Towards Heterogeneous Automatic MT Error Analysis

Jesús Giménez and Lluís Màrquez

TALP Research Center, LSI Department
Universitat Politécnica de Catalunya
Jordi Girona Salgado 1–3. 08034, Barcelona
{jgimenez,lluism}@lsi.upc.edu

Abstract

This work studies the viability of performing heterogeneous automatic MT error analyses. Error analysis is, undoubtly, one of the most crucial stages in the development cycle of an MT system. However, often not enough attention is paid to this process. The reason is that performing an accurate error analysis requires intensive human labor. In order to speed up the error analysis process, we suggest partially automatizing it by having automatic evaluation metrics play a more active role. For that purpose, we have compiled a large and heterogeneous set of features at different linguistic levels and at different levels of granularity. Through a practical case study, we show how these features provide an effective means of ellaborating interpretable and detailed automatic reports of translation quality.

1. Introduction

Error analysis plays a very important role in the development cycle of current MT systems. In each loop of the cycle, prior to suggesting new improvement mechanisms, system developers must first have a clear idea of the kind of errors their system commits. However, performing an accurate error analysis is a slow and delicate process which requires intensive human labor. Part of the effort is devoted to high-level analysis which involves a precise knowledge of the architecture of the system under development, but there is also a heavily time-consuming low-level part of the process related to the linguistic analysis of translation quality, which we believe that could be partially automatized.

Our proposal consists in having automatic evaluation metrics play a more active role in this part of the work. In our opinion, in the current error analysis scheme, evaluation metrics are only minimally exploited. They are used as quantitative measures, i.e., so as to identify low/high quality translations, but not as genuine qualitative measures which allow developers to automatically obtain detailed linguistic interpretations of the translation quality attained. This limited usage of automatic metrics for error analysis is a direct consequence of the shallow similarity assumptions commonly utilized for metric development. Until very recently, most metrics were based only on lexical similarity. However, in the last few years, there have been several approaches based on similarity assumptions at deeper linguistic levels. For instance, we may find syntax-based metrics (Liu and Gildea, 2005; Amigó et al., 2006; Mehay and Brew, 2007; Owczarzak et al., 2007), which compute similarities over dependency or constituency trees, metrics at the level of shallow-semantics, e.g., over semantic roles and named entities (Giménez and Màrquez, 2007), and metrics at the properly semantic level, e.g., over discourse representations (Giménez and Màrquez, 2008).

We suggest taking advantage of this recent progress achieved in automatic MT evaluation so as to conduct *heterogeneous* MT error analyses, i.e., analyses which take into account features at different linguistic levels (e.g., lexical, syntactic, and semantic), and at different levels of granularity —from very fine aspects of quality, related to how well certain linguistic structures are transferred, to coarser ones, related to how well the translation under evaluation complies with the expected overall lexical/syntactic/semantic reference structure.

Relying on a rich variety of automatic partial quality features would allow developers to analyze the performance of their systems with respect to different quality dimensions and from different viewpoints, and, consequently, to have a more precise idea of what quality aspects require improvement. Besides, in this manner, they would be allowed to concentrate on high-level decisions.

The work described in this paper is ongoing. We have compiled a large and heterogeneous set of partial features based on which we are currently conducting semiautomatic processes of error analysis over several test beds. Preliminary results show that heterogeneous error analyses provide an effective means of elaborating interpretable and detailed reports both at the sentence and document levels. In addition, all the metrics used in this work have been made publicly available inside the IQ_{MT} framework for MT evaluation (Jesús Giménez, 2007)[1].

2. A Heterogeneous Metric Set

For our experiments, we have compiled a rich set of metric variants at 5 different linguistic levels (lexical, shallow-syntactic, syntactic, shallow-semantic and semantic). Although from different viewpoints, and based on different similarity assumptions, in all cases, translation quality is measured by comparing automatic translations against a set of human reference translations. Below, we provide only a brief description. Extensive details may be found in the IQ_{MT} technical manual (Jesús Giménez, 2007).

2.1. Lexical Similarity

WER. Word Error Rate (Nießen et al., 2000).

PER. Position-independent Word Error Rate (Tillmann et al., 1997).

BLEU. Precision oriented (Papineni et al., 2001).

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

NIST. Modified BLEU (Doddington, 2002).

GTM. F-measure (Melamed et al., 2003).

ROUGE. Recall oriented (Lin and Och, 2004a).

METEOR. F-measure based on unigram alignment (Banerjee and Lavie, 2005).

TER. Translation Edit Rate (Snover et al., 2006).

2.2. Shallow Syntactic Similarity

- **On Shallow Parsing (SP)**

  SP-O_{p}-t Lexical overlapping according to the part-of-speech 't'. For instance, 'SP-O_p-NN' roughly reflects the proportion of correctly translated singular nouns. We also use a coarser metric, 'SP-O_p-$\star$' which computes the average lexical overlapping over all parts-of-speech.

  SP-O_{c}-t Lexical overlapping according to the base phrase chunk type 't'. For instance, 'SP-O_c-NP' roughly reflects the proportion of successfully translated noun phrases. We also use the 'SP-O_c-$\star$' metric, which computes the average lexical overlapping over all chunk types.

  At a more abstract level, we also use the NIST metric to compute accumulated/individual scores over sequences of:

  SP-NIST(i)$_{\mathrm{l}}$-n Lemmas.

  SP-NIST(i)$_{\mathrm{p}}$-n Parts-of-speech.

  SP-NIST(i)$_{\mathrm{c}}$-n Base phrase chunks.

  SP-NIST(i)$_{\mathrm{iob}}$-n Chunk IOB labels[2]

2.3. Syntactic Similarity

- **On Dependency Parsing (DP)**

  DP-HWC(i)-l These metrics correspond to variants of the head-word chain matching (HWCM) metric presented by Liu and Gildea (2005) slightly modified so as to consider different head-word chain types:

  DP-HWC(i)$_{\mathrm{w}}$-l words.

  DP-HWC(i)$_{\mathrm{c}}$-l grammatical categories.

  DP-HWC(i)$_{\mathrm{r}}$-l grammatical relations.

  DP-$O_{\mathrm{l}}|O_{\mathrm{c}}|O_{\mathrm{r}}$ These metrics correspond exactly to the LEVEL, GRAM and TREE metrics introduced by Amigó et al. (2006).

  DP-O_{l}-l Overlapping between words hanging at level 'l', or deeper.

  DP-O_{c}-t Overlapping between words *directly hanging* from terminal nodes (i.e. grammatical categories) of type 't'.

  DP-O_{r}-t Overlapping between words ruled by non-terminal nodes (i.e. grammatical relationships) of type 't'.

  Node types are determined by grammatical categories and relationships as defined by the dependency parser. For instance, 'DP-O_r-s' reflects lexical overlapping between subtrees of type 's' (subject). Additionally, we consider three coarser metrics, ('DP-O_l-$\star$', 'DP-O_c-$\star$' and 'DP-O_r-$\star$') which correspond to the uniformly averaged values over all levels, categories, and relationships, respectively.

- **On Constituency Parsing (CP)**

  CP-STM(i)-l These metrics correspond to variants of the syntactic tree matching (STM) metric presented by Liu and Gildea (2005).

  CP-O_{p}-t Similarly to the 'SP-O_p-t' metrics, these metrics compute lexical overlapping according to the part-of-speech 't'.

  CP-O_{c}-t These metrics compute lexical overlapping according to the phrase constituent type 't'. The difference between these metrics and 'SP-O_c-t' variants is in the phrase scope. In contrast to base phrase chunks, constituents allow for phrase embedding and overlapping.

2.4. Shallow-Semantic Similarity

- **On Named Entities (NE)**

  NE-O_{e}-t Lexical overlapping between NEs according to their type t. For instance, 'NE-O_e-PER' reflects lexical overlapping between NEs of type 'PER' (i.e., person), which provides a rough estimate of the successfully translated proportion of person names. We also use the 'NE-O_e-$\star$' metric, which considers average lexical overlapping over all NE types. This metric focus only on actual NEs. We use also another variant, 'NE-O_e-$\star\star$', which includes overlapping among items of type 'O' (i.e., Not-a-NE).

  NE-M_{e}-t Lexical matching between NEs according to their type t. For instance, 'NE-M_e-LOC' reflects the proportion of fully translated locations. The 'NE-M_e-$\star$' metric considers average lexical matching over all NE types, excluding type 'O'.

- **On Semantic Roles (SR)**

  SR-O_{r}-t Lexical overlapping between SRs according to their type t. For instance, 'SR-O_r-A0' reflects lexical overlapping between 'A0' arguments. 'SR-O_r-$\star$' considers the average lexical overlapping over all SR types.

  SR-M_{r}-t Lexical matching between SRs according to their type t. For instance, the metric 'SR-M_r-AM-MOD' reflects the proportion of fully translated modal adjuncts. The 'SR-M_r-$\star$' metric considers the average lexical matching over all SR types.

[2] IOB labels are used to denote the position (Inside, Outside, or Beginning of a chunk) and, if applicable, the type of chunk.

SR-O_{r} This metric reflects 'role overlapping', i.e., overlapping between semantic roles independently from their lexical realization.

We also use more restrictive versions of these metrics ('SR-M_{rv}-t', 'SR-O_{rv}-t', and 'SR-O_{rv}'), which require SRs to be associated to the same verb.

2.5. Semantic Similarity

- **On Discourse Representations (DR)**

  We have designed a family of metrics operating over semantic trees (Giménez and Màrquez, 2008). These trees are based on the Discourse Representation Theory (Kamp, 1981). Three kinds of metrics are defined:

  DR-STM(i)-l These metrics are similar to the *'CP-STM'* variants discussed above, in this case applied to DR structures instead of constituency trees.

  DR-O_{r}-t These metrics compute lexical overlapping between discourse representations structures (i.e., discourse referents and discourse conditions) according to their type 't'. For instance, 'DR-O_r-pred' roughly reflects lexical overlapping between the referents associated to predicates (i.e., one-place properties), We also use the 'DR-O_r-$\star$' metric, which computes average lexical overlapping over all DRS types.

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

3. Metrics at Work

In this section, we show how the rich set of metrics described in the previous section may be applied to different types of error analysis.

3.1. Types of Error Analysis

Error analyses may be classified, from the perspective of the system developer, according to two different criteria. First, according to the level of abstraction:

- **document-level analysis,** i.e., over a representative set of test cases. Such type of analysis allows developers to quantify the overall system performance. For that reason, it is often also referred to as analysis at the system level.
- **sentence-level analysis,** i.e., over individual test cases. This type of analysis allows developers to identify translation problems over particular instances.

Second, according to the evaluation referent:

- **isolated analysis,** i.e., with no referent other than human translations. This type of analysis allows developers to evaluate the individual performance of their MT system, independently from other MT systems.

#human-references	5
#system-outputs	7
#system-outputs-assessed	6
#sentences	1,056
#sentences-assessed per-system	266

Table 1: Test bed description

	A	F	A+F	BLEU
LinearB	3.69	3.66	7.35	0.47
Best SMT	3.15	2.69	5.84	0.51

Table 2: The 'LinearB vs. SMT' puzzle on BLEU

- **contrastive analysis,** i.e., on the performance of MT systems in comparison to other MT systems. This type of analysis is crucial for the MT research community so as to advance together, by allowing system developers to borrow successful mechanisms from each other.

3.2. Data Set

We have applied our approach to several evaluation test beds from different MT evaluation campaigns. In the following, we exemplify the application of heterogeneous MT error analyses through the case of the Arabic-to-English exercise from the 2005 NIST MT evaluation campaign (Le and Przybocki, 2005). This test bed presents the particularity of providing automatic translations produced by *heterogeneous* MT systems (i.e., systems belonging to different paradigms). Specifically, all systems are statistical except one, *LinearB*, which is human-aided. A brief numerical description of this test bed is available in Table 1.

This data set was used by Callison-Burch et al. (2006) to discuss the strong tendency of BLEU to favor statistical systems. Table 2 illustrates this fact by showing overall scores for LinearB and the best statistical system according to BLEU and human assessments of adequacy (A) and fluency (F). Indeed, we found out that all n-gram based metrics exhibit a similar behavior, and that metrics operating at deeper linguistic levels are able to produce more reliable system rankings (Giménez and Màrquez, 2007). This result evinces that, in order to perform a rigorous error analysis of heterogeneous systems, several quality dimensions must be taken into account. Therefore, this case constitutes an excellent material in order to test the applicability of our approach. For that purpose, we have focused on the automatic outputs by LinearB and the best statistical system we had access to (from now on referred to as 'Best SMT'). We have performed isolated and contrastive analyses, both at the document and sentence levels.

3.3. Error Analysis at the Document Level

In first place, we analyze MT quality at the document level Assisted by the heterogeneous metric set, we study system performance over a number of partial aspects of quality. Table 3 shows evaluation results for several representatives from each linguistic level. Metrics are grouped according to the linguistic level at which they operate.

It can be observed (columns 2-3) that, as we progress from the lexical level to deeper linguistic aspects, the difference

Level	Metric	KING	Linear B	Best SMT
Lexical	1-PER	0.63	0.65	**0.70**
	1-TER	**0.70**	0.53	**0.58**
	1-WER	0.67	0.49	**0.54**
	BLEU	0.65	0.47	**0.51**
	GTM (e=2)	0.66	0.31	**0.32**
	NIST	0.69	10.63	**11.27**
	ROUGE$_W$	0.68	0.31	**0.33**
	METEOR$_{wnsyn}$	0.68	0.64	**0.68**
Shallow Syntactic	SP-O$_p$-⋆	0.64	0.52	**0.55**
	SP-O$_p$-J	0.26	0.53	**0.59**
	SP-O$_p$-N	0.53	0.57	**0.63**
	SP-O$_p$-V	0.43	0.39	**0.41**
	SP-O$_c$-⋆	0.63	0.54	**0.57**
	SP-O$_c$-NP	0.60	0.59	**0.63**
	SP-O$_c$-PP	0.38	0.63	**0.66**
	SP-O$_c$-VP	0.41	0.49	**0.51**
	SP-NIST$_l$-5	0.69	10.78	**11.44**
	SP-NIST$_p$-5	**0.71**	8.74	**9.04**
	SP-NIST$_{iob}$-5	0.65	6.81	6.91
	SP-NIST$_c$-5	0.57	6.13	6.18
Syntactic	DP-HWC$_w$-4	0.59	0.14	0.14
	DP-HWC$_c$-4	0.48	**0.42**	0.41
	DP-HWC$_r$-4	0.52	**0.33**	0.31
	DP-O$_l$-⋆	0.58	0.41	**0.43**
	DP-O$_c$-⋆	0.60	0.50	**0.51**
	DP-O$_c$-aux	0.14	**0.56**	0.54
	DP-O$_c$-det	0.35	**0.75**	0.73
	DP-O$_r$-⋆	**0.66**	0.36	0.36
	DP-O$_r$-fc	0.21	**0.26**	0.24
	DP-O$_r$-i	0.60	**0.44**	0.43
	DP-O$_r$-obj	0.43	**0.36**	0.35
	DP-O$_r$-s	0.47	**0.52**	0,45
	CP-O$_c$-⋆	0.63	0.50	**0.53**
	CP-O$_c$-VP	*0.59*	0.49	**0.52**
	CP-STM-9	0.58	0.35	0.35
Shallow Semantic	NE-M$_e$-⋆	0.32	0.53	**0.56**
	NE-M$_e$-ORG	0.11	0.27	**0.29**
	NE-M$_e$-PER	0.13	0.34	0.34
	SR-M$_r$-⋆	0.50	**0.19**	0.18
	SR-M$_r$-A0	0.33	**0.31**	0.30
	SR-M$_r$-A1	0.28	0.14	0.14
	SR-O$_r$	0.41	**0.64**	0.63
	SR-O$_r$-⋆	**0.53**	0.36	**0.37**
	SR-O$_r$-AM-TMP	0.13	**0.39**	0.38
Semantic	DR-O$_r$-⋆	**0.59**	**0.36**	0.34
	DR-O$_r$-card	0.12	**0.49**	0.45
	DR-O$_r$-dr	0.56	**0.43**	0.40
	DR-O$_r$-eq	0.12	**0.17**	0.16
	DR-O$_r$-named	0.38	**0.48**	0.45
	DR-O$_r$-pred	0.55	**0.38**	0.36
	DR-O$_r$-prop	0.39	**0.27**	0.24
	DR-O$_r$-rel	0.56	**0.38**	0.34
	DR-STM-9	0.40	0.26	0.26

Table 3: Document level analysis

in favor of the Best SMT system diminishes and, indeed, ends reversing in favor of the LinearB system when we enter the syntactic and semantic levels.

Our heterogeneous set of metrics also allows us to analyze very specific aspects of quality. For instance, lexical metrics tell us that the LinearB system does not match well the expected reference lexicon. This is corroborated by analyzing shallow-syntactic similarities. For instance, observe how, while Best SMT is better than LinearB according to 'SP-O$_p$-J|N|V' metrics, which compute lexical overlapping respectively over adjectives, nouns and verbs, LinearB is better than Best SMT at translating determiners ('DP-O$_c$-det') and auxiliary verbs ('DP-O$_c$-aux'), closed grammatical categories which are, therefore, presumably less prone to suffer the effects of a biased lexical selection.

At the syntactic level, differences between both systems are rather small. Metrics based on dependency parsing assign the LinearB system a higher quality, both overall ('DP-HWC$_r$-4' and 'DP-O$_r$-⋆') and with respect to finer aspects such as the translation of finite complements ('DP-O$_r$-fc'), clause relations ('DP-O$_r$-i'), verb objects ('DP-O$_r$-obj'), and specially surface subjects ('DP-O$_r$-s'). In contrast, metrics based on constituent analysis tend to prefer the Best SMT system except for the 'CP-STM-9' metric which assigns both systems the same quality.

As to shallow-semantic metrics, it can be observed that LinearB has more problems than Best SMT to translate NEs, except for the case of person names. In the case of semantic arguments and adjuncts the two systems exhibit a very similar performance with a slight advantage on the side of LinearB, both overall ('SR-M$_r$-⋆' and 'SR-O$_r$') and for fine aspects such as the translation of agent roles ('SR-M$_r$-A0') and temporal adjuncts ('SR-M$_r$-AM-TMP'). Also, it can be observed that both systems have difficulties to translate theme roles ('SR-M$_r$-A1').

At the properly semantic level (i.e., over discourse representations), observe how there is not a single metric which ranks the Best SMT system first. LinearB is consistently better at translating basic discourse representation structures ('DR-O$_r$-dr'), cardinal expressions ('DR-O$_r$-card'), NEs ('DR-O$_r$-named'), equality conditions ('DR-O$_r$-eq'), predicates ('DR-O$_r$-pred'), relations ('DR-O$_r$-rel') and propositional attitudes ('DR-O$_r$-prop'), and overall ('DR-O$_r$-⋆'). It can also be observed that both systems have problems to translate equality conditions. Finally, both systems are assigned the same quality according to semantic tree matching ('DR-STM-9').

Meta-Evaluation

Metric quality has been evaluated on the basis of human likeness, i.e., in terms of the metric ability to discern between manual and automatic translations (Corston-Oliver et al., 2001; Lin and Och, 2004b; Kulesza and Shieber, 2004; Amigó et al., 2005; Gamon et al., 2005). We have computed human likeness through the KING measure defined inside the QARLA Framework (Amigó et al., 2005)[3]. Given a metric x, a set of human references R, and a set of automatic translations A, KING$_{R,A}(x)$ represents the probability, estimated over all sentence test cases, that a human reference in R does not receive a lower x score than the x score attained by *any* automatic translation in A. Broadly speaking, KING is a measure of discriminative power. For

[3]For KING computation we have used only the automatic outputs provided by the LinearB and Best SMT systems. We did not limit to segments counting on human assessments. All segments were used.

Reference 1:	Over 1000 monks and nuns , observers and scientists from over 30 countries and the host country attended the religious summit held for the first time in Myanmar which started today , Thursday .
2:	More than 1000 monks , nuns , observers and scholars from more than 30 countries , including the host country , participated in the religious summit which Myanmar hosted for the first time and which began on Thursday .
3:	The religious summit , staged by Myanmar for the first time and began on Thursday , was attended by over 1,000 monks an nuns , observers and scholars from more than 30 countries and host Myanmar .
4:	More than 1,000 monks , nuns , observers and scholars from more than 30 countries and the host country Myanmar participated in the religious summit , which is hosted by Myanmar for the first time and which began on Thursday .
5:	The religious summit , which started on Thursday and was hosted for the first time by Myanmar , was attended by over 1,000 monks and nuns , observers and scholars from more than 30 countries and the host country Myanmar .
Information:	**(1) → subject:** over/more_than 1,000 monks and nuns, observers and scientists/scholars from over/more_than 30 countries , and/including the host country **action:** attended/participated_in **object:** the religious summit **(2) → subject:** the religious summit **action:** began/started **temporal:** on Thursday **(3) → object:** the religious summit **action:** hosted **subject:** by Myanmar **mode:** for the first time
LinearB:	1000 monks from more than 30 States and the host State Myanmar attended the Summit , which began on Thursday , hosted by Myanmar for the first time .
Best SMT:	Religious participated in the summit , hosted by Myanmar for the first time began on Thursday , as an observer and the world of the 1000 monk nun from more than 30 countries and the host state Myanmar .

Table 4: Test case #637

instance, if a metric obtains a KING of 0.6, it means that in 60% of the test cases, it is able to explain by itself the difference in quality between manual and automatic translations.

In the context of error analysis, KING serves as an estimate of the impact of specific quality aspects on the system performance. In that respect, it can be observed (Table 3, column 1) that metrics at the lexical, shallow-syntactic and syntactic levels attain slightly higher KING values than metrics based on semantic similarities. We speculate that a possible explanation may be found in the performance of linguistic processors whose effectiveness suffers a significant decrease for deeper levels of analysis. Also, observe that finer grained metrics such as 'SP-O_p-J' (i.e., lexical overlapping over adjectives), 'NE-M_e-ORG' (i.e., lexical matching over organization names) or 'DR-O_r-card' (i.e., lexical overlapping over cardinal expressions) exhibit a much lower discriminative power. The reason is that they cover very partial aspects of quality.

3.4. Error Analysis at the Sentence Level

The heterogeneous set of metrics allows us to analyze different dimensions of translation quality over individual test cases. In this manner, we can better search for problematic cases according to different criteria. For instance, we could seek translations lacking of subject ('DP-O_r-s') and/or agent role ('SR-O_r-A0'). Or, at a more abstract level, by simultaneously relying on syntactic and semantic metrics, we could, for instance, locate a subset of possibly well-formed translations (i.e., high syntactic similarity) which somehow do not match well the reference semantic structure (i.e., low semantic similarity).

Level	Metric	Linear B	Best SMT
Human	Adequacy	**3**	2
	Fluency	**3.5**	2
Lexical	1-TER	**0.53**	0.51
	BLEU	0.44	**0.45**
	METEOR$_{wnsyn}$	0.59	**0.64**
Shallow Syntactic	SP-O$_p$-⋆	**0.52**	0.51
	SP-O$_p$-NN	**0.67**	0.38
	SP-O$_p$-NNP	0.60	**0.75**
	SP-O$_p$-V	0.40	**0.75**
Syntactic	DP-HWC$_w$-4	**0.17**	0.16
	DP-O$_r$-⋆	**0.46**	0.44
	DP-O$_r$-mod	**0.62**	0.41
	DP-O$_r$-obj	**0.29**	0.00
	DP-O$_r$-pcomp-n	**0.71**	0.39
	DP-O$_r$-rel	**0.33**	0.00
	CP-O$_c$-⋆	**0.59**	0.48
	CP-O$_c$-NP	**0.59**	0.55
	CP-O$_c$-PP	**0.57**	0.54
	CP-O$_c$-SB	**0.73**	0.00
	CP-O$_c$-VP	**0.64**	0.42
	CP-STM-9	**0.34**	0.23
Shallow Semantic	SR-O$_r$	**0.84**	0.25
	SR-O$_r$-⋆	**0.56**	0.18
	SR-O$_r$-A0	**0.44**	0.10
	SR-O$_r$-A1	**0.57**	0.28
Semantic	DR-O$_r$-⋆	**0.45**	0.34
	DR-O$_r$-dr	**0.57**	0.40
	DR-O$_r$-nam	**0.75**	0.24
	DR-O$_r$-pred	0.44	**0.45**
	DR-O$_r$-rel	**0.51**	0.32
	DR-STM-9	**0.32**	0.29

Table 5: Analysis of test case #637

A Case of Analysis

We have inspected particular cases. For instance, Table 4 presents the case of sentence 637 in which according to BLEU the translation by Best SMT is ranked first, whereas according to human assessments the translation by LinearB is judged of a superior quality both in terms of adequacy and fluency. This case is deeply analyzed in Table 5. In spite of its ill-formedness the translation by Best SMT deceives all lexical metrics. Particularly interesting is the case of 'METEOR$_{wnsyn}$', a metric designed to deal with differences in lexical selection, by allowing for morphological variations through stemming, and synonyms through dictionary lookup. METEOR is in this case, however, unable

LinearB:	You should cooperate and support one another .
Best SMT:	You that you will be more and more cooperative unit some of you and support each other .
Reference 1:	You must be more united and more cooperative and you must support each other .
2:	You must be more united and cooperative and supportive of each other .
3:	You must be more united and cooperative and supportive of each other .
4:	You have to be more united and more cooperative , and support each other .
5:	You have to be more united and more cooperative and you have to support each other .

Table 6: Translation Case #149.

Level	Metric	Linear B	Best SMT
Human	Adequacy	4	1.5
	Fluency	5	1.5
Lexical	1-PER	0.36	0.62
	1-TER	0.36	0.49
	BLEU	0.00	0.37
	NIST	1.64	9.42
	METEOR$_{wnsyn}$	0.32	0.67
Shallow Syntactic	SP-O_p-$\star$	0.25	0.46
	SP-O_p-V	0.17	0.40
	SP-O_c-$\star$	0.19	0.43
	SP-O_c-NP	0.43	0.50
	SP-O_c-VP	0.14	0.40
Syntactic	DP-HWC$_w$-4	0.07	0.12
	DP-HWC$_c$-4	0.32	0.19
	DP-HWC$_r$-4	0.32	0.25
	CP-STM-4	0.33	0.36
Shallow Semantic	SR-M_r-$\star$	0.14	0.67
	SR-O_r-$\star$	0.10	0.75
Semantic	DR-O_r-$\star$	0.17	0.26
	DR-O_{rp}-$\star$	0.24	0.26
	DR-O_{rp}-drs	0.27	0.30
	DR-O_{rp}-pred	0.29	0.40
	DR-O_{rp}-rel	0.30	0.24
	DR-STM-4	0.25	0.45

Table 7: Analysis of test case #149

to deal with differences in word ordering.

In contrast, scores conferred by metrics at deeper linguistic levels reveal, in agreement with human evaluation, that LinearB produced a more fluent (syntactic similarity) and adequate (semantic similarity) translation. Overall syntactic and semantic scores (e.g., 'DP-O$_r$-$\star$, 'CP-STM-9', 'SR-O$_r$-$\star$, 'DR-O$_r$-$\star$ and 'DR-STM-9'), all lower than 0.6, also indicate that important pieces of information were not captured or only partially captured.

Getting into details, by analyzing fine shallow-syntactic similarities, it can be observed, for instance, that, while LinearB successfully translated a larger proportion of singular nouns, Best SMT translated more proper nouns and verb forms. Analysis at the syntactic level reports that LinearB captured more dependency relations of several types (e.g., word adjunct modifiers, verb objects, nominal complements of prepositions, and relative clauses), and translated a larger proportion of different verb phrase types (e.g., noun, prepositional and verb phrases, and subordinated clauses). As to shallow-semantic similarity, it can be observed that the level of lexical overlapping over verb subjects and objects attained by LinearB is significantly higher. At the semantic level, the discourse representation associated to LinearB is, in general, more similar to the reference discourse representations. Only in the case of predicate conditions, both systems exhibit a similar performance.

Difficult Cases

One of the main problems of current automatic MT evaluation methods is that their reliability depends very strongly on the representativity of the set of reference translations available. In other words, if reference translations cover only a small part of the space of valid solutions, the predictive power of automatic metrics will decrease. This may be particularly dangerous in the case of n-gram based metrics, which are not able to deal with differences in lexical selection. For instance, Tables 6 presents a case in which the LinearB is unfairly penalized by lexical metrics for its strong divergence with respect to reference translations while the Best SMT system is wrongly favored for the opposite reason[4].

Metrics at deeper linguistic levels allow for partially overcoming this problem by inspecting syntactic and semantic structures. However, as it can be observed in the case selected, these structures may also exhibit a great variability. For instance, the translation by LinearB is considerably shorter than expected according to human references. Besides, while reference translations use "you must" or "you have", the LinearB translation uses "you should". Also, LinearB selected the verb form "cooperate" instead of "be more united and cooperative", etc. Table 7 shows the scores conferred by several metrics. It can be observed how lexical metrics completely fail to reflect the actual quality of the LinearB output. Indeed, only some dependency-based metrics are able to capture its quality (e.g., 'DP-HWC$_c$').

In the case depicted in Table 8, differences are mostly related to the sentence structure. Table 9 shows the scores conferred by several metrics. It can be observed, for instance, that several lexical metrics are able to capture the superior quality of the LinearB translation. In contrast, metrics at deeper linguistic levels do not reflect, in general, this difference in quality. Interestingly, only some syntax-based metrics confer a slightly higher score to LinearB (e.g., 'SP-O_p-$\star$' 'DP-HWC$_w$-4' 'CP-O_p-$\star$' 'CP-O_c-$\star$', etc.). All these metrics share the common property of computing lexical overlapping/matching over syntactic structures or grammatical categories.

[4]LinearB translation receives high scores from human assessors, but a null BLEU score. In contrast, the Best SMT system attains a high BLEU score, but receives low scores from human assessors.

LinearB:	It is important to analyze and address these problems properly .
Best SMT:	It should be to analyze these problems and take them up properly .
Reference 1:	We must analyze these problems and handle them correctly .
2:	So we must analyze these problems and take them in the right way .
3:	We must correctly analyze and properly handle these problems .
4:	And so it is imperative that we analyze these problems and deal with them properly .
5:	And so we must correctly analyze and properly handle these problems .

Table 8: Translation Case #728.

Level	Metric	Linear B	Best SMT
Human	Adequacy	4.5	2.5
	Fluency	5	2.5
Lexical	1-PER	0.63	0.48
	1-TER	0.55	0.48
	BLEU	0.00	0.46
	NIST	7.82	9.97
	ROUGE$_W$	0.25	0.29
	METEOR$_{wnsyn}$	0.54	0.44
Shallow Syntactic	SP-O_p-$\star$	0.44	0.39
	SP-O_p-PRP	0.50	0.33
	SP-O_c-$\star$	0.28	0.38
Syntactic	DP-O_c-$\star$	0.48	0.47
	DP-HWC$_w$-4	0.23	0.16
	DP-HWC$_c$-4	0.31	0.42
	DP-HWC$_r$-4	0.21	0.43
	DP-O_r-$\star$	0.25	0.36
	DP-O_r_i	0.44	0.43
	DP-O_r_mod	0.11	0.33
	DP-O_r_s	0.50	0.50
	CP-O_p-$\star$	0.45	0.41
	CP-O_p-RB	0.50	0.50
	CP-O_c-$\star$	0.43	0.38
	CP-O_c-VP	0.42	0.38
	CP-STM-4	0.48	0.59
Shallow Semantic	SR-O_r-$\star$	0.42	0.44
	SR-O_r	0.88	0.86
Semantic	DR-O_r-$\star$	0.20	0.36
	DR-O_{rp}-$\star$	0.52	0.60
	DR-O_r-drs	0.22	0.37
	DR-O_r-pred	0.25	0.33
	DR-O_r-rel	0.20	0.45
	DR-STM-4	0.25	0.33

Table 9: Analysis of test case #728

In order to deal with divergences between system outputs and reference translations, other authors have suggested taking advantage of paraphrasing support so as to extend the reference material (Russo-Lassner et al., 2005; Zhou et al., 2006; Kauchak and Barzilay, 2006; Owczarzak et al., 2006). We believe the two approaches could be combined.

4. Conclusions and Future Work

We have presented a valid path towards heterogeneous automatic MT error analysis. Our approach allows developers to rapidly obtain detailed automatic linguistic reports on their system's capabilities. Thus, human efforts may concentrate on high-level analysis.

Still, our proposal presents a limitation. It relies on a rich set of evaluation measures, most of which are based on language dependent automatic linguistic processors which may not be always available and whose quality may vary[5]. However, in our opinion, although these resources may be expensive to produce, unlike manual evaluations, they offer the important advantage of being reusable along the development cycle as systems and metrics improve. Besides, they are useful for NLP applications in general.

For future work, we plan to enhance the interface of the evaluation tool, currently in text format, so as to allow for a fast and elegant visual access from different viewpoints corresponding to the different dimensions of quality. Besides, evaluation measures generate, as a by-pass product, syntactic and semantic analyses which could be displayed. This would allow users to separately analyze the translation of different types of linguistic elements (e.g., constituents, relationships, arguments, adjuncts, discourse representation structures, etc.). For instance, missing or partially translated elements could appear highlighted in different colors.

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 NIST MT Evaluation Campaign organizers, and participants who agreed to share their system outputs for the purpose of this research.

5. References

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

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*.

[5]Currently, most metrics are only available for English. Some are also available for Spanish and Catalan.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of EACL*.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147.

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.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-Level MT evaluation without reference translations: beyond language modeling. In *Proceedings of EAMT*, pages 103–111.

Jesús Giménez and Lluís Màrquez. 2008. On the Robustness of Linguistic Features for Automatic MT Evaluation. (Under submission).

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.

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

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. Mathematisch Centrum, address = Amsterdam.

David Kauchak and Regina Barzilay. 2006. Paraphrasing for Automatic Evaluation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 455–462.

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

Audrey Le and Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. In *Official release of automatic evaluation scores for all submissions, August*.

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, Declan Groves, Josef Van Genabith, and Andy Way. 2006. Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 148–155.

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.

Grazia Russo-Lassner, Jimmy Lin, and Philip Resnik. 2005. A Paraphrase-Based Approach to Machine Translation Evaluation. Technical report, University of Maryland, College Park.

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*.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy. 2006. Re-evaluating Machine Translation Results with Paraphrase Support. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 77–84.