

MACHINE TRANSLATION SYSTEM DEVELOPMENT BASED ON HUMAN LIKENESS

Patrik Lambert*, Jesús Giménez*, Marta R. Costa-jussà*
Enrique Amigó†, Rafael E. Banchs*, Lluís Màrquez* and J.A. R. Fonollosa*

*TALP Research Center. Universitat Politècnica de Catalunya.
Jordi Girona Salgado 1–3. 08034, Barcelona.

†Departamento de Lenguajes y Sistemas Informáticos.
Universidad Nacional de Educación a Distancia. Juan del Rosal, 16. 28040, Madrid.
{lambert, mruiz, rbanchs, adrian}@gps.tsc.upc.edu
{jgimenez, lluis}@lsi.upc.edu, enrique@lsi.uned.es

ABSTRACT

We present a novel approach for parameter adjustment in Empirical Machine Translation systems. Instead of relying on a single evaluation metric, or in an ad-hoc linear combination of metrics, our method works over metric combinations with maximum descriptive power, aiming to maximise the Human Likeness of the automatic translations. We apply it to the problem of optimising decoding stage parameters of a state-of-the-art Statistical Machine Translation system. By means of a rigorous manual evaluation, we show how our methodology provides more reliable and robust system configurations than a tuning strategy based on the BLEU metric alone.

1. INTRODUCTION

Parameter adjustment is one of the most crucial issues in the development stage of a Statistical Machine Translation (SMT) system. Particularly critical is the tuning of parameters that govern the decoding (search) step. Commonly, a Minimum Error Rate iterative strategy is followed [1]. At each iteration the MT system is run over a so-called development set under a certain parameter configuration. At the end of the process, the configuration exhibiting the lowest error rate is selected to translate new text. Error rate is typically measured by comparing the system output against a set of human references, according to an evaluation metric at choice.

By far, the most widely used metric in the recent literature is BLEU, which computes lexical matching accumulated precision for n-grams up to length four [2]. However, it presents several deficiencies which cast serious doubts on its usefulness, both for sentence-level error analysis [3] and for system-level comparison [4]. Moreover, optimising over an error measure based on a single metric presents a major

drawback. The system may end strongly biased towards configurations which maximise this metric score but may not necessarily maximise the scores conferred by other metrics [5]. We refer to this problem as system *over-tuning*. Some authors have tried to overcome its negative effects by defining error measures over linear combinations of metrics [6, 7]. However, in these cases metric combinations are selected arbitrarily, or based on uncertain or ad-hoc criterion.

In our work, we suggest a tuning procedure based on a robust and stable criterion. We aim to maximise the ‘*Human Likeness*’ of automatic translations [8]. Translations are evaluated in terms of the probability that they could have been generated by a human translator, instead of the probability that they could look acceptable to human judges (‘*Human Acceptability*’). We approach this target with the QARLA Framework [9]. We apply our methodology to optimise a state-of-the-art SMT system [10]. We show, through a rigorous manual evaluation process, how tuning based on Human Likeness provides more reliable parameter configurations.

The rest of the paper is organised as follows. Section 2 describes the fundamentals of QARLA and its application to MT. In Section 3 the SMT system used is described. Experimental work is deployed in Section 4. Conclusions and further work are briefly outlined in Section 5.

2. QARLA FOR MACHINE TRANSLATION

Inside the QARLA Framework, metrics are ranked according to their descriptive power, i.e. their capability to discern between human and automatic translations [9]. Given a set of test cases A , a set of similarity metrics X , and sets of human references R , QARLA provides three measures:

- **QUEEN** $_{X,R}(A)$, a measure to evaluate the quality of a translation using a set of similarity metrics. QUEEN operates under the assumption that a good translation must be similar to all human references according to all metrics. QUEEN is defined as the probability, over

This work has been partly funded by the European Union under the integrated project TC-STAR (IST-2002-FP6-506738), the European Social Fund, and the Spanish Ministry of Science and Technology, projects R2D2 (TIC-2003-7180) and ALIADO (TIC-2002-04447-C02).

$R \times R \times R$, that for every metric in X the automatic translation $a \in A$ is closer to a reference than two other references to each other.

- **KING** $_{A,R}(X)$, a measure to evaluate the quality of a set of similarity metrics. KING represents the probability that, for a given set of human references R , and a set of metrics X , the QUEEN quality of a human reference is greater than the QUEEN quality of *any* automatic translation in A . Thus, KING accounts for the proportion of cases in which a set of metrics has been able to fully distinguish between automatic and manual translations.
- **JACK** (A, R, X) , a measure to evaluate the reliability of a test set, defined as the probability over all human references $r \in R$ of finding a couple of automatic translations a, a' which are (i) close to all human references (QUEEN > 0) and (ii) closer to r than to each other, according to all metrics. In other words, JACK measures the heterogeneity of system outputs with respect to human references. A high JACK value means that most references are closely and heterogeneously surrounded by automatic translations. Thus, it ensures that R and A are not biased.

The QARLA Framework for MT Evaluation is publicly and freely available under the name of IQ_{MT}¹ (Inside Qarla Machine Translation Evaluation Framework) [11]. IQ_{MT} provides a useful mechanism for MT evaluation based on ‘Human Likeness’ [8]. We use QARLA in two complementary manners. First, we determine the set of metrics with highest descriptive power by maximising the KING measure. Second, we use QUEEN to measure MT quality according to the optimal metric set. Furthermore, for completeness, we estimate test set reliability by means of the JACK measure.

3. TRANSLATION SYSTEM

Although the method described in this paper is valid for any Empirical MT system, we briefly present the models which constitute our system [10], and whose respective weights are tuned. The translation model is based on a 4-gram language model of bilingual units which are referred to as tuples. Tuples are extracted from Viterbi alignments and can be formally defined as the set of shortest phrases that provides a monotonic segmentation of the bilingual corpus.

In addition to the translation model, the translation system implements a log-linear combination of six additional feature functions: a 5-gram language model of the target language (denoted TM); a 5-gram language model of target POS-tags (TTM), a 5-gram language model of reordered source POS-tags (TSM), used to support a pattern-based reordering

¹<http://www.lsi.upc.edu/~nlp/IQMT>.

strategy; a word bonus feature (WB); and finally, two lexicon models (L1 and L2) that implement, for a given tuple, the IBM-1 translation probability estimate between the source and target (or target and source, respectively) sides of it.

4. EXPERIMENTAL WORK

4.1. Settings

Experiments were performed in both English-to-Spanish and Spanish-to-English translation directions. 33 variants from 7 families of metrics (BLEU, NIST, WER and PER, GTM, ROUGE, and METEOR)² were considered.

We have used the Spanish-English EPPS parallel corpus distributed under the TC-STAR OpenLab on Speech Translation³. It contains the proceedings of the European Parliament debates from 1996 to May 2005. The training set contains over 34 million running words in both languages. Table 1 shows statistics of the development and test data used.

		sent	words	vocab.	avg len
Dev. (3 refs)	English	1008	26070	3173	25.9
	Spanish		25778	3895	25.6
Test (2 refs)	English	1094	26917	3958	24.6
	Spanish		840	22774	4081

Table 1. Development and test sets statistics.

4.2. Procedure

We optimised the contribution of each feature function in the SMT system⁴ using a tool based on the Downhill Simplex method [13]. This algorithm adjusts the log-linear weights so as to maximise an objective function. Note that in this problem, only a local optimum is usually found. Tuning was performed according to two different MT quality measures, evaluated over development data: (i) BLEU and (ii) QUEEN.

To reduce the possibility of having an initial set of weights which would happen to be particularly bad for one of the two objective functions (leading to a particularly poor local optimum), optimisations were started from three initial parameter sets: 1) all free parameters are set to 1; 2) they are all set to 0; and 3) they are alternatively set to 1 and 0. Thus, for the objective function corresponding to each metric, we got three sets of final parameters. Between these three, we chose the final set which corresponded to the best local optimum in the development set.

In both cases (BLEU and QUEEN), optimal parameters were used to translate the test data, and a manual comparison of the resulting two sets of translations was performed

²A detailed list of the variants incorporated may be found in [12]

³<http://www.tc-star.org/openlab2006>

⁴In the log-linear combination, weights can be rescaled to set one of the parameters to some value, so the translation model was set to 1 and kept fixed during optimisation.

		TM	TTM	TSM	WB	L1	L2
E→S	B	0.49	0.24	0.96	1.12	0.58	0.41
	Q	0.65	0.23	1.6	1.58	0.97	0.88
S→E	B	0.38	0.22	1.0	0.9	0.76	0.4
	Q	0.31	0.25	0.72	1.9	0.25	0.76

Table 2. Final parameters obtained in Spanish-to-English (S→E) and English-to-Spanish (E→S) directions. The translation model weight is set to 1 and kept fixed. B and Q stand for system optimised respectively with BLEU and QUEEN.

by 4 different human evaluators. Each evaluator compared 150 randomly extracted translation pairs, and assessed in each case whether one system produced a better translation, or whether both were of equivalent quality. Strictly equal outputs were removed before choosing the 150 pairs. Each judge evaluated a different set of (possibly overlapping) sentences. In order to avoid any bias in the evaluation, the respective position in the display of the sentences corresponding to each system was also random.

4.3. Results

As described in Section 2, the first step deals with finding the optimal metric set, based on the KING measure optimisation. In the case of Spanish-to-English the optimal metric set is: $\{MTR_{wnsyn}, MTR_{stem}$ and $RG_{W.1.2}\}$ (KING = 0.1472), where MTR refers to METEOR and RG to ROUGE. Whereas for the English-to-Spanish the optimal metric set is: $\{MTR_{exact}, MTR_{stem}$ and $RG_{W.1.2}\}$ (KING = 0.2593). These metric sets are used to compute the QUEEN measure.

The systems optimised with BLEU and QUEEN are then compared at various levels. The final model weights obtained from tuning are indicated in Table 2. According to this table, the main characteristic of QUEEN optimisation is its tendency to favour the word bonus model with respect to the translation model and the word and POS tags target language models. Thus, the QUEEN measure rated long sentences more favourably than BLEU.

Automatic results are presented in Table 3. According to all metrics, both English-to-Spanish systems are equivalent, whereas the Spanish-to-English system optimised with BLEU achieves better translations (1.7% absolute BLEU and nearly 2% absolute WER above the other system). This was expected. After all, conventional metrics have been developed on the basis of Human Acceptability.

In order to clarify this scenario a manual evaluation has been conducted as described in Subsection 4.2. Table 4 shows, for each evaluator, the results of its manual comparison, along with the results of the comparison of the same sentences with respect to WER scores. Manual comparisons are in strong disagreement with conventional automatic evaluation metrics. For example, in negative sentences the negation was sometimes omitted by the system tuned with BLEU

		BLEU	WER	PER	MTR	RG
E→S	B	0.486	40.3	31.4	0.7004	0.3974
	Q	0.480	40.2	31.2	0.7000	0.3972
S→E	B	0.562	33.3	25.3	0.7084	0.4310
	Q	0.545	35.4	26.6	0.7154	0.4330

Table 3. Automatic translation evaluation results. MTR and RG stand respectively for METEOR and ROUGE. We had only 2 references so QUEEN was not measured (see below).

		EVAL 1		EVAL 2		EVAL 3		EVAL 4	
		H	W	H	W	H	W	H	W
E→S	B>Q	33	55	37	72	56	52	32	54
	Q>B	41	57	57	51	78	65	60	57
	B=Q	76	38	56	29	16	33	52	39
S→E	B>Q	35	79	31	83	46	91	37	85
	Q>B	41	36	52	37	36	33	46	33
	B=Q	74	35	67	30	68	26	67	32

Table 4. Number of sentences that the system optimised with BLEU has translated better (B>Q), worse (Q>B) or with equivalent quality (B=Q) as that optimised with QUEEN, according to Human Experts (H) and WER scores (W). Evaluators of translation into Spanish were different from those of translation into English.

but not by that tuned with QUEEN. The omission of a word present in *all* references implies indeed a stronger penalty in QUEEN. Table 6 shows automatic scores for the translations in Table 5. Although the translation of the system tuned with BLEU is more fluent, it has the opposite meaning as the source sentence. When the correct meaning is restored (‘B corrected’), BLEU slightly worsen, and QUEEN improves.

Evaluators have clearly considered that the English-to-Spanish system optimised by QUEEN performed better. For translation into English, human judges have globally preferred the system optimised on QUEEN, but with less contrast. The fact that QUEEN favoured longer sentences may provide an explanation, since English is denser than Spanish. As a second possible explanation, translation into Spanish is far more difficult than into English. This difficulty would benefit QUEEN. First, because in that case metrics would become more expressive, i.e. there would be more features to capture in order to distinguish automatic from human translations. Second, because the English-to-Spanish test set would exhibit a higher degree of heterogeneity. Fortunately, we may test this hypothesis by inspecting the reliability of the test sets according to the JACK measure. The JACK measure for the Spanish-to-English test set is 0.2189, whereas for English-to-Spanish the JACK value is significantly higher, 0.3122. This confirms our intuitions.

Nevertheless, notice that in both cases the level of reliability is low. This was expected. All systems are indeed different parameterisations of the same original system.

Source	Creo que no se puede pensar que lo que gana una institución lo pierde la otra .
Translation Q	I believe that it is not possible to think that what wins a institution what loses the other .
Translation B	I believe that it is possible to think that what wins a institution loses the other .
B corrected	I believe that it is not possible to think that what wins a institution loses the other .
Ref. 1	I do not believe there is any mileage in imagining that what the one institution gains , the other loses .
Ref. 2	I believe that we should not think that what one institution wins , the other loses .
Ref. 3	I think that we can not think that what an institution wins is lost by the other .

Table 5. Example from development data: source, translation of systems tuned with QUEEN (Q), BLEU (B), and references.

Translation	BLEU	WER	QUEEN
Q	0.203	48.2	0.222
B	0.213	48.2	0.056
B corrected	0.204	48.2	0.222

Table 6. Evaluation of the translations of Table 5.

Finally, we must note some limitations of IQ_{MT}: (i) at least three human references per sentence must be available for the purpose of QUEEN computation, (ii) QUEEN computations depend cubically on the size of reference sets, and linearly on the size of test and metric sets, thus in our current experimental setup there is a severe associated time overhead.

5. CONCLUSIONS AND FURTHER WORK

We have provided an effective methodology for MT system development based on ‘Human Likeness’. We have shown that this approach provides more reliable and robust system configurations than a tuning strategy based on BLEU alone. The disagreement between conventional metric scores and manual evaluation has shown one more evidence of the need for this type of methodology.

Problems caused by the minor reliability of the Spanish-to-English test set could be alleviated by enriching it with outputs by different MT systems implementing other approaches (e.g. rule-based, or word-based SMT), and by working on more sophisticated metrics which discriminate to a greater extent between human and automatic translations. We are currently working on a wider set of partial metrics working at different linguistic levels further than lexical, i.e. at the syntactic and shallow semantic levels. We are also performing similar experiments on other data sets and language pairs, such as Chinese-English and Arabic-English, in the framework of NIST evaluations.

Acknowledgements

Authors are grateful to David Vilar for providing the software for conducting human evaluation, and to Adrià de Gispert and Josep Maria Crego for participating in the evaluation itself.

6. REFERENCES

[1] F.J. Och, “Minimum error rate training in statistical machine translation,” in *ACL*, 2003, pp. 160–167.

[2] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” IBM Research Report, RC22176, 2001.

[3] Joseph P. Turian, Luke Shen, and I. Dan Melamed, “Evaluation of machine translation and its evaluation,” in *Proteus technical report 03-005*, 2003.

[4] C. Callison-Burch, M. Osborne, and P. Koehn, “Re-evaluating the role of BLEU in machine translation research,” in *EACL*, Trento, Italy, 2006, pp. 249–256.

[5] J.M. Crego, A. de Gispert, and J.B. Mariño, “The TALP Ngram-based SMT System for IWSLT’05,” in *IWSLT*, 2005, pp. 191–198.

[6] S. Hewavitharana, B. Zhao, A.S. Hildebrand, M. Eck, C. Hori, S. Vogel, and A. Waibel, “The CMU SMT System for IWSLT 2005,” in *IWSLT*, 2005, pp. 63–70.

[7] B. Chen, R. Cattoni, N. Bertoldi, M. Cettolo, and M. Federico, “The ITC-irst SMT System for IWSLT-2005,” in *IWSLT*, 2005, pp. 98–104.

[8] E. Amigó, J. Giménez, J. Gonzalo, and L. Màrquez, “MT evaluation: Human-like vs. human acceptable,” in *COLING-ACL*, Sydney, Australia, 2006, pp. 17–24.

[9] E. Amigó, J. Gonzalo, A. Peñas, and F. Verdejo, “QARLA: A framework for the evaluation of text summarization systems,” in *ACL*, Ann Arbor, Michigan, 2005, pp. 280–289.

[10] J.M. Crego, A. de Gispert, P. Lambert, M.R. Costajussà, M. Khalilov, R. Banchs, J.B. Mariño, and J.A. Fonollosa, “N-gram-based SMT system enhanced with reordering patterns,” in *HLT-NAACL Workshop on Statistical MT*, New York, USA, 2006, pp. 162–165.

[11] J. Giménez and E. Amigó, “IQMT: A framework for automatic MT evaluation,” in *LREC*, 2006, pp. 685–690.

[12] J. Giménez, E. Amigó, and C. Hori, “MT evaluation inside QARLA,” in *IWSLT*, 2005, pp. 199–206.

[13] W.H. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C++: the Art of Scientific Computing*, Cambridge University Press, 2002.