Machine Translation Evaluation Inside QARLA

Jesús Giménez † , Enrique Amigó ‡ and Chiori Hori §

† TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
‡ Departamento de Lenguajes y Sistemas Informáticos
Universidad Nacional de Educación a Distancia
§ InterACT Language Technologies Institute
Carnegie Mellon University
jgimenez@lsi.upc.edu, enrique@lsi.uned.es, chiori@cs.cmu.edu

Abstract

In this work we present the fundamentals of the IQ_{MT} framework for MT evaluation. IQ_{MT} offers a common workbench on which existing evaluation metrics can be utilized. We suggest the IQ measure and test it on the Chinese-to-English data from the IWSLT 2004 Evaluation Campaign. We show how the correlation with human assessments at the system level improves substantially for most individual metrics. Moreover, IQ_{MT} allows to robustly combine several metrics avoiding scaling problems and metric weightings. Several metric combinations were tried, but correlations did not further improve significantly.

1. Introduction

At the current level of improvement in a couple of years there will probably exist Machine Translation (MT) systems that perform better than humans according to existing MT evaluation metrics. By then, these metrics, as they are currently applied, will become useless and more sophisticated metrics will be needed (Franz Och, talk at the ACL 2005 Workshop on "Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond"). We refer to this problem as the '2008 MT Evaluation Challenge'.

In this work we present the fundamentals of IQ_{MT}[1] (Inside QARLA MT evaluation), a framework for MT evaluation which intends to overcome the MT evaluation challenge by offering a common workbench on which existing evaluation metrics can be used and combined.

Inside QARLA [1], automatic evaluation of translations is interpreted as the application of similarity metrics between a set of candidate translations and a set of reference translations. In this context, one of the main issues is to determine how similar a machine-produced translation must be to a set of human references to certify that it is a good translation. That is, how the scale properties of the similarity metrics must be interpreted.

Another important issue is how to combine information from different metrics into a single measure of quality. In the last years, it has been repeatedly argued that current MT evaluation metrics do not capture well possible improvements attained by means of incorporating linguistic knowledge [2]. One of the possible reasons for that is that most of the current metrics are based on rewarding lexical similarity, thus not taking into account any additional syntactic or semantic information. We believe that new metrics should be investigated and combined with current ones. The question then would be how to ponderate new similarity metrics with respect to existing ones.

The QARLA framework has been successfully applied to the automatic evaluation of summaries [3]. Its probabilistic model is not affected by the scale properties of individual metrics. It allows also to combine evaluation metrics in a single measure, QUEEN, such that it is non-dependent on individual metric scales. Our goal is to adapt the QUEEN measure to MT evaluation. For that purpose, we have defined the IQ (Innovated QUEEN) measure.

We have applied IQ_{MT} to the task of evaluating the IWSLT 2004 results [4]. We show how existing metrics can be used inside QARLA exhibiting higher levels of correlation with human assessments at the system level. We also worked on combinations of metrics but did not achieve any further improvement.

The rest of the paper is organized as follows. In Section 2 the QARLA framework is described. In Section 3 we discuss current trends in MT evaluation. Our approach to MT evaluation inside QARLA is described in Section 4. Experimental work is deployed in Section 5. In Section 6 we present a preliminary evaluation of the results of the IWSLT 2005. Finally, some conclusions and further work are drawn in Section 7.

[1]The IQ_{MT} package is publically available, released under the GNU Lesser General Public License (LGPL) of the Free Software Foundation, and may be freely downloaded at `http://www.lsi.upc.edu/~nlp/IQMT`.

2. The QARLA Framework

The QARLA framework was originally defined for automatic evaluation of summaries. QARLA uses similarity to models (human references) as a building block for the evaluation of automatic summarisation systems. The input for QARLA, in a summarisation task, is a set of test cases, a set of similarity metrics X, and sets of models M for each test case. With such a testbed, QARLA provides a measure QUEEN which combines assorted similarity metrics to estimate the quality of automatic summarisers.

QUEEN operates under the assumption that a good summary must be similar to all model summaries according to all metrics. QUEEN is defined as the probability, over $M \times M \times M$, that for every metric in X the automatic summary a is closer to a model than two other models to each other:

$$\text{QUEEN}_X(a, M) \equiv Prob(\forall x \in X : x(a, m) \geq x(m', m''))$$

where a is the automatic summary being evaluated, $\langle m, m', m'' \rangle$ are three models in M, and $x(a, m)$ stands for the similarity of m to a. QUEEN is stated as a probability, and therefore its range of values is $[0, 1]$.

We can think of the QUEEN measure as using a set of tests (every similarity metric in X) to test the hypothesis that a given summary a is a model. Given $\langle a, m, m', m'' \rangle$, we test $x(a, m) \geq x(m', m'')$ for each metric x. a is accepted as a model only if it passes the test for every metric. QUEEN(a) is, then, the probability of acceptance for a in the sample space $M \times M \times M$.

This measure has some interesting properties:

(i) it is able to combine different similarity metrics into a single evaluation measure;

(ii) it is not affected by the scale properties of individual metrics, i.e. it does not require metric normalisation and it is not affected by metric weighting.

(iii) Peers (automatic summaries) which are very far from the set of models all receive QUEEN=0. In other words, QUEEN does not distinguish between very poor summarisation strategies.

(iv) The value of QUEEN is maximised for peers that "merge" with the models under all metrics in X.

(v) The universal quantifier on the metric parameter x implies that adding redundant metrics does not bias the result of QUEEN.

3. The '2008 MT Evaluation Challenge'

In the last years, many efforts have been devoted to including linguistic information further than lexical units in the parameter estimation of translation models in Statistical Machine Translation.

However, to our knowledge, no significant improvement has been reported so far. An exception is the case of [5] who presented a syntax-based language model based upon that described by [6], which combined with the syntax based translation model described by [7], achieved a notable improvement in grammaticality. However, they measured this improvement by means of human evaluation.

At this point, one may argue that evaluation metrics are not well suited to capture improvements attained. Most of the existing metrics work only at the lexical level. This is the case of metrics such as BLEU [8], NIST [9], WER and PER [10], and GTM [11]. We may find some notable exceptions such as METEOR [12], ROUGE [13], and WNM [14], which consider additional information. For instance, ROUGE and METEOR consider stemming, and allow for WordNet [15] lookup . METEOR performs a synonym search in WordNet. As to WNM, this metric is a variant of BLEU which weights n-grams according to their statistical salience estimated out from a monolingual corpus.

Further than that, we may find the approach by [16] who introduce a series of syntactic features such as constituent/dependency precision and recall and head-word chain matching. They show how adding syntactic information to the evaluation metric improves both sentence-level and system-level correlation with human judgements.

In a recent work, [17] tried to combine some aspects of different metrics. They applied machine learning techniques to build a classifier that distinguished between human-generated (good) and machine-generated (bad) translations. They used features inspired in metrics like BLEU, NIST, WER and PER, obtaining higher levels of correlation with human judgements. Similarly, the IQ_{MT} framework permits metric combinations, with the singularity that there is no need to perform any training or adjustment of parameters.

4. QARLA for MT Evaluation

QUEEN operates under the assumption that there exists a set of similarity metrics which are capable of grouping models as opposite to low quality elements. That is, QUEEN assumes that a good element must be close to all models. The question is whether it is possible to find such a set of metrics in the context of translations. In a first experiment we tried to apply the original QUEEN to MT, but we did not obtain significant improvements in correlation with human judgements for most of the metrics. See details in Subsection 5.4.

A possible reason is that translations are shorter than summaries. While a summary may contain around 100 words, tipically, sentences are much shorter. For instance, we are working on translations with an average length of 8 words. Less information implies more difficulties to find metrics which characterise the properties of models. That is, models are not grouped separate from incorrect translations. This means that current MT evaluation metrics do not satisfy the QUEEN conjectures.

Therefore, we defined a new metric IQ (Innovated

QUEEN) derived from QUEEN. The first change is to assume that a good translation should be similar to just one of the models, and not necessarily to all models. Formally, if an automatic translation a is equal to one of the models, then $IQ_X(a, M)$ is maximum. For this, we consider the distance from a to the nearest model in M:

$$IQ_X(a, M) \equiv max_{m \in M}\, iq_{X,M}(a, m)$$

$$iq_{X,M}(a, m) \equiv Prob(\forall x \in X : x(a, m) \geq x(m', m''))$$

In order to estimate the similarity to one model IQ considers the distribution of distances between pairs of models (m' and m'' in the formula). However, we work under the assumption that the metrics are not capable to group all models. Moreover, we do not know which model pairs should be chosen. Therefore, we define the following criterion: "a good translation must be at least as similar to one of the models as the rest of model pairs are to each other". In order to introduce this idea into the IQ definition we universally quantify the variables m' and m'':

$$iq_{X,M}(a, m) = \begin{cases} 1 & \text{if } \forall x \in X : \forall m', m'' \in M : \\ & \quad x(a, m) \geq x(m', m'') \\ 0 & otherwise \end{cases}$$

This IQ definition satisfies the QUEEN properties described in Section 2. The main disadvantage of IQ with respect to QUEEN is that IQ considers only the similarity to the nearest model. Furthermore, it does not consider the distribution of distances between models. Therefore, IQ becomes a binary value (zero or one). That is, IQ assumes that there exist just 'correct' or 'incorrect' translations.

5. Experimental Work

5.1. Data

In order to test our approach we utilized the data and results from the IWSLT04 evaluation campaign. We focused on the evaluation of the Chinese-to-English (CE) translation task, in which a set of 500 short sentences from the Basic Travel Expressions Corpus (BTEC) [4] were translated.

For purposes of automatic evaluation, 16 reference translations and outputs by 20 different MT systems were available for each sentence. Moreover, each of these outputs was evaluated by three judges on the basis of adequacy and fluency [18].

5.2. Metric Set

We considered a set of 26 different metrics from 7 metric families:

BLEU [2] accumulated BLEU scores for several n-gram levels ($n = 1, 2, 3, 4$).

NIST [3] accumulated NIST scores for several n-gram levels ($n = 1, 2, 3, 4, 5$).

GTM [4] for several values of the e parameter ($e = 1, 2, 3$).

mWER (default).

mPER (default).

METEOR [5] We used 4 variants.

- **METEOR.exact** running "exact" module only.
- **METEOR.porter** (default) running "exact" and "porter_stem" modules, in that order.
- **METEOR.wn1** running "exact", "porter_stem" and "wn_stem" modules, in that order.
- **METEOR.wn2** running "exact", "porter_stem", "wn_stem" and "wn_synonymy" modules, in that order.

ROUGE [6] for several n-grams ($n = 1, 2, 3, 4$), and 4 other variants at the 4-gram level, always with stemming:

- **ROUGE-L** longest common subsequence (LCS).
- **ROUGE-S*** skip bigrams with no max-gap-length.
- **ROUGE-SU*** skip bigrams with no max-gap-length, including unigrams.
- **ROUGE-W** weighted longest common subsequence (WLCS) with weighting factor $w = 1.2$.

5.3. Automatic Evaluation Metrics outside QARLA

First, we studied the performance of individual metrics outside the QARLA framework. System-level scores for 5 different metrics (i.e. BLEU, NIST, mWER, mPER, and GTM) were available. Additionaly, we computed the rest of metrics described in Subsection 5.2.

Table 1 shows Pearson Correlation between individual metrics and human assessments. The first two columns, 'Adequacy' and 'Fluency', respectively refer to the correlation with adequacy and fluency outside the QARLA framework. ROUGE variants outperform the rest of metrics both in adequacy and fluency. The highest correlation in adequacy is obtained by ROUGE-S*, whereas for fluency ROUGE.n3 obtains the highest correlation. BLEU and METEOR variants achieve also high levels of correlation.

[2] We used mteval-kit-v10/mteval-v11b.pl for BLEU calculation.
[3] We used mteval-kit-v10/mteval-v11b.pl for NIST calculation.
[4] We used GTM version 1.2.
[5] We used METEOR version 0.4.3.
[6] We used ROUGE version 1.5.5. Options are `"-z SPL -2 -1 -U -m -r 1000 -n 4 -w 1.2 -c 95 -d"`.

Metric	Adequacy	Fluency	Adequacy$_Q$	Fluency$_Q$	Adequacy$_{IQ}$	Fluency$_{IQ}$
BLEU.n1	0.7623	0.6380	0.6781	0.5933	0.0529	0.0802
BLEU.n2	0.8442	0.8002	0.8770	0.8215	0.2567	0.2788
BLEU.n3	0.8449	0.8326	0.8499	0.8212	0.3923	0.4064
BLEU.n4	0.7407	0.8600	0.8569	0.8063	0.3156	0.3434
GTM.e1	0.5136	0.5214	0.6204	0.5452	0.8293	0.7715
GTM.e2	0.6784	0.6566	0.6687	0.6140	0.8015	0.8126
GTM.e3	0.7022	0.6906	0.6590	0.6094	0.7775	0.8213
METEOR.exact	0.8899	0.7463	0.7836	0.6888	0.9358	0.8593
METEOR.porter	0.8837	0.7265	0.7800	0.6706	0.9494	0.8599
METEOR.wn1	0.8784	0.7147	0.7886	0.6709	0.9420	0.8554
METEOR.wn2	0.8725	0.6923	0.7784	0.6513	0.8942	0.8094
NIST.n1	0.4077	0.2323	0.7837	0.6150	0.5124	0.4845
NIST.n2	0.5245	0.3629	0.8385	0.6934	0.7945	0.7063
NIST.n3	0.5745	0.4222	0.8421	0.7000	0.7277	0.6952
NIST.n4	0.5965	0.4497	0.8438	0.7030	0.8466	0.8136
NIST.n5	0.6820	0.5950	0.8440	0.7036	0.8768	0.8650
ROUGE.n1	0.8582	0.6590	0.9028	0.7303	0.9695	0.8876
ROUGE.n2	0.9287	0.8435	0.9238	0.8421	0.9673	0.9142
ROUGE.n3	0.9190	**0.8646**	0.9076	**0.8630**	0.9588	**0.9180**
ROUGE.n4	0.9010	0.8527	0.8756	0.8156	0.9492	0.9008
ROUGE-L	0.9153	0.7644	0.9325	0.8112	**0.9713**	0.8979
ROUGE-S*	**0.9376**	0.8164	**0.9357**	0.8119	0.9663	0.9062
ROUGE-SU*	0.9328	0.8114	0.9317	0.8096	0.9656	0.9064
ROUGE-W	0.9219	0.7737	0.8918	0.7899	0.9234	0.8503
mPER/1-PER	-0.5779	-0.6010	0.4212	0.3662	0.0242	0.0421
mWER/1-WER	-0.6427	-0.7214	0.4507	0.4209	0.0880	0.0770

Table 1: Adequacy and Fluency correlation coefficients for individual automatic evaluation metrics for the IWSLT'04 CE Supplied Data track. 'Adequacy' and 'Fluency' refer to correlation outside QARLA. 'Adequacy$_Q$' and 'Fluency$_Q$' refer to correlation inside QARLA, using the QUEEN measure. 'Adequacy$_{IQ}$' and 'Fluency$_{IQ}$' refer to correlation inside QARLA, using the IQ measure.

5.4. Automatic Evaluation Metrics inside QARLA

First, we computed the QUEEN measure based on each metric individually. See correlation results in Table 1, columns 3 and 4, 'Adequacy$_Q$' and, 'Fluency$_Q$', respectively. For most of the metrics there is no significant improvement. Only in the case of the NIST family of metrics , there is a consistent and very substantial improvement with respect both to adequacy and fluency. The highest levels of correlation are again achieved by ROUGE.n3 and ROUGE-S* metrics, but at the same degree than outside the QARLA framework. The combination of these two metrics, {ROUGE.n3, ROUGE-S*}, does not report any significant improvement.

Next, we computed IQ measure based on each metric individually. See correlation results in Table 1, columns 5 and 6, 'Adequacy$_{IQ}$' and, 'Fluency$_{IQ}$', respectively. All metrics but BLEU-based, WER and PER, obtain higher levels of correlation both with respect to adequacy and fluency when applied inside QARLA. Again, ROUGE variants attain the highest levels of correlation in adequacy and fluency. METEOR variants obtain also high levels of correlation. The highest correlation in adequacy is obtained by ROUGE-L, whereas for fluency ROUGE.n3 achieves the highest correlation. The combination of these two metrics, {ROUGE.n3, ROUGE-L}, does not report any significant improvement.

The extremely low levels of correlation attained by BLEU, WER and PER deserve further analysis. By inspecting results, we observe that these metrics generate very low IQ values. A possible explanation is that while most of the current metrics are able to exploit multiple references simultaneously, QARLA works with similarities on a single-reference basis. Each translation is contrasted with each reference independently, so there is a decrease in the reliability of automatic metric scores. The QUEEN measure is not affected because it considers the similarity to all references whereas the IQ measure considers only the similarity to the closest reference.

BLEU, WER and PER seem to be specially sensitive to this problem. BLEU looks for high precision over any of the models. We conjecture that BLEU is specially useful when it works over a set of models (multiple references), which is not the case in QARLA. Regarding WER and PER, we think that these metrics are possibly capturing non-relevant differ-

ences between translations. Thus, they are placing models too close to each other. Recall the IQ definition in Section 4. Good translations must be at least as similar to one of the models as the rest of model pairs are to each other. WER and PER are therefore obliging candidate translations to be extremely similar to one of the references in order to be considered correct.

5.5. Metric Combinations

One of the main features of QARLA is that it allows to robustly combine several evaluation metrics. We study several combinations. Due to the computational complexity of exhaustively trying all metric combinations[7] we performed a clustering as described in [3] so as to detect metrics that behave similarly. This clustering process is based on the behaviour of metrics over samples $\{a, m, m', m''\}$. We consider that two sets of metrics behave similarly if the automatic translation a is as close to the model m as m', m'' are to each other for both sets of metrics. We applied the k-means algorithm [19].

Clustering results are shown in Table 2. Very interestingly, clusters 1 to 4 group some metric variants at the same level of granularity (from 1-gram to 4-gram). WER and PER remain together in cluster 5. Clusters 6 to 9 put together several variants of METEOR, NIST, GTM, and ROUGE, respectively.

From each cluster we selected a representative based on the level of correlation between the IQ measure and human assessments, as reported in Table 1 (columns 5 and 6). Actually, a representative for adequacy and a representative for fluency were chosen. We did not use cluster 5. Therefore, we limited our exploration to 510 metric combinations, 255 for fluency and 255 for adequacy.

Table 3 and Table 4 show correlation with adequacy and fluency, respectively, for some combinations of metrics. In the case of adequacy we did not find a combination exhibiting a higher correlation than ROUGE-L alone. In the case of fluency 4 combinations outperformed ROUGE.n3, although not very significantly. The best combination is {ROUGE.n3, ROUGE-SU*}.

We suspect that the benefits of combining metrics are hidden by the very high levels of correlation already achieved by single metrics. We further discuss this problem in Section 7.

6. IQ_{MT} for IWSLT 2005

We present preliminary results on the evaluation of the Chinese-to-English Supplied Data track of the IWSLT 2005 Evaluation Campaign [20]. The test set consists of 506 very short sentences (average length of 6 words). 16 reference translations and 11 system outputs were available for each sentence. Human assessments, based on adequacy, fluency and meaning maintenance at the system level were available.

[7] There are $2^{26} - 1$ possible combinations if we take into account all metrics.

We studied the behaviour of individual metrics outside QARLA. Very high levels of correlation (over 0.95) are achieved. METEOR variants and ROUGE.n1 are the metrics that obtain the highest levels of correlation with respect to adequacy (0.98) and meaning maintenance (0.99). For fluency, BLEU.n4 and GTM.e3 obtain the highest correlation (0.95).

In spite of the very high levels of correlation already achieved outside QARLA, we tested the behaviour of these metrics inside QARLA. Levels of correlation attained for adequacy and meaning maintenance are also very high inside QARLA. NIST.n1 is the highest scoring metric for adequacy (0.98) and meaning maintenance (0.97). All metrics exhibit very high levels of correlation for adequacy (over 0.82), and meaning maintenance (over 0.85). As in the case of the IWSLT 2004, ROUGE variants obtain very competitive results.

However, for fluency, a significant drop is observed. The levels of correlation range from 0.56 to 0.85, being the highest correlation value achieved by BLEU.n4. Although most metrics correlate better with fluency inside QARLA, metrics such as BLEU.n4, GTM.e2, GTM.3 or ROUGE.n4, which reward longer matches, exhibit a substantial decrease. We suspect that our framework is not well suited to measure the fluency over translations that are so short (6 words). In fact, we argue whether it makes sense to do so. By working on very short translations we are practically forcing candidate translations to match exactly one of the references.

Finally, we tried some metric combinations. Again, due to time constraints, we performed a clustering, obtaining similar clusters to those derived from the IWSLT 2004 data. We arbitrarily explored some combinations by selecting the six most promising metrics. For adequacy and meaning maintenance we explored 63 combinations determined by the set {BLEU.n1, GTM.e1, METEOR.wn2, NIST.n1, ROUGE.n1, 1-PER}. For fluency we explored the 63 combinations in the set {BLEU.n4, GTM.e2, METEOR.exact, NIST.n5, ROUGE.n4, 1-WER}. Table 5 shows Pearson correlation values with respect to adequacy, fluency and meaning maintenance, for the best combinations. Consistently to the results on the IWSLT 2004 data, no significant improvements are reported when combining different metrics.

7. Conclusions

The most important conclusion in this work is that most individual metrics improve when they are applied inside the QARLA framework. The reason for that improvement is that IQ takes as reference similarities between models, normalising the scale of the metric regarding to the models set distribution.

We observed that improvements obtained in the case of the IWSLT 2004 are more significant than in the case of the IWSLT 2005. We believe that the sentence average length is a key factor to explain this fact.

Moreover, one of the motivations for our work was to

Cluster_id	Metrics
1	{BLEU.n1, **GTM.e1**}
2	{BLEU.n2 **ROUGE.n2**}
3	{BLEU.n3, **ROUGE.n3**}
4	{BLEU.n4, **ROUGE.n4**}
5	{1-WER, 1-PER}
6	{METEOR.exact, **METEOR.porter**, METEOR.wn1, METEOR.wn2}
7	{NIST.n1, NIST.n2, NIST.n3, NIST.n4, **NIST.n5**}
8	{**GTM.e2**, **GTM.e3**}
9	{ROUGE.n1, **ROUGE-L**, ROUGE-S*, **ROUGE-SU***, ROUGE-W}

Table 2: Clusters of Metrics. See highlighted the cluster representatives selected for the study of metric combinations.

Metric Combination	Adequacy$_{IQ}$
ROUGE-L	**0.9713**
ROUGE.n2 ROUGE-L	**0.9701**
ROUGE.n3 ROUGE-L	0.9681
ROUGE.n2 ROUGE.n3 ROUGE-L	0.9661
ROUGE.n3 ROUGE.n4 ROUGE-L	0.9621
ROUGE.n2 ROUGE.n4 ROUGE-L	0.9619
ROUGE.n2 ROUGE.n3	0.9608
ROUGE.n2 ROUGE.n3 ROUGE.n4 ROUGE-L	0.9593
ROUGE.n4 ROUGE-L	0.9584
ROUGE.n2 ROUGE.n4	0.9570
ROUGE.n2 ROUGE.n3 ROUGE.n4	0.9538
ROUGE.n3 ROUGE.n4	0.9528
ROUGE.n3 METEOR.porter ROUGE-L	0.9525
ROUGE.n3 METEOR.porter	0.9522
ROUGE.n3 ROUGE.n4 METEOR.porter ROUGE-L	0.9495

Table 3: Adequacy correlation coefficients for some combinations of automatic evaluation metrics inside the QARLA Framework, using the IQ measure, for the IWSLT'04 CE Supplied Data track.

study how to improve MT evaluation by combining different metrics. However, our results show that the correlation with human judgements does not improve when metric combinations are considered. We point some possible reasons. First, we are calculating Pearson correlations with human assessments over only 20 systems, and the levels of correlation achieved by individual metrics are already very high. With so very few samples and these high levels of correlation, one could perhaps argue that improvements are not very significant. This problem could be solved by testing correlation at the sentence level. We would then have thousands of samples. Correlations at this level would also tend to be lower.

A second reason for the lack of success in the combination of metrics is that we have used metrics that capture similar features. In future works, new metrics centered in partial features that capture linguistic aspects of translation further than lexical will be included.

Furthermore, a main drawback of the IQ measure is that it requires several reference translations, when actually in most cases a single reference is available. Others, like [21], avoid the use of references by building classifiers that learn to distinguish between human-produced and machine-produced translations. In the short term, we plan to apply IQ$_{MT}$ to other working sets so as to study its behaviour when fewer reference translations are available. That would allow us to test also our approach over longer sentences.

A final remark, IQ$_{MT}$ is not yet properly a framework because it does not allow for meta-evaluation yet. Further work involves dealing with the two other QARLA components, namely KING and JACK, which measure the quality of a set of metrics, and the quality of a test set with respect to a set of metrics, respectively.

8. Acknowledgements

This research has been funded by the Spanish Ministry of Science and Technology, projects R2D2 (TIC-2003-7180) and ALIADO (TIC-2002-04447-C02). The TALP Research Center is recognized as a Quality Research Group (2001 SGR 00254) by DURSI, the Research Department of the Catalan Government. Authors are thankful to Michael Gamon, Julio Gonzalo and Lluís Màrquez for their valuable comments and suggestions.

Metric Combination	Fluency_{IQ}
ROUGE.n3 ROUGE-SU*	**0.9251**
ROUGE.n3 ROUGE-L	0.9244
ROUGE.n2 ROUGE.n3	0.9206
ROUGE.n2 ROUGE-SU*	0.9184
ROUGE.n3	**0.9180**
ROUGE.n4 ROUGE-SU*	0.9124
ROUGE.n2 ROUGE-L	0.9121
ROUGE.n3 ROUGE.n4 ROUGE-SU*	0.9096
ROUGE.n2 ROUGE.n4	0.9090
ROUGE.n2 ROUGE.n3 ROUGE.n4	0.9056
ROUGE.n3 ROUGE.n4	0.9031
ROUGE.n3 METEOR.porter	0.8951
ROUGE.n3 METEOR.porter ROUGE-SU*	0.8916
ROUGE.n2 ROUGE.n3 METEOR.porter	0.8895
ROUGE.n3 ROUGE.n4 METEOR.porter	0.8875
ROUGE.n4 METEOR.porter ROUGE-SU*	0.8870
ROUGE.n2 METEOR.porter ROUGE-SU*	0.8841
ROUGE.n2 ROUGE.n4 METEOR.porter	0.8813

Table 4: Fluency correlation coefficients for some combinations of automatic evaluation metrics inside the QARLA Framework, using the IQ measure, for the IWSLT'04 CE Supplied Data track.

Metric Combination	Correlation
$\text{Best}_{Adequacy}$ {NIST.n1}	0.9826
$\text{Best}_{fluency}$ {BLEU.n4}	0.8549
$\text{Best}_{meaning}$ {NIST.n1 1-PER}	0.9766

Table 5: Adequacy, Fluency and Meaning Maintenance correlation coefficients for best combinations of automatic evaluation metrics inside the QARLA Framework, for the IWSLT'05 CE Supplied Data track.

9. References

[1] Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo, "QARLA: a Framework for the Evaluation of Automatic Sumarization", Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics, 2005.

[2] Franz Josef Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin and Dragomir Radev, "Final Report of the Summer Workshop on Syntax for Statistical Machine Translation", Johns Hopkins University, 2003.

[3] Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo, "Evaluating DUC 2004 with QARLA Framework", Proceedings of the ACL'05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005.

[4] Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul and Jun'ichi Tsujii, "Overview of the IWSLT04 Evaluation Campaign", Proceedings of the International Workshop on Spoken Language Translation, 2004.

[5] Eugene Charniak, Kevin Knight and Kenji Yamada, "Syntax-based Language Models for Machine Translation", Proceedings of MT SUMMIT IX, 2003.

[6] Eugene Charniak, "Immediate-Head Parsing for Language Models", Proceedings of ACL, 2001.

[7] Kenji Yamada and Kevin Knight, "A Syntax-based Statistical Translation Model", Proceedings of ACL, 2001.

[8] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation, IBM Research Report, RC22176", IBM T.J. Watson Research Center, 2001.

[9] George Doddington, "Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics", Proc. of the 2nd Internation Conference on Human Language Technology, 2002.

[10] C. Tillmann, S. Vogel, H. Ney, A. Zubiaga and H. Sawaf, "Accelerated DP based Search for Statistical

Translation", Proceedings of European Conference on Speech Communication and Technology, 1997.

[11] I. Dan Melamed, Ryan Green and Joseph P. Turian, "Precision and Recall of Machine Translation", Proceedings of HLT/NAACL, 2003.

[12] Satanjeev Banerjee and Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005.

[13] Chin-Yew Lin and Franz Josef Och, "Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics", Proceedings of ACL, 2004.

[14] Bogdan Babych and Tony Hartley, "Extending the BLEU MT Evaluation Method with Frequency Weightings", Proceedings of ACL, 2004.

[15] C. Fellbaum, "WordNet. An Electronic Lexical Database", The MIT Press, 1998.

[16] Ding Liu and Daniel Gildea, "Syntactic Features for Evaluation of Machine Translation", Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005.

[17] Alex Kulesza and Stuart M. Shieber, "A learning approach to improving sentence-level MT evaluation", Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, 2004.

[18] LDC, "Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Chinese-English Translations Revision 1.0", Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/-TIDES/Translation/TransAssess02.pdf, 2002.

[19] J. B. MacQueen, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, 1:281-297, 1967.

[20] Matthias Eck and Chiori Hori, "Overview of the IWSLT 2005 Evaluation Campaign", Proceedings of the International Workshop on Spoken Language Translation, 2005.

[21] Michael Gamon, Anthony Aue and Martine Smets, "Sentence-Level MT evaluation without reference translations: beyond language modeling", Proceedings of EAMT, 2005.