D4.3bis: Addendum to the deliverable D4.3

Germán Sanchis-Trilles, Mauro Cettolo, Mara Chinea-Rios, Francisco Casacuberta

Distribution: Public

CASMACAT
Cognitive Analysis and Statistical Methods
for Advanced Computer Aided Translation

ICT Project 287576 Deliverable D4.3bis

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	none
Actual date of delivery	November 7, 2014
Date of last update	November 7, 2014
Deliverable number	D4.3bis
Deliverable title	Addendum to the deliverable D4.3
Type	Report
Status & version	Final
Number of pages	8
Contributing WP(s)	WP4
WP / Task responsible	UPVLC
Other contributors	FBK
Internal reviewer	Philipp Koehn
Author(s)	Germán Sanchis-Trilles, Mauro Cettolo, Mara Chinea-Rios, 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 is an extension of D4.3 detailing work done in Task 4.3 in collaboration with the MATECAT project. Such work was in progress when Task 4.3 finished officially, and hence could not be included into D4.3.

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. Specifically, Task 4.3, which is the one we report on in this document, is devoted to the development of novel techniques for enabling the system to adapt to new domains and/or users.

This was mainly tackled during the duration of the task by means of Bayesian adaptation, sentence selection strategies, and language model interpolation. Given that the MATECAT FP7 project was also working on model adaptation via sentence selection, we decided that it would be in the benefit of both projects to tighten the collaboration in this specific subject. To this end, one of the CASMACAT members of the UPVLC team spent one week in Trento, Italy, in the premises of the Fondazione Bruno Kessler (FBK), the MATECAT project co-ordinator. The main purpose of the collaboration undertaken was to analyse to what extent infrequent n-grams, the sentence selection strategy adopted in the CASMACAT project, could lead to complementary results with the adaptation strategy adopted in the MATECAT project.

In the rest of this document, we will first briefly review the infrequent n-grams selection strategy (Section 1), then the adaptation framework of the MATECAT project (Section 2) in which we integrated infrequent n-grams, and finally the results achieved in the collaboration (Section 3).

Contents

1 Sentence selection in CASMACAT **4**

2 Adaptation in the MATECAT project **4**
2.1 Data selection strategy 5
2.2 Model adaptation 5

3 Experimental results **6**

4 Conclusions **7**

Bibliography **8**

1 Sentence selection in CASMACAT

Within the CASMACAT project, sentence selection has been used for selecting, from very large bilingual corpora, which sentences are the best for training test-specific translation and language models. Specifically, we used the infrequent n-grams selection strategy. Proposed by [3], 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:

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

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 of too many sentences with the same infrequent n-gram. First, out-of-domain sentences are scored using Equation (1). 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.

2 Adaptation in the MATECAT project

Note that the data selection strategy described in this section was also analysed in the CASMACAT project, and we determined, in traditional SMT experiments, that the n-gram sentence selection

strategy presented in Section 1 led to better results (see deliverable D4.3). Nevertheless, the difference between the experiments reported in D4.3 and the ones carried out in the MATECAT project is the use that is given to the selected sentences: while in the CASMACAT project we only used the sentences selected for the purpose of training the SMT system (and only those sentences selected were used), the use given in the MATECAT project was to generate adapted mixture models. Such strategy will be described in the second part of this section.

2.1 Data selection strategy

The MATECAT project had been applying the sentence selection strategy developed by [4], which was then extended by [1] to work on bi-texts and implemented in the public tool XenC [6]. 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 [4], but is left out here for simplicity reasons and because it does not provide further practical information.

2.2 Model adaptation

Within the MATECAT project, translation model adaptation is performed by means of the fill-up technique described initially in [5], and then extended by [2]. The main intuition behind this idea is to merge a rather generic background phrase table (obtained from all the data available) with a specific foreground table (obtained, in this case, from the topic-specific data selected by means of a sentence selection strategy). This fill-up is performed by complementing the foreground table with those entries of the background table that do not appear in the foreground table. This means that the phrase table obtained in such a way ensures the same coverage as the generic phrase table, but with the scores biased towards topic specific data, whenever such data is available. In addition, to keep track of the origin of each phrase, a binary feature is added, that is one in case a specific phrase pair origins from the background table, or zero otherwise. Such binary feature is assigned a scaling factor whose value is estimated as usual by means of MERT. The same procedure can be applied to the lexicalised reordering models, which are standard in Moses.

As for the language model, two separate language models are obtained: one for the generic data, and one for the topic-specific data selected by means of a sentence selection strategy. Then such language models are interpolated linearly so as to maximise the log-likelihood of an in-domain development set. Similarly to the translation model, the purpose of this interpolation is to provide the resulting language model with as much coverage as possible, while still biasing the probability distribution towards the in-domain data.

test	baseline	XenC	infreq.
D1	47.6/36.9	47.6/36.7	48.4/36.1
D2	46.8/35.3	48.5/34.0	48.4/34.4
D3	45.5/38.7	50.7/35.4	50.0/36.1
D4	48.8/37.6	49.8/36.5	49.9/35.9
avg.	47.3/37.2	49.4/35.7	49.4/35.6

Table 1: Model adaptation by using both cross entropy (XenC) and infrequent n-grams (infreq.) for sentence selection for the foreground model. Results given in BLEU/TER.

3 Experimental results

In the framework of the collaboration with MATECAT, the first step was to compare the sentence selection strategy MATECAT was using, i.e., the one implemented within XenC, and the sentence selection strategy developed within CASMACAT, i.e., infrequent n-gram selection, for the specific purpose of model adaptation within the MATECAT adaptation framework.

Initial experiments were conducted with a portion of the English-Italian JRC-Acquis corpus, which belongs to the legal domain. The test set was divided into five different subsets, each one with approximately 3500 words, which is the estimated amount of words a human post-editor can post-edit per day.

Preliminary experiments with perplexity were conducted so as to determine which options were best for each one of the two different sentence selection strategies, at it was established that the best-performing setup for the XenC selection was to select the bilingual sentence pairs using as seed data the source side of all the test data (i.e., all five days), and the target data of the previous days. For instance, selection for day 4 would be performed according to the source side of all five days, plus the target side of days 1 to 3. However, in the case of infrequent n-gram selection, the best setup according to the preliminary experiments implied selecting bilingual sentences only according to the source side of the current day, i.e., in day 4 the selection would happen only according to the source side of day 4.

With the setup described in the previous paragraph, SMT systems were built by means of Moses, and were evaluated on the different test subsets representing each of the days. Table 1 reports on the results obtained. Note that day 0 is omitted because there is no adaptation data available at that point. As shown, results for both sentence selection strategies were quite positive. This was quite expected, since this experiment was performed on a setup which the MATECAT team had already been experimenting with, and had already reported positive results with the cross entropy sentence selection strategy. However, more interesting is that the infrequent n-gram sentence selection strategy seems to be able to provide improvements to the same degree as cross entropy selection. In some cases (e.g. D3) the translation quality delivered falls below the translation quality delivered by the XenC setup, but in other cases it is just the other way round (e.g. D1), averaging very similar results after all test subsets have been processed.

After observing these results, the first question posed was whether both strategies were selecting approximately the same sentences. However, this was not the case: for instance, for Day 3 the XenC setup selected approximately $131k$ sentences, whereas the infrequent n-grams strategy selected only $19k$ sentences. In addition, and most interestingly, for Day 3 there was no overlap at all between the sentences selected by XenC and those selected by infrequent n-grams, and the same fact could be observed for the rest of the test subsets.

We understood that the lack of overlap between XenC and infrequent n-gram selection evidenced that a combination of both strategies would lead to better results. In this direction,

	English→Spanish		English→German	
system	BLEU	TER	BLEU	TER
baseline	46.9	37.8	29.4	54.1
XenC	49.2	36.4	30.3	56.1
Infreq.	48.4	36.5	30.8	55.6
Pooled	48.9	36.4	29.3	57.8

Table 2: BLEU and TER results using different sentence selection approaches for two different language pairs.

we built a system in which the foreground model was built using all the sentences selected by both selection strategies, pooled together. This system achieved average scores, after processing all of the test subsets, of 50.2 BLEU and 35.4 TER, i.e., 0.8 BLEU and 0.2 TER points better than the best of the systems built using only one of the sentence selection strategies.

With such encouraging results, we pursued experiments with further languages of the same corpus. Such results are shown in Table 2. In contrast to the results obtained with the English-Italian language pair, in this case results are rather mixed: infrequent n-gram selection seems to perform better than XenC for the English-German language pair, but worse for the English-Spanish language pair. In addition, in this case the pooling results did not yield better results than the best performing individual sentence selection strategy, and in the case of English-German the pooled system even performed slightly worse than the baseline system. The reasons for this behaviour are at the time of writing still not clear, and we are conducting different analysis with the purpose elucidating possible reasons. Additional experiments with the CASMACAT corpora (i.e., Europarl as training corpus and News-Commentary as domain-specific corpus) also lead to mixed results, with the infrequent n-grams strategy achieving 24.3 BLEU and XenC 25.4 BLEU, with a baseline of 24.4 BLEU.

4 Conclusions

Work in this direction is still ongoing, since we still need to analyse further why such mixed results are being obtained, and whether it is possible to know beforehand whether a given corpus or test set will yield improvements or not. For this reason, and given that the results obtained are still very preliminary, this adaptation approach has not been implemented into the CASMACAT workbench. Nevertheless, infrequent n-gram selection was used to train some of the systems that have been used in the field trials, and adaptation is performed effectively in the CASMACAT workbench by means of online learning of both log-linear weights and features, with strategies that have proven to be successful in different scenarios (see deliverables D4.1, D4.2 and D4.3).

References

[1] Amittai Axelrod, Xiaodong He, and Jianfeng Gao. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 355–362, 2011.

[2] Arianna Bisazza, Nick Ruiz, and Marcello Federico. Fill-up versus interpolation methods for phrase-based smt adaptation. In *Proceedings of the International Workshop on Spoken Language Translation*, 2011.

[3] 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.

[4] 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.

[5] Preslav Nakov. Improving english-spanish statistical machine translation: experiments in domain adaptation, sentence paraphrasing, tokenization, and recasing. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 147–150, 2008.

[6] Anthony Rousseau. Xenc: An open-source tool for data selection in natural language processing. *The Prague Bulletin of Mathematical Linguistics*, 100:73–82, 2013.