Mach Translat (2009) 23:181–193
DOI 10.1007/s10590-009-9060-y

Measuring machine translation quality as semantic equivalence: A metric based on entailment features

Sebastian Padó · Daniel Cer · Michel Galley · Dan Jurafsky · Christopher D. Manning

Received: 10 May 2009 / Accepted: 15 October 2009 / Published online: 8 November 2009

Abstract Current evaluation metrics for machine translation have increasing difficulty in distinguishing good from merely fair translations. We believe the main problem to be their inability to properly capture *meaning*: A good translation candidate *means* the same thing as the reference translation, regardless of formulation. We propose a metric that assesses the quality of MT output through its semantic equivalence to the reference translation, based on a rich set of match and mismatch features motivated by *textual entailment*. We first evaluate this metric in an evaluation setting against a combination metric of four state-of-the-art scores. Our metric predicts human judgments better than the combination metric. Combining the entailment and traditional features yields further improvements. Then, we demonstrate that the entailment metric can also be used as learning criterion in minimum error rate training (MERT) to improve parameter estimation in MT system training. A manual evaluation of the resulting translations indicates that the new model obtains a significant improvement in translation quality.

S. Padó (✉)
Stuttgart University, Stuttgart, Germany
e-mail: pado@ims.uni-stuttgart.de; Sebastianpado@gmail.com

D. Cer · M. Galley · D. Jurafsky · C. D. Manning
Stanford University, Stanford, USA

D. Cer
e-mail: cerd@stanford.edu

M. Galley
e-mail: mgalley@stanford.edu

D. Jurafsky
e-mail: jurafsky@stanford.edu

C. D. Manning
e-mail: manning@stanford.edu

Keywords MT evaluation · Automated metric · MERT · Semantics · Entailment · Linguistic analysis · Paraphrase

1 Introduction

Since human evaluation is costly and difficult to do reliably, *automatic* measures of translation quality that accurately mirror human judgments are vital to the progress of machine translation (MT).

Pioneer measures such as BLEU (Papineni et al. 2002) and NIST (Doddington 2002) measure MT quality cheaply and objectively through the strong correlation between human judgments and the *n*-gram overlap between a system translation and one or more reference translations. While more advanced measures are emerging, popular second generation metrics such as *translation edit rate* (TER) (Snover et al. 2006) and *Metric for Evaluation of Translation with Explicit ORdering* (METEOR) (Banerjee and Lavie 2005) still largely focus on word sequence overlap or, in the latter case, overlap of semantically related lexical items. Since these evaluation metrics are used for parameter optimization during *minimum error rate training* (MERT, Och (2003)), they can directly influence the quality of the resulting translation models.

With the improving state-of-the-art in machine translation, however, BLEU and related surface-based metrics have come under scrutiny. Studies such as Callison-Burch et al. (2006) have identified a number of problems: (1) BLEU-like metrics are unreliable at the level of individual sentences due to the small number of n-grams involved; (2) BLEU metrics can be "gamed" by permuting word order; (3) for some corpora and language pairs, the correlation to human ratings is very low even at the system level; (4) BLEU scores are biased towards statistical MT systems; (5) the gap in quality between MT systems and human translators is not reflected in equally large BLEU differences.

Given the important role evaluation plays in MT research, this is problematic, but not surprising: These metrics treat *any* divergence from the reference as negative. This is clearly an oversimplification. There is a long tradition of work in (computational) linguistics on *paraphrase*, linguistic variation *that preserves the meaning*, such as in Example 1:

(1) *Hypothesis:* This was declared terrorism by observers and witnesses.
Reference: Commentators as well as eyewitnesses are terming it terrorism.

A number of metrics have aimed at accounting for linguistic variation, either by making the matching more intelligent (Snover et al. 2006) or by integrating linguistic information, mostly lexical or structural (Banerjee and Lavie 2005; Owczarzak et al. 2008). Unfortunately, almost all metrics, with the exception of Giménez and Márquez (2008), concentrate on only one type of linguistic information[1] and therefore lack robustness.

In this article, we describe a metric that takes advantage the relationship between MT evaluation and the related task of *recognizing textual entailment* (RTE)

[1] BLEU and TER focus on exact word sequence overlap, while METEOR primarily scores translations according to the semantic relatedness of individual lexical items.

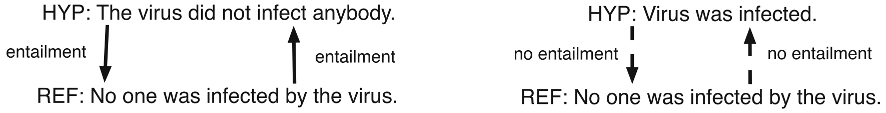
HYP: The virus did not infect anybody.
entailment
entailment
REF: No one was infected by the virus.
HYP: Virus was infected.
no entailment
no entailment
REF: No one was infected by the virus.

Fig. 1 Status of entailment between an MT system hypothesis and a reference translation for good translations (*left*) and bad translations (*right*)

(Dagan et al. 2005). Both tasks aim at assessing semantic equivalence, and need to distinguish meaning-preserving variation (paraphrase) from true changes in meaning. We predict the quality of MT hypotheses with a rich RTE feature set incorporating matches and mismatches between system output and reference on all linguistic levels. Extending beyond existing semantics-aware metrics such as METEOR, which deal with word relations, our metric examines more global phenomena such as multiword paraphrases, argument and modification relations, and phrase reorderings. We show that in addition to evaluation, the resulting metric can be directly integrated into system tuning via MERT and performs significantly better than BLEU/TER-trained models.

2 The relation of textual entailment and MT evaluation

Textual entailment (Dagan et al. 2005) was introduced as a concept of inference that corresponds better to "common sense" reasoning patterns than strict logical entailment. It is defined as a relation between two natural language sentences (a premise P and a hypothesis H) that holds if "a human reading P would infer that H is most likely true".

Figure 1 illustrates the (idealized) relation between the recognition of textual entailment (RTE) and MT evaluation: Very good MT output and the reference translation entail each other. Missing and additional hypothesis material breaks forward and backward entailment, respectively. For bad translations, entailment fails in both directions.

The examples show the common denominator between RTE and MT evaluation, namely determining the degree of *semantic equivalence*. Both tasks require the identification of different types of meaning-preserving reformulations. Example 1 above involves synonymy, semantic relatedness (*observers/commentators*), phrasal replacements (*and/as well as*), and a voice alternation implying structural change (*is declared/are terming*).

The tasks differ in that RTE is a binary task, while MT evaluation involves a real-valued prediction. Fortunately, many models developed for RTE use a classification architecture, whose features constitute evidence for matches and mismatches on different linguistic levels (cf. previous paragraph). Thus, the approach we follow in this article is to build a regression model over RTE features for MT evaluation.

Another difference is that RTE assumes well-formed sentences. This is not generally true in MT, and could lead to degraded linguistic features. However, MT judgments are more robust than entailment since they are not as sensitive to the contribution of individual words. Thus, we expect RTE features to be predictive in MT evaluation as well, provided that the analysis degrades gracefully on ungrammatical sentences.

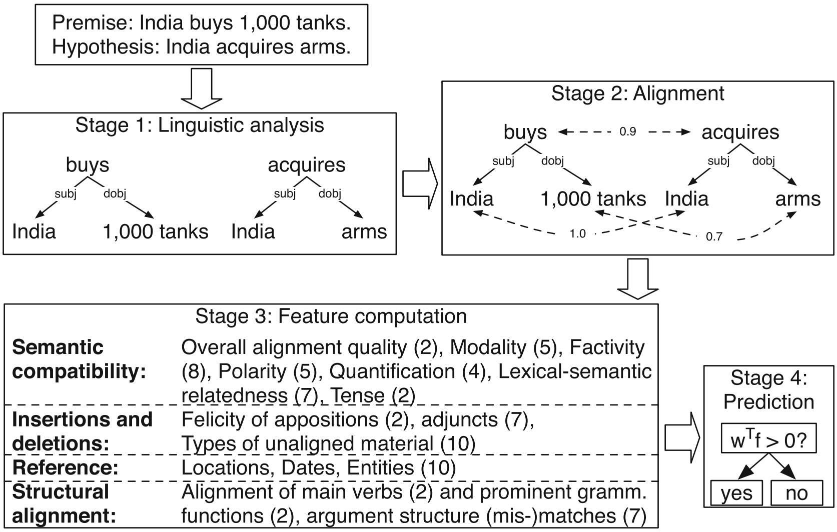
Premise: India buys 1,000 tanks.
Hypothesis: India acquires arms.
Stage 1: Linguistic analysis
buys
subj
dobj
India
1,000 tanks
acquires
subj
dobj
India
arms
Stage 2: Alignment
buys
0.9
acquires
subj
dobj
India
1,000 tanks
India
arms
1.0
0.7
Stage 3: Feature computation
Semantic compatibility:
Overall alignment quality (2), Modality (5), Factivity (8), Polarity (5), Quantification (4), Lexical-semantic relatedness (7), Tense (2)
Insertions and deletions:
Felicity of appositions (2), adjuncts (7), Types of unaligned material (10)
Reference:
Locations, Dates, Entities (10)
Structural alignment:
Alignment of main verbs (2) and prominent gramm. functions (2), argument structure (mis-)matches (7)
Stage 4: Prediction
w^Tf > 0?
yes
no

Fig. 2 The Stanford entailment recognizer

2.1 The stanford entailment recognizer

We predict MT quality with a regression model over features obtained from the Stanford Entailment Recognizer (MacCartney et al. 2006), which computes match and mismatch features for premise-hypothesis pairs.

Figure 2 shows the system architecture. It first constructs dependency graphs of the two sentences in the Stanford Dependencies representation (de Marneffe et al. 2006). These graphs contain one node per word and labeled edges representing grammatical relations. Contiguous collocations (“take off”) and named entities (“1000 tanks” in Fig. 2) are combined into single nodes. Stage 2 computes the highest-scoring alignment between hypothesis and premise nodes. Alignment scores are composed of local word and edge alignment scores, which use about ten lexical similarity resources, including WordNet (Miller et al. 1990), InfoMap (Takayama et al. 1999), and Dekang Lin’s thesaurus (Lin 1998). Exhaustive search being intractable, the system uses a stochastic search technique based on Gibbs sampling. See de Marneffe et al. (2007) for more details.

In the third stage, the system produces roughly 70 features for the aligned premise-hypothesis pair. A small number of them are real-valued (mostly quality scores), but most are binary implementations of small linguistic theories whose activation indicates syntactic and semantic (mis-)matches of different types. Figure 2 groups the features into four classes. *Semantic compatibility* assesses to what extent the aligned material has the same meaning and preserves semantic dimensions such as modality and factivity, taking a limited amount of context into account. *Insertions/Deletions* explicitly addresses material that remains unaligned and assesses to what extent these edits preserve meaning. The *Reference* features ascertain that the same events and participants are described. Finally, *Structural alignment* considers larger-scale structure by checking, e.g., whether grammatically prominent features of the text, such as the main verb, can be mapped onto the hypothesis.

To predict the quality of a reference/hypothesis pair, we compute entailment features in both directions (see Fig. 2), avoiding biases towards hypotheses that are systematically more specific (longer) or less specific (shorter) than their references. The final prediction, called RTER, is a simple linear combination of all features.[2] The feature weights are learnable from reasonably small training datasets (see Section 3).

The use of deep linguistic analysis makes our metric more heavyweight than traditional metrics, with an average per-sentence runtime of 4 s on an AMD 2.6 GHz core. This is efficient enough to perform regular evaluations on development and test sets, and Sect. 4 will show that RTER can already be profitably used for minimum error rate training.

3 Experiment 1: Predicting human judgments

We first assess the performance of RTER on an evaluation task, namely the prediction of human judgments for MT output on a 7-point Likert scale (Likert 1932; Fabrigar et al. 2005).

3.1 Setup

We consider five baselines. The first four combine component scores of four widely used MT metrics (BLEU, NIST, METEOR v.0.7, and TER) using the same linear combination approach as RTER.[3] We refer to these regression metrics with their original name, plus the suffix -R. To alleviate possible nonlinearity, we add all features in linear and log space. The fifth baseline, MTR, combines the features of all four baselines into a strong ensemble of traditional MT scores. Table 1 provides more details on the individual baselines. Finally, we test a combination model, MTRTER, that uses all surface (MTR) and entailment (RTER) features.

Our first experiment evaluates the entailment-based metric on the task of predicting human judgments on a seven-point Likert scale. We use the NIST MT 2008 corpus, which contains English translations of newswire text from three source languages: Arabic (Ar), Chinese (Ch), Urdu (Ur). Each language consists of 1500–2800 sentence pairs produced by 7–15 MT systems. We adopt a “round robin” scheme: We optimize the weights of our regression models on two languages and then predict human scores on the third language. This gauges performance of our models when training and test data come from the same genre, but from different languages, which is a setup of practical interest.

We evaluate both on the sentence and on the system level. At the sentence level, we can correlate predictions directly with human judgments using Spearman’s ρ,

[2] The linear regression prediction we use can be replaced easily; see Padó et al. (2009) for a logistic regression version of our metric that predicts pairwise preferences.

[3] Note that these regression models are strictly more powerful than the individual component scores, which can be simulated by the regression model.

Table 1 Baseline regression models: Number of features and description

Model	Features	Description
BLEUR	36	BLEU-n (Papineni et al. 2002) and n-gram precision scores ($1 \leq n \leq 4$); BLEU brevity penalty (BP); BLEU score divided by BP. To counteract BLEU's sentence-level brittleness, we also smooth BLEU-n and n-gram precision as in Lin and Och (2004)
NISTR	32	NIST-n (Doddington 2002) scores ($1 \leq n \leq 10$) and information-weighted n-gram precision scores ($1 \leq n \leq 4$); NIST brevity penalty (BP); and NIST score divided by BP
TERR	100	Standard TER (Snover et al. 2006) score and the number of each edit operation, using the default uniform cost as well as 9 non-uniform edit costs, with insertion cost close to 0
METR	2	The METEOR (Banerjee and Lavie 2005) metric
MTR	170	BLEUR + NISTR + TERR + METR

Features are represented twice (in linear space and in logarithmic space)

a rank correlation coefficient appropriate for non-normally distributed data. ρ ranges between -1 and 1. -1 means perfect inverse correlation, 0 no correlation, and 1 perfect correlation.

At the system level, we follow Callison-Burch et al. (2008) in computing quality as the percentage of sentences for which a system provides the best translation. We extend this procedure because real-valued metrics cannot predict ties, while human raters decide for a significant portion of sentences to "tie" two systems for first place. To account for this behavior, we compute a system's "tie-aware" quality as the percentage of sentences where the system's hypothesis was assigned a score *better or at most ϵ worse than the best system*. We set ϵ to make the relative frequency of ties equal to that observed in the training set. We find that this leads to a "confidence interval" for ϵ of 0.3–0.5 points on a 7-point scale.

3.2 Results

Table 2 shows the results (Spearman's ρ values). At the sentence level (upper half of the table), all correlations between model predictions and human judgments are highly significant ($p < 0.01$). Nevertheless, we see differences in robustness between the metrics. METR achieves the best correlation for Chinese and Arabic, but fails for Urdu, apparently the most difficult language. TERR shows the best result for Urdu, but does worse than METR for Arabic and even worse than BLEUR for Chinese. The MTR combination metric alleviates this problem to some extent by improving the "worst-case" performance on Urdu to the level of the best individual metric. The entailment metric (RTER) outperforms MTR on each language. In particular, it improves on MTR's correlation with Urdu. Even though METR still does somewhat better than MTR and RTER, this indicates the usefulness and robustness of entailment features as a complementary measure of translation quality.

In addition, the joint MTRTER model is best for all three languages, outperforming METR for each language pair. Interestingly, it performs considerably better than

Table 2 Experiment 1: Spearman's ρ between human ratings and model scores (OpenMT 2008)

Data		Metrics						
Train	Test	BleuR	MetR	NistR	TerR	MtR	RteR	MtRteR
Sentence								
ArZh	Ur	0.499	0.491	0.495	0.501	0.501	0.545	**0.556**
ArUr	Zh	0.539	0.611	0.531	0.503	0.573	0.580	**0.627**
ZhUr	Ar	0.525	0.601	0.504	0.545	0.552	0.599	**0.611**
System								
ArZh	Ur	0.739	0.684	0.500	0.900*	**0.927***	0.774*	0.810*
ArUr	Zh	0.385	0.443	0.400	0.590*	0.518*	0.477	**0.573***
ZhUr	Ar	0.597*	**0.863***	0.619*	0.421	0.481	0.597*	0.617*

Sentence level: All correlations significant ($p < 0.01$)

System level: $^*p < 0.05$

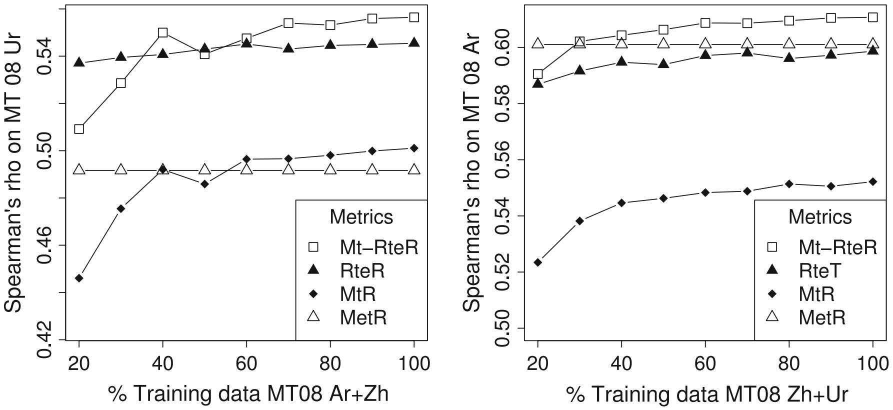

Fig. 3 Experiment 1: Learning curves for Urdu (*left*) and Arabic (*right*)

either MtR or RteR. We see this as a second promising result: the types of evidence provided by MtR and RteR are *complementary* and can be combined into a superior joint model.

At the system level, there is high variance due to the small number of datapoints, and only a few predictions show significant correlation. BleuR, MetR, and NistR significantly predict one language (Arabic); TerR, MtR, and RteR predict two languages. MtRteR is the only significant model for all three languages. These results correspond well to the sentence-level analysis, and are further supported by the WMT 2009 shared evaluation task results (Padó et al. 2009).

Next, we consider the role of training data. Figure 3 shows average correlations on Urdu and Arabic test data, using subsets of the training data (10% increments, 10 random draws per step). The relative performance of the metrics (MtR < RteR < MtRteR) remains largely the same throughout. Most of the learning takes place in the first 40% (800 sentence pairs) of training data. RteR shows a surprisingly

Table 3 Experiment 1: Reference and hypothesis translations (Urdu)

Sentence pairs	RTER features
Reference: I shall face that fact today *Hypothesis:* Today I will face this reality METR: 2.8 RTER: 6.1 Gold: 6	Only function words unaligned (*will, this*) Fact/reality: hypernym licenses alignment
Reference: What does BBC's Haroon Rasheed say after a visit to Lal Masjid Jamia Hafsa complex? There are no underground tunnels in Lal Masjid or Jamia Hafsa. The presence of the foreigners could not be confirmed as well *Hypothesis: BBC Haroon Rasheed Lal Masjid, Jamia Hafsa after* his *visit to* Auob Medical *Complex says Lal Masjid* and seminary in under a land mine, *not also been confirmed the presence of foreigners could not be*? METR: 4.5 RTER: 1.2 Gold: 1	Hypothesis root node unaligned Missing subject alignments Important entities in hypothesis cannot be aligned Reference, hypothesis differ in polarity

Scores are out of 7

flat learning curve despite the large number of free parameters, presumably because most features are binary. The performance of METR remains the same as the training data increases, which is to be expected from a two-feature combination, but differs substantially between across languages.

Finally, we perform a qualitative analysis, comparing the output of the RTER metric to METR, which we found to be generally the best individual MT metric at the sentence level. Table 3 illustrates two frequently observed patterns in cases where RTER outperforms METR. In the top example, a good translation is erroneously assigned a low score by METR because it cannot align *fact* and *reality* and punishes the change in word order. RTER correctly assigns a high score, based on the (mostly semantic) features shown on the right. Generally, RTER accounts for more valid variation in good translations because (a) it judges alignments by context; (b) incorporates a broader range of semantic relations; (c) weighs mismatches based on the word's status. The bottom example shows a very bad translation that is scored highly by METR since almost all of the reference words appear either literally or as synonyms in the translation hypothesis (shown in italics). In combination with METR's concentration on recall, this results in a moderately high score. RTER's features indicate semantic incompatibilities with the reference, indirectly pointing toward the hypothesis' ill-formedness.

4 Experiment 2: MT model optimization via MERT

Minimum error rate training (MERT) is the standard technique for tuning a machine translation model by varying its parameters to maximize performance achieved on a specific evaluation metric (Och 2003), traditionally BLEU or TER. However, the success of MERT depends highly on the evaluation metric. Translations produced by a MERT model are likely to exhibit properties that the metric rewards, but will be largely blind to aspects of translation quality that are not captured by the metric. As discussed above, almost all existing metrics capture either just local surface phenomena (like BLEU and TER) or do not have a good grasp on sentence coherence (like METEOR, cf. Table 3). In contrast, the large feature set used by RTER spanning a number of linguistic levels is a good candidate for use in MERT.

4.1 Experimental setup

We compare the performance of a statistical MT system trained with MERT using the popular BLEU and TER metrics against a system trained using RTER. Our MT system is a log-linear translation model with 14 real-valued features: the eight standard baseline features available in Moses (Hoang et al. 2007), plus six lexicalized re-ordering features. These MT system features are completely independent from the RTE features described in Sect. 2.1.

The MT system features are optimized with MERT on the Chinese-English data provided for NIST MT 2002. As test set, we use the NIST MT Chinese-English 2003 and 2005 datasets. We use Phrasal, Stanford's left-to-right beam search decoder (Cer et al. 2008), which is very similar to Moses. Our phrase table was built from 1.1M bi-sentences sampled from the NIST constrained track[4] and GALE Y2 training data. The Chinese data was word segmented using the Stanford CRF segmenter (Tseng et al. 2005). We extracted phrases as in Koehn et al. (2003) by running GIZA++ (Och and Ney 2003) in both directions and merging alignments with the grow-diag-final heuristic. We also produced a bidirectional lexical reordering model conditioned on source and target phrases. A 5-gram language model was created using the SRI language modeling toolkit (Stolcke 2002) and trained using the Gigaword corpus and English sentences from the parallel data.

For the MERT runs with an entailment-based metric, we optimized the parameters of MTRTER on the NIST MT 2006 and 2008 datasets (source languages: Arabic, Chinese, and Urdu). We introduced extensive caching for the linguistic analysis process. However, each iteration of MERT learning using 100-best lists still took on average 4 days. Since 10 iterations or more are often performed prior to convergence, MERT training with MTRTER would still take well over a month to complete. We therefore broke training down into two manageable steps: (1) train an initial MT model using one of the traditional MT evaluation metrics (BLEU and TER); (2) use the resulting model weights as the starting point for MERT training using MTRTER with small n-best lists (e.g., with $n = 10$). While such small n-best lists would not be sufficient to reliably estimate MT features from scratch, we found that they do allow for successful model adaptation while minimizing the computation cost. With this change, each iteration of MERT requires approximately 11 hours. We call the resulting models BLEU → MTRTER and TER→MTRTER, respectively.

[4] http://www.itl.nist.gov/iad/mig/tests/mt/2008/doc/mt08_constrained.html.

Table 4 Human pairwise preference for translation models

System pairs	MtRteR preferred (%)	Level of significance
BLEU → MtRteR vs. BLEU	52.9	$p < 0.001$
TER→MtRteR vs. TER	51.7	$p < 0.01$

4.2 Results

We evaluated the output of the four models (BleuR, TerR, BLEU → MtRteR and TER→MtRteR) against the three automatic metrics involved (BLEU, TER, and MtRteR). MERT training could always maximize the objective function. The BLEU-trained model performed best on BLEU, the TER-trained model on TER. According to MtRteR, the best model was TER→MtRteR, followed by BLEU → MtRteR, TER, and finally BLEU. This shows a systematic advantage of TER over BLEU: TER-trained models beat BLEU-trained models (cf. Table 2); TER is also a better starting point for tuning the model with MtRteR.

Then, we validated that the tendencies indicated by MtRteR correspond to true translation quality. We used Amazon's Mechanical Turk service, following Snow et al. (2008) in obtaining expert-level labeling by averaging ratings performed by a small number of human raters. Raters were presented with two system outputs and a single reference translation and asked to choose the better translation. The elicitation of MT quality judgments as pairwise preferences is particularly suitable for the comparatively small differences in translation that we expect, but has also been adopted more generally in the ACL SMT workshops (Callison-Burch et al. 2008). Inter-annotator agreement was 0.632, $\kappa = 0.26$ (fair agreement), comparable to the 2009 SMT results. Even though far from perfect, we find it promising that naive human annotators can produce agreement comparable to professionals. As gold standard, we adopted the label that received the majority of votes. They are fairly reliable, with an average of 2.4 times as many votes for the majority than for the minority label.

Table 4 shows that use of the entailment-based metric during training does in fact result in translations that tend to be preferred by human judges. While the magnitude of the effect is still modest, the results are highly statistically significant. We expect that more substantial gains can be obtained in the future by providing the translation model with a richer feature set that addresses more of the phenomena measured by the entailment-based evaluation metric. In an analysis of MT02, we found that sentences produced by MtRteR tend to exhibit better translations of structurally important words and are in particular better at preserving argument structure. Table 5 shows a typical case, where both translations miss information. Hyp (BLEU) scores better in BLEU-2,[5] but its deviance is more severe: It does not mention that it is the interior minister who provides the information, and falsely claims that he employed

[5] We report BLEU-2, that is the BLEU score calculated using only unigram and bigram counts, because BLEU with higher ns becomes zero for this sentence.

Table 5 Development set errors made by RTE- and BLEU-trained MT systems

Reference: Interior minister confirms senior adviser to the Italian government was shot by the "red brigade"	BLEU-2	MTRTER
Hypothesis (BLEU): The senior adviser to the interior minister was shot dead	0.4289	3.73
*Hypothesis (TER→*MTRTER*):* Interior minister said the senior adviser was shot	0.2124	4.15

the adviser. RTER tries to avoid such errors by penalizing translations that involve actions by or on the wrong entities.

5 Related work

Researchers have exploited various ways to enable matching between non-identical words or n-grams. Banerjee and Lavie (2005) and Chan and Ng (2008) use WordNet, and Zhou et al. (2006) and Kauchak and Barzilay (2006) exploit large collections of automatically extracted paraphrases. These approaches reduce the risk that a good (but free) translation is rated poorly, but do not address the problem of translations that contain long matches while lacking coherence and grammaticality (cf. Table 3). This issue has been addressed by a line of research on incorporating syntactic knowledge. Amigó et al. (2006) use the degree of dependency overlap between reference and hypothesis as a quality predictor. Similar ideas have been applied by Owczarzak et al. (2008) to LFG parses, and by Liu and Gildea (2005) to phrase-structure trees.

The most comparable work to ours is Giménez and Márquez (2008). Our studies agree on the crucial point that the use of a wide range of linguistic knowledge in MT evaluation is desirable and important. However, Giménez and Márquez propose a rather different approach to the actual integration of this knowledge. They advocate the bottom-up combination of "heterogeneous", independent metrics each of which measures overlap with respect to one linguistic level. In contrast, our aim is to provide a "top-down" motivation for the features we integrate through the entailment recognition task.

6 Conclusion and outlook

In this article, we have proposed an automatic metric for MT evaluation that exploits the close relationship between MT evaluation and the recognition of textual entailment (RTE) which allows the use of features developed for RTE (covering lexical, syntactic, and syntacto-semantic phenomena) for the assessment of MT hypotheses.

We have shown that our entailment-based metric correlates better with human judgments of MT quality than traditional metrics, demonstrating that deep linguistic features are sufficiently robust. Furthermore, we find that traditional and entailment features can be combined into a superior joint system. Reassuringly, the amount of data necessary for weight optimization is fairly small, and learned weights generalize well. Furthermore, we have demonstrated that the entailment-based metric can be used directly in MT parameter optimization, resulting in translations that are significantly preferred by human judges.

Further data analysis has confirmed the benefits of entailment-based MT evaluation (it abstracts away from valid variation such as word order or lexical substitution, but detects semantic divergences such as mismatches in argument structure) and clarified the relationship between MT evaluation and textual entailment: The majority of phenomena (but not all) that are relevant for RTE are also informative for MT evaluation. In the future, linguistically meaningful RTE features (cf. Fig. 2) may also be helpful in uncovering shortcomings of MT systems.

Acknowledgements This article is based on work funded by DARPA through IBM. The content does not necessarily reflect the views of the U.S. Government, and no official endorsement should be inferred.

References

Amigó E, Giménez J, Gonzalo J, Màrquez L (2006) MT evaluation: human-like vs. human acceptable. In: Proceedings of COLING/ACL 2006, pp 17–24

Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on evaluation measures, pp 65–72

Callison-Burch C, Fordyce C, Koehn P, Monz C, Schroeder J (2008) Further meta-evaluation of machine translation. In: Proceedings of the ACL workshop on statistical machine translation, pp 70–106

Callison-Burch C, Osborne M, Koehn P (2006) Re-evaluating the role of BLEU in machine translation research. In: Proceedings of EACL. pp 249–256

Cer D, Jurafsky D, Manning CD (2008) Regularization and search for minimum error rate training. In: Proceedings of the third workshop on statistical machine translation, Columbus, Ohio, pp 26–34

Chan YS, Ng HT (2008) MAXSIM: a maximum similarity metric for machine translation evaluation. In: Proceedings of ACL-08/HLT, pp 55–62

Dagan I, Glickman O, Magnini B (2005) The PASCAL recognising textual entailment challenge. In: Proceedings of the PASCAL RTE workshop, pp 177–190

de Marneffe M-C, Grenager T, MacCartney B, Cer D, Ramage D, Kiddon C, Manning CD (2007) Aligning semantic graphs for textual inference and machine reading. In: Proceedings of the AAAI spring symposium on machine reading, pp 36–42

de Marneffe M-C, MacCartney B, Manning CD (2006) Generating typed dependency parses from phrase structure parses. In: Fifth international conference on language resources and evaluation (LREC 2006), pp 449–454

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In: Proceedings of HLT, pp 128–132

Fabrigar LR, Krosnick JA, MacDougall BL (2005) Attitude measurement: techniques for measuring the unobservable. In: Brock T, Green M (eds) Persuasion: psychological insights and perspectives, Chap 2. 2nd edn. Sage, Thousand Oaks

Giménez J, Márquez L (2008) Heterogeneous automatic MT evaluation through non-parametric metric combinations. In: Proceedings of IJCNLP, pp 319–326

Hoang H, Birch A, Callison-Burch C, Zens R, Aachen R, Constantin A, Federico M, Bertoldi N, Dyer C, Cowan B, Shen W, Moran C, Bojar O (2007) Moses: open source toolkit for statistical machine translation. In: Proceedings of ACL, pp 177–180

Kauchak D, Barzilay R (2006) Paraphrasing for automatic evaluation. In: Proceedings of HLT-NAACL, pp 455–462

Koehn P, Och F, Marcu D (2003) Statistical Phrase-Based Translation. In: Proceedings of HLT-NAACL. pp 127–133

Likert R (1932) A technique for the measurement of attitudes. Arch Psychol 22(140):1–55

Lin C-Y, Och FJ (2004) ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In: Proceedings of COLING. pp. 501–507

Lin D (1998) Extracting collocations from text corpora. In: First workshop on computational terminology, pp 57–63

Liu D, Gildea D (2005) Syntactic features for evaluation of machine translation. In: Proceedings of the ACL workshop on evaluation measures, pp 25–32

MacCartney B, Grenager T, de Marneffe M-C, Cer D, Manning CD (2006) Learning to recognize features of valid textual entailments. In: Proceedings of NAACL, pp 41–48

Miller GA, Beckwith R, Fellbaum C, Gross D, Miller K (1990) WordNet: an on-line lexical database. Int J Lexicogr 3:235–244

Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of ACL, pp 160–167

Och FJ, Ney H (2003) A systematic comparison of various statistical alignment models. Comput Linguist 29(1):19–51

Owczarzak K, van Genabith J, Way A (2008) Evaluating machine translation with LFG dependencies. Mach Transl 21(2):95–119

Padó S, Galley M, Jurafsky D, Manning C (2009) Textual entailment features for machine translation evaluation. In: Proceedings of the EACL workshop on machine translation, pp 37–41

Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of ACL, pp 311–318

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of AMTA, pp 223–231

Snow R, O'Connor B, Jurafsky D, Ng A (2008) Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP, pp 254–263

Stolcke A (2002) SRILM—an extensible language modeling toolkit. In: Proceedings of the international conference on spoken language processing, pp 901–904

Takayama Y, Flournoy R, Kaufmann S, Peters S (1999) Information retrieval based on domain-specific word associations. In: Proceedings of PACLING, pp 155–161

Tseng H, Chang P-C, Andrew G, Jurafsky D, Manning C (2005) A conditional random field word segmenter for the SIGHAN bakeoff 2005. In: Proceedings of the SIGHAN workshop on chinese language processing, pp 32–39

Zhou L, Lin C-Y, Hovy E (2006) Re-evaluating machine translation results with paraphrase support. In: Proceedings of EMNLP, pp 77–84