Linguistic Features for Automatic Evaluation of Heterogenous MT Systems

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

Evaluation results recently reported by Callison-Burch et al. (2006) and Koehn and Monz (2006), revealed that, in certain cases, the BLEU metric may not be a reliable MT quality indicator. This happens, for instance, when the systems under evaluation are based on different paradigms, and therefore, do not share the same lexicon. The reason is that, while MT quality aspects are diverse, BLEU limits its scope to the lexical dimension. In this work, we suggest using metrics which take into account linguistic features at more abstract levels. We provide experimental results showing that metrics based on deeper linguistic information (syntactic/shallow-semantic) are able to produce more reliable system rankings than metrics based on lexical matching alone, specially when the systems under evaluation are of a different nature.

1 Introduction

Most metrics used in the context of Automatic Machine Translation (MT) Evaluation are based on the assumption that *'acceptable'* translations tend to share the lexicon (i.e., word forms) in a predefined set of manual reference translations. This assumption works well in many cases. However, several results in recent MT evaluation campaigns have cast some doubts on its general validity. For instance, Callison-Burch et al. (2006) and Koehn and Monz (2006) reported and analyzed several cases of strong disagreement between system rankings provided by human assessors and those produced by the BLEU metric (Papineni et al., 2001). In particular, they noted that when the systems under evaluation are of a different nature (e.g., rule-based vs. statistical, human-aided vs. fully automatical, etc.) BLEU may not be a reliable MT quality indicator. The reason is that BLEU favours MT systems which share the expected reference lexicon (e.g., statistical systems), and penalizes those which use a different one.

Indeed, the underlying cause is much simpler. In general, lexical similarity is nor a sufficient neither a necessary condition so that two sentences convey the same meaning. On the contrary, natural languages are expressive and ambiguous at different levels. Consequently, the similarity between two sentences may involve different dimensions. In this work, we hypothesize that, in order to 'fairly' evaluate MT systems based on different paradigms, similarities at more abstract linguistic levels must be analyzed. For that purpose, we have compiled a rich set of metrics operating at the lexical, syntactic and shallow-semantic levels (see Section 2). We present a comparative study on the behavior of several metric representatives from each linguistic level in the context of some of the cases reported by Koehn and Monz (2006) and Callison-Burch et al. (2006) (see Section 3). We show that metrics based on deeper linguistic information (syntactic/shallow-semantic) are able to produce more reliable system rankings than those produced by metrics which limit their scope to the lexical dimension, specially when the systems under evaluation are of a different nature.

Proceedings of the Second Workshop on Statistical Machine Translation, pages 256–264,
Prague, June 2007.

2 A Heterogeneous Metric Set

For our experiments, we have compiled a representative set of metrics[1] at different linguistic levels. We have resorted to several existing metrics, and we have also developed new ones. Below, we group them according to the level at which they operate.

2.1 Lexical Similarity

Most of the current metrics operate at the lexical level. We have selected 7 representatives from different families which have been shown to obtain high levels of correlation with human assessments:

BLEU We use the default accumulated score up to the level of 4-grams (Papineni et al., 2001).

NIST We use the default accumulated score up to the level of 5-grams (Doddington, 2002).

GTM We set to 1 the value of the e parameter (Melamed et al., 2003).

METEOR We run all modules: 'exact', 'porter_stem', 'wn_stem' and 'wn_synonymy', in that order (Banerjee and Lavie, 2005).

ROUGE We used the ROUGE-S* variant (skip bigrams with no max-gap-length). Stemming is enabled (Lin and Och, 2004a).

mWER We use $1 -$ mWER (Nießen et al., 2000).

mPER We use $1 -$ mPER (Tillmann et al., 1997).

Let us note that ROUGE and METEOR may consider stemming (i.e., morphological variations). Additionally, METEOR may perform a lookup for synonyms in WordNet (Fellbaum, 1998).

2.2 Beyond Lexical Similarity

Modeling linguistic features at levels further than the lexical level requires the usage of more complex linguistic structures. We have defined what we call *'linguistic elements'* (LEs).

2.2.1 Linguistic Elements

LEs are linguistic units, structures, or relationships, such that a sentence may be partially seen as a 'bag' of LEs. Possible kinds of LEs are: word forms, parts-of-speech, dependency relationships, syntactic phrases, named entities, semantic roles, etc. Each LE may consist, in its turn, of one or more LEs, which we call 'items' inside the LE. For instance, a 'phrase' LE may consist of 'phrase' items, 'part-of-speech' (PoS) items, 'word form' items, etc. Items may be also combinations of LEs. For instance, a 'phrase' LE may be seen as a sequence of 'word-form:PoS' items.

2.2.2 Similarity Measures

We are interested in comparing linguistic structures, and linguistic units. LEs allow for comparisons at different granularity levels, and from different viewpoints. For instance, we might compare the semantic structure of two sentences (i.e., which actions, semantic arguments and adjuncts exist) or we might compare lexical units according to the semantic role they play inside the sentence. For that purpose, we use two very simple kinds of similarity measures over LEs: *'Overlapping'* and *'Matching'*. We provide a general definition:

Overlapping between items inside LEs, according to their type. Formally:

$$\text{Overlapping}(t) = \frac{\sum_{i \in items_t(hyp)} count'_{hyp}(i,t)}{\sum_{i \in items_t(ref)} count_{ref}(i,t)}$$

where t is the LE type[2], $items_t(s)$ refers to the set of items occurring inside LEs of type t in sentence s, $count_{ref}(i,t)$ denotes the number of times item i appears in the reference translation inside a LE of type t, and $count'_{hyp}(i,t)$ denotes the number of times i appears in the candidate translation inside a LE of type t, limited by the number of times i appears in the reference translation inside a LE of type t. Thus, 'Overlapping' provides a rough measure of the proportion of items inside elements of a certain type which have been 'successfully' translated. We also introduce a coarser metric, **'Overlapping(*)'**, which considers the uniformly averaged 'overlapping' over all types:

$$\text{Overlapping}(\star) = \frac{1}{|T|} \sum_{t \in T} \text{Overlapping}(t)$$

where T is the set of types.

[1] All metrics used in this work are publicly available inside the IQMT Framework (Giménez and Amigó, 2006). http://www.lsi.upc.edu/~nlp/IQMT

[2] LE types vary according to the specific LE class. For instance, in the case of Named Entities types may be 'PER' (i.e., person), 'LOC' (i.e., location), 'ORG' (i.e., organization), etc.

Matching between items inside LEs, according to their type. Its definition is analogous to the 'Overlapping' definition, but in this case the relative order of the items is important. All items inside the same element are considered as a single unit (i.e., a sequence in left-to-right order). In other words, we are computing the proportion of 'fully' translated elements, according to their type. We also introduce a coarser metric, **'Matching(*)'**, which considers the uniformly averaged 'Matching' over all types.

notes:

- 'Overlapping' and 'Matching' operate on the assumption of a single reference translation. The extension to the multi-reference setting is computed by assigning the maximum value attained over all human references individually.

- 'Overlapping' and 'Matching' are general metrics. We may apply them to specific scenarios by defining the class of linguistic elements and items to be used. Below, we instantiate these measures over several particular cases.

2.3 Shallow Syntactic Similarity

Metrics based on shallow parsing (*'SP'*) analyze similarities at the level of PoS-tagging, lemmatization, and base phrase chunking. Outputs and references are automatically annotated using state-of-the-art tools. PoS-tagging and lemmatization are provided by the SVMT*ool* package (Giménez and Màrquez, 2004), and base phrase chunking is provided by the P*hreco* software (Carreras et al., 2005). Tag sets for English are derived from the Penn Treebank (Marcus et al., 1993).

We instantiate 'Overlapping' over parts-of-speech and chunk types. The goal is to capture the proportion of lexical items correctly translated, according to their shallow syntactic realization:

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 introduce a coarser metric, **'SP-O_p-*'** which computes average overlapping over all parts-of-speech.

SP-O_c-t Lexical overlapping according to the chunk type 't'. For instance, 'SP-O_c-NP' roughly reflects the successfully translated proportion of noun phrases. We also introduce a coarser metric, **'SP-O_c-*'** which considers the average overlapping over all chunk types.

At a more abstract level, we use the NIST metric (Doddington, 2002) to compute accumulated/individual scores over sequences of:

Lemmas – **SP-NIST(i)$_l$-n**
Parts-of-speech – **SP-NIST(i)$_p$-n**
Base phrase chunks – **SP-NIST(i)$_c$-n**

For instance, **'SP-NIST$_l$-5'** corresponds to the accumulated NIST score for lemma n-grams up to length 5, whereas 'SP-NISTi$_p$-5' corresponds to the individual NIST score for PoS 5-grams.

2.4 Syntactic Similarity

We have incorporated, with minor modifications, some of the syntactic metrics described by Liu and Gildea (2005) and Amigó et al. (2006) based on dependency and constituency parsing.

2.4.1 On Dependency Parsing (DP)

'DP' metrics capture similarities between dependency trees associated to automatic and reference translations. Dependency trees are provided by the MINIPAR dependency parser (Lin, 1998). Similarities are captured from different viewpoints:

DP-HWC(i)-l This metric corresponds to the HWC metric presented by Liu and Gildea (2005). All head-word chains are retrieved. The fraction of matching head-word chains of a given length, 'l', is computed. We have slightly modified this metric in order to distinguish three different variants according to the type of items head-word chains may consist of:

Lexical forms – **DP-HWC(i)$_w$-l**
Grammatical categories – **DP-HWC(i)$_c$-l**
Grammatical relationships – **DP-HWC(i)$_r$-l**

Average accumulated scores up to a given chain length may be used as well. For instance, 'DP-HWCi$_w$-4' retrieves the proportion of matching length-4 word-chains, whereas **'DP-HWC$_w$-4'** retrieves average accumulated proportion of matching word-chains up to length-4. Analogously, **'DP-HWC$_c$-4'**, and **'DP-HWC$_r$-4'** compute average accumulated proportion of category/relationship chains up to length-4.

DP-$O_l|O_c|O_r$ 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 defined by MINIPAR. For instance, 'DP-O_r-s' reflects lexical overlapping between subtrees of type 's' (subject). 'DP-O_c-A' reflects lexical overlapping between terminal nodes of type 'A' (Adjective/Adverbs). 'DP-O_l-4' reflects lexical overlapping between nodes hanging at level 4 or deeper. Additionally, we consider three coarser metrics (**'DP-O_l-*'**, **'DP-O_c-*'** and **'DP-O_r-*'**) which correspond to the uniformly averaged values over all levels, categories, and relationships, respectively.

2.4.2 On Constituency Parsing (CP)

'CP' metrics capture similarities between constituency parse trees associated to automatic and reference translations. Constituency trees are provided by the Charniak-Johnson's Max-Ent reranking parser (Charniak and Johnson, 2005).

CP-STM(i)-l This metric corresponds to the STM metric presented by Liu and Gildea (2005). All syntactic subpaths in the candidate and the reference trees are retrieved. The fraction of matching subpaths of a given length, 'l', is computed. For instance, 'CP-STMi-5' retrieves the proportion of length-5 matching subpaths. Average accumulated scores may be computed as well. For instance, **'CP-STM-9'** retrieves average accumulated proportion of matching subpaths up to length-9.

2.5 Shallow-Semantic Similarity

We have designed two new families of metrics, 'NE' and 'SR', which are intended to capture similarities over Named Entities (NEs) and Semantic Roles (SRs), respectively.

2.5.1 On Named Entities (NE)

'NE' metrics analyze similarities between automatic and reference translations by comparing the NEs which occur in them. Sentences are automatically annotated using the *BIOS* package (Surdeanu et al., 2005). BIOS requires at the input shallow parsed text, which is obtained as described in Section 2.3. See the list of NE types in Table 1.

Type	Description
ORG	Organization
PER	Person
LOC	Location
MISC	Miscellaneous
O	Not-a-NE
DATE	Temporal expressions
NUM	Numerical expressions
ANGLE_QUANTITY DISTANCE_QUANTITY SIZE_QUANTITY SPEED_QUANTITY TEMPERATURE_QUANTITY WEIGHT_QUANTITY	Quantities
METHOD MONEY LANGUAGE PERCENT PROJECT SYSTEM	Other

Table 1: Named Entity types.

We define two types of metrics:

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. The **'NE-O_e-*'** metric considers the average lexical overlapping over all NE types. This metric includes the NE type 'O' (i.e., Not-a-NE). We introduce another variant, **'NE-O_e-**'**, which considers only actual NEs.

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 NEs of type 'LOC' (i.e., location). The **'NE-M_e-*'**

metric considers the average lexical matching over all NE types, this time excluding type 'O'.

Other authors have measured MT quality over NEs in the recent literature. In particular, the **'NE-M_e-*'** metric is similar to the **'NEE'** metric defined by Reeder et al. (2001).

2.5.2 On Semantic Roles (SR)

'SR' metrics analyze similarities between automatic and reference translations by comparing the SRs (i.e., arguments and adjuncts) which occur in them. Sentences are automatically annotated using the SwiRL package (Màrquez et al., 2005). This package requires at the input shallow parsed text enriched with NEs, which is obtained as described in Section 2.5.1. See the list of SR types in Table 2.

Type	Description
A0 A1 A2 A3 A4 A5	arguments associated with a verb predicate, defined in the PropBank Frames scheme.
AA	Causative agent
AM-ADV	Adverbial (general-purpose) adjunct
AM-CAU	Causal adjunct
AM-DIR	Directional adjunct
AM-DIS	Discourse marker
AM-EXT	Extent adjunct
AM-LOC	Locative adjunct
AM-MNR	Manner adjunct
AM-MOD	Modal adjunct
AM-NEG	Negation marker
AM-PNC	Purpose and reason adjunct
AM-PRD	Predication adjunct
AM-REC	Reciprocal adjunct
AM-TMP	Temporal adjunct

Table 2: Semantic Roles.

We define three types of metrics:

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-*'** 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-*'** metric considers the average lexical matching over all SR types.

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

Note that in the same sentence several verbs, with their respective SRs, may co-occur. However, the metrics described above do not distinguish between SRs associated to different verbs. In order to account for such a distinction we introduce a more restrictive version of these metrics ('SR-M_{rv}-t', 'SR-O_{rv}-t', **'SR-M_{rv}-*'**, **'SR-O_{rv}-*'**, and **'SR-O_{rv}'**), which require SRs to be associated to the same verb.

3 Experimental Work

In this section, we study the behavior of some of the metrics described in Section 2, according to the linguistic level at which they operate. We have selected a set of coarse-grained metric variants (i.e., accumulated/average scores over linguistic units and structures of different kinds)[3]. We analyze some of the cases reported by Koehn and Monz (2006) and Callison-Burch et al. (2006). We distinguish different evaluation contexts. In Section 3.1, we study the case of a single reference translation being available. In principle, this scenario should diminish the reliability of metrics based on lexical matching alone, and favour metrics based on deeper linguistic features. In Section 3.2, we study the case of several reference translations available. This scenario should alleviate the deficiencies caused by the shallowness of metrics based on lexical matching. We also analyze separately the case of *'homogeneous'* systems (i.e., all systems being of the same nature), and the case of *'heterogenous'* systems (i.e., there exist systems based on different paradigms).

As to the metric meta-evaluation criterion, the two most prominent criteria are:

Human Acceptability Metrics are evaluated on the basis of correlation with human evaluators.

Human Likeness Metrics are evaluated in terms of descriptive power, i.e., their ability to distinguish between human and automatic translations (Lin and Och, 2004b; Amigó et al., 2005).

In our case, metrics are evaluated on the basis of 'Human Acceptability'. Specifically, we use Pearson correlation coefficients between metric scores

[3] When computing 'lexical' overlapping/matching, we use lemmas instead of word forms.

and the average sum of adequacy and fluency assessments at the document level. The reason is that meta-evaluation based on 'Human Likeness' requires the availability of heterogenous test beds (i.e., representative sets of automatic outputs and human references), which, unfortunately, is not the case of all the tasks under study. First, because most translation systems are statistical. Second, because in most cases only one reference translation is available.

3.1 Single-reference Scenario

We use some of the test beds corresponding to the *"NAACL 2006 Workshop on Statistical Machine Translation" (WMT 2006)* (Koehn and Monz, 2006). Since linguistic features described in Section 2 are so far implemented only for the case of English being the target language, among the 12 translation tasks available, we studied only the 6 tasks corresponding to the Foreign-to-English direction. A single reference translation is available. System outputs consist of 2000 and 1064 sentences for the 'in-domain' and 'out-of-domain' test beds, respectively. In each case, human assessments on adequacy and fluency are available for a subset of systems and sentences. Table 3 shows the number of sentences assessed in each case. Each sentence was evaluated by two different human judges. System scores have been obtained by averaging over all sentence scores.

	in	**out**	**sys**
French-to-English	2,247	1,274	11/14
German-to-English	2,401	1,535	10/12
Spanish-to-English	1,944	1,070	11/15

Table 3: WMT 2006. 'in' and 'out' columns show the number of sentences assessed for the 'in-domain' and 'out-of-domain' subtasks. The 'sys' column shows the number of systems counting on human assessments with respect to the total number of systems which presented to each task.

Evaluation of Heterogeneous Systems

In four of the six translation tasks under study, all the systems are statistical except *'Systran'*, which is rule-based. This is the case of the German/French-to-English in-domain/out-of-domain tasks. Table 4 shows correlation with human assessments for some metric representatives at different linguistic levels.

Level	**Metric**	**fr2en in**	**fr2en out**	**de2en in**	**de2en out**
Lexical	1-PER	0.73	0.64	0.57	0.46
	1-WER	0.73	0.73	0.32	0.38
	BLEU	0.71	0.87	0.60	0.67
	NIST	0.74	0.82	0.56	0.63
	GTM	0.84	0.86	0.12	0.70
	METEOR	**0.92**	**0.95**	**0.76**	**0.81**
	ROUGE	0.85	0.89	0.65	**0.79**
Shallow Syntactic	SP-O_p-*	**0.81**	0.88	0.64	0.71
	SP-O_c-*	**0.81**	0.89	0.65	0.75
	SP-NIST$_l$-5	0.75	0.81	0.56	0.64
	SP-NIST$_p$-5	0.75	**0.91**	**0.77**	**0.77**
	SP-NIST$_c$-5	0.73	0.88	0.71	0.54
Syntactic	DP-HWC$_w$-4	0.76	0.88	0.64	0.74
	DP-HWC$_c$-4	**0.93**	**0.97**	0.88	0.72
	DP-HWC$_r$-4	**0.92**	**0.96**	**0.91**	0.76
	DP-O_l-*	0.87	0.94	0.84	0.84
	DP-O_c-*	0.91	0.95	0.88	**0.87**
	DP-O_r-*	0.87	**0.97**	**0.91**	**0.88**
	CP-STM-9	**0.93**	0.95	**0.93**	**0.87**
Shallow Semantic	NE-M_e-*	0.80	0.79	**0.93**	0.63
	NE-O_e-*	0.79	0.76	**0.91**	0.59
	NE-O_e-**	0.81	0.87	0.63	0.70
	SR-M_r-*	0.83	**0.95**	**0.92**	0.84
	SR-O_r-*	0.89	**0.95**	0.88	0.90
	SR-O_r	**0.95**	0.85	0.80	0.75
	SR-M_{rv}-*	0.77	0.92	0.72	0.85
	SR-O_{rv}-*	0.81	0.93	0.76	**0.94**
	SR-O_{rv}	0.84	0.93	0.81	**0.92**

Table 4: WMT 2006. Evaluation of Heterogeneous Systems. French-to-English (fr2en) / German-to-English (de2en), in-domain and out-of-domain.

Although the four cases are different, we have identified several regularities. For instance, BLEU and, in general, all metrics based on lexical matching alone, except METEOR, obtain significantly lower levels of correlation than metrics based on deeper linguistic similarities. The problem with lexical metrics is that they are unable to capture the actual quality of the 'Systran' system. Interestingly, METEOR obtains a higher correlation, which, in the case of French-to-English, rivals the top-scoring metrics based on deeper linguistic features. The reason, however, does not seem to be related to its additional linguistic operations (i.e., stemming or synonymy lookup), but rather to the METEOR matching strategy itself (unigram precision/recall).

Metrics at the shallow syntactic level are in the same range of lexical metrics. At the properly syntactic level, metrics obtain in most cases high correlation coefficients. However, the 'DP-HWC$_w$-4' metric, which, although from the viewpoint of dependency relationships, still considers only lexical matching, obtains a lower level of correlation. This reinforces the idea that metrics based on rewarding long n-grams matchings may not be a reliable quality indicator in these cases.

At the level of shallow semantics, while 'NE' metrics are not equally useful in all cases, 'SR' metrics prove very effective. For instance, correlation attained by 'SR-O_r-*' reveals that it is important to translate lexical items according to the semantic role they play inside the sentence. Moreover, correlation attained by the 'SR-M_r-*' metric is a clear indication that in order to achieve a high quality, it is important to 'fully' translate 'whole' semantic structures (i.e., arguments/adjuncts). The existence of all the semantic structures ('SR-O_r'), specially associated to the same verb ('SR-O_{rv}'), is also important.

Evaluation of Homogeneous Systems

In the two remaining tasks, Spanish-to-English in-domain/out-of-domain, all the systems are statistical. Table 5 shows correlation with human assessments for some metric representatives. In this case, BLEU proves very effective, both in-domain and out-of-domain. Indeed, all metrics based on lexical matching obtain high levels of correlation with human assessments. However, still metrics based on deeper linguistic analysis attain in most cases higher correlation coefficients, although not as significantly higher as in the case of heterogeneous systems.

Level	Metric	es2en in	es2en out
Lexical	1-PER	0.82	0.78
	1-WER	0.88	0.83
	BLEU	**0.89**	**0.87**
	NIST	0.88	0.84
	GTM	0.86	0.80
	METEOR	0.84	0.81
	ROUGE	**0.89**	0.83
Shallow Syntactic	SP-O_p-*	0.88	0.80
	SP-O_c-*	**0.89**	0.84
	SP-NIST$_l$-5	0.88	0.85
	SP-NIST$_p$-5	0.85	**0.86**
	SP-NIST$_c$-5	0.84	0.83
Syntactic	DP-HWC$_w$-4	**0.94**	0.83
	DP-HWC$_c$-4	0.91	0.87
	DP-HWC$_r$-4	0.91	**0.88**
	DP-O_l-*	0.91	0.84
	DP-O_c-*	0.88	0.83
	DP-O_r-*	0.88	0.84
	CP-STM-9	0.89	0.86
Shallow Semantic	NE-M_e-*	0.75	0.76
	NE-O_e-*	0.71	0.71
	NE-O_e-**	0.88	0.80
	SR-M_r-*	0.86	0.82
	SR-O_r-*	**0.92**	**0.92**
	SR-O_r	0.91	**0.92**
	SR-M_{rv}-*	0.89	0.88
	SR-O_{rv}-*	0.91	**0.92**
	SR-O_{rv}	0.91	0.91

Table 5: WMT 2006. Evaluation of Homogeneous Systems. Spanish-to-English (es2en), in-domain and out-of-domain.

3.2 Multiple-reference Scenario

We study the case reported by Callison-Burch et al. (2006) in the context of the Arabic-to-English exercise of the *"2005 NIST MT Evaluation Campaign"*[4] (Le and Przybocki, 2005). In this case all systems are statistical but *'LinearB'*, a human-aided MT system (Callison-Burch, 2005). Five reference translations are available. System outputs consist of 1056 sentences. We obtained permission[5] to use 7 system outputs. For six of these systems we counted on a subjective manual evaluation based on adequacy and fluency for a subset of 266 sentences (i.e., 1596 sentences were assessed). Each sentence was evaluated by two different human judges. System scores have been obtained by averaging over all sentence scores.

Table 6 shows the level of correlation with human assessments for some metric representatives (see 'ALL' column). In this case, lexical metrics obtain extremely low levels of correlation. Again, the problem is that lexical metrics are unable to capture the actual quality of 'LinearB'. At the shallow syntactic level, only metrics which do not consider any lexical information ('SP-NIST$_p$-5' and 'SP-NIST$_c$-5') attain a significantly higher quality. At the properly syntactic level, all metrics attain a higher correlation. At the shallow semantic level, again, while 'NE' metrics are not specially useful, 'SR' metrics prove very effective.

On the other hand, if we remove 'LinearB' (see

[4] http://www.nist.gov/speech/tests/summaries/2005/mt05.htm

[5] Due to data confidentiality, we contacted each participant individually and asked for permission to use their data. A number of groups and companies responded positively: University of Southern California Information Sciences Institute (ISI), University of Maryland (UMD), Johns Hopkins University & University of Cambridge (JHU-CU), IBM, University of Edinburgh, MITRE and LinearB.

Level	Metric	ar2en ALL	ar2en SMT
Lexical	1-PER	-0.35	0.75
	1-WER	-0.50	0.69
	BLEU	**0.06**	0.83
	NIST	0.04	0.81
	GTM	0.03	**0.92**
	ROUGE	-0.17	0.81
	METEOR	0.05	0.86
Shallow Syntactic	SP-O_p-*	0.05	0.84
	SP-O_c-*	0.12	0.89
	SP-NIST$_l$-5	0.04	0.82
	SP-NIST$_p$-5	0.42	0.89
	SP-NIST$_c$-5	0.44	0.68
Syntactic	DP-HWC$_w$-4	0.52	0.86
	DP-HWC$_c$-4	0.80	0.75
	DP-HWC$_r$-4	**0.88**	0.86
	DP-O_l-*	0.51	0.94
	DP-O_c-*	0.53	0.91
	DP-O_r-*	0.72	0.93
	CP-STM-9	0.74	**0.95**
Shallow Semantic	NE-M_e-*	0.33	0.78
	NE-O_e-*	0.24	0.82
	NE-O_e-**	0.04	0.81
	SR-M_r-*	**0.72**	**0.96**
	SR-O_r-*	0.61	0.87
	SR-O_r	0.66	0.75
	SR-M_{rv}-*	0.68	**0.97**
	SR-O_{rv}-*	0.47	0.84
	SR-O_{rv}	0.46	0.81

Table 6: NIST 2005. Arabic-to-English (ar2en) exercise. 'ALL' refers to the evaluation of all systems. 'SMT' refers to the evaluation of statistical systems alone (i.e., removing 'LinearB').

'SMT' column), lexical metrics attain a much higher correlation, in the same range of metrics based on deeper linguistic information. However, still metrics based on syntactic parsing, and semantic roles, exhibit a slightly higher quality.

4 Conclusions

We have presented a comparative study on the behavior of a wide set of metrics for automatic MT evaluation at different linguistic levels (lexical, shallow-syntactic, syntactic, and shallow-semantic) under different scenarios. We have shown, through empirical evidence, that linguistic features at more abstract levels may provide more reliable system rankings, specially when the systems under evaluation do not share the same lexicon.

We strongly believe that future MT evaluation campaigns should benefit from these results, by including metrics at different linguistic levels. For instance, the following set could be used:

{ *'DP-HWC$_r$-4', 'DP-O_c-*', 'DP-O_l-*', 'DP-O_r-*', 'CP-STM-9', 'SR-O_r-*', 'SR-O_{rv}'* }

All these metrics are among the top-scoring in all the translation tasks studied. However, none of these metrics provides, in isolation, a *'global'* measure of quality. Indeed, all these metrics focus on *'partial'* aspects of quality. We believe that, in order to perform *'global'* evaluations, different quality dimensions should be integrated into a single measure of quality. With that purpose, we are currently exploring several metric combination strategies. Preliminary results, based on the QUEEN measure inside the QARLA Framework (Amigó et al., 2005), indicate that metrics at different linguistic levels may be robustly combined.

Experimental results also show that metrics requiring linguistic analysis seem very robust against parsing errors committed by automatic linguistic processors, at least at the document level. That is very interesting, taking into account that, while reference translations are supposedly well formed, that is not always the case of automatic translations. However, it remains pending to test the behaviour at the sentence level, which could be very useful for error analysis. Moreover, relying on automatic processors implies two other important limitations. First, these tools are not available for all languages. Second, usually they are too slow to allow for massive evaluations, as required, for instance, in the case of system development. In the future, we plan to incorporate more accurate, and possibly faster, linguistic processors, also for languages other than English, as they become publicly available.

Acknowledgements

This research has been funded by the Spanish Ministry of Education and Science, projects OpenMT (TIN2006-15307-C03-02) and TRANGRAM (TIN2004-07925-C03-02). We are recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government. Authors are thankful to the WMT organizers for providing such valuable test beds. Authors are also thankful to Audrey Le (from NIST), and to the 2005 NIST MT Evaluation Campaign participants who agreed to share their system

outputs and human assessments for the purpose of this research.

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 COLING-ACL06*.

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 Machine Translation and/or Summarization*.

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

Chris Callison-Burch. 2005. Linear B system description for the 2005 NIST MT evaluation exercise. In *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Xavier Carreras, Lluís Márquez, and Jorge Castro. 2005. Filtering-ranking perceptron learning for partial parsing. *Machine Learning*, 59:1–31.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of ACL*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd IHLT*.

C. Fellbaum, editor. 1998. *WordNet. An Electronic Lexical Database*. The MIT Press.

Jesús Giménez and Enrique Amigó. 2006. IQMT: A Framework for Automatic Machine Translation Evaluation. In *Proceedings of the 5th LREC*.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In *Proceedings of 4th LREC*.

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.

Audrey Le and Mark Przybocki. 2005. NIST 2005 machine translation evaluation official results. Technical report, NIST, 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 ACL*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of COLING*.

Dekang Lin. 1998. Dependency-based Evaluation of MINIPAR. In *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

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 Machine Translation and/or Summarization*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of HLT/NAACL*.

Lluís Màrquez, Mihai Surdeanu, Pere Comas, and Jordi Turmo. 2005. Robust Combination Strategy for Semantic Role Labeling. In *Proceedings of HLT/EMNLP*.

S. Nießen, F.J. Och, G. Leusch, and H. Ney. 2000. Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd LREC*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation, rc22176, ibm. Technical report, IBM T.J. Watson Research Center.

Florence Reeder, Keith Miller, Jennifer Doyon, and John White. 2001. The Naming of Things and the Confusion of Tongues: an MT Metric. In *Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at MT Summit VIII*, pages 55–59.

Mihai Surdeanu, Jordi Turmo, and Eli Comelles. 2005. Named Entity Recognition from Spontaneous Open-Domain Speech. In *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*.

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*.