Incremental Adaptation Using Translation Information and Post-Editing Analysis

Frédéric Blain[*†], *Holger Schwenk*[*] *Jean Senellart*[†]

[*] LIUM
Université du Maine, Avenue Laennec
72085 Le Mans, France
lastname@lium.univ-lemans.fr

†Systran SA
5, rue Feydeau
75002 Paris, France
lastname@systran.fr

Abstract

It is well known that statistical machine translation systems perform best when they are adapted to the task. In this paper we propose new methods to quickly perform incremental adaptation without the need to obtain word-by-word alignments from GIZA or similar tools. The main idea is to use an automatic translation as pivot to infer alignments between the source sentence and the reference translation, or user correction. We compared our approach to the standard method to perform incremental re-training. We achieve similar results in the BLEU score using less computational resources. Fast retraining is particularly interesting when we want to almost instantly integrate user feed-back, for instance in a post-editing context or machine translation assisted CAT tool. We also explore several methods to combine the translation models.

1. Introduction

Due to multiplication of resources and the diversity of languages, Machine Translation (MT) systems are widely used as a precious help for human translators. Most of the systems used today are based on the statistical approach. Those systems extract all the knowledge from the provided data. Nevertheless, these systems have some limits: first, the specific resources available at *t* time could be less appropriate at *t+1*. Consequently, they need to be regularly re-trained in order to be updated, which is usually computationally demanding. The goal of incremental adaptation is then twofold: to adapt the system on the fly when new resources are available without re-training the entire system.

Post-Editing (PE) the output of SMT systems is widely used, amongst others, by professional translators of localization services which need for example to translate technical data in specific domains into several languages. However, the use of PE is restricted by some aspects that must be taken into consideration. As resumed by [1], the time spent by the post-editor is a commonly used measure of the PE effort, which should not to be, in case of poor translation quality, more important than translation from scratch. Even if this temporal aspect can be see as the most important, PE effort can be evaluated using automatic metrics based on the edit distance. These metrics commonly use the number of required edits of the MT system output to reach a reference translation. From then, the combination of PE and incremental adaptation can be seen as a way to reduce the task effort by allowing a MT system to gradually learn from its own errors. Especially considering the repetitive nature of the task highlighted by [2].

However, incremental adaptation is still a tricky task: how to adapt the system correctly? Adaptation should not degrade system performance and accuracy. Some approaches are possible and we will try to see the impact of several of them in the second part of this article.

First of all, we present a new experimental approach for incremental adaptation of a MT system using PE analysis. Starting from a generic baseline, we have gradually adapted our system by translating an in-domain corpora which was split beforehand. Each part of the corpora was translated using the translation model adapted at the previous step, *i.e.* updated with new extracted phrases. These phrases are the result of a word-to-word alignment combination we present afterward.

1.1. Similar work

The most similar approach in the literature is proposed in [3] who present an incremental re-training algorithm to simulate a post-editing situation. It is proposed to extract new phrases from *approximate alignments* which were obtained by a *modified* version of Giza-pp [4]. An initial alignment with one-to-one links between the same sentence positions is created and then iteratively updated as long as improvements are observed. In practice, a greedy search algorithm is used to find the locally optimal word alignment. All source positions carrying only one link are tried, and the single link change which produces the highest probability increase according to the Giza-pp model 4 is kept. The resulting alignment is improved with two simple post-processing steps. First, each unknown word in source side is aligned with the first non-aligned unknown word on the target side. Second, unaligned pairs of positions surrounded by corresponding alignments are automatically aligned.

In this paper, we present a very fast word-to-word align-

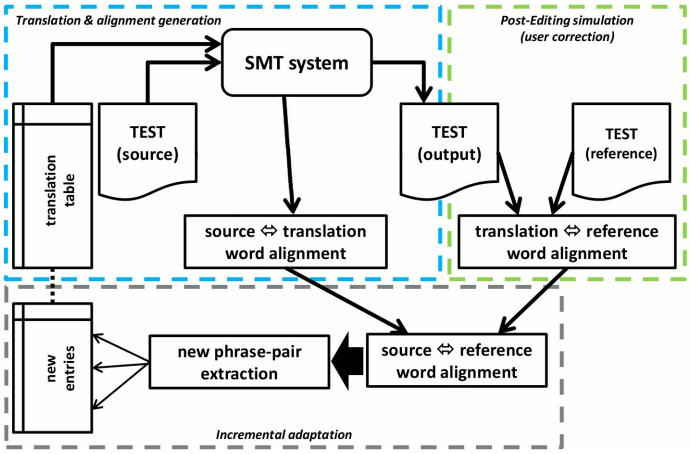

Figure 1: Incremental adaptation workflow in three steps protocol: 1. Translation and source-translation alignment: source sentences are translated using the SMT system Moses. Alignment links are generated during the translation step; 2. Edit distance on translation-reference: MT system output and its reference translation are aligned using edit distance algorithm of TER; 3. Source-reference alignment: the alignment links are deduced from combination of alignments of both step 1 and 2. Phrase pairs are then extracted, scored and added to translation model which is finally re-trained.

ment algorithm which is partially based on the edit-distance algorithm. As argued in [3], *"to be practical, incremental retraining must be performed in less than one second"*. For comparison, our entire alignment process takes few hundredths of second for 1500 sentences, in comparison to several seconds per sentences as reported in [3].

[5] present stream based incremental adaptation using an on-line version of the EM algorithm. This approach designed for large amounts of incoming data is not really adapted for the post-editing context. Like [3], we propose an incremental adaptation workflow that is more oriented to real time processing.

As part of our experiments, we have compared our approach with the use of the freely available tool named Inc-Giza-pp,[1] an incremental version of Giza-pp. It is precisely intended to inject new data into an SMT system without having to restart the entire word alignment procedure. To our knowledge, this is the standard method currently used in the field. In our experiments, we achieve similar results with respect to the BLEU score using less time.

The reminder of this paper is organized as follows. In the next section we first describe our incremental adaptation workflow and more particularly the word-to-word alignment methodology based on the edit distance. Section 3 is dedicated to the experimental protocols and compares the performance of our approach with the standard method using Inc-Giza-pp. The paper concludes with a discussion of perspectives of this work.

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

2. Incremental Adaptation Workflow

In this paper, we present a new methodology to perform incremental training and domain adaptation. Starting with a generic phrase-based MT baseline system (PBMT), we have sequentially translated the source side of an in-domain corpus. At each step, like [3], we have simulated a human post-editing the translations by using the corresponding reference translations of the data. At the sentence level, the source and its reference translation are aligned in order to subsequently retrieve the corresponding phrase pairs. The extracted phrase pairs are then scored and used to retrain (i.e. adapt) the translation model of our PBMT system.

We have developed an aligning protocol which operates in three steps, named *"translation"*, *"analysis"* and *"adaptation"*. These three steps are linked together by a word-to-word alignment algorithm which allows us to align a source and its reference translation and then, to extract new phrase pairs with which the MT system will be adapted. This algorithm is illustrated in Figure 1 and explained in details in the next section.

2.1. Word-to-word alignment combination

Our approach to align the source and its corresponding reference translation could be seen as a combination of the source to hypothesis word alignments and an analysis of the edit distance between the hypothesis and the reference. The central element of this approach is an automatic translation of the source sentence into the target language. The principle of this idea is illustrated in Figure 2.

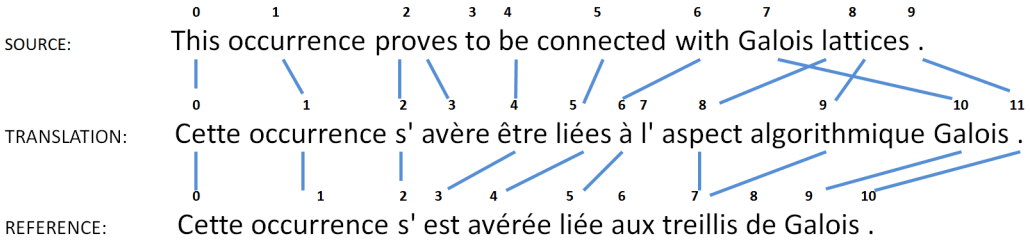

Figure 2: Example of a source-to-reference alignment using using the automatic translation as pivot. The alignment links between the source sentence and the translation are generated by the MT system. Those between the translation and its post-edited version (i.e. the reference) are calculated by TER. Finally, the source-to-reference alignment links are deduced by an alignment combination based on both alignment sets computed before.

2.1.1. Translation: source to translation alignment

The SMT system used to translate the source sentences is based on the Moses SMT toolkit [6]. Moses can provide the word-to-word alignments between the source sentence and the translation hypothesis. This aligning information represents the first part of our alignment combination. This automatic translation is "compared" with the reference translation using an edit distance algorithm.

2.1.2. Analysis: edit distance alignment

In this paper, we use the Translation Error Rate (TER) algorithm as proposed in [7]. TER is an extension of the Word Error Rate (WER) which is more suitable for machine translation since it can take into account word reorderings. TER uses the following edit types: *insertion, deletion, substitution* and *shift*.

The TER is computed between the output of our SMT system and the corresponding reference translation, and the word-to-word alignments are inferred. We only keep the aligned and substituted edit types in order to extract what we consider as the most interesting phrase pairs. Indeed, we argue that what is aligned correspond to what our system knows, while what is substituted correspond to what our system does not know.

Our approach can be extended to use TER-Plus [8], an extension of TER using paraphrases, stemming and synonyms in order to obtain better word-to-word alignments.

2.1.3. Adaptation: source to reference alignment

Considering the SMT translation hypothesis as a "pivot" for aligning both source and its reference sentence, we have designed the word-to-word alignment algorithm shown by Algorithm 1. It combines source-to-translation and translation-to-reference alignments, and then deduces the source-to-reference alignment path. From this path, the translation model is finally updated using the standard training phrase extraction and scoring script provided with Moses.

Data: src-to-tgt word alignments, tgt-to-ref edit-path
foreach *src-to-tgt word alignment* **do**
    alignment(src-word, tgt-word) = 1;
end
if *edit-path has shift* **then**
    foreach *shift* **do**
        updateWordPosition(tgt, shift);
    end
end
foreach *edit-type of edit-path* **do**
    if *edit-type is 'align' or 'substitution'* **then**
        alignment(tgt-word, ref-word) = 1;
    end
end
foreach *ref-word of ref* **do**
    foreach *tgt-word aligned to ref-word* **do**
        if *isAligned?(src-word, tgt-word)* **then**
            alignment(src-word, ref-word) = 1;
        end
    end
end

Algorithm 1: Source-to-reference alignment algorithm at word level. Using both source-to-translation alignments and translation-to-reference edit-path, the source-to-reference alignments path are build.

3. Experimental evaluation

The approach described in the previous section is compared to inc-Giza-pp which is considered as the state-of-the-art tool for incremental training. In our first experiments, each system uses a single translation model which is updated and entirely retrained after each iteration. For the results we present hereinafter, the system with inc-Giza-pp will be called "inc-Giza-pp" and the system with our approach will be called "noGizapp".

3.1. Training data

The experiments were performed on data which was made available by the French COSMAT project. The goal of this project is to provide task-specific automatic translations of scientific texts on the French HAL archive.[2] This archive contains a large amount of scientific publications and PhD Thesis. The MT system is closely integrated into the workflow of the HAL archive. In particular, the author has the possibility to correct the provided automatic translations. These translations will be then used to improve the system. In this paper, we consider the automatic translation from English into French.

Three corpora of parallel data are available to train the translation model: two generic corpora and an in-domain corpus for adaptation. The two first corpora are Europarl and News Commentary with 50 million and 3 million words, respectively. They were used to train our SMT baseline systems. The third corpus, named "absINFO", contains 500 thousand words randomly selected from abstracts of scientific papers in the domain of Computer Science. Information on the sub-domains is also available (networks, AI, data base, theoretical CS, ...), but was not used in this study. The corpus if freely available to support research in domain adaptation and was already used by the 2012 JHU summer workshop on this topic. A detailed description of this corpus can be found in [9].

This in-domain corpus was split into three sub-corpora:

- **absINFO.corr.train** is composed of 350k words and is used to simulate the user post-editing or corrective training.
- **absINFO.dev** is a set of 75k words and used for development.
- **absINFO.test** another set of 75k words used as a test corpus to monitor the performance of our adaptation workflow.

Moreover, in order to better simulate a sequential post-editing process, the absINFO.corr.train corpus was split into 10 sub-sets (about 1.5k sentences with 35k words each). This corresponds quite well to the update of an MT system after a post-correction of an entire document.

3.2. Baseline Training

The baseline SMT systems were constructed using the standard Moses pipeline and Giza-pp for word alignment. In order to later use Inc-Giza-pp, the incremental version of Giza-pp, we had to train a specific baseline system using the Hidden Markov Model (HMM) word alignment model option. However, to make a fair comparison of the two adaptation techniques, the baseline and following systems were trained on the same data and tuned with MERT [10] with the same initial parametrization. The inc-Giza-pp and noGizapp baseline SMT systems achieve a BLEU score of 35.27 and 35.32 BLEU points on the development corpus respectively, and 31.89 and 32.27 BLEU points on the test corpus.

3.3. Analysis of processing time and alignment quality

The two incremental training approaches are compared with respect to the BLEU score obtained by adding the additional aligned data. We also report the time needed to perform the word alignments. For inc-Giza-pp, the alignment protocol is composed of several steps (for more details, see "Incremental Training" of the "Advanced Features" section in Moses user documentation.[3]) First, one has to preprocess the data for use by Giza-pp. This involves updating the vocab files, converting the sentences into the *snt* format of Giza-pp, and then, updating the co-occurrence file. Then, Giza-pp is executed to update and compute the alignments for the new data. This is performed in both directions, source-to-translation and translation-to-source. For each iteration of our experiment, this process takes about 14 minutes.

For the noGizapp system, the required time to perform the source-to-translation alignment can be considered as null because it is implicitly achieved during the translation. The TER between the SMT translation and the reference translation is computed using a fast and freely available C++ implementation.[4] This tool can align about 35k words in about three seconds (corresponding to 1.5k sentences in the 10% subset of the absINFO.corr.train corpus). The alignment combination of the source and reference translation, described in algorithm 1, takes less than a second. Overall, we can obtain the source-to-reference alignments of 35k words in a few seconds only.

The BLEU scores on the development (left part) and test data (right part) are compared in Figure 3. The following systems were built:

Gizapp for each subcorpus of the absINFO.corr.train training data (10%, 20%, 30%...100%), all the available training data is concatenated and the full training pipeline is performed, including a new word alignment which considers **all** the training data. We consider this as the upper limit of the performance we could achieve by incremental training. This procedure is very time consuming.

inc-Giza-pp the subcorpora of of the absINFO.corr.train training data are added using the incremental version of Giza. This resulted in a slight decrease of the BLEU score on the development data and a quite unstable performance on the Test data.

noGizapp incremental training using the new approach described in this paper. We always used the same base-

[2] http://hal.archives-ouvertes.fr/

[3] Available online: http://www.statmt.org

[4] http://sourceforge.net/projects/tercpp/

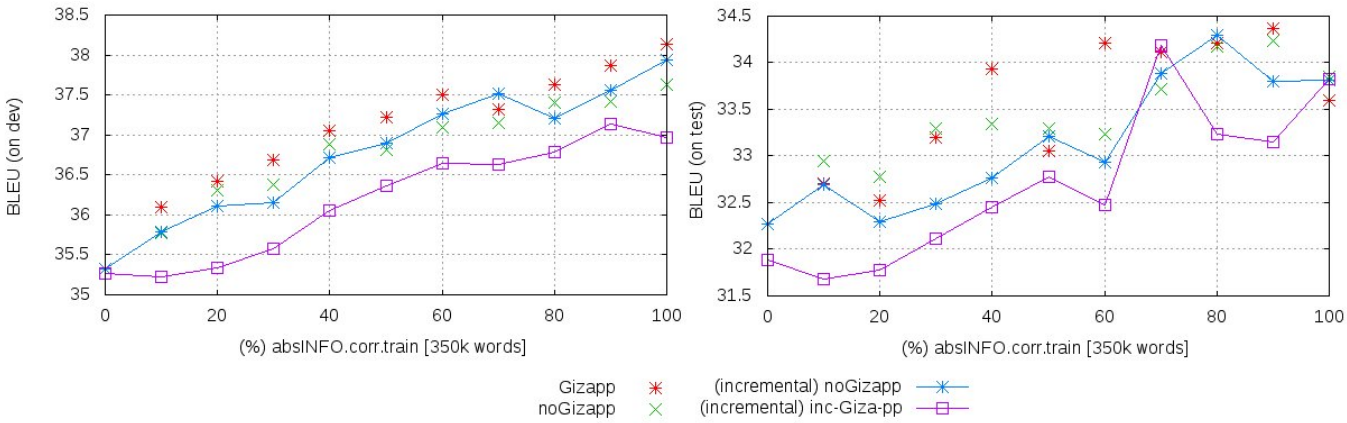

Figure 3: Incremental adaptation in BLEU score for our two PBMT systems on both development and test corpora. The Inc-Giza-pp system uses incremental version of Giza-pp for aligning sentence pair, while noGizapp system uses the approach we present in this paper, which is based on translation information and edit distance combination. The 'Gizapp' and 'noGizapp' curves represent the BLEU score obtained with a in-domain adaptation of our baseline systems, without incremental approach. While the curves 'Inc-Giza-pp' and '(incremental) noGizapp' represent the in-domain adaptation scores over an incremental process.

line SMT system to translate the additional adaptation data.

inc-noGizapp like *noGizapp*, but using the system adapted in the previous step to translate the additional adaptation data.

The proposed approach to obtain incremental word alignments achieves slightly better BLEU scores on both the development and the test corpus, but performs much faster.

The large variations on the test corpus could be explained by two potential reasons. The first one could be the characteristics of the absINFO.corr.train corpus. It was created from abstracts of (Computer Science) sub-domains which were randomly selected. Consequently, a sub-corpus predominantly represented in a sub-corpus of absINFO corpus could be not represented in the test corpus. The second reason could be the use of only one translation model. As explained above, this translation model is updated with new phrase pairs extracted from each iteration. Because we are only interested by edit types corresponding to 'align' and 'substitution' edit type during the edit distance analysis (see Section 2.1.2), the extracted phrase pairs could be generic or in-domain. Added to all entries already in the translation model, these new phrases disturb the probability distribution. This could also explain why our incremental systems are performing worse than the non incremental systems (what we have called "oracle systems") for which, the probability distribution is tuned in better way.

Another possibility could be to use two translation models like [3]. In this way, we can quickly create a phrase-table from the word alignments of the additional data.

3.4. Combination of translation models

In this section, we present results achieved by combining several translation models. The techniques described in the previous sections can significantly speed-up the word-alignment process, in comparison to running incremental Giza-pp, but we still need to create a new phrase table on all the data. Therefore, we propose to create a new phrase table on the newly added data only and to combine it with the original unadapted phrase table.

3.4.1. Back-off Models

Moses support several modes to use multiple phrase tables. We first explored the back-off mode which favors the principal phrase table: the second phrase table is only considered if the word or phrase is not found in the first one. Figure 4. The dotted curve represents the use of the incrementally trained in-domain translation model with the generic one as back-off. The crossed curve represents the use of these same models but in reverse order.

As we can see, we got very different results depending on which translation model is used first, but this can be easily explained by the nature of the back-off models. Our in-domain translation model is built with the incrementally added data only, i.e. very small amounts of data, in particular during the first iterations.

Figure 5 presents when jointly using both translation models. In this configuration, separate translation options are created for each occurrence, the score being combined if the same translation option is found in both translation models. Compared to the use of only one translation model, we

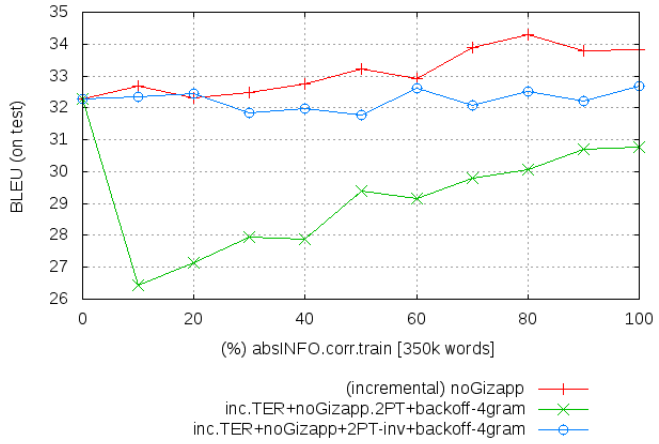

Figure 4: Results for use of "back-off" models. The crossed curve represents our PBMT system using only one translation model while the dotted and third curves represent respectively the impact of use two back-off models but in different order.

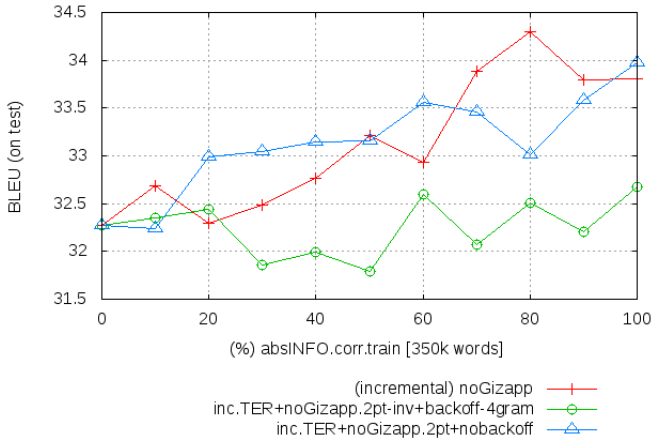

Figure 5: Comparison between use of back-off (dotted curve) and non back-off models. The crossed curve represents our PBMT system using only one translation model. The third curve represents a PBMT system using its both translation models for the decoding path while the dotted curve shows our results for using our translation models in back-off mode.

can observe a significant degradation near 80% of adaptation data before finally achieving a similar final BLEU score (up to +0.2 points) compared to inc-Giza-pp and noGizapp.

Once again, we believe that the nature of our absINFO corpus may explain the evolution of our score. When our SMT systems has to translate more generic sentences, it is likely that the translation options were provided by our generic translation rather than our in-domain model.

Based on this observation, we tried to limit edit distance analysis to substitutions only.

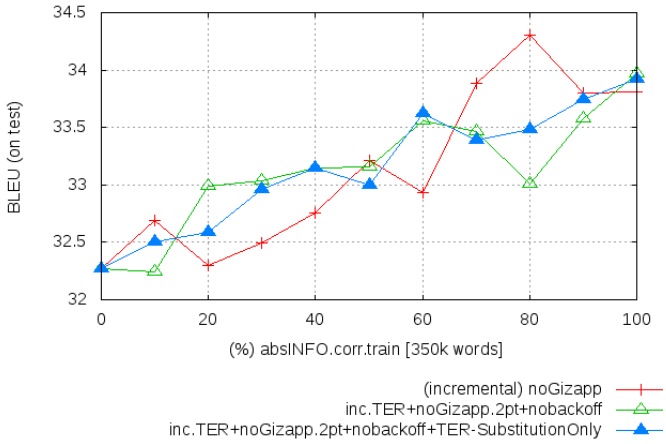

Figure 6: Use of 2 translations models with noback-off and only substitution were kept, or not.

3.4.2. Filtering by edit-distance type

The Figure 6 shows the results obtained with an in-domain translation model only trained from substitutions which were detected during the edit distance analysis. As we argued in section 2.1.2, we consider that the "substitution" edit type corresponds to what the MT system does not know since it was necessary to fix its output.

As we can see, the previous degradation is less important.Overall, the evolution of the BLEU score is smoother than for the other approaches tested so far. By keeping the phrase pairs corresponding to substitutions only (in the edit-path), we have also limited the contextual phrases in our in-domain translation model. It should also take into account the alignment errors that would have a more important impact in this configuration on the quality of the translation model.

3.4.3. N-best alignment generation

One of the key points presented in this paper is the use of the translations to generate the alignment links between a source sentence and its translation generated by the system. By default, our MT system returns the best translation candidate after decoding. This means that this translation has obtained the highest decoding score, but that does not necessarily mean that the alignment associated with it is the best one.

Based on this observation, we tried to explore the n most likely translations hypothesis (n-best list). Indeed, a source sentence could be translated into the same translation using different segmentations into phrase-pairs. With our approach, for the same sentence-translation pair, if we have multiple alignment candidates, we can generate more source-to-reference alignments and then, potentially reinforce our in-domain translation model. Using only the two best non distinct translation candidates, we obtained the results shown

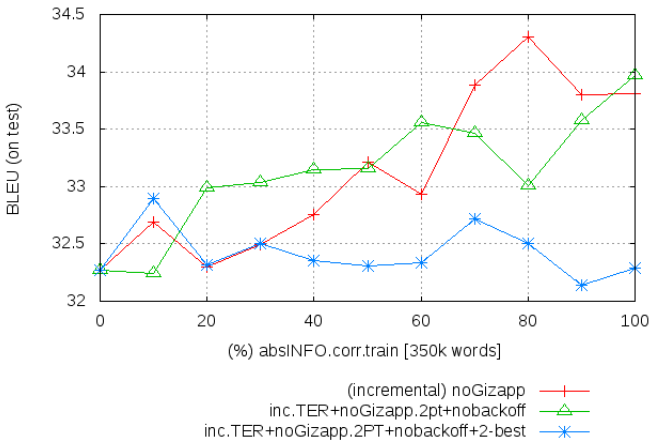

Figure 7: Use of n-best translation candidate to reinforce alignment possibilities and then, extend our phrase-pair generation. The starred curve presents our PBMT system for which we used the two first translation candidates in order to extract phrase pairs, while the second curve represents the same system but only the 1-best translation candidate is used.

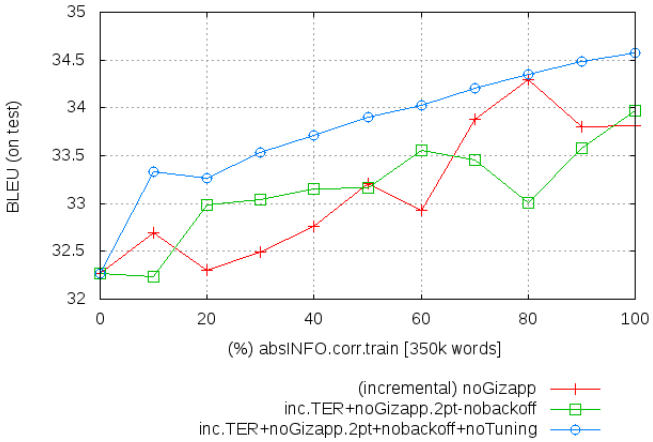

Figure 8: Results for incremental adaptation with no tuning step. The squared curve represents a PBMT system with normal tuning process achieved at each adaptation iteration, while the dotted curve represents the same system for which the tuning weights obtained at 10% level remain stable throughout the entire adaptation.

in Figure 7. Unfortunately, the results are worse than expected. In future work, we will investigate other options to use the information in the n-best lists.

3.4.4. No tuning step

In the final part of the paper, results from an incremental adaptation of a PBMT system without tuning step are presented. This procedure is very time-efficient and stable since we do not apply tuning at every adaptation step. We argue that we do not need to re-tune our models since adaptation only adds small amounts of information. Tuning is only applied at the creation of the model, and the resulting parameters are maintained during the adaptation process. The results of this procedure are shown in Figure 8.

First, we can observe a clear difference between the squared and the dotted curves for the 10% adaptation level, even though they result from the same approach. This is due to the baseline that we applied: By default, our PBMT system is a translation model using only one phrase table. We need to tune however on a "new baseline system" using two phrase tables (the one at the 10% level), for which the tuning weights obtained remain stable throughout adaptation.

Second, the resulting curve is rather smooth, indicating the instability of the tuning process.
To sum up, by applying our incremental adaptation, we obtain a clear improvement in BLEU scores (+0.5 points), however without the need to retune at every adaptation. Tuning can be performed in larger time intervals, for example - in an industrial post-editing context - every night or as soon as processing resources become available.

4. Conclusion and Future Work

In this paper, we have presented a new word-to-word alignment methodology for incremental adaptation using a phrase-based MT system. This method uses the information generated during the translation step and then relies on an analysis of a (simulated) post-editing step to infer a source-to-reference alignment at the word level.

Compared to incremental Giza, the standard method currently used in the field, the first part of our experiments show that our approach allows us to obtain similar performance in the BLEU score at an significantly improved speed. Incremental Giza needs several minutes to align two corpora of about 35k words while the approach proposed in this paper runs in some seconds. Our approach could be therefore integrated into an interface dedicated to post-editing which would exploit user feedback in real time.

The second part of this article was dedicated to experiments on translation model combination. These experiments show that we can get better results by jointly using two translation models instead of only one. The results of these experiments suggest some directions for future research. For example, the use of the TER algorithm for analyzing the post-editing result could be reinforced by the notion of "Post Edit Actions" introduced by [2], in order to better identify errors of the SMT system.

5. Acknowledgment

This research was partially financed by the DGA and the ANRT under CIFRE-Defense 7/2009, the french ANR project COSMAT under ANR-09-CORD-004, and the European Commission under the project MATECAT, ICT-

2011.4.2 – 287688.

6. References

[1] M. Koponen, "Comparing human perceptions of post-editing effort with post-editing operations," *Proceedings of the Seventh Workshop on Statistical Machine Translation*, p. 181–190, June 2012.

[2] F. Blain, J. Senellart, H. Schwenk, M. Plitt, and J. Roturier, "Qualitative analysis of post-editing for high quality machine translation," in *Machine Translation Summit XIII*, A.-P. A. for Machine Translation (AAMT), Ed., Xiamen (China), 19-23 sept. 2011.

[3] D. Hardt and J. Elming, *Incremental Re-training for Post-editing SMT.*, 2010.

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

[5] A. Levenberg, C. Callison-Burch, and M. Osborne, "Stream-based translation models for statistical machine translation," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 394–402.

[6] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Annual meeting-association for computational linguistics*, vol. 45, no. 2, 2007, p. 2.

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

[8] M. Snover, N. Madnani, B. Dorr, and R. Schwartz, "Fluency, adequacy, or hter? exploring different human judgments with a tunable mt metric," in *Proceedings of the Fourth Workshop on Statistical Machine Translation*, vol. 30. Association for Computational Linguistics, 2009, pp. 259–268.

[9] L. Patrik, H. Schwenk, and F. Blain, "Automatic translation of scientific documents in the hal archive," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012, pp. p.3933–3936.

[10] F. Och, "Minimum error rate training in statistical machine translation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 160–167.