D4.3: Final Report on Adaptive Translation Models

Daniel Ortiz-Martínez, Germán Sanchis-Trilles, Jesús González-Rubio, Francisco Casacuberta

Distribution: Public

CASMACAT
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D4.3

Project funded by the European Community
under the Seventh Framework Programme for
Research and Technological Development.

Project ref no.	ICT-287576
Project acronym	CASMACAT
Project full title	Cognitive Analysis and Statistical Methods for Advanced Computer Aided Translation
Instrument	STREP
Thematic Priority	ICT-2011.4.2 Language Technologies
Start date / duration	01 November 2011 / 36 Months

Distribution	Public
Contractual date of delivery	April 30, 2013
Actual date of delivery	April 30, 2013
Date of last update	April 30, 2013
Deliverable number	D4.3
Deliverable title	Final Report on Adaptive Translation Models
Type	Report
Status & version	Final
Number of pages	15
Contributing WP(s)	WP4
WP / Task responsible	UPVLC
Other contributors	CS
Internal reviewer	Philipp Koehn
Author(s)	Daniel Ortiz-Martínez, Germán Sanchis-Trilles, Jesús González-Rubio, Francisco Casacuberta
EC project officer	Aleksandra Wesolowska
Keywords	Adaptation, Active Learning, On-line Learning, Interactive Machine Translation

The partners in CASMACAT are:

University of Edinburgh (UEDIN)
Copenhagen Business School (CBS)
Universitat Politècnica de València (UPVLC)
Celer Soluciones (CS)

For copies of reports, updates on project activities and other CASMACAT related information, contact:

The CASMACAT Project Co-ordinator
Philipp Koehn, University of Edinburgh
10 Crichton Street, Edinburgh, EH8 9AB, United Kingdom
pkoehn@inf.ed.ac.uk
Phone +44 (131) 650-8287 - Fax +44 (131) 650-6626

Copies of reports and other material can also be accessed via the project's homepage: http://www.casmacat.eu/

Executive Summary

This document contains information about the progress made on WP4 for the duration of the CASMACAT project. The main objective of this WP was to deal with adaptation in interactive translation prediction (ITP). Domain adaptation fits naturally in ITP scenarios, since the user validates the translation of each source sentence after a series of interactions with the system. As a result, ITP inherently generates new training pairs that can be used to feed the statistical models involved in the translation process. Work within WP4 has been focused on the study of different techniques for domain and user adaptation for ITP. The specific objectives were: (1) To develop algorithms for an efficient model adaptation; (2) To develop techniques for efficient sentence selection for adaptation; (3) To explore approaches for an effective adaptation to new domains and users.

Work in this work-package has been split into three different tasks:

- **Task 4.1: On-line Learning for Interactive Translation Prediction** (month 1-24)

  The aim of the work performed in this task is to develop, implement, and integrate into the CASMACAT workbench self-learning capabilities, which enable the underlying machine translation system to learn dynamically from the corrections provided by the user. In this direction, two different lines were explored: first, the online learning of the features of the log-linear models (i.e., phrase table scores and language model probabilities), implemented by means of an online version of the EM algorithm; and second, online learning of the log-linear weights that adjust the discriminative power of the previously mentioned features, implemented by means of a novel strategy based on the concept of Ridge regression. In this direction, good progress was made, and the CASMACAT system now features online learning capabilities which benefit from the research performed in this task.

- **Task 4.2: Active Learning in Interactive Translation Prediction** (month 13-24)

  This task aimed to develop novel strategies for selecting which sentences, that have been already been revised or that are yet to be revised, are best suitable for being integrated into the ITP system, i.e., which are the sentences from which the ITP system would benefit the most. In this direction, experimental results reported according to standard SMT evaluation metrics reveal that the techniques implemented are successful at achieving a good compromise between user effort and final translation quality achieved.

- **Task 4.3: Domain and User Adaptation** (month 13-30)

  This task was devoted to the development of new techniques for enabling the system to adapt to new domains and/or users. This was mainly tackled by means of Bayesian adaptation, sentence selection strategies, and language model interpolation. Even though this task is already finished according to the initial planning, work is still ongoing with the purpose of establishing a close collaboration with the MATECAT FP7 project, in which there has also been a very big effort regarding adaptation strategies.

As it is shown above, Task 4.3 is the only active task during this reporting period. In spite of this, in this deliverable we have also included summaries of already completed work for the other two tasks.

Contents

1 On-line learning for Interactive Translation Prediction

In this section we will briefly summarise the effort made in Task 4.1 on online learning for interactive translation prediction (ITP) during the project. Task 4.1 was scheduled for months 1 to 24, and hence it is not active for this reporting period.

During the first year of the project, we extended previous work on online learning for statistical machine translation (SMT) presented at an international conference [28]. The proposal is based on maintaining a set of incrementally updateable sufficient statistics for the feature functions that compose a state-of-the-art log-linear model for SMT. One of the key points of our proposal is the use of the incremental EM algorithm to estimate HMM-based word alignments, which are required to generate the word alignment matrices from which phrase pairs are extracted. We provided a mathematical derivation of the generative models involved in the translation process as well as a detailed explanation of the update rules that are used to extend such models. In addition to this, we also investigated techniques to adjust the values of the weights of the log-linear combination, proposing a total of four different update rules. The specific details are given in two different research papers [22, 9].

To demonstrate the feasibility of online learning to learn the parameters associated to the feature functions of the log-linear model, lab experiments were carried out using two translation tasks that have been previously used to report ITP results, namely, the Xerox and the EU tasks. In all cases, the proposed ITP system was able to incrementally update its statistical parameters from scratch of from previously estimated models in real time.

Additionally, we also tested the above mentioned update rules for the log-linear weights using the well-known Europarl and News-Commentary corpora. The majority of the experiments were carried out in an SMT scenario where the translations are generated in a fully-automatic way. In this case, mixed results were obtained, with two of the proposed update rules providing significant positive results. By contrast, preliminary experiments with ITP were not positive.

During the second year of the project, work on the online estimation of the feature function parameters was focused on clarifying its experimental properties. As it was explained above, previous experimentation was focused on demonstrating the feasibility of the application of online learning to the SMT framework. However, some crucial aspects of online learning were not studied, such as its performance with respect to a conventional batch learning algorithm or the impact of update frequency in the system performance. The new experiments allowed us to clarify such aspects. Specifically, experiments carried out using the Europarl corpus showed that online learning is able to achieve comparable or even slightly better results than those obtained by means of batch learning. Additionally, the presented results also showed the strong impact of update frequency in the performance of the ITP system. Moreover, according to the results, the best system should be capable of updating its models on a per sentence basis, highlighting the great potential of online learning to reduce the user effort in translation tasks.

Experiments reported during the first year of the project were executed on corpora of a small or medium size, mainly due to the computational requirements of processing larger corpora. A significant part of the work carried out within Task 4.1 during the second year of CASMACAT was focused on creating the software tools that are required to process large corpora as well as the integration of such tools in the CASMACAT workbench.

Work concerning the online learning of the log-linear weights was also extended during this period. As it was mentioned above, initial ITP experiments did not produce the same positive results obtained in a conventional SMT setup. To deal with this problem, the previously proposed algorithms for online learning of the weights were redefined, and some encouraging results were observed by obtaining weight samples from a Gaussian distribution and hence yielding different wordgraphs [9, 5]. Work in this direction led to two different M.Sc. theses [21, 4].

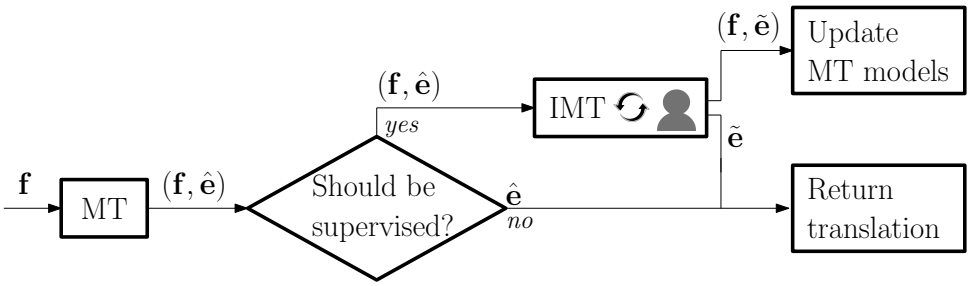

Figure 1: Dataflow of the proposed AL scenario for ITP.

Ongoing work that is being carried out on online learning has been focused on incorporating the related software in a new public release of the Thot toolkit [27] for SMT. The details are more carefully explained in Deliverable 5.4. In addition to this, an article on online learning for SMT has also been submitted to the Computational Linguistics journal.

For the future we plan to study ways to give more preference to newly acquired data with respect to previously acquired knowledge (domain adaptation) and to incorporate bounds to the data structures used to store the incremental models.

2 Active learning in Interactive Translation Prediction

In this section, we summarise the work carried out in Task 4.2 on active learning for interactive translation prediction during the project. Task 4.2 was scheduled for months 13 and 24, thus it is not active for this reporting period.

In the conventional ITP scenario, a human translator and an MT system collaborate in order to obtain the translation the user has in mind [8, 18, 7, 19, 3]. We identify two potential practical limitations for the deployment of ITP systems:

1. The human expert is supposed to supervise all the translations suggested by the MT system. This may be a limitation in real world-scenarios where translation agencies usually have to deal with tight restrictions in terms of manpower, money or time.

2. The MT system is able to respond to user feedback improving its predictions, but it is unable to "learn" from such information. This is an important limitation in practice because, after being corrected, the MT system may repeat its errors and the user will be justifiably disappointed.

Here, we propose an *active learning* (AL) [1, 2, 6, 20] scenario for ITP specifically developed to address these limitations. In our approach, the MT system not only responds to user interactions but it becomes an active entity in the interaction process. Our ultimate goal is to make the most effective use of the expensive human effort. That is, we want to maximise the translation quality obtained per unit of user supervision effort.

The proposed AL scenario boosts the productivity of ITP technology by addressing its two potential limitations. On the one hand, we do not require the human expert to exhaustively supervise all translations. Instead, we propose a selective interaction protocol where the user only supervises a subset of *informative* translations. This selective supervision is similar to the active interaction protocol studied in Task 2.3 in WP2 but applied at the sentence-level. On the other hand, we replace the batch MT model typically used by ITP systems by an incremental SMT model [28].

Figure 1 displays a diagram of the proposed AL scenario. Given a source sentence $\mathbf{f}$, the system first obtains its automatic translation $\hat{\mathbf{e}}$ and decides if it should be supervised by a human

expert or not. If the automatic translation requires supervision, the user is asked to perform a conventional ITP translation session to obtain the corresponding user-supervised translation $\tilde{\mathbf{e}}$[1]. The new translation pair $(\mathbf{f}, \tilde{\mathbf{e}})$ is then used to update the MT models, and the user-supervised translation $\tilde{\mathbf{e}}$ is returned. If the translation does not require user supervision then it automatic translation $\hat{\mathbf{e}}$ is directly returned. Note that the output of our scenario is thus a mixture of user-supervised and automatic translations.

The potential productivity improvements of our proposal are twofold. On the one hand, user effort is focused on those translations whose supervision is considered most "informative". Thus, we maximise the utility of each user interaction. Additionally, we also define an implicit upper bound to the amount of human supervision effort which is a very useful feature in practice. On the other hand, MT models are continually updated with user feedback. Thus, they are able to learn new translations and to adapt their output to match the users preferences; which prevents the user from making repeatedly the same corrections. The specific implementation details of the proposed AL scenario have been published in the following articles [13, 12].

Results of our AL scenario in comparison to a similar scenario with no MT model updating showed a large leap in productivity. In fact, we were able to obtain twice the translation quality for the same human effort. Moreover, the proposed AL scenario allows us to adjust the behaviour of the system between a conventional *automatic translation* scenario where the human expert supervise zero translations and an *online learning* scenario where all sentences are supervised by the user. Thus, we can adapt our system according to the requirements of the task or the amount of available resources so to reach an adequate trade-off between human effort and expected quality of the generated translations.

The work in this task has shown that the AL scenario proposed for ITP has the potential to largely improve the quality of the generated translations, or similarly to largely reduce the human effort required to generate translations of a certain quality. Among the tested setups, the one using coverage augmentation raking provided the best results improving productivity in comparison to any other setup.

This AL scenario has been incorporated into the CASMACAT taking into account the strict response-time constraints due to interactivity. This scenario is therefore fully-functional but it was decided not to test this feature in the final field-test given the small size of the test corpora to be used.

Even though Task 4.2 is officially finished according to the DOW, we consider that there are many different research directions that can be further explored. Probably, the main research direction involves the development, and efficient implementation, of a more formal AL approach as depicted in [11].

3 Domain and User Adaptation

In the first year of this task, extensive work was conducted in terms of developing both the theoretical framework and the practical implementation of Bayesian predictive adaptation. Under this framework, model parameters are viewed as random variables having some kind of a priori distribution. Observing these random variables leads to a posterior density, which typically peaks at the optimal values of these parameters. The benefits of this approach are that the parameters are biased towards the optimal values according to the adaptation set, while avoiding over-training towards such set by not forgetting the generality provided by the training set. Furthermore, re-estimating the system's parameters from scratch may imply a computational overhead which may not be acceptable in certain environments, such as SMT systems configured

[1]Other CAT approaches can be used. For example, to use post-edition instead of ITP, we only have to modify the corresponding module in the diagram.

for post-editing or ITP, in which the final human user is waiting for the translations to be produced.

In the bulk of the work conducted, the main focus was to adapt the log-linear weights present in every state-of-the-art SMT system. In this direction, experimental results analysing the effectiveness of such adaptation procedures were reported. Such results showed that Bayesian predictive adaptation is able to provide consistent improvements in translation quality over the baseline systems, as measured by TER, with as few as 10 adaptation samples, and up to an amount of adaptation data that allows a complete re-estimation of the model parameters. In addition, BPA proved to be more stable than most re-estimation strategies, which rely heavily on the amount of adaptation data. From a computational point of view, the Bayesian adaptation technique presented does not imply a significant computational overhead, and most terms can be precomputed using a heuristic sampling strategy. Work in this direction led to part of a Ph.D. thesis [29].

Despite the positive results obtained, implementing Bayesian predictive adaptation within the CASMACAT workbench was found to be more complex than expected: the current theoretical formulation requires a given n-best list to be readily available for the development set. Even though there are possible ways to bypass this, the improvements in terms of translation quality achieved were quite comparable to the ones obtained by DRR (see Task 4.1). Since DRR has a much more straight-forward implementation within a regular phrase-based decoder, it was decided to integrate DRR into CASMACAT instead of Bayesian predictive adaptation.

In the remaining 6 months of activity of this task, we explored different adaptation strategies which have been reported as successful in the literature, such as language model interpolation and bilingual sentence selection. The progress on these two lines is detailed in the following two subsections.

3.1 Language model interpolation

One possible way to assign more importance to in-domain data is to train separate models for out-of-domain and in-domain texts and linearly interpolate the resulting models. More specifically, the linear interpolation is performed at word level:

$$p(\mathbf{e}) = \prod_{i=1}^{|\mathbf{e}|} \left(\sum_j \lambda_j p_j(e_i|h_i) \right) \tag{1}$$

where $\mathbf{e} \equiv e_1 ... e_{|\mathbf{e}|}$, is a sentence in the target language composed of $|\mathbf{e}|$ words, $p_j(e_i|h_i)$ represents the probability given by the j'th language model in the log-linear combination for word e_i given history h_i, and λ_j is the weight assigned for the previous probability.

The weights of the linear combination can be trained by minimising the perplexity of the interpolated model for a development set [17]. The perplexity minimisation process can be implemented by means of the downhill simplex algorithm [24].

During the last months, we have started to include the above described language model interpolation functionality in the Thot toolkit [27]. Within the ongoing implementation process, a set of static language models can be linearly combined with an in-domain specific language model that can be incrementally updated by means of online learning techniques [28].

3.2 Sentence selection

3.2.1 Strategies studied

In the work conducted during the last semester of task 4.3, we analysed and compared two state-of-the-art sentence selection strategies. These two strategies, which have shown in the literature to be the most promising ones, are infrequent n-gram recovery [10], and cross-entropy selection [23]. The main framework underlying both is one in which there is a small in-domain corpus, which is sufficient to train a baseline SMT system. In addition, there is also a larger out-of-domain corpus, whose size is around one or two orders of magnitude larger than the in-domain corpus. In this scenario, we need to make the best possible use of such large amount of data. Hence, the problem is which portion of the out-of-domain data to incorporate into the system so as to improve the final translation quality, while avoiding to overwhelm the in-domain data with too much out-of-domain data.

The rest of this section provides a short review of both techniques.

Infrequent n-gram recovery. Proposed by [10], the main intuition underlying this sentence selection strategy is to choose sentences that provide information not seen in the in-domain training data. Such selection will be done according to the source side of the data to be translated.

The performance of phrase-based machine translation systems strongly relies in the quality of the phrases extracted from the training samples. In most of the cases, the inference of such phrases or rules is based on word alignments, which cannot be computed accurately when appearing rarely in the training corpus. The extreme case are the out-of-vocabulary words: words that do not appear in the training set, cannot be translated. Moreover, this problem can be extended to sequences of words (n-grams). Consider a 2-gram $f_i f_j$ appearing few or no times in the training set. Although f_i and f_j may appear separately in the training set, the system might not be able to infer the translation of the 2-gram $f_i f_j$, which may be different from the concatenation of the translations of both words separately.

When selecting sentences from the out-of-domain corpus it is important to choose sentences that contain source n-grams that have never been seen (or have been seen just a few times) in the in-domain set. Such source n-grams will be henceforth referred to as *infrequent n-grams*. An n-gram is considered infrequent when it appears less times than an infrequent threshold t. If the source language sentences to be translated are known beforehand, the set of infrequent n-grams can be reduced to those present in such sentences. Then, the technique consists in selecting from the out-of-domain data those sentences which contain infrequent n-grams present in the source sentences to be translated.

Out-of-domain sentences are sorted by their infrequency score in order to select first the most informative. Let $\mathcal{X}$ the set of n-grams that appear in the sentences to be translated and $\mathbf{w}$ one of them; $C(\mathbf{w})$ the counts of $\mathbf{w}$ in the source language in-domain set; and $N(\mathbf{w})$ the counts of $\mathbf{w}$ in the source sentence $\boldsymbol{x}$ to be scored. The infrequency score of $\boldsymbol{x}$ is:

$$i(\boldsymbol{x}) = \sum_{\mathbf{w} \in \mathcal{X}} \min(1, N(\mathbf{w})) \max(0, t - C(\mathbf{w})) \tag{2}$$

In order to avoid giving a high score to noisy sentences with a lot of occurrences of the same infrequent n-gram, only one occurrence of each n-gram is taken into account to compute the score. In addition, the score gives more importance to the n-grams with lowest counts in the training set. Although it could be possible to select the highest scored sentences, we updated the scores each time a sentence is selected. This decision was taken to avoid the selection

		Spanish	English
Europarl training	Nr. sentences	1.9M	
	Run. words	51.5M	49.1M
	Voc. size	422k	308k

Table 1: Statistics of the Europarl data.

of too many sentences with the same infrequent n-gram. First, out-of-domain sentences are scored using Equation (2). Then, in each iteration, the sentence $\boldsymbol{x}^*$ with the highest score is selected, added to the in-domain set and removed from the set of candidate sentences. In addition, the counts of the n-grams present in $\boldsymbol{x}^*$ are updated and, hence, the scores of the rest of out-of-domain sentences. Since re-scoring the whole out-of-domain data would incur in a very high computational cost, a suboptimal search strategy was followed, in which the search was constrained to a given set of highest scoring sentences. Here it was set to one million.

Cross-entropy selection. Proposed by [23], the main intuition behind this technique is to select new training data according to the perplexity score assigned by a language model trained on the in-domain data. Since perplexity and cross-entropy are monotonically related, selecting according to perplexity is equivalent to selecting according to cross-entropy. However, instead of selecting out-of-domain sentences according to cross-entropy directly, the difference in entropy when computing it with respect to the in-domain data or the out-of-domain data is considered. Let be $H_I(\boldsymbol{x})$ the cross-entropy of $\boldsymbol{x}$ according to the in-domain data I, and $H_G(\boldsymbol{x})$ the cross-entropy of $\boldsymbol{x}$ according to the out-of-domain data. Then, sentences with a score $H_I(\boldsymbol{x}) - H_G(\boldsymbol{x})$ will be selected for inclusion into $\boldsymbol{x}$ if the score exceeds a certain threshold T. Alternatively, and since $H_I(\boldsymbol{x}) - H_G(\boldsymbol{x})$ establishes an ordering of the out-of-domain data, we could select only the desired portion of the out-of-domain data, i.e. rather than fixing a threshold value, fixing a number of desired sentences to be included. A formal proof concerning why $H_I(\boldsymbol{x}) - H_G(\boldsymbol{x})$ is a theoretically sound score to use is explained in [23], but is left out here for simplicity reasons and because it does not provide further practical information.

3.2.2 Experiments

In this section we detail the experiments conducted with the purpose of establishing which one of both techniques works best in the CASMACAT framework. Even though the results presented here are conducted following a typical SMT framework, preliminary experimentation seems to show that the improvements obtained in an SMT framework will carry on nicely to an ITP setup. However, further experimentation is required before being certain about this point.

The experiments were performed using two different corpora: the Europarl corpus [16] and the News Commentary corpus[2]. The Europarl corpus is built from the proceedings of the European Parliament, whereas the News Commentary corpus is a compilation of a collection of news editorials freely available in the web. In the present work, we focused on English–Spanish translation. The Europarl corpus was used as out-of-domain (or general-domain) corpus, and the News-Commentary corpus was considered in-domain. Table 1 provides a brief overview of the Europarl data, and Table 2 provides the corresponding overview of the News Commentary data. The development data chosen was the one used as test data in the 2008 workshop on SMT of the ACL, and the actual test data was the one used as test data in the 2013 edition of that same workshop.

Experiments were conducted by means of the open source SMT toolkit Moses [15], in its standard non-monotonic setup. The phrase tables were generated by means of word alignments

[2]Available from http://www.statmt.org

		Spanish	English
Train	Nr. Sentences	149k	
	Run. words	4.5M	3.9M
	Voc. size	178k	143k
Devel.	Nr. Sentences	2051	
	Run. words	47.1k	43.4k
	Voc. size	141k	123k
Test	Nr. Sentences	3000	
	Run. words	62.6k	56.9k
	Voc. size	158k	134k

Table 2: Statistics of the News-Commentary data.

using the GIZA++ toolkit [25], and the language model used was a 5-gram language model with modified Kneser-Ney smoothing [14], built by means of the SRILM toolkit [30]. The log-linear weights present in every state-of-the-art SMT system were optimised by means of Minimum Error Rate Training (MERT) [26] on the development partition of the corpus. The weights were only optimised once, i.e., in the experiment involving the SMT system trained with the both in-domain and out-of-domain data, and the set of weights obtained was carried on to further experiments with the purpose of avoiding possible noise due to fluctuations in the MERT procedure.

In this work we report on two different baselines: the first one involves using only the in-domain data as training data for the machine translation and language models, and the second one implies using all the data available, i.e., training partitions of both Europarl and News-Commentary concatenated.

Both sentence selection strategies were used to select data from the Europarl training set, and according to the source side of the test data. In the case of the infrequent n-gram recovery technique, the meta-parameter t was set to 10.

3.2.3 Results

The results concerning the comparison of both sentence selection strategies are shown in Figure 2. There are several things which should be noted in this figure:

1. Both sentence selection techniques are able to improve over the in-domain baseline from the very beginning.
2. Both sentence selection techniques are able to improve over the all-data baseline when the amount of data added is only 400k sentences, i.e., one forth of the original size of the Europarl data.
3. From both techniques, the infrequent n-gram recovery strategy appears to work best, improving over the entropy-based strategy by about one BLEU point.

3.2.4 Conclusions and future work

The work presented here has shown that both sentence selection strategies presented are able to deliver improvements over both the in-domain baseline and the all-data baseline. In addition,

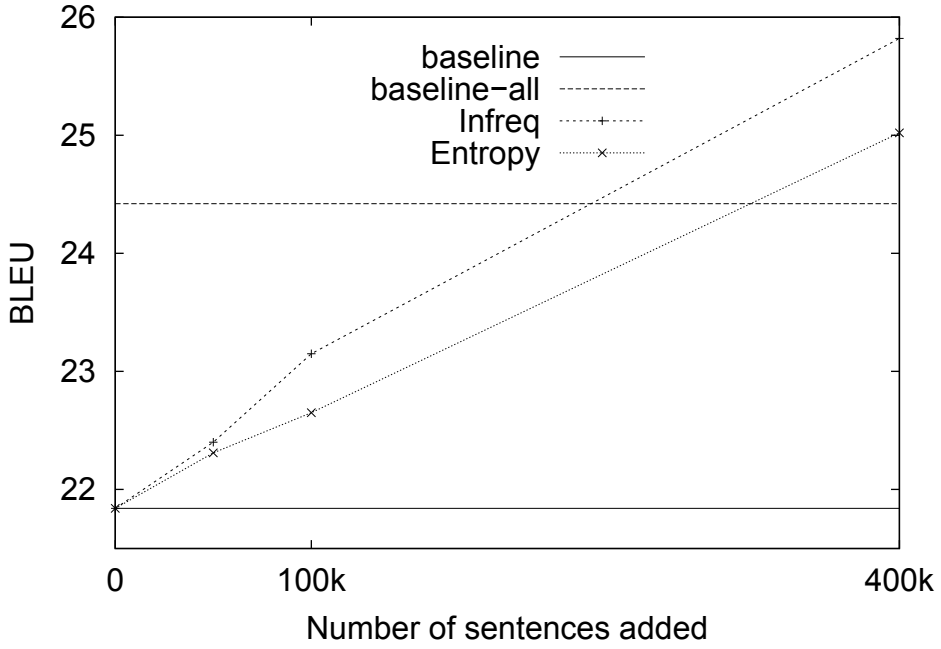

Figure 2: Comparison of the two sentence selection strategies presented here

we have also shown that the infrequent n-gram recovery technique is the one that provides the best improvements.

Even though Task 4.3 is officially finished according to the DOW, we feel that sentence selection is a promising enough avenue for further research. Hence, we plan to continue working on this direction for the remaining of the project. Scheduled extensions include different combinations of both strategies presented here, as well as selecting sentences according to a development set (instead of the actual source side of the test data).

3.2.5 Collaboration with MATECAT

In addition, part of this work will be done in close collaboration with the MATECAT[3] project, since such project has also led an extensive amount of work in the field of sentence selection for translation model adaptation. In this particular direction, research has already begun regarding how to integrate the phrase-table adaptation strategy developed in the MATECAT project. For this purpose, one of the members of the UPVLC CASMACAT team stayed one week at the FBK research centre in Trento, the leading team of MATECAT, and a joint research paper is being prepared and will be submitted to a first-level conference.

The adaptation strategy developed within MATECAT implies selecting those sentences from the in-domain corpus which are considered to be the most similar either to the current data to be translated, or to the data translated in the previous days. Then, a *foreground* translation and reordering model is learnt, and the probabilities of the original translation and reordering models (i.e., the ones estimated on the complete data) are only used in case the foreground model is not able to provide an appropriate translation for the sentence being considered. Preliminary experiments show that the improvements obtained by building the foreground model with the samples retrieved by either of the two sentence selection strategies described above are comparable.

Hence, work in this direction concerns mainly studying ways in which the two sentence selection strategies described can be combined so as to obtain the best possible foreground model.

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

Preliminary results show that there is much to be gained when combining both selection strategies. For instance, simply pooling the sentences obtained by both techniques report interesting gains in terms of automatic translation quality metrics such as BLEU or TER, although further analysis is still required. In addition, a preliminary analysis of the sentences selected by both strategies report very little overlap both in terms of which sentences are being selected and in terms of which n-grams are present in the selected sentences. This implies that both strategies are providing very disjoint bilingual sentences pairs for the foreground model, and hence a selection strategy that combines both should be able to yield improvements over only using one of them.

References

[1] Dana Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, April 1988.

[2] Les Atlas, David Cohn, Richard Ladner, M. A. El-Sharkawi, and R. J. Marks, II. Advances in neural information processing systems 2. In David S. Touretzky, editor, *NIPS*, chapter Training connectionist networks with queries and selective sampling, pages 566–573. Morgan Kaufmann Publishers Inc., 1990.

[3] Sergio Barrachina, Oliver Bender, Francisco Casacuberta, Jorge Civera, Elsa Cubel, Shahram Khadivi, Antonio Lagarda, Hermann Ney, Jesús Tomás, Enrique Vidal, and Juan-Miguel Vilar. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35:3–28, 2009.

[4] Mara Chinea-Rios. Estrategias de aprendizaje online de los pesos del modelo log-lineal en traduccin automtica interactiva. Master's thesis, Universitat Politècnica de València, 2014. Advisors: Germán Sanchis-Trilles and Francisco Casacuberta.

[5] Mara Chinea-Rios, Germán Sanchis Trilles, Daniel Ortiz-Martnez, and Francisco Casacuberta. Online optimisation of log-linear weights in interactive machine translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, May 2014.

[6] David Cohn, Les Atlas, and Richard Ladner. Improving generalization with active learning. *Machine Learning*, 15:201–221, 1994.

[7] George Foster. *Text Prediction for Translators*. PhD thesis, Université de Montréal, may 2002.

[8] George Foster, Pierre Isabelle, and Pierre Plamondon. Target-text mediated interactive machine translation. *Machine Translation*, 12(1/2):175–194, 1998.

[9] Francisco Casacuberta Francisco-Javier López-Salcedo, Germán Sanchis-Trilles. Online learning of log-linear weights in interactive machine translation. In *Proceedings of Advances in Speech and Language Technologies for Iberian Languages (iberSPEECH)*, pages 277–286, 2012.

[10] Guillem Gascó, Martha-Alicia Rocha, Germán Sanchis-Trilles, Jesús Andrés-Ferrer, and Francisco Casacuberta. Does more data always yield better translations? In *Proceedings of the 13th European Chapter of the Association for Computational Linguistics*, pages 152–161, 2012.

[11] Jesús González-Rubio. *On the Effective Deployment of Current Machine Translation Technology*. PhD thesis, Universitat Politècnica de València, 2014. Advisors: Daniel Ortiz-Martínez and Francisco Casacuberta.

[12] Jesús González-Rubio and Francisco Casacuberta. Cost-sensitive active learning for computer-assisted translation. *Pattern Recognition Letters*, 37:124–134, 2014.

[13] Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. Active learning for interactive machine translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254, 2012.

[14] Reinhard Kneser and Hermann Ney. Improved backing-off for m-gram language modeling. *Proceedings of the International Conference on Acoustic, Speech and Signal Processing*, II:181–184, 1995.

[15] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christie Moran, Richard Zens, Chris Dyer, Ontraj Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, 2007.

[16] Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121, 2006.

[17] Philipp Koehn and Josh Schroeder. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the 2nd ACL Workshop on Statistical Machine Translation*, pages 224–227, 2007.

[18] Philippe Langlais, George Foster, and Guy Lapalme. Unit completion for a computer-aided translation typing system. *Machine Translation*, 15(4):267–294, 2000.

[19] Philippe Langlais and Guy Lapalme. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 15(4):77–98, 2002.

[20] David Lewis and William Gale. A sequential algorithm for training text classifiers. In *Proceedings of the ACM SIGIR conference on Research and development in information retrieval*, pages 3–12, 1994.

[21] Francisco Javier López-Salcedo. Aprendizaje online de los pesos del modelo log-lineal en traducción automática interactiva. Master's thesis, Universitat Politècnica de València, 2012. Advisors: Germán Sanchis-Trilles and Francisco Casacuberta.

[22] Pascual Martínez-Gómez, Germán Sanchis-Trilles, and Francisco Casacuberta. Online adaptation strategies for statistical machine translation in post-editing scenarios. *Pattern Recognition*, 45(9):3193–3203, 2012.

[23] Robert C. Moore and William Lewis. Intelligent selection of language model training data. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 220–224, 2010.

[24] John A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4):308–313, 1965.

[25] Franz J. Och and Hermann Ney. A systematic comparison of various statistical alignment models. 29(1):19–51, 2003.

[26] Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41th Annual Conference of the Associations for Computational Linguistics*, pages 160–167, 2003.

[27] D. Ortiz-Martínez and F. Casacuberta. The new Thot toolkit for fully automatic and interactive statistical machine translation. In *14th Annual Meeting of the European Association for Computational Linguistics*, pages 45–48, April 2014.

[28] Daniel Ortiz-Martínez, Ismael García-Varea, and Francisco Casacuberta. Online learning for interactive statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 546–554, 2010.

[29] Germán Sanchis-Trilles. *Building task-oriented machine translation systems.* PhD thesis, Universitat Politècnica de València, 2012. Advisor: Francisco Casacuberta.

[30] Andreas Stolcke. SRILM – an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904, 2002.