On the Robustness of Syntactic and Semantic 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

Linguistic metrics based on syntactic and semantic information have proven very effective for Automatic MT Evaluation. However, no results have been presented so far on their performance when applied to heavily ill-formed low quality translations. In order to glean some light into this issue, in this work we present an empirical study on the behavior of a heterogeneous set of metrics based on linguistic analysis in the paradigmatic case of speech translation between non-related languages. Corroborating previous findings, we have verified that metrics based on deep linguistic analysis exhibit a very robust and stable behavior at the system level. However, these metrics suffer a significant decrease at the sentence level. This is in many cases attributable to a loss of recall, due to parsing errors or to a lack of parsing at all, which may be partially ameliorated by backing off to lexical similarity.

1 Introduction

Recently, there is a growing interest in the development of automatic evaluation metrics which exploit linguistic knowledge at the syntactic and semantic levels. For instance, we may find metrics which compute similarities over shallow syntactic structures/sequences (Giménez and Màrquez, 2007; Popovic and Ney, 2007), constituency trees (Liu and Gildea, 2005) and dependency trees (Liu and Gildea, 2005; Amigó et al., 2006; Mehay and Brew, 2007; Owczarzak et al., 2007). We may also find metrics operating over shallow semantic structures, such as named entities and semantic roles (Giménez and Màrquez, 2007).

Linguistic metrics have been proven to produce more reliable system rankings than metrics limiting their scope to the lexical dimension, in particular when applied to test beds with a rich system typology, i.e., test beds in which there are automatic outputs produced by systems based on different paradigms, e.g., statistical, rule-based and human-aided (Giménez and Màrquez, 2007). The reason is that they are able to capture deep MT quality distinctions which occur beyond the shallow level of lexical similarities.

However, these metrics have the limitation of relying on automatic linguistic processors, tools which are not equally available for all languages and whose performance may vary depending on the type of analysis conducted and the application domain. Thus, it could be argued that linguistic metrics should suffer a significant quality drop when applied to a different translation domain, or to ill-formed sentences. Clearly, metric scores computed on partial or wrong syntactic/semantic structures will be less informed. But, should this necessarily lead to less reliable evaluations? In this work, we have analyzed this issue by conducting a contrastive empirical study on the behavior of a heterogeneous set of metrics over several evaluation scenarios of decreasing translation quality. In particular, we have studied the case of Chinese-to-English speech translation, which is a paradigmatic example of low quality and heavily ill-formed output.

The rest of the paper is organized as follows. In Section 2, prior to presenting experimental work, we describe the set of metrics employed in our experiments. We also introduce a novel family of metrics which operate at the properly semantic level by analyzing similarities over discourse representations. Experimental work is then presented in Section 3. Metrics are evaluated both in terms of human likeness and human acceptability (Amigó et al., 2006). Finally, in Section 4, main conclusions are summarized and future work is outlined.

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 250–258,
Athens, Greece, 30 March – 31 March 2009. ©2009 Association for Computational Linguistics

2 A Heterogeneous Metric Set

We have used a heterogeneous set of metrics selected out from the metric repository provided with the IQ_{MT} evaluation package (Giménez and Màrquez, 2007)[1]. We have considered several metric representatives from different linguistic levels (lexical, syntactic and semantic). A brief description of the metric set is available in Appendix A.

In addition, taking advantage of newly available semantic processors, we have designed a novel family of metrics based on the Discourse Representation Theory, a theoretical framework offering a representation language for the examination of contextually dependent meaning in discourse (Kamp, 1981). A discourse is represented in a discourse representation structure (DRS), which is essentially a variation of first-order predicate calculus —its forms are pairs of first-order formulae and the free variables that occur in them.

2.1 Exploiting Semantic Similarity for Automatic MT Evaluation

'DR' metrics analyze similarities between automatic and reference translations by comparing their respective DRSs. These are automatically obtained using the C&C Tools (Clark and Curran, 2004)[2]. Sentences are first parsed on the basis of a combinatory categorial grammar (Bos et al., 2004). Then, the BOXER component (Bos, 2005) extracts DRSs. As an illustration, Figure 1 shows the DRS representation for the sentence *"Every man loves Mary."*. The reader may find the output of the BOXER component (top) together with the equivalent first-order formula (bottom).

DRS may be viewed as semantic trees, which are built through the application of two types of DRS conditions:

basic conditions: one-place properties (predicates), two-place properties (relations), named entities, time-expressions, cardinal expressions and equalities.

complex conditions: disjunction, implication, negation, question, and propositional attitude operations.

Three kinds of metrics have been defined:

DR-STM-*l* (Semantic Tree Matching) These metrics are similar to the *Syntactic Tree Matching* metric defined by Liu and Gildea (2005), in this case applied to DRSs instead of constituency trees. All semantic subpaths in the candidate and the reference trees are retrieved. The fraction of matching subpaths of a given length, $l \in [1..9]$, is computed. Then, average accumulated scores up to a given length are retrieved. For instance, 'DR-STM-4' corresponds to the average accumulated proportion of matching subpaths up to length-4.

DR-$\boldsymbol{O_r}$-$\boldsymbol{t}$ These metrics compute lexical overlapping[3] between discourse representation structures (i.e., discourse referents and discourse conditions) according to their type 't'. For instance, 'DR-$\boldsymbol{O_r}$-pred' roughly reflects lexical overlapping between the referents associated to predicates (i.e., one-place properties), whereas 'DR-$\boldsymbol{O_r}$-imp' reflects lexical overlapping between referents associated to implication conditions. We also introduce the **'DR-$\boldsymbol{O_r}$-$\star$'** metric, which computes average lexical overlapping over all DRS types.

DR-$\boldsymbol{O_{rp}}$-$\boldsymbol{t}$ These metrics compute morphosyntactic overlapping (i.e., between parts of speech associated to lexical items) between discourse representation structures of the same type t. We also define the **'DR-$\boldsymbol{O_{rp}}$-$\star$'** metric, which computes average morphosyntactic overlapping over all DRS types.

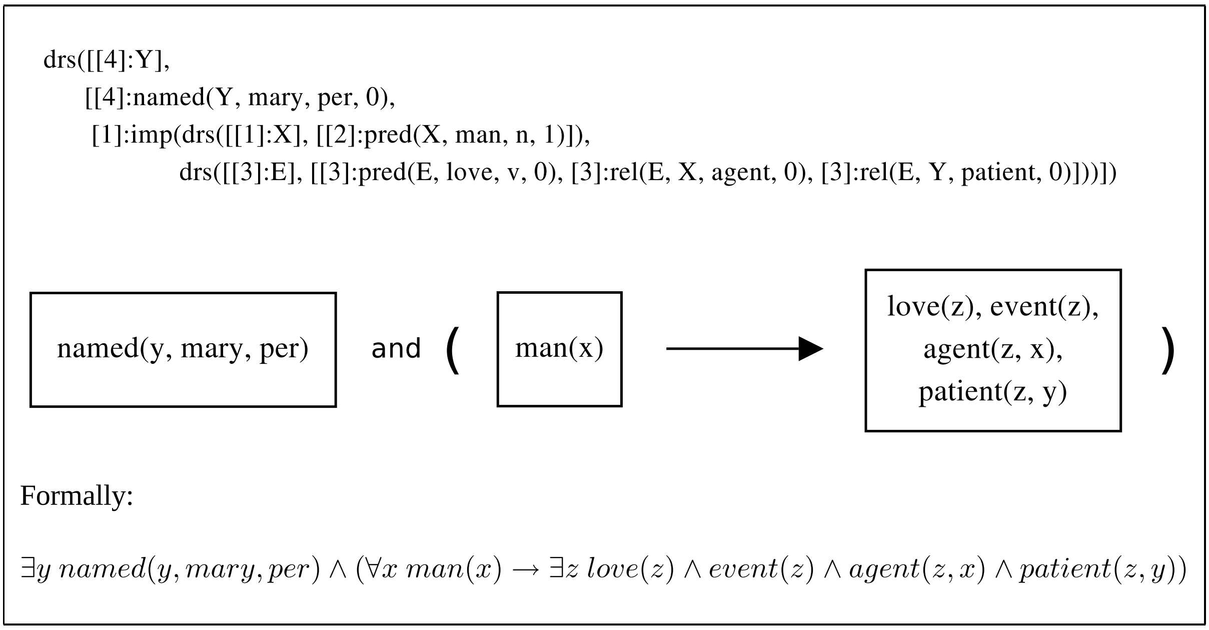
Note that in the case of some complex conditions, such as implication or question, the respective order of the associated referents in the tree is important. We take this aspect into account by making order information explicit in the construction of the semantic tree. We also make explicit the type, symbol, value and date of conditions when these are applicable (e.g., predicates, relations, named entities, time expressions, cardinal expressions, or anaphoric conditions).

Finally, the extension to the evaluation setting based on multiple references is computed by assigning the maximum score attained against each individual reference.

[1] http://www.lsi.upc.edu/~nlp/IQMT

[2] http://svn.ask.it.usyd.edu.au/trac/candc

[3] Overlapping is measured following the formulae and definitions by Giménez and Màrquez (2007). A short definition may be found in Appendix A.



```
drs([[4]:Y],
    [[4]:named(Y, mary, per, 0),
     [1]:imp(drs([[1]:X], [[2]:pred(X, man, n, 1)]),
             drs([[3]:E], [[3]:pred(E, love, v, 0), [3]:rel(E, X, agent, 0), [3]:rel(E, Y, patient, 0)]))])
```

Formally:

$$\exists y\ named(y, mary, per) \wedge (\forall x\ man(x) \rightarrow \exists z\ love(z) \wedge event(z) \wedge agent(z, x) \wedge patient(z, y))$$

Figure 1: DRS representation for *"Every man loves Mary."*

3 Experimental Work

In this section, we present an empirical study on the behavior of a heterogeneous set of metrics based on linguistic analysis in the case of speech translation between non-related languages.

3.1 Evaluation Scenarios

We have used the test bed from the Chinese-to-English translation task at the *"2006 Evaluation Campaign on Spoken Language Translation"* (Paul, 2006)[4]. The test set comprises 500 translation test cases corresponding to simple conversations (question/answer scenario) in the travel domain. In addition, there are 3 different evaluation subscenarios of increasing translation difficulty, according to the translation source:

CRR: Translation of correct recognition results (as produced by human transcribers).

ASR read: Translation of automatic read speech recognition results.

ASR spont: Translation of automatic spontaneous speech recognition results.

For the purpose of automatic evaluation, 7 human reference translations and automatic outputs by 14 different MT systems for each evaluation subscenario are available. In addition, we count on the results of a process of manual evaluation. For each subscenario, 400 test cases from 6 different system outputs were evaluated, by three human assessors each, in terms of adequacy and fluency on a 1-5 scale (LDC, 2005). A brief numerical description of these test beds is available in Table 1. It includes the number of human references and system outputs available, as well as the number of sentences per output, and the number of system outputs and sentences per system assessed. For the sake of completeness, we report the performance of the Automatic Speech Recognition (ASR) system, in terms of accuracy, over the source Chinese utterances, both at the word and sentence levels. Also, in order to give an idea of the translation quality exhibited by automatic systems, average adequacy and fluency scores are also provided.

3.2 Meta-Evaluation

Our experiment requires a mechanism for evaluating the quality of evaluation metrics, i.e., a meta-evaluation criterion. The two most prominent are:

- *Human Acceptability:* Metrics are evaluated in terms of their ability to capture the degree of acceptability to humans of automatic translations, i.e., their ability to emulate human assessors. The underlying assumption is that *good* translations should be acceptable to human evaluators. Human acceptability is usually measured on the basis of *correlation* between automatic metric scores and human assessments of translation quality.

[4] `http://www.slc.atr.jp/IWSLT2006/`

	CRR	ASR read	ASR spont
#human-references	7	7	7
#system-outputs	14	14	13
#sentences	500	500	500
#outputs$_{\text{assessed}}$	6	6	6
#sentences$_{\text{assessed}}$	400	400	400
Word Recognition Accuracy	—	0.74	0.68
Sentence Recognition Accuracy	—	0.23	0.17
Average Adequacy	1.40	1.02	0.93
Average Fluency	1.16	0.98	0.98

Table 1: IWSLT 2006 MT Evaluation Campaign. Chinese-to-English test bed description

- *Human Likeness:* Metrics are evaluated in terms of their ability to capture the features which distinguish human from automatic translations. The underlying assumption is that *good* translations should resemble human translations. Human likeness is usually measured on the basis of *discriminative power* (Lin and Och, 2004b; Amigó et al., 2005).

In this work, metrics are evaluated both in terms of human acceptability and human likeness. In the case of human acceptability, metric quality is measured on the basis of correlation with human assessments both at the sentence and document (i.e., system) levels. We compute Pearson correlation coefficients. The sum of adequacy and fluency is used as a global measure of quality. Assessments from different judges have been averaged.

In the case of human likeness, we use the probabilistic KING measure defined inside the QARLA Framework (Amigó et al., 2005). KING represents the probability, estimated over the set of test cases, that the score attained by a human reference is equal or greater than the score attained by *any* automatic translation. Although KING computations do not require human assessments, for the sake of comparison, we have limited to the set of test cases counting on human assessments.

3.3 Results

Table 2 presents meta-evaluation results for a set of metric representatives from different linguistic levels over the three subscenarios defined ('CRR', 'ASR read' and 'ASR spont'). Highest scores in each column have been highlighted. Lowest scores appear in italics.

System-level Behavior

At the system level (R_{sys}, columns 7-9), the highest quality is in general attained by metrics based on deep linguistic analysis, either syntactic or semantic. Among lexical metrics, the highest correlation is attained by BLEU and the variant of GTM rewarding longer matchings (e = 2).

As to the impact of sentence ill-formedness, while most metrics at the lexical level suffer a significant variation across the three subscenarios, the performance of metrics at deeper linguistic levels is in general quite stable. However, in the case of the translation of automatically recognized spontaneous speech (ASR spont) we have found that the 'SR-O_r-$\star$' and 'SR-M_r-$\star$' metrics, respectively based on lexical overlapping and matching over semantic roles, suffer a very significant decrease far below the performance of most lexical metrics. Although 'SR-O_r-$\star$' has performed well on other test beds (Giménez and Màrquez, 2007), its low performance over the BTEC data suggests that it is not fully portable across all kind of evaluation scenarios.

Finally, it is highly remarkable the degree of robustness exhibited by semantic metrics introduced in Section 2.1. In particular, the metric variants based on lexical and morphosyntactic overlapping over discourse representations ('DR-O_r-$\star$' and 'DR-O_{rp}-$\star$', respectively), obtain a high system-level correlation with human assessments across the three subscenarios.

Sentence-level Behavior

At the sentence level (KING and R_{snt}, columns 1-6), highest quality is attained in most cases by metrics based on lexical matching. This result was expected since all MT systems are statistical and the test set is in-domain, that is it belongs to the

		Human Likeness			**Human Acceptability**					
		KING			R_{snt}			R_{sys}		
Level	**Metric**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**
Lexical	**1-WER**	0.63	0.69	0.71	0.47	0.50	0.48	0.50	*0.32*	0.52
	1-PER	0.71	0.79	0.79	0.44	0.48	0.45	0.67	*0.39*	0.60
	1-TER	0.69	0.75	0.77	0.49	0.52	0.50	0.66	*0.36*	0.62
	BLEU	0.69	0.72	0.73	0.54	0.53	0.52	0.79	0.74	0.62
	NIST	0.79	0.84	0.85	0.53	0.54	0.53	*0.12*	*0.26*	*-0.02*
	GTM ($e = 1$)	0.75	0.81	0.83	0.50	0.52	0.52	*0.35*	*0.10*	*-0.09*
	GTM ($e = 2$)	0.72	0.78	0.79	**0.62**	**0.64**	**0.61**	0.78	0.65	0.62
	METEOR$_{wnsyn}$	**0.81**	**0.86**	**0.86**	0.44	0.50	0.48	0.55	*0.39*	*0.08*
	ROUGE$_{W\_1.2}$	0.74	0.79	0.81	0.58	0.60	0.58	0.53	0.69	0.43
	O_l	0.74	0.81	0.82	0.57	0.62	0.58	0.77	0.51	0.34
Shallow Syntactic	**SP-O_p-$\star$**	0.75	0.80	0.82	0.54	0.59	0.56	0.77	0.54	0.48
	SP-O_c-$\star$	0.74	0.81	0.82	0.54	0.59	0.55	0.82	0.52	0.49
	SP-NIST$_l$	0.79	0.84	0.85	0.52	0.53	0.52	*0.10*	*0.25*	*-0.03*
	SP-NIST$_p$	0.74	0.78	0.80	0.44	0.42	0.43	*-0.02*	*0.24*	*0.04*
	SP-NIST$_{iob}$	0.65	0.69	0.70	*0.33*	*0.32*	*0.35*	*-0.09*	*0.17*	*-0.09*
	SP-NIST$_c$	*0.55*	*0.59*	*0.59*	*0.24*	*0.22*	*0.25*	*-0.07*	*0.19*	*0.08*
Syntactic	**CP-O_p-$\star$**	0.75	0.81	0.82	0.57	0.63	0.59	**0.84**	0.67	0.52
	CP-O_c-$\star$	0.74	0.80	0.82	**0.60**	**0.64**	**0.61**	0.71	0.53	0.43
	DP-O_l-$\star$	0.68	0.75	0.76	0.48	0.50	0.50	**0.84**	0.77	0.67
	DP-O_c-$\star$	0.71	0.76	0.77	0.41	0.46	0.43	0.76	0.65	0.71
	DP-O_r-$\star$	0.75	0.80	0.81	0.51	0.53	0.51	0.81	0.75	0.62
	DP-HWC$_w$	*0.54*	*0.57*	*0.57*	*0.29*	*0.32*	*0.28*	0.73	0.74	0.37
	DP-HWC$_c$	*0.48*	*0.51*	*0.52*	*0.17*	*0.18*	*0.22*	0.73	0.64	0.67
	DP-HWC$_r$	*0.44*	*0.49*	*0.48*	*0.20*	*0.21*	*0.25*	0.71	0.58	0.56
	CP-STM	0.71	0.77	0.80	0.53	0.56	0.54	0.65	0.58	0.47
Shallow Semantic	**SR-M_r-$\star$**	*0.40*	*0.43*	*0.45*	*0.29*	*0.28*	*0.29*	0.52	0.60	*0.20*
	SR-O_r-$\star$	*0.45*	*0.49*	*0.51*	*0.35*	*0.35*	*0.36*	0.56	0.58	*0.14*
	SR-O_r	*0.31*	*0.33*	*0.35*	*0.16*	*0.15*	*0.18*	0.68	0.73	0.53
	SR-M_{rv}-$\star$	*0.38*	*0.41*	*0.42*	*0.33*	*0.34*	*0.34*	0.79	**0.81**	0.42
	SR-O_{rv}-$\star$	*0.40*	*0.44*	*0.45*	*0.36*	*0.38*	*0.38*	0.64	0.72	0.72
	SR-O_{rv}	*0.36*	*0.40*	*0.40*	*0.27*	*0.31*	*0.29*	*0.34*	0.78	0.38
Semantic	**DR-O_r-$\star$**	0.67	0.73	0.75	0.48	0.53	0.50	**0.86**	0.74	0.77
	DR-O_{rp}-$\star$	0.59	0.64	0.65	0.34	0.35	0.33	**0.84**	0.78	**0.95**
	DR-STM	0.58	0.63	0.65	*0.23*	*0.26*	*0.26*	0.75	0.62	0.67

Table 2: Meta-evaluation results for a set of metric representatives from different linguistic levels

same domain in which systems have been trained. Therefore, translation outputs have a strong tendency to share the sublanguage (i.e., word selection and word ordering) represented by the predefined set of human reference translations.

Metrics based on lexical overlapping and matching over shallow syntactic categories and syntactic structures (‘SP-O_p-$\star$’, ‘SP-O_c-$\star$’, ‘CP-O_p-$\star$’, ‘CP-O_c-$\star$’, ‘DP-O_l-$\star$’, ‘DP-O_c-$\star$’, and ‘DP-O_r-$\star$’) perform similarly to lexical metrics. However, computing NIST scores over base phrase chunk sequences (‘SP-NIST$_{iob}$’, ‘SP-NIST$_c$’) is not as effective. Metrics based on head-word chain matching (‘DP-HWC$_w$’, ‘DP-HWC$_c$’, ‘DP-HWC$_r$’) suffer also a significant decrease. Interestingly, the metric based on syntactic tree matching (‘CP-STM’) performed well in all scenarios.

Metrics at the shallow semantic level suffer also a severe drop in performance. Particularly significant is the case case of the ‘SR-O_r’ metric, which does not consider any lexical information. Interestingly, the ‘SR-O_{rv}’ variant, which only differs in that it distinguishes between SRs associated to different verbs, performs slightly better.

At the semantic level, metrics based on lexical and morphosyntactic overlapping over discourse representations (‘DR-O_r-$\star$’ and ‘DR-O_{rp}-$\star$’) suffer only a minor decrease, whereas semantic tree matching (‘DR-STM’) reports as a specially bad predictor of human acceptability (R_{snt}).

However, the most remarkable result, in relation to the goal of this work, is that the behavior of syntactic and semantic metrics across the three evaluation subscenarios is, in general, quite stable —the three values in each subrow are in a very similar range. Therefore, answering the question posed in the introduction, *sentence ill-formedness is not a limiting factor in the performance of linguistic metrics*.

		Human Likeness			Human Acceptability					
		KING			R_{snt}			R_{sys}		
Level	**Metric**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**
Lexical	**NIST**	0.79	0.84	0.85	0.53	0.54	0.53	*0.12*	*0.26*	*-0.02*
	GTM ($e = 2$)	0.72	0.78	0.79	**0.62**	**0.64**	**0.61**	0.78	0.65	0.62
	METEOR$_{wnsyn}$	0.81	**0.86**	**0.86**	0.44	0.50	0.48	0.55	*0.39*	*0.08*
	O_l	0.74	0.81	0.82	0.57	0.62	0.58	0.77	0.51	0.34
Syntactic	**CP-O_p-$\star$**	0.75	0.81	0.82	0.57	0.63	0.59	0.84	0.67	0.52
	CP-O_c-$\star$	0.74	0.80	0.82	**0.60**	**0.64**	**0.61**	0.71	0.53	0.43
	DP-O_l-$\star$	0.68	0.75	0.76	0.48	0.50	0.50	0.84	0.77	0.67
Shallow Semantic	**SR-M_r-$\star$**	*0.40*	*0.43*	*0.45*	*0.29*	*0.28*	*0.29*	0.52	0.60	*0.20*
	SR-M_r-$\star_b$	0.68	0.72	0.73	0.31	0.30	0.31	0.52	0.60	0.20
	SR-M_r-$\star_i$	**0.84**	**0.86**	**0.88**	0.34	0.34	0.34	0.56	0.63	0.25
	SR-O_r-$\star$	*0.45*	*0.49*	*0.51*	*0.35*	*0.35*	*0.36*	0.56	0.58	*0.14*
	SR-O_r-$\star_b$	0.71	0.75	0.78	0.38	0.38	0.38	0.56	0.58	0.14
	SR-O_r-$\star_i$	**0.84**	**0.88**	**0.89**	0.41	0.41	0.41	0.62	0.60	0.22
	SR-O_r	*0.31*	*0.33*	*0.35*	*0.16*	*0.15*	*0.18*	0.68	0.73	0.53
	SR-O_{rb}	0.54	0.58	0.60	0.19	0.18	0.20	0.68	0.73	0.53
	SR-O_{ri}	0.72	0.77	0.79	0.26	0.26	0.27	0.80	0.73	0.67
	SR-M_{rv}-$\star$	*0.38*	*0.41*	*0.42*	*0.33*	*0.34*	*0.34*	0.79	0.81	0.42
	SR-M_{rv}-$\star_b$	0.70	0.73	0.74	0.34	0.35	0.34	0.79	0.81	0.42
	SR-M_{rv}-$\star_i$	**0.88**	**0.90**	**0.92**	0.36	0.38	0.37	0.81	**0.82**	0.45
	SR-O_{rv}-$\star$	*0.40*	*0.44*	*0.45*	*0.36*	*0.38*	*0.38*	0.64	0.72	0.72
	SR-O_{rv}-$\star_b$	0.72	0.76	0.77	0.38	0.40	0.39	0.64	0.72	0.72
	SR-O_{rv}-$\star_i$	**0.88**	**0.90**	**0.91**	0.40	0.42	0.41	0.69	0.74	0.74
	SR-O_{rv}	*0.36*	*0.40*	*0.40*	*0.27*	*0.31*	*0.29*	*0.34*	0.78	0.38
	SR-O_{rvb}	0.66	0.70	0.71	0.29	0.32	0.30	0.34	0.78	0.38
	SR-O_{rvi}	**0.83**	**0.86**	**0.88**	0.33	0.36	0.33	0.49	**0.82**	0.56
Semantic	**DR-O_r-$\star$**	0.67	0.73	0.75	0.48	0.53	0.50	0.86	0.74	0.77
	DR-O_r-$\star_b$	0.69	0.75	0.77	0.50	0.53	0.50	**0.90**	0.69	0.56
	DR-O_r-$\star_i$	**0.83**	**0.87**	**0.89**	0.53	0.57	0.53	**0.88**	0.70	0.61
	DR-O_{rp}-$\star$	0.59	0.64	0.65	0.34	0.35	0.33	0.84	0.78	**0.95**
	DR-O_{rp}-$\star_b$	0.61	0.65	0.67	0.35	0.36	0.34	0.86	0.71	0.57
	DR-O_{rp}-$\star_i$	**0.80**	**0.84**	**0.85**	0.43	0.46	0.43	**0.90**	0.75	0.70
	DR-STM	0.58	0.63	0.65	0.23	0.26	0.26	0.75	0.62	0.67
	DR-STM-b	0.64	0.68	0.71	0.23	0.26	0.27	0.75	0.62	0.67
	DR-STM-i	**0.83**	**0.87**	**0.87**	0.33	0.36	0.36	0.84	0.63	0.66

Table 3: Meta-evaluation results. Improved sentence-level evaluation of SR and DR metrics

Improved Sentence-level Behavior

By inspecting particular instances, we have found that linguistic metrics are, in many cases, unable to produce any evaluation result. The number of unscored sentences is particularly significant in the case of SR metrics. For instance, the 'SR-O_r-$\star$' metric is unable to confer an evaluation score in 57% of the cases. Several reasons explain this fact. The first and most important is that linguistic metrics rely on automatic processors trained on out-of-domain data, which are, thus, prone to error. Second, we argue that the test bed itself does not allow for fully exploiting the capabilities of these metrics. Apart from being based on a reduced vocabulary (2,346 distinct words), test cases consist mostly of very short segments (14.64 words on average), which in their turn consist of even shorter sentences (8.55 words on average)[5].

A possible solution could be to back off to a measure of lexical similarity in those cases in which linguistic processors are unable to produce any linguistic analysis. This should significantly increase their recall. With that purpose, we have designed two new variants for each of these metrics. Given a linguistic metric x, we define:

- $\boldsymbol{x_b}$ → by backing off to lexical overlapping, O_l, only when the linguistic processor was not able to produce a parsing. Lexical scores are conveniently scaled so that they are in a similar range to x scores. Specifically, we multiply them by the average x score attained over all other test cases for which the parser succeeded. Formally, given a test case t belonging to a set of test cases T:

$$x_b(t) = \begin{cases} x(t) & \text{if t} \in ok(T) \\ O_l(t)\frac{\sum_{j \in ok(T)} x(j)}{|ok(T)|} & \text{otherwise} \end{cases}$$

[5] Vocabulary size and segment/sentence average lengths have been computed over the set of reference translations.

where $ok(T)$ is the subset of test cases in T which were successfully parsed.

- $\boldsymbol{x_i}$ $\longrightarrow$ by linearly interpolating x and O_l scores for all test cases, via arithmetic mean:

$$x_i(t) = \frac{x(t) + O_l(t)}{2}$$

In both cases, system-level scores are calculated by averaging over all sentence-level scores.

Table 3 shows meta-evaluation results on the performance of these variants for several representatives from the SR and DR families. For the sake of comparison, we also show the scores attained by the base versions, and by some of the top-scoring metrics from other linguistic levels.

The first observation is that in all cases the new variants outperform their respective base metric, being linear interpolation the best alternative. The increase is particularly significant in terms of human likeness. New variants even outperform lexical metrics, including the O_l metric, which suggests that, in spite of its simplicity, this is a valid combination scheme. However, in terms of human acceptability, the gain is only moderate, and still their performance is far from top-scoring metrics.

Sentence-level improvements are also reflected at the system level, although to a lesser extent. Interestingly, in the case of the translation of automatically recognized spontaneous speech (ASR spont, column 9), mixing with lexical overlapping improves the low-performance 'SR-O_r' and 'SR-O_{rv}' metrics, at the same time that it causes a significant drop in the high-performance 'DR-O_r' and 'DR-O_{rp}' metrics. Still, the performance of linguistic metrics at the sentence level is under the performance of lexical metrics. This is not surprising. After all, apart from relying on automatic processors, linguistic metrics focus on very partial aspects of quality. However, since they operate at complementary quality dimensions, their scores are suitable for being combined.

4 Conclusions and Future Work

We have presented an empirical study on the robustness of a heterogeneous set of metrics operating at different linguistic levels for the particular case of Chinese-to-English speech translation of basic travel expressions. As an additional contribution, we have presented a novel family of metrics which operate at the semantic level by analyzing discourse representations.

Corroborating previous findings by Giménez and Màrquez (2007), results at the system level, show that metrics guided by deeper linguistic knowledge, either syntactic or semantic, are, in general, more effective and stable than metrics which limit their scope to the lexical dimension.

However, at the sentence level, results indicate that metrics based on deep linguistic analysis are not as reliable overall quality estimators as lexical metrics, at least when applied to low quality translations, as it is the case. This behavior is mainly attributable a drop in recall due to parsing errors. By inspecting particular sentences we have observed that in many cases these metrics are unable to produce any result. In that respect, we have showed that backing off to lexical similarity is a valid and effective strategy so as to improve the performance of these metrics.

But the most remarkable result, in relation to the goal of this work, is that syntactic and semantic metrics exhibit a very robust behavior across the three evaluation subscenarios of decreasing translation quality analyzed. Therefore, sentence ill-formedness is not a limiting factor in the performance of linguistic metrics. The quality drop, when moving from the system to the sentence level, seems, thus, more related to a shift in the application domain.

For future work, we are currently studying the possibility of further improving the sentence-level behavior of present evaluation methods by combining the outcomes of metrics at different linguistic levels into a single measure of quality (citation omitted for the sake of anonymity).

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, project OpenMT (TIN2006-15307-C03-02). Our NLP group has been recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government. We are grateful to the SLT Evaluation Campaign organizers and participants for providing such valuable test beds.

References

Enrique Amigó, Julio Gonzalo, Anselmo Pe nas, and Felisa Verdejo. 2005. QARLA: a Framework for the Evaluation of Automatic Sumarization. In *Pro-*

ceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL).

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez. 2006. MT Evaluation: Human-Like vs. Human Acceptable. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.

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

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. 2004. Wide-Coverage Semantic Representations from a CCG Parser. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 1240–1246.

Johan Bos. 2005. Towards Wide-Coverage Semantic Interpretation. In *Proceedings of the Sixth International Workshop on Computational Semantics*, pages 42–53.

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.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Proceedings of the 2nd Internation Conference on Human Language Technology*, pages 138–145.

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

Hans Kamp. 1981. A Theory of Truth and Semantic Representation. In J.A.G. Groenendijk, T.M.V. Janssen, , and M.B.J. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322, Amsterdam. Mathematisch Centrum.

LDC. 2005. Linguistic Data Annotation Specification: Assessment of Adequacy and Fluency in Translations. Revision 1.5. Technical report, Linguistic Data Consortium. http://www.ldc.upenn.edu/-Projects/TIDES/Translation/TransAssess04.pdf.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

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*.

Dennis Mehay and Chris Brew. 2007. BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-Based Automatic Evaluation for Machine Translation. In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87.

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

Michael Paul. 2006. Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15.

Maja Popovic and Hermann Ney. 2007. Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic, June. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of AMTA*, pages 223–231.

C. Tillmann, S. Vogel, H. Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*.

A Metric Set

Metrics are grouped according to the linguistic dimension at which they operate:

- **Lexical Similarity**

  WER (Nießen et al., 2000).

  PER (Tillmann et al., 1997).

  BLEU (Papineni et al., 2001).

  NIST (Doddington, 2002).

  GTM (Melamed et al., 2003).

  ROUGE (Lin and Och, 2004a).

  METEOR. (Banerjee and Lavie, 2005).

  TER (Snover et al., 2006).

  $\boldsymbol{O_l}$ (Giménez and Màrquez, 2007). O_l is a short name for lexical overlapping. Automatic and reference translations are considered as unordered sets of lexical items. O_l is computed as the cardinality of the intersection of the two sets divided into the cardinality of their union.

- **Shallow Syntactic Similarity (SP)**

  SP-$\boldsymbol{O_p}$-$\star$. Average lexical overlapping over parts-of-speech.

  SP-$\boldsymbol{O_c}$-$\star$. Average lexical overlapping over base phrase chunk types.

  SP-NIST. NIST score over sequences of:

  SP-NIST$_l$ Lemmas.

  SP-NIST$_p$ Parts-of-speech.

  SP-NIST$_c$ Base phrase chunks.

  SP-NIST$_{iob}$ Chunk IOB labels.

- **Syntactic Similarity**

  On Dependency Parsing (DP)

  DP-HWC Head-word chain matching (HWCM), as presented by Liu and Gildea (2005), but slightly modified so as to consider different head-word chain types:

  DP-HWC$_w$ words.

  DP-HWC$_c$ categories.

  DP-HWC$_r$ relations.

  In all cases only chains up to length 4 are considered.

  DP-$\boldsymbol{O_l|O_c|O_r}$ These metrics correspond exactly to the LEVEL, GRAM and TREE metrics introduced by Amigó et al. (2006):

  DP-$\boldsymbol{O_l}$-$\star$ Average overlapping between words hanging at the same level of the tree.

  DP-$\boldsymbol{O_c}$-$\star$ Average overlapping between words assigned the same grammatical category.

  DP-$\boldsymbol{O_r}$-$\star$ Average overlapping between words ruled by the same type of grammatical relations.

  On Constituency Parsing (CP)

  CP-STM Syntactic tree matching (STM), as presented by Liu and Gildea (2005), i.e., limited up to length-4 subtrees.

  CP-$\boldsymbol{O_p}$-$\star$ Average lexical overlapping over parts-of-speech, similarly to 'SP-O_p-$\star$', except that parts-of-speech are now consistent with the full parsing.

  CP-$\boldsymbol{O_c}$-$\star$ Average lexical overlapping over phrase constituents. The difference between this metric and 'SP-O_c-$\star$' is in the phrase scope. In contrast to base phrase chunks, constituents allow for phrase embedding and overlapping.

- **Shallow-Semantic Similarity**

  On Semantic Roles (SR)

  SR-$\boldsymbol{O_r}$-$\star$ Average lexical overlapping between SRs of the same type.

  SR-$\boldsymbol{M_r}$-$\star$ Average lexical matching between SRs of the same type.

  SR-$\boldsymbol{O_r}$ Overlapping between semantic roles independently from their lexical realization.

  We also consider a more restrictive version of these metrics (**'SR-M_{rv}-$\star$'**, **'SR-O_{rv}-$\star$'**, and **'SR-O_{rv}'**), which require SRs to be associated to the same verb.

- **Semantic Similarity**

  On Discourse Representations (DR)

  DR-STM Average semantic tree matching considering semantic subtrees up to length 4.

  DR-$\boldsymbol{O_r}$-$\star$ Average lexical overlapping between DRSs of the same type.

  DR-$\boldsymbol{O_{rp}}$-$\star$ Average morphosyntactic overlapping between DRSs of the same type.