# A Smorgasbord of Features for Automatic MT Evaluation

**Jesús Giménez** and **Lluís Màrquez**
TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

## Abstract

This document describes the approach by the NLP Group at the Technical University of Catalonia (UPC-LSI), for the shared task on Automatic Evaluation of Machine Translation at the ACL 2008 Third SMT Workshop.

## 1 Introduction

Our proposal is based on a rich set of individual metrics operating at different linguistic levels: lexical (i.e., on word forms), shallow-syntactic (e.g., on word lemmas, part-of-speech tags, and base phrase chunks), syntactic (e.g., on dependency and constituency trees), shallow-semantic (e.g., on named entities and semantic roles), and semantic (e.g., on discourse representations). Although from different viewpoints, and based on different similarity assumptions, in all cases, translation quality is measured by comparing automatic translations against human references. Extensive details on the metric set may be found in the IQMT technical manual (Giménez, 2007).

Apart from individual metrics, we have also applied a simple integration scheme based on uniformly-averaged linear metric combinations (Giménez and Màrquez, 2008a).

## 2 What is new?

The main novelty, with respect to the set of metrics presented last year (Giménez and Màrquez, 2007), is the incorporation of a novel family of metrics at the properly semantic level. *DR* metrics analyze similarities between automatic and reference translations by comparing their respective discourse representation structures (DRS), as provided by the the C&C Tools (Clark and Curran, 2004). DRS are essentially a variation of first-order predicate calculus which can be seen as semantic trees. We use three different kinds of metrics:

**DR-STM** Semantic Tree Matching, a la Liu and Gildea (2005), but over DRS instead of over constituency trees.

**DR-$O_r$-$\star$** Lexical overlapping over DRS.

**DR-$O_{rp}$-$\star$** Morphosyntactic overlapping on DRS.

Further details on DR metrics can be found in (Giménez and Màrquez, 2008b).

### 2.1 Improved Sentence Level Behavior

Metrics based on deep linguistic analysis rely on automatic processors trained on out-domain data, which may be, thus, prone to error. Indeed, we found out that in many cases, metrics are unable to produce a result due to the lack of linguistic analysis. For instance, in our experiments, for SR metrics, we found that the semantic role labeler was unable to parse 14% of the sentences. In order to improve the recall of these metrics, we have designed two simple variants. Given a linguistic metric $x$, we define:

- $x_b$ $\longrightarrow$ by backing off to lexical overlapping, $O_l$, only when the linguistic processor is not able to produce a linguistic analysis. Otherwise, $x$ score is returned. Lexical scores are conveniently scaled so that they are in a similar range to scores of $x$. Specifically, we multiply



*Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198,
Columbus, June 2008. 

them by the average $x$ score attained over all other test cases for which the parser succeeded.

- $x_i$ $\longrightarrow$ by linearly interpolating $x$ and $O_l$ scores for all test cases, via the arithmetic mean.

In both cases, system scores are calculated by averaging over all sentence scores. Currently, these variants are applied only to SR and DR metrics.

### 2.2 Uniform Linear Metric Combinations

We have simulated a non-parametric combination scheme based on human acceptability by working on uniformly averaged linear combinations (ULC) of metrics (Giménez and Màrquez, 2008a). Our approach is similar to that of Liu and Gildea (2007) except that in our case the contribution of each metric to the overall score is not adjusted.

Optimal metric sets are determined by maximizing the correlation with human assessments, either at the document or sentence level. However, because exploring all possible combinations was not viable, we have used a simple algorithm which performs an approximate search. First, metrics are ranked according to their individual quality. Then, following that order, metrics are added to the optimal set only if in doing so the global quality increases.

## 3 Experimental Work

We use all into-English test beds from the 2006 and 2007 editions of the SMT workshop (Koehn and Monz, 2006; Callison-Burch et al., 2007). These include the translation of three different language-pairs: German-to-English (de-en), Spanish-to-English (es-en), and French-to-English (fr-en), over two different scenarios: in-domain (European Parliament Proceedings) and out-of-domain (News Commentary Corpus)[1]. In all cases, a single reference translation is available. In addition, human assessments on adequacy and fluency are available for a subset of systems and sentences. Each sentence has been evaluated at least by two different judges. A brief numerical description of these test beds is available in Table 1.

[1]We have not used the out-of-domain Czech-to-English test bed from the 2007 shared task because it includes only 4 systems, and only 3 of them count on human assessments.

| | **WMT 2006** | | | |
|---|---|---|---|---|
| | **in-domain** | | **out-of-domain** | |
| | 2,000 cases | | 1,064 cases | |
| | **#snt** | **#sys** | **#snt** | **#sys** |
| **de-en** | 2,281 | 10/12 | 1,444 | 10/12 |
| **es-en** | 1,852 | 11/15 | 1,008 | 11/15 |
| **fr-en** | 2,268 | 11/14 | 1,281 | 11/14 |

| | **WMT 2007** | | | |
|---|---|---|---|---|
| | **in-domain** | | **out-of-domain** | |
| | 2,000 cases | | 2,007 cases | |
| | **#snt** | **#sys** | **#snt** | **#sys** |
| **de-en** | 956 | 7/8 | 947 | 5/6 |
| **es-en** | 812 | 8/10 | 675 | 7/9 |
| **fr-en** | 624 | 7/8 | 741 | 7/7 |

Table 1: Test bed description. '#snt' columns show the number of sentences assessed (considering all systems). '#sys' columns shows the number of systems counting on human assessments with respect to the total number of systems which participated in each task.

Metrics are evaluated in terms of human acceptability, i.e., according to their ability to capture the degree of acceptability to humans of automatic translations. We measure human acceptability by computing Pearson correlation coefficients between automatic metric scores and human assessments of translation quality both at document and sentence level. We use the sum of adequacy and fluency to simulate a global assessment of quality. Assessments from different judges over the same test case are averaged into a single score.

### 3.1 Individual Performance

In first place, we study the behavior of individual metrics. Table 2 shows meta-evaluation results, over into-English WMT 2007 test beds, in-domain and out-of-domain, both at the system and sentence levels, for a set of selected representatives from several linguistic levels.

At the system level (columns 1-6), corroborating previous findings by Giménez and Màrquez (2007), highest levels of correlation are attained by metrics based on deep linguistic analysis (either syntactic or semantic). In particular, two kinds of metrics, respectively based on head-word chain matching over grammatical categories and relations ('DP-

| Level | Metric | System Level | | | | | | Sentence Level | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | de-en | | es-en | | fr-en | | de-en | | es-en | | fr-en | |
| | | in | out | in | out | in | out | in | out | in | out | in | out |
| **Lexical** | 1-TER | 0.64 | 0.41 | 0.83 | 0.58 | 0.72 | 0.47 | 0.43 | 0.29 | 0.23 | 0.23 | 0.29 | 0.20 |
| | BLEU | 0.87 | 0.76 | 0.88 | 0.70 | 0.74 | 0.54 | 0.46 | 0.27 | 0.33 | 0.20 | 0.20 | 0.12 |
| | GTM ($e = 2$) | 0.82 | 0.69 | 0.93 | 0.71 | 0.76 | 0.60 | **0.56** | 0.36 | **0.43** | 0.33 | 0.27 | 0.18 |
| | ROUGE$_W$ | 0.87 | 0.91 | 0.96 | 0.78 | 0.85 | 0.83 | **0.58** | **0.40** | **0.43** | **0.35** | 0.30 | **0.31** |
| | METEOR$_{wn}$ | 0.83 | 0.92 | 0.96 | 0.74 | 0.91 | 0.86 | 0.53 | **0.41** | 0.35 | 0.28 | **0.33** | **0.32** |
| | $O_l$ | 0.79 | 0.75 | 0.91 | 0.55 | 0.81 | 0.66 | 0.48 | 0.33 | 0.35 | 0.30 | 0.30 | 0.21 |
| **Syntactic** | CP-$O_c$-$\star$ | 0.84 | 0.88 | 0.95 | 0.62 | 0.84 | 0.76 | 0.49 | 0.37 | 0.38 | 0.33 | **0.32** | 0.25 |
| | DP-HWC$_w$-4 | 0.85 | 0.93 | 0.96 | 0.68 | 0.84 | 0.80 | 0.31 | 0.26 | 0.33 | 0.07 | 0.10 | 0.14 |
| | DP-HWC$_c$-4 | **0.91** | **0.98** | 0.96 | **0.90** | **0.98** | **0.95** | 0.30 | 0.25 | 0.23 | 0.06 | 0.13 | 0.12 |
| | DP-HWC$_r$-4 | 0.89 | **0.97** | 0.97 | **0.92** | **0.97** | **0.95** | 0.33 | 0.28 | 0.29 | 0.08 | 0.16 | 0.16 |
| | DP-$O_r$-$\star$ | 0.88 | 0.96 | 0.97 | 0.84 | 0.89 | 0.89 | **0.57** | **0.41** | **0.44** | **0.36** | **0.33** | **0.30** |
| | CP-STM-4 | 0.88 | **0.97** | 0.97 | 0.79 | 0.89 | 0.89 | 0.49 | 0.39 | 0.40 | **0.37** | **0.32** | 0.26 |
| **Shallow Semantic** | NE-$M_e$-$\star$ | *-0.13* | 0.79 | 0.95 | 0.68 | 0.87 | 0.92 | *-0.03* | *0.07* | *0.07* | *-0.05* | *0.05* | *0.06* |
| | NE-$O_e$-$\star\star$ | *-0.18* | 0.78 | 0.95 | 0.58 | 0.81 | 0.71 | 0.32 | 0.26 | 0.37 | 0.26 | 0.31 | 0.20 |
| | SR-$O_r$-$\star$ | 0.55 | 0.96 | 0.94 | 0.69 | 0.89 | 0.85 | 0.26 | 0.14 | 0.30 | 0.11 | 0.08 | 0.19 |
| | SR-$O_r$-$\star_b$ | 0.24 | **0.98** | 0.94 | 0.68 | 0.92 | 0.87 | 0.33 | 0.21 | 0.35 | 0.15 | 0.18 | 0.24 |
| | SR-$O_r$-$\star_i$ | 0.51 | 0.95 | 0.93 | 0.67 | 0.88 | 0.83 | 0.37 | 0.26 | 0.38 | 0.19 | 0.24 | 0.27 |
| | SR-$M_r$-$\star$ | 0.38 | 0.95 | 0.96 | 0.83 | 0.79 | 0.75 | 0.32 | 0.18 | 0.28 | 0.18 | 0.08 | 0.14 |
| | SR-$M_r$-$\star_b$ | 0.14 | **0.98** | 0.97 | 0.82 | 0.84 | 0.79 | 0.37 | 0.23 | 0.32 | 0.21 | 0.15 | 0.17 |
| | SR-$M_r$-$\star_i$ | 0.38 | 0.94 | 0.96 | 0.80 | 0.79 | 0.74 | 0.40 | 0.27 | 0.36 | 0.24 | 0.20 | 0.20 |
| | SR-$O_r$ | 0.73 | **0.99** | 0.94 | 0.66 | 0.97 | 0.93 | 0.12 | 0.09 | 0.16 | 0.07 | -0.04 | 0.17 |
| | SR-$O_{ri}$ | 0.66 | **0.99** | 0.94 | 0.64 | **0.95** | 0.89 | 0.29 | 0.25 | 0.29 | 0.19 | 0.15 | 0.28 |
| **Semantic** | DR-$O_r$-$\star$ | 0.87 | 0.89 | 0.96 | 0.71 | 0.78 | 0.75 | 0.50 | **0.40** | 0.37 | **0.35** | 0.27 | 0.28 |
| | DR-$O_r$-$\star_b$ | 0.91 | 0.93 | 0.97 | 0.72 | 0.83 | 0.80 | 0.52 | **0.41** | 0.38 | **0.34** | 0.28 | 0.27 |
| | DR-$O_r$-$\star_i$ | 0.87 | 0.87 | 0.96 | 0.68 | 0.79 | 0.74 | 0.53 | **0.42** | 0.39 | **0.35** | 0.30 | 0.28 |
| | DR-$O_{rp}$-$\star$ | **0.92** | **0.98** | **0.99** | 0.81 | 0.91 | 0.89 | 0.42 | 0.32 | 0.29 | 0.25 | 0.21 | **0.30** |
| | DR-$O_{rp}$-$\star_b$ | **0.93** | **0.98** | **0.99** | 0.81 | **0.94** | **0.91** | 0.45 | 0.34 | 0.32 | 0.22 | 0.22 | **0.30** |
| | DR-$O_{rp}$-$\star_i$ | **0.91** | 0.95 | **0.98** | 0.75 | 0.89 | 0.85 | 0.50 | 0.38 | 0.36 | 0.28 | 0.27 | **0.33** |
| | DR-STM-4 | 0.89 | 0.95 | **0.98** | 0.79 | 0.85 | 0.87 | 0.28 | 0.29 | 0.25 | 0.21 | 0.15 | 0.22 |
| | DR-STM-4$_b$ | **0.92** | **0.97** | **0.98** | 0.80 | 0.90 | **0.91** | 0.36 | 0.31 | 0.29 | 0.21 | 0.19 | 0.23 |
| | DR-STM-4$_i$ | 0.91 | 0.94 | 0.97 | 0.74 | 0.87 | 0.86 | 0.43 | 0.35 | 0.34 | 0.26 | 0.24 | 0.27 |
| **ULC** | Optimal$_{07}$ | 0.93 | 1.00 | 0.99 | 0.92 | 0.98 | 0.95 | 0.60 | 0.46 | 0.47 | 0.42 | 0.36 | 0.39 |
| | Optimal$_{06}$ | *0.01* | 0.95 | 0.96 | 0.75 | 0.97 | 0.87 | 0.50 | 0.41 | 0.40 | 0.20 | 0.27 | 0.30 |
| | Optimal$\star_{07}$ | 0.93 | 0.98 | 0.99 | 0.81 | 0.94 | 0.91 | 0.58 | 0.45 | 0.46 | 0.39 | 0.35 | 0.34 |
| | Optimal$\star_{06}$ | *0.34* | 0.96 | 0.98 | 0.82 | 0.92 | 0.93 | 0.54 | 0.41 | 0.42 | 0.32 | 0.32 | 0.34 |
| | Optimal$_h$ | 0.87 | 0.98 | 0.97 | 0.79 | 0.91 | 0.89 | 0.56 | 0.44 | 0.43 | 0.32 | 0.31 | 0.35 |

Table 2: Meta-evaluation results based on human acceptability for the WMT 2007 into-English translation tasks

HWC$_c$-4', 'DP-HWC$_r$-4'), and morphosyntactic overlapping over discourse representations ('DR-$O_{rp}$-$\star$'), are consistently among the top-scoring in all test beds. At the lexical level, variants of ROUGE and METEOR attain the best results, close to the performance of syntactic and semantic features. It can also be observed that metrics based on semantic roles and named entities have serious troubles with the German-to-English in-domain test bed (column 1).

At the sentence level, the highest levels of correlation are attained by metrics based on lexical similarity alone, only rivaled by lexical overlapping over dependency relations ('DP-$O_r$-$\star$') and discourse representations ('DR-$O_r$-$\star$'). We speculate the underlying cause might be on the side of parsing errors. In that respect, lexical back-off strategies report in all cases a significant improvement.

It can also be observed that, over these test beds, metrics based on named entities are completely useless at the sentence level, at least in isolation. The reason is that they capture a very partial aspect of quality which may be not relevant in many cases. This has been verified by computing the 'NE-$O_e$-$\star\star$' variant which considers also lexical overlapping over regular items. Observe how this metric attains a much higher correlation with human assessments.

### 3.2 Metric Combinations

We also study the behavior of metric combinations under the ULC scheme. Last 5 rows in Table 2 shows meta-evaluation results following 3 different optimization strategies:

**Optimal:** the metric set is optimized for each test bed (language-pair and domain) individually.

**Optimal$\star$:** the metric set is optimized over the union of all test beds.

**Optimal$_h$:** the metric set is heuristically defined so as to include several of the top-scoring representatives from each level: Optimal$_h$ = { ROUGE$_W$, METEOR$_{wnsyn}$, DP-HWC$_c$-4, DP-HWC$_r$-4, DP-$O_r$-$\star$, CP-STM-4, SR-$M_r$-$\star_i$, SR-$O_r$-$\star_i$, SR-$O_{ri}$, DR-$O_r$-$\star_i$, DR-$O_{rp}$-$\star_b$ }.

We present results optimizing over the 2006 and 2007 data sets. Let us provide, as an illustration, Optimal$\star_{07}$ sets. For instance, at the system level, no combination improved the isolated global performance of the 'DR-$O_{rp}$-$\star_b$' metric (R=0.94). In contrast, at the sentence level, the optimal metric set contains several metrics from each linguistic level: Optimal$\star_{07}$ = { ROUGE$_W$, DP-$O_r$-$\star$, CP-STM-4, SR-$O_r$-$\star_i$, SR-$M_r$-$\star_i$, DR-$O_r$-$\star_i$ }. A similar pattern is observed for all test beds, both at the system and sentence levels, although with different metrics.

The behavior of optimal metric sets is in general quite stable, except for the German-to-English in-domain test bed which presents an anomalous behavior when meta-evaluating WMT 2006 optimal metric sets at the system level. The reason for this anomaly is in the 'NE-$M_e$-$\star$' metric, which is included in the 2006 optimal set: { 'NE-$M_e$-$\star$', 'SR-$O_{ri}$' }. 'NE-$M_e$-$\star$' is based on lexical matching over named entities, and attains in the 2006 German-to-English in-domain test bed a very high correlation of 0.95 with human assessments. This partial aspect of quality seems to be of marginal importance in the 2007 test bed. We have verified this hypothesis by computing optimal metrics sets without considering NE variants. Correlation increases to more reasonable values (e.g., from 0.01 to 0.66 and from 0.34 to 0.91. This result suggests that more robust metric combination schemes should be pursued.

For future work, we plan to apply parametric combination schemes based on human likeness classifiers, as suggested by Kulesza and Shieber (2004). We must also further investigate the impact of parsing errors on the performance of linguistic metrics.

## Acknowledgments

This research has been funded by the Spanish Ministry of Education and Science (OpenMT, TIN2006-15307-C03-02). Our group is recognized by DURSI as a Quality Research Group (2005 SGR-00130).

## References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2007. (Meta-) Evaluation of Machine Translation. In *Proceedings of the ACL Second SMT Workshop*, pages 136–158.

Stephen Clark and James R. Curran. 2004. Parsing the WSJ using CCG and Log-Linear Models. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 104–111.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the ACL Second SMT Workshop*, pages 256–264.

Jesús Giménez and Lluís Màrquez. 2008a. Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. In *Proceedings of IJCNLP*, pages 319–326.

Jesús Giménez and Lluís Màrquez. 2008b. On the Robustness of Linguistic Features for Automatic MT Evaluation. To be published.

Jesús Giménez. 2007. IQMT v 2.1. Technical Manual (LSI-07-29-R). Technical report, TALP Research Center. LSI Department. http://www.lsi.upc.edu/ nlp/IQMT/IQMT.v2.1.pdf.

Philipp Koehn and Christof Monz. 2006. Manual and Automatic Evaluation of Machine Translation between European Languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121.

Alex Kulesza and Stuart M. Shieber. 2004. A learning approach to improving sentence-level MT evaluation. In *Proceedings of the 10th TMI*, pages 75–84.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Ding Liu and Daniel Gildea. 2007. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of NAACL*, pages 41–48.