Building task-oriented machine translation systems

Germán Sanchis-Trilles
Advisor: Francisco Casacuberta

Pattern Recognition and Human Language Technologies Group
Departamento de Sistemas Informáticos y Computación
Universitat Politècnica de València
gsanchis@dsic.upv.es

Abstract. This PhD dissertation, written by Germán Sanchis-Trilles and supervised by Francisco Casacuberta, was defended on June 20^{th}, 2012, at the Universitat Politècnica de València. The committee members were Enrique Vidal (UPV), Ismael García-Varea (UCM), Hermann Ney (RWTH-Aachen), José Oncina-Carratalá (UA), and Nicola Cancedda (XRCE). The qualification obtained was "sobresaliente cum laude". The main goal is to develop computer assisted translation and machine translation systems which present a more robust synergy with their potential users. Hence, the main purpose is to make current state-of-the-art systems more ergonomic, intuitive and efficient, so that the human expert feels more comfortable when using them. For doing this, different techniques are presented, focusing on improving the adaptability and response time of the underlying statistical machine translation systems, as well as a strategy aiming at enhancing human-machine interaction within an interactive machine translation setup. All of this with the ultimate purpose of filling in the existing gap between the state of the art in machine translation and the tools that are usually made available to the final human translators.

Keywords: Machine translation, adaptation, human-computer interaction

1 Introduction

In the last years, machine translation (MT) and computer assisted translation (CAT) is being increasingly embraced by human translators, who often find that post-editing the output of an MT system is an efficient way of producing high quality translations. Nevertheless, there are yet several problems which need to be dealt with before the usage of statistical machine translation (SMT) systems within CAT systems finds a more widespread usage. These problems concern mainly the time-efficiency, the adaptability, and the usability of the SMT systems. In this PhD dissertation, these three problems are confronted, yielding different degrees of success.

In the first place, state-of-the-art statistical machine translation (SMT) systems are often unable to yield real-time performance. This problem is even worse when the system has been trained on very large amounts of data, which is always desirable given that more data usually implies higher model coverage. When the amount of translation options and bilingual data made available to the system increases, translation throughput is necessarily affected, and model pruning strategies need to be applied with the

purpose of not having the human translator waiting too long for the system to produce its output, which would be on the one hand exasperating, and on the other hand economically inefficient. In this thesis, we focused on proposing a model pruning strategy which proves to be able to decrease system response time drastically, while keeping translation quality within state-of-the-art ranges.

Another topic tackled in this thesis is system adaptability. There is extensive work in SMT which proves that the translation quality produced by a typical machine translation system drops significantly when the text to be translated stems from a different topic than the data which has been used to train the system. In addition, different human translators may have different styles when translating a document, which implies that lexical choice or sentence length may be required to vary even when working within one single domain. Furthermore, from a user point of view it is mentally exhausting for a human translator to correct the same mistakes over and over again, while having the impression that those same mistakes will keep on appearing because the system is not learning from its own errors. For these reasons, system adaptability is unveiled as a key feature within a machine translation system that is setup within a human-machine collaborative framework. In the present thesis, two different model adaptation techniques are presented. The first one deals with the problem of language model adaptation, i.e., adapting the specific part of the translation system that controls word ordering and structure in the hypotheses produced. The second one deals with the adaptation of the translation model itself, which is the part of the translation system that will account for lexical choice and sentence length, among other features.

Lastly, usability of interactive machine translation systems is also a very important topic when attempting to build systems that are to be used by humans, whose expertise when using computers should not always be assumed. Hence, it is important to take special care when designing the interaction scheme, so that the human translator feels as comfortable as possible when using the translation interface. In this context, it is important to realise that the keyboard is not the only input device that the human user may use, but rather that richer interaction schemes might boost productivity. Nevertheless, it is also important to keep the interaction interface simple, so that the human expert is not overburdened. In this thesis, we propose a very simple and intuitive extension to the classical interactive machine translation interaction scheme, which takes into account the actions that the user performs before correcting a given hypothesis.

2 State of the art before this thesis

At the time of beginning this thesis, translation models were (and are still) often too large to fit into the memory of most table-top machines (let alone portable devices). This led to the widely-accepted practice of filtering the translation model according to the current test set to be translated. Even after doing so, state-of-the-art translation systems often took about 20 seconds to a whole minute when translating a single sentence, depending on its lenght. This fact led to the development of different test-independant translation model pruning strategies, some of which [1, 2] were published simultaneously to the first experiments reported in the present thesis. The technique presented in this thesis follows a similar direction to the one pointed out in [2], but instead of relying

on additional data for extracting usage statistics, it relies on the idea of re-estimating the current model parameters directly on the training data, thus discarding more parameters than reported in [2], while keeping translation quality mostly unaltered.

Regarding adaptation in SMT systems, there had been already some work in this direction. Specially concerning language model adaptation, the work presented in this thesis extends previous work already performed in [3, 4]. In these works, the training data is also partitioned in different ways with the purpose of extracting different sub-domains and building smaller topic-specific language models. The technique developed in this dissertation extends such work by presenting and comparing several different corpus subdivision strategies. In addition, the resulting language models are interpolated dynamically in translation time so as to optimise the perplexity of the input sentence, which had not been done as of yet. Regarding translation model adaptation, work performed before the beginning of this thesis mainly focused on examining different ways of combining the available training data [5, 6], did not confront the adaptation of the final translation model parameters directly [7], and the adaptation problem was seldom confronted from a purely statistical point of view. In this sense, the Bayesian adaptation strategy presented in this thesis confronts the adaptation of the model parameters directly, and has a solid statistical background.

Finally, user-machine interaction in interactive machine translation (IMT) had been remaining the same for the last years preceding the beginning of this thesis, i.e., the system waited for a keyboard interaction before performing any prediction [8]. With the introduction of the mouse as an alternative in the present thesis, it was proven that the classical IMT paradigm accepts numerous extensions, and since then there have been other works that have also attempted to extend such paradigm in different ways [9, 10].

3 Main contributions of the thesis

The main contributions of this thesis can be split into the three problems it tackles:

- A novel translation model pruning technique has been developed. Performing an unsmoothed re-estimation of the model parameters on the training data leads to much smaller models (about 3-5% of the original size), while translation quality remains mostly unaltered. The technique developed relies on the idea of translating the source side of the training data, selecting a set of best translations $G(\boldsymbol{x})$ for each input sentence, and then re-estimating the model parameters on those translations.
- Two novel adaptation techniques have been developed, one of them dealing with the adaptation of the language model, while the other one involves the adaptation of the translation model. The language model adaptation strategy first splits the training data so as to conform different smaller language models, which are then dynamically interpolated in translation time according to the input data. Results are not reported in the present paper, but are shown in the manuscript of the dissertation. The translation model adaptation strategy relies on the theoretical framework of Bayesian learning. Considering the model parameters as random variables whose prior distribution is influenced by the adaptation data leads to an effective adaptation framework. Since the Bayesian paradigm often implies computing an intractable integral, several different sampling strategies were implemented. In this

paper, only the best performing sampling methods are shown, namely Markov chain Monte-Carlo (MCMC) and a heuristic strategy (more details about this strategy in the document of the dissertation).

– The classical interactive machine translation framework is expanded by considering the mouse as an additional and valuable information source for the system. Two different pointer actions (PAs) were considered. The first one relies on the fact fact, before typing in a word, the user needs to position the cursor first. At this point, the system may already realise that the user intends to change the word located directly after the cursor (say y_i), so it is already able to provide a different translation completion, in which word y_i has been changed. This kind of pointer action was named anticipated proposal. In addition, the possibility of performing an explicit pointer action that explicitly refuses the system's proposal was also added. In this scenario, the user would perform a pointer action in front of a given word, thus explicitly asking the system to change it, together with the rest of the translation. This kind of pointer action was called partial refusal.

4 Corpus and methodology

With the purpose of giving soundness and replicability to the results reported in this thesis, all experiments were conducted by applying state-of-the-art SMT systems whenever possible, and using standard corpora that have been widely used in different SMT workshops. However, since the results presented correspond to a period of time that spans through five years $(2007 - 2012)$, some corpora were updated in the meantime.

Confidence intervals were computed for each one of the results reported, with the purpose of pointing out clearly which results are statistically significant, following the bootstrap strategy described in [11]. In the case of the experiments concerning adaptation, confidence intervals were computed by repeatedly extracting different random adaptation sets from the data available.

All the results presented in the present paper were obtained by using the open source SMT toolkit Moses [12], in its most standard setup, i.e., with direct and inverse translation models, lexical weights, a lexicalised re-ordering model [13] and a 5-gram language model estimated with the SRILM [14] toolkit.

Concerning the corpora used, all of the experiments reported on this paper were performed by using the Europarl corpus [15] as training data, in the partitions established in the WMT07 [16] and WMT10 [17] workshops on MT. The use of one or the other depended on the date in which the experiments were performed. Hence, the experiments concerning the parameter pruning strategy and effects of Bayesian adaptation are reported on the WMT10 partition, whereas the experiments conducted to assess the effectiveness of using pointer actions within an IMT framework are reported on the WMT07 partition. For the experiments concerning adaptation, the News-Commentary (NC) corpus was additionaly considered, also in the partition established in the WMT10 workshop. Unless stated otherwise, the log-linear weights of the translation model were always estimated on the `Devel.` set of the Europarl corpus. Different statistics of these corpora can be seen in Table 1.

System evaluation was performed by means of TER [18] and BLEU [19], whenever the experiments involved an SMT system, and WSR [8] for those involving an IMT

		De	En	Fr	En
WMT07 training	Sentences	751k		688k	
	Running words	15.3M	16.1M	15.6M	13.8M
	Vocabulary size	195k	66k	80k	62k
WMT10 training	Sentences	1219k		1251k	
	Running words	24.9M	26.1M	28.1M	25.6M
	Vocabulary size	255k	82k	101k	81k
Devel.	Sentences	2000		2000	
	Running words	55k	59k	67k	59k
	OoV 07/10	432/348	125/103	144/99	138/104
Devtest	Sentences	2000		2000	
	Running words	54k	58k	66k	58k
	OoV 07/10	377/310	127/111	139/114	133/112
NC training	Sentences	86.9k		67.6k	
	Running words	1.8M	1.8M	1.6M	1.4M
	Vocabulary size	86.7k	40.8k	43.3k	35.6k
NC 09 test	Sentences	2525		2525	
	Running words	62.7k	65.6k	72.6k	65.6k
	OoV NC/10	3629/2410	1853/1247	2478/1446	2035/1247

Table 1. Characteristics of Europarl and News-Commentary, for each of the sub-corpora. OoV stands for "Out of Vocabulary" words with respect to the WMT07 training data (07), the WMT10 training data (10), or the NC training data (NC). Devel. stands for Development, k for thousands of elements and M for millions of elements.

system. TER is an error metric (i.e., the lower the better) that measures the minimum amount of edits required to transform the system's hypothesis into the reference sentence. BLEU is a precision metric (i.e., the higher the better) that measures n-gram precision, with a penalty for sentences that are too short. WSR measures the ratio of words that a human translator would need to type before achieving the sentence he has in mind (i.e., the lower the better).

5 Experimental results

5.1 Model size reduction

For assessing the effectiveness of the model pruning strategy described in Section 3, a baseline model was estimated by means of Moses, as described in the previous section, by using the WMT10 training data for training the initial models and the `Devtest` set as evaluation set. Results for different language pairs are available in the document of the dissertation, but in this paper we only report results on French–English. Different sizes of $G(\boldsymbol{x})$ (see Section 3 were considered, both within a SMT and an IMT setting.

The results for the SMT case are shown in Figure 1, whereas the results for IMT are shown in Figure 2. It can be seen that the reduction strategy implemented has a very important impact in system speed, while mantaining translation quality almost unaltered in the case of SMT. It is worth noting that, for $|G(\boldsymbol{x})| = 1$, only about 3% of the original paremeters are retained.

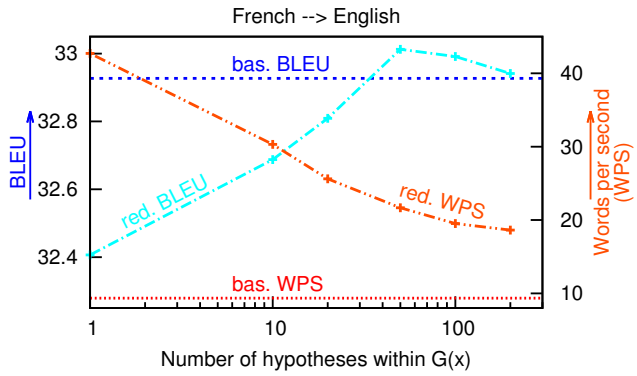

Fig. 1. Model pruning within in SMT. `bas.` stands for baseline, and `red.` for the pruned system.

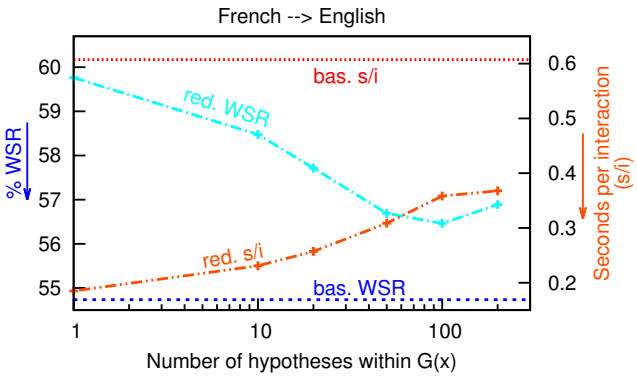

Fig. 2. Effect of model pruning within an IMT scenario.

In the case of IMT, the results suggest that a certain trade-off is required. Although the speed increase is also very significative, the increase in the amount of interactions required by the human translator may not be acceptable for small sizes of $G(\boldsymbol{x})$. However, it must be considered that response time in IMT is critical specially when the suffix to be produced is long, which implies that the technique implemented might be useful in such cases, but not when the suffix to be produced is short.

5.2 Bayesian translation model adaptation

The effectiveness of the Bayesian translation model adaptation strategy (Section 3) was tested by training a baseline SMT system on the German–English WMT10 training data. Then, the NC training data was used for randomly extracting adaptation sets of different sizes, with the purpose of analysing the behaviour when increasing the amount of adaptation data, and the adapted systems were evaluated on the NC09 test set. The results are shown in Figure 3, for the two sampling strategies that yielded the best results. As shown, the two Bayesian adaptation strategies perform better than the complete re-estimation of the log-linear weights. Although not shown here, confidence intervals

were also found to be much smaller for the two Bayesian approaches than for MERT (although not for MERT+), implying more relyiable results.

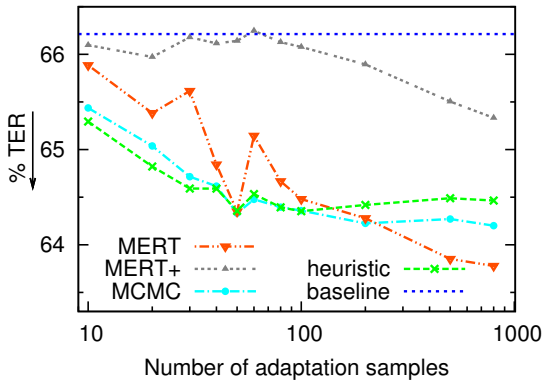

Fig. 3. Translation quality comparison between re-estimating $\boldsymbol{\lambda}$ from scratch and Bayesian adaptation for different sampling strategies. MERT stands for re-estimating the log-linear weights from scratch on the adaptation data, and MERT+ for re-estimating them on the concatenation of adaptation data and development set.

Additionally, a temporal analysis was also conducted, and is shown in Figure 4. As shown, the two Bayesian approaches perform about half an order of magnitude faster than the MERT approach, and two orders of magnitude faster than the MERT+ approach. When computing the time required by each system, all of the steps were considered (e.g., computing the sentence-level TER counts for the Bayesian approaches). We find that these results are very encouraging, since when a user is sitting waiting for the translation to be produced, computational time is critical. Although not presented here for space reasons, online adaptation experiments are also shown in the document of the dissertation, reporting encouraging results.

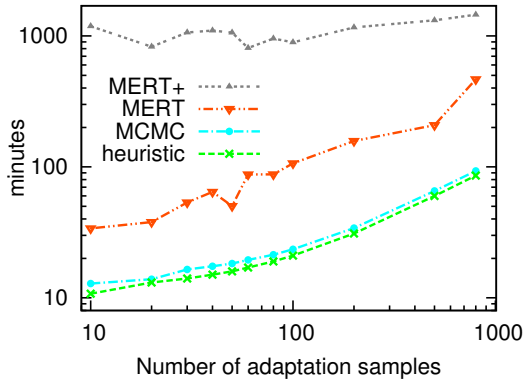

Fig. 4. Temporal comparison between re-estimating $\boldsymbol{\lambda}$ from scratch and Bayesian adaptation for different sampling strategies.

5.3 Enriching user-machine interaction

For assessing the effectiveness of the novel interaction scheme described in Section 3, an IMT scenario was simmulated by considering that the sentence the user would want to obtain is the reference present in the test set. The French–English WMT07 training data was used to train the SMT translation models, and system performance was evaluated on the `Devtest` set (Figure 5). Experiments concerning other language pairs lead to similar results, and are omitted here for space reasons.

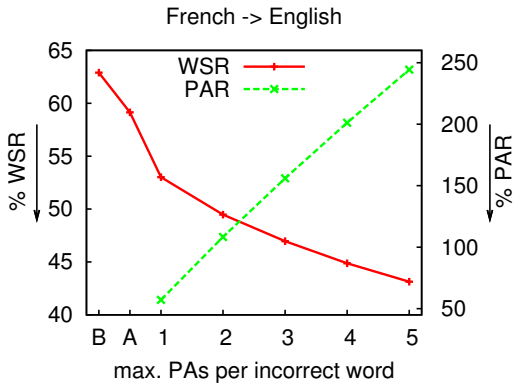

Fig. 5. IMT system performance when introducing pointer actions. `B` stands for baseline (i.e., the classical interaction scheme), `A` for teh system that incorporates the anticipated proposal, and considering up to n partial refusal pointer actions. `PAR` is the ratio of pointer actions performed.

As shown in Figure 5, considering the anticipated proposal pointer action already leads to a reduction of about 5% in the amount of words typed by a human translator. Considering only one partial refusal pointer actions leads to a further drop of more than 5%. However, considering further partial refusal pointer actions has less effect, and the benefit tends to fade for larger amounts of pointer actions. Hence, it seems quite unlikely that someone would perform more than 2 or 3 pointer actions. However, by just performing 3, the potential user would already spare typing in about 10% of words.

6 Conclusions

As a result of this thesis, several contributions to the human-machine interaction framework were made. The response time of state-of-the-art SMT and IMT systems was improved, without any loss of translation quality in the former case, requiring a certain compromise between time efficiency and human effort in the latter. System adaptability was also improved in two key aspects of a typical SMT system, such as the translation model and the language model. Results in the first case achieved very promissing improvements with respect to the baseline system. In addition, the classical user-machine interface of an IMT system was extended with very promissing results. All these strategies and ideas were tested within state-of-the-art SMT and IMT systems, trained on publicly available corpora which have been used in recent SMT workshops, and yielded satisfactory results, as shown by the experiments reported.

7 Publications

Preliminary work regarding the parameter pruning strategy described in Section 3 was published in an international workshop [20] and an international conference [21]. A more refined version of such work, and re-formulated as a parameter re-estimation technique, was presented later on in an international conference [22]. The language model adaptation technique lead to two publications in an international workshop [23, 24], and a publication in an international conference [25]. Related work was also published in another international workshop [26]. The Bayesian translation model adaptation strategy lead to one publication in an international workshop [27] and another publication in an international conference [28]. The online variant was published in the Pattern Recognition Journal [29], and the batch variant was used within an SMT system presented in an international MT competition [30]. Finally, the interaction framework described in Section 3 was accepted for publication in an international workshop [31] and in an international conference [32]. Four of these publications [32, 28, 27, 26] are ranked as high impact by the Computer Research and Education Association of Australasia (CORE).

Acknowledgements

Work partially supported by the EU 7th Framework Programme (FP7/2007-2013) under grant Nr. 287576 (CasMaCat), by the EC (FEDER/FSE) and the Spanish MICINN under projects MIPRCV "Consolider Ingenio 2010" (CSD2007-00018) and iTrans2 (TIN2009-14511), and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

1. Johnson, H., Martin, J., Foster, G., Kuhn, R.: Improving translation quality by discarding most of the phrasetable. In: Proc. of EMNLP/CoNLL. (June 2007) 967–975
2. Eck, M., Vogel, S., Waibel, A.: Translation model pruning via usage statistics for statistical machine translation. In: Proc. of NAACL/HLT. (April 2007) 21–24
3. Zhao, B., Eck, M., Vogel, S.: Language model adaptation for statistical machine translation with structured query models. In: Proc. of COLING. (August 2004) 411–417
4. Lü, Y., Huang, J., Liu, Q.: Improving statistical machine translation performance by training data selection and optimization. In: Proc. of EMNLP/CoNLL. (June 2007) 343–350
5. Koehn, P., Schroeder, J.: Experiments in domain adaptation for statistical machine translation. In: Proc. of the ACL second workhop on SMT. (June 2007) 224–227
6. Bertoldi, N., Federico, M.: Domain adaptation in statistical machine translation with monolingual resources. In: Proc. of the EACL fourth workshop on SMT. (March 2009) 182–189
7. Civera, J., Juan, A.: Domain adaptation in statistical machine translation with mixture modelling. In: Proc. of the ACL second workhop on SMT. (June 2007) 177–180
8. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Lagarda, A., Ney, H., Tomás, J., Vidal, E.: Statistical approaches to computer-assisted translation. Computational Linguistics **35**(1) (2009) 3–28
9. Alabau, V., Ortiz-Martnez, D., Sanchis, A., Casacuberta, F.: Multimodal interactive machine translation. In: Proc. of ICMI/MLMI. (2010) DOI: 10.1145/1891903.1891960.
10. Alabau, V., et. al: On multimodal interactive machine translation using speech recognition. In: Proc of ICMI. (2011) 129–136

11. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Proc. of EMNLP. (July 2004) 388–395
12. Koehn, P., et. al: Moses: open source toolkit for statistical machine translation. In: Proc. of ACL: Demo and Poster Sessions. (June 2007) 177–180
13. Koehn, P., et. al: Edinburgh system description for the 2005 iwslt speech translation evaluation. In: Proc. of IWSLT. (October 2005)
14. Stolcke, A.: SRILM – an extensible language modeling toolkit. In: Proc. of the Int. Conf. on Spoken Language Processing. (September 2002) 901–904
15. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proc. of the MT Summit X. (September 2005) 79–86
16. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-) evaluation of machine translation. In: Proc. of the ACL workshop on SMT. (June 2007) 136–158
17. Callison-Burch, C., et al.: Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In: Proc. of the ACL workshop on SMT and Metrics MATR. (July 2010) 17–53
18. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A study of translation edit rate with targeted human annotation. In: Proc. of AMTA. (August 2006) 223–231
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: A method for automatic evaluation of machine translation. In: Technical Report RC22176 (W0109-022). (2001)
20. Sanchis-Trilles, G., Casacuberta, F.: Increasing translation speed in phrase-based models via suboptimal segmentation. In: Proc. of PRIS. (June 2008)
21. González, J., Sanchis-Trilles, G., Casacuberta, F.: Learning finite state transducers using bilingual phrases. In: Proc. of CICLing. (February 2008)
22. Sanchis-Trilles, G., Ortiz-Martínez, D., González-Rubio, J., González, J.: Bilingual segmentation for phrasetable pruning in statistical machine translation. In: Proc. of EAMT. (May 2011)
23. Sanchis-Trilles, G., Cettolo, M., Bertoldi, N., Federico, M.: Online language model adaptation for spoken dialog translation. In: Proc. of IWSLT. (December 1 – 2 2009) 160–167
24. Bertoldi, N., Bisazza, A., Cettolo, M., Sanchis-Trilles, G., Federico, M.: Fbk @ iwslt 2009. In: Proc of IWSLT. (December 2009) 37–44
25. Sanchis-Trilles, G., Cettolo, M.: Online language model adaptation via n-gram mixtures for statistical machine translation. In: Proc. of EAMT. (May 2010)
26. Andrs-Ferrer, J., Sanchis-Trilles, G., Casacuberta, F.: Similarity word-sequence kernels for sentence clustering. In: Proc. of S+SSPR. (August 2010)
27. Sanchis-Trilles, G., Casacuberta, F.: Bayesian adaptation for statistical machine translation. In: Proc. of S+SSPR. (August 2010)
28. Sanchis-Trilles, G., Casacuberta, F.: Log-linear weight optimisation via bayesian adaptation in statistical machine translation. In: Proc. of COLING. (August 2010)
29. Martínez-Gómez, P., Sanchis-Trilles, G., Casacuberta, F.: Online adaptation strategies for statistical machine translation in post-editing scenarios. Pattern Recognition **45**(9) (2012) 3193–3203
30. Gascó, G., Alabau, V., Andrés-Ferrer, J., González-Rubio, J., Rocha, M.A., Sanchis-Trilles, G., Casacuberta, F., González, J., Sánchez, J.A.: Iti-upv system description for iwslt 2010. In: Proc. of IWSLT. (December 2010)
31. Sanchis-Trilles, G., González, M., Casacuberta, F., Vidal, E., Civera, J.: Introducing additional input information into imt systems. In: Proc. of MLMI. LNCS. Volume 5237. (September 2008) 284–295
32. Sanchis-Trilles, G., Ortiz-Martínez, D., Civera, J., Casacuberta, F., Vidal, E., Hoang, H.: Improving interactive machine translation via mouse actions. In: Proc. of EMNLP. (October 2008)