplexity. Green bars (resp. red bars) show the number of documents per bins for which `PPL+all` outperformed `moses` (resp. the other way round) in the out-of-domain tuning condition for French to English translation.

French to English out-of-domain condition; however, this translates as 43.42% of documents for which `IR+all` outperforms `moses` (by 0.5 BLEU point or more), and 34.21% for the opposite. Computing those values on a large set of test documents would provide us with some probability that a given system would perform better at translating a new document than some other system, while corpus-based BLEU would give higher importance to longer documents, thus introducing a bias to their respective adaptation situation.

6. Discussion

Our experiments have shown that on-the-fly contextual adaptation could lead to significant improvements over several baselines, including one that exploits translation models derived from different domains. These results shed a new light on the complexity of the adaptation problem and provided concrete examples to illustrate the complexities of conditions under which adaptation can be successful. Furthermore, the oracle self-tuning condition demonstrated the sub-optimality of using large supposedly "in-domain" tuning sets, and our experiments more generally have provided arguments in favor of a document-level adaptation.

Our most salient result in relation to our target scenario of *any-text* translation is that when no well-adapted tuning set is available, i.e. in the out-of-domain tuning condition, the proposed instance-weighting schemes significantly improved over both a `moses` baseline and an adaptation baseline at the corpus-level (`2-tables`) in French to English translation. This condition is illustrated by the histogram on Figure 1, where it is clearly apparent that adaptation at the document-level was very successful in this case, using both BLEU and TER metrics (we note, for instance, that `PPL` improved the translation of 30 documents by a 2 or more TER points decrease compared to `moses`, while the opposite case was found for only 2 documents).

As the synthesis of all test documents shows significant improvement, we may question whether this result would be due to the length of the document, with the intuition that longer documents would allow for better adaptation[14], or to the similarity of documents, with the intuition that documents that have close matches in the training corpus should be translated better. Figure 2 displays results of a document-level contrastive comparison between the same systems of Figure 1 for document length and perplexity values intervals. If we obtain a very clear advantage for our adapted system for documents over 1,000 tokens, this result is also true (though based on a somewhat limited number of documents) for the shortest documents. Likewise, our adapted system clearly performs best for both test documents of low perplexity values, and test documents of high perplexity values.

The question remains of why the English to French language pair resulted in such a different set of observations. We have a number of hypotheses to account for this:

- For this language pair, the advantages of in-domain tuning vs. out-of-domain tuning were non-apparent for all systems, including our `moses` baseline, a fact that seems counter-intuitive.
- The perplexity values of both the in-domain and out-of-domain tuning sets w.r.t. to the training corpus are much higher on English than on French (see Table 1), suggesting that the English texts in our sets use a more "complicated" language. Note also that in the case of our test corpus and in-domain tuning corpus, English texts have significantly more out-of-vocabulary (**oov**) tokens. As the same texts are available in both languages, the differences cannot be attributed to thematic differences w.r.t. the training set.
- It may also be the case that English as an original language, resulting in a more complex language as opposed to when English is the result of translation (i.e. translationese), is less present in our training data. In fact, considering our Europarl data only (`epps`),

[14] Recall that in our settings documents from the training set were limited to single sentences, something we plan to improve on.

	$moses_{out}$	$moses_{out}^{2tables}$	IR_{out}^{+all}	PPL_{out}^{+all}	$moses_{in}$	$moses_{in}^{2tables}$	IR_{in}^{+all}	PPL_{in}^{+all}	$moses_{oracle}$	IR_{oracle}^{+all}	PPL_{oracle}^{+all}
$moses_{out}$	-	39.47	34.21	25.00	28.95	26.32	18.42	25.00	11.84	10.53	13.16
$moses_{out}^{2tables}$	38.16	-	28.95	27.63	25.00	22.37	14.47	23.68	10.53	11.84	10.53
IR_{out}^{+all}	43.42	39.47	-	18.42	30.26	32.89	14.47	26.32	9.21	10.53	10.53
PPL_{out}^{+all}	52.63	57.89	39.47	-	39.47	36.84	19.74	34.21	11.84	10.53	11.84
$moses_{in}$	57.89	55.26	50.00	38.16	-	36.84	26.32	30.26	10.53	9.21	9.21
$moses_{in}^{2tables}$	52.63	56.58	50.00	39.47	32.89	-	27.63	31.58	14.47	11.84	9.21
IR_{in}^{+all}	64.47	63.16	61.84	50.00	50.00	47.37	-	39.47	11.84	13.16	10.53
PPL_{in}^{+all}	57.89	61.84	52.63	44.74	43.42	42.11	21.05	-	10.53	11.84	10.53
$moses_{oracle}$	84.21	84.21	86.84	85.53	85.53	84.21	82.89	82.89	-	31.58	38.16
IR_{oracle}^{+all}	84.21	81.58	82.89	81.58	82.89	80.26	78.95	81.58	48.68	-	38.16
PPL_{oracle}^{+all}	81.58	85.53	82.89	80.26	80.26	81.58	80.26	81.58	48.68	39.47	-

Table 4: Document-level contrastive evaluation for French to English translation experiments. Numbers indicate the percentage of documents for which the system of the row outperformed the system of the column by more than the specified margin (BLEU > 0.5). Green background indicates that the system of the row outperformed the system of the column, while red indicates the opposite, and darker colors indicates larger differences.

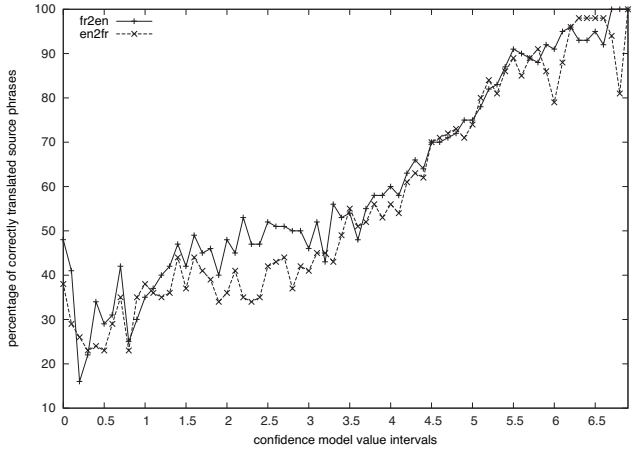

Figure 3: Percentage of correctly translated source phrases in the trace of the decoder for the `PPL+all` systems against score value intervals of the confidence model (`conf`).

> which correspond to the large majority of our training data, previous studies have shown that French as an original language is significantly more represented than English as an original language [22]. Experimenting with other corpora in which original language is known may help us to confirm this hypothesis.

Our adapted systems have recourse to sampling, and consequently do not use a reverse translation model [9], thus resulting in systems that may be built very efficiently, even for large data set conditions. Most previously published domain adaptation techniques cannot be applied directly to our studied scenario, as the availability of an in-domain training corpus is almost always assumed. Note that the `newsco` part of our training corpus was in fact "in-domain" w.r.t. our test documents. However, this corpus part was not identified as such, and our sampling strategies had no means to specifically access these data. The `2-tables` baseline system [1] is the only setting where we perform translation where subparts of the whole training data are known, identifying in particular an in-domain corpus: this situation obtained lower results than our systems under all conditions, indicating that the granularity of training corpus used was not appropriate and should be adapted.

Lastly, we assess whether our confidence model (section 4) is a good predictor of translation quality. Figure 3 plots the percentage of correctly translated source phrases in the trace of the decoder (counted as such when their target phrase matches the reference translation) against score intervals of the model. For our `PPL+all` systems, we observe a clear tendency to provide better translations for test phrases with higher confidence. This result clearly calls for a better handling of low-confidence phrases, e.g. by source-side paraphrasing [10].

7. Conclusion and future work

In this paper, we have studied a new scenario for Machine Translation that we called "any-text" translation, in which no in-domain training or development corpora can be identified in the general case. We have described an adaptation strategy that adapts translation models at the level of each input document by sampling and weighting training examples, and adds information about translation unit goodness and translation confidence. We found that our on-the-fly contextual adaptation significantly improves the results of French to English translation (up to 1.15 BLEU point improvement over `moses` and 1.02 BLEU over a corpus-level adaptation baseline (`2-tables`)). In comparison, results for the English to French pair do not reveal any clear gains. Some of our observations and hypotheses may pave the way to future experiments to determine under what conditions adaptation techniques can improve translation results. In particular, it turned out that our English documents were less similar to our training corpus than our French documents. The precise reasons for this situation should be investigated further.

We have introduced a document-level contrastive evaluation scheme (see Table 4), which offers a straightforward way to interpret and analyze the difference between any two

systems. Each reported value can be understood as the probability that one system would translate a document better (by some pre-defined margin using some evaluation metrics) than the other. The more input documents, the more accurate such probabilities will be. Those figures exhibit interesting conclusions: for instance, using a perfect tuning set at the document level allows to improve translation performance for more than 75% of documents for `moses` or our adapted systems over using a supposedly in-domain tuning set.

Given the large improvements obtained with the oracle tuning condition, we intend to study document-level adaptation schemes [18]. A better method of scoring the examples in the training corpus should be explored, for instance by taking more document context into account. More generally, we would like to recast the issue of instance weighting into one of determining the probability that a given training example is appropriate to translate a given test example in context: in this respect, textual similarity metrics such as `tf.idf` and perplexity values can only be used as features, in conjunction to other relevant features possibly indicating translation equivalence.

8. Acknowledgments

This work was partly funded by the European Union under the FP7 project META-NET (T4ME), Contract No. 249119.

9. References

[1] P. Koehn and J. Schroeder, "Experiments in Domain Adaptation for Statistical Machine Translation," in *WMT*, Prague, Czech Republic, 2007.

[2] G. Foster and R. Kuhn, "Mixture-model adaptation for SMT," in *WMT*, Prague, Czech Republic, 2007.

[3] G. Foster, C. Goutte, and R. Kuhn, "Discriminative Instance Weighting for Domain Adaptation in Statistical Machine Translation," in *EMNLP*, Cambridge, USA, 2010.

[4] R. Sennrich, "Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation," in *EACL*, Avignon, France, 2012.

[5] B. Haddow and P. Koehn, "Analysing the Effect of Out-of-Domain Data on SMT Systems," in *WMT*, Montréal, Canada, 2012.

[6] S. Matsoukas, A.-V. I. Rosti, and B. Zhang, "Discriminative Corpus Weight Estimation for Machine Translation," in *EMNLP*, Singapore, 2009.

[7] K. Duh, K. Sudoh, and H. Tsukada, "Analysis of Translation Model Adaptation in Statistical Machine Translation," in *IWLST*, Paris, France, 2010.

[8] C. Callison-burch, C. Bannard, and J. Schroeder, "Scaling Phrase-Based Statistical Machine Translation to Larger Corpora and Longer Phrases," in *ACL*, Ann Arbor, USA, 2005.

[9] A. Lopez, "Tera-Scale Translation Models via Pattern Matching," in *COLING*, Manchester, UK, 2008.

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

[11] A. Axelrod, X. He, and J. Gao, "Domain Adaptation via Pseudo In-Domain Data Selection," in *WMT*, Edinburgh, UK, 2011.

[12] G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta, "Does more data always yield better translations?" in *EACL*, Avignon, France, 2012.

[13] U. Manber and G. Myers, "Suffix arrays: A new method for on-line string searches," *SIAM Journal of Computing*, vol. 22, no. 5, pp. 935–948, 1993.

[14] P. Koehn, F. J. Och, and D. Marcu, "Statistical Phrase-Based Translation," in *NAACL*, Edmonton, Canada, 2003.

[15] N. Tomeh, M. Turchi, G. Wisniewski, A. Allauzen, and F. Yvon, "How Good Are Your Phrases? Assessing Phrase Quality with Single Class Classification," in *IWSLT*, San Francisco, USA, 2011.

[16] J. Wuebker, A. Mauser, and H. Ney, "Training Phrase Translation Models with Leaving-One-Out," in *ACL*, Upsala, Sweden, 2010.

[17] M. Carpuat and D. Wu, "Context-Dependent Phrasal Translation Lexicons for Statistical Machine Translation," in *MT Summit*, Copenhagen , Denmark, 2007.

[18] L. Liu, H. Cao, T. Watanabe, T. Zhao, M. Yu, and C. Zhu, "Locally Training the Log-Linear Model for SMT," in *EMNLP*, Jeju Island, Korea, 2012.

[19] A. S. Hildebrand, M. Eck, S. Vogel, and A. Waibel, "Adaptation of the Translation Model for Statistical Machine Translation based on Information Retrieval," in *EAMT*, Budapest, Hungary, 2005.

[20] K. Papineni, S. Roukos, T. Ward, and W.-j. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *ACL*, Philadelphia, USA, 2002.

[21] M. Snover, B. J. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," in *AMTA*, Cambridge, USA, 2006.

[22] B. Cartoni and T. Meyer, "Extracting directional and comparable corpora from a multilingual corpus for translation studies," in *LREC*, Istanbul, Turkey, 2012.