Evaluating MT output with entailment technology

Sebastian Padó, Michel Galley, Dan Jurafsky, Chris Manning
Stanford University
{pado,mgalley,jurafsky,manning}@stanford.edu

September 30, 2008

1 Introduction

Constant evaluation is vital to the progress of machine translation. However, human evaluation is costly, time-consuming, and difficult to do reliably. On the other hand, automatic measures of machine evaluation performance (such as BLEU, NIST, TER, and METEOR), while cheap and objective, have increasingly come under suspicion as to whether they are satisfactory measuring instruments. Recent work (e.g., Callison-Burch et al. (2006)) has demonstrated that for current state-of-the-art MT systems, the correlation between BLEU scores and human adequacy and fluency ratings is often low; BLEU scores tend to favor statistical over rule-based systems; and BLEU-like measures tend to perform worse at the segment level than at the corpus level.

The core of the problem is that BLEU (Papineni et al., 2001), and to a first approximation other automatic measures, work at the surface level, looking for shared word sequences between a system translation and one or more reference translations. This evaluation ignores many known facts about linguistic semantics, whereby the same meaning can be conveyed in many different ways, whether by the use of syntactic rearrangements or by exploiting lexical semantics (synonyms, etc.) and larger semantic paraphrases. Consider the two (real-world) example sentences in Figure 1, which are largely equivalent, but differ substantially on the surface. The equivalence of the two sentences hinges not only on the synonymy of individual words (*practice/policy*), but also on phrasal replacements (*promote/make statements in favor of*) and lexical-semantic properties of words (such as the "built-in" negation of *barring*).

In this paper, we present a study whose goal is to improve the prediction of adequacy judgments for MT system translations by accounting for such semantic phenomena. To do so, we model MT evaluation as an instance of the "recognition of textual entailment" (RTE) task (Dagan et al., 2005). RTE was introduced as a "practical" inference procedure that determines the existence of a causal relation between two short segments of text, the *premise* and the *hypothesis*: Is the hypothesis entailed by the premise, or not? Textual entailment has been found to be beneficial for a range of applications, for example in answer validation in Question Answering or in word sense disambiguation (Dagan et al., 2006; Harabagiu and Hickl, 2006). Our intuition is that the evaluation of MT output for *adequacy* can also be seen as an entailment task: A candidate translation (i.e., MT

The company had a	strict practice	of not paying	outside commentators	to promote	its products
	firm policy	barring payments to		for making statements in favor of	its interests

Figure 1: An equivalent sentence pair with little surface overlap

system output) should entail the reference translation, and vice versa. To the extent that this is not true, the candidate translation is not adequate.

Phrasing MT evaluation as an entailment task has at least two major benefits. The first one is that entailment decisions are naturally made at the segment level, and not at the corpus level, which gives reason to hope for more accurate per-sentence judgments. The second benefit is that an entailment-based framework for MT evaluation is a priori designed to account for the variability of linguistic realization that presents a problem to BLEU-based approaches, and can profit from the experiences gained from RTE over the course of successive RTE workshops. One such result is that entailment recognition can benefit from the integration of deeper linguistic syntactico-semantic analysis, and attention to global semantic features. This lesson has informed the development of the Stanford RTE system (described in Section 2) which we will apply to MT evaluation.

The results of our pilot study indicate that an entailment recognition system with limited customization already provides results for MT evaluation which are competitive with the best existing metrics. Crucially, we find that entailment scores provide complementary information to surface-based metrics: They can be combined to obtain hybrid evaluation systems that outperform both individual metric types.

2 Entailment technology for MT evaluation

2.1 The Stanford RTE system

The basis for our experiments is the Stanford entailment recognition system (MacCartney et al., 2006). The input to the system is a pair of a premise and a hypothesis. The system uses a three-stage architecture that (a) conducts linguistic analysis, (b) builds an alignment between dependency graphs of the two segments, and (c) performs inference to determine entailment.

Analysis stage. Our goal at this stage is to compute linguistic representations of the premise and the hypothesis that contain as much information as possible about their semantic content. We use typed dependency graphs, which contain a node for each word and labeled edges representing the grammatical relations between words. Named entities are identified by a CRF-based NER system, and contiguous collocations which appear in WordNet are identified and grouped.

Alignment stage. The second stage identifies a good partial alignment between the graphs representing the hypothesis and the premise. In the current version of the system, the alignment is constructed as the highest-scoring mapping from each node in the hypothesis graph to a single node in the premise graph, or to null. We choose a locally decomposable scoring function, such that the score of an alignment is the sum of the local node and edge alignment scores. The node alignment score is a lexical similarity score that combines lexical similarity judgments from about ten resources, including WordNet, InfoMap, Dekang Lin's thesaurus, and gazetteers.

Nevertheless, the large number of possible alignments (exponential in the number of hypothesis words) makes exhaustive search intractable. Instead, we use a stochastic search technique based on Gibbs sampling, a well-known Markov Chain Monte Carlo technique (see de Marneffe et al. (2007) for details). This Gibbs sampler is guaranteed to give us samples from the posterior distribution over alignments as defined by the scoring function.

Entailment stage. The final inference stage determines if the hypothesis is entailed by the premise. We construct a set of features based on the output of the previous stages. These features model a wide range of syntactic, lexical, and semantic phenomena, including factivity; polarity; antonymy; felicity of adjuncts; modality; quantification; matches and mismatches of names, dates and numbers; compatibility of syntactic structure; and the quality of the alignment. The final score for each premise-hypothesis pair is computed as the dot product of the feature values and a weight vector. The feature weights, as well as the decision boundary between entailment and non-entailment, can either be set by hand, or learned from a development dataset with (regularized) logistic regression.

The disassociation between an alignment and an entailment stage is in contrast to most systems developed for RTE, which determine entailment directly from (graph) alignment quality. We have found that alignment scoring and deciding entailment are two conceptually distinct tasks, whose separation is important to deal with hypotheses that align well but are still not entailed (e.g., for reasons of polarity – see MacCartney et al. (2006) for details).

2.2 Extensions for MT evaluation

In an MT evaluation context, we replace the logistic regression training by linear regression training to estimate feature weights from MT datasets with graded human adequacy judgments. Consequently, we interpret the real-valued final score of the RTE system as an adequacy prediction for the given premise-hypothesis pair. Two other changes over the vanilla RTE system involved directionality and robustness.

Directionality. The standard entailment recognition task is asymmetric, with one designated premise and one hypothesis. In contrast, MT evaluation is at least conceptually a symmetric task: entailment must hold in both directions for the system translation to be fully adequate. In practice, it could be surmised that translation is often lossy and thus it is more important to treat the system translation as the premise that must be strong

enough to entail the hypothesis (the reference translation). Nevertheless, we found in preliminary experiments that there is a considerable gain from testing for entailment in both directions, and combining the entailment-stage features of both runs to predict adequacy.

Improving robustness. The Stanford RTE system was developed for processing basically well-formed English text, which is a reasonable expectation for almost all English NLP tasks. However, the output of MT systems can clearly be ungrammatical. Instead of redesigning the RTE system to be more robust, we decided to address this problem by supplementing the RTE features with a number of state-of-the-art MT evaluation scores that are known to be robust to noise. Specifically, our feature set included segment-level BLEU-4 (Papineni et al., 2001), NIST (Doddington, 2002), and translation error rate (TER) (Snover et al., 2006) scores. In the case of BLEU-4 and NIST, we treated the brevity penalty and individual n-gram ($n \leq 4$) precision scores as additional features. Since BLEU-4 tends to be brittle when computed at the segment level[1] , we smoothed n-gram precision counts as in Lin and Och (2004). In the case of TER, we computed multiple scores by assigning different costs to the four kinds of translation errors identified by TER (i.e., insertion, deletion, substitution, shift). All MT scores are provided to the linear regression model both in linear and in log space.

3 Experimental results

In this paper, we assess the performance of an entailment-based method for predicting human adequacy ratings for MT system translations. We use a single reference translation.[2]

We have evaluated our method on corpora from two MT datasets with English as target language for which human adequacy judgments are available: The NIST MT dataset (rated for adequacy on a seven-point scale), and the ACL SMT dataset (rated for adequacy on a five-point scale).

NIST06 The NIST MT-06 Arabic-to-English dataset provided as MetricsMATR development data (2000 segment pairs)

NIST08A The NIST MT-08 Arabic-to-English dataset (2800 segment pairs)

NIST08C The NIST MT-08 Chinese-to-English dataset (1800 segment pairs)

NIST08U The NIST MT-08 Urdu-to-English dataset (1500 segment pairs)

SMT06E The NAACL 2006 SMT workshop EUROPARL dataset (source languages: German, French, Spanish, 6400 segment pairs)

SMT06C The NAACL 2006 SMT workshop Czech-English News Commentary dataset (3800 segment pairs)

[1] BLEU-4 score is zero if the candidate translation fails to match any 4-gram of the reference translation.

[2] It is possible to generalize the setting to multiple reference translations. We did not follow this path since the human effort involved in producing reference translations is typically high.

	BLEU	MT	RTE	RTE+MT
NIST06	59.7	68.6	67.4	72.6
NIST08 Arabic	56.2	63.8	64.1	66.9
NIST08 Chinese	54.3	59.3	62.5	63.7
NIST08 Urdu	47.6	56.3	54.7	59.4
SMT06 Europarl	35.3	40.9	44.6	47.1
SMT06 Commentary	35.5	41.1	47.0	49.8
SMT07 Europarl	36.0	43.0	44.4	47.2
SMT07 Commentary	31.2	40.5	40.2	43.7

Table 1: Experiment 1: Performance of MT evaluation systems on individual corpora (Spearman's ρ averaged over 10-fold cross validation)

SMT07E The ACL 2007 SMT workshop EUROPARL dataset (source languages: German, French, Spanish, 2400 segment pairs)

SMT07C The ACL 2007 SMT workshop Czech-English News Commentary dataset (3100 segment pairs)

To assess the performance of the entailment features and the MT features in the system independently, and to compare against a simple baseline, we have set up four systems. The first system, our baseline, uses only BLEU-4. The second, MT, uses the committee of traditional MT metrics described in Section 2.2. The third one, RTE, predicts adequacy on the basis of the standard RTE features. The last system, RTE+MT, uses the union of both feature sets to combine robustness with accurate semantic analysis. This hybrid system is our primary MetricsMATR submission.

3.1 Experiment 1: Individual corpora

In a first experiment, we compare the performance of the three systems on all individual corpora. We perform 10-fold cross-validation, optimizing the feature weights on 90% of the training data and testing the correlation on the remaining 10%. We use Spearman's ρ, a non-parametric correlation measure that is appropriate for these datasets which are not normally distributed. The averages over the 10 folds are shown in Table 1.

We observe that all systems (MT, RTE, RTE+MT) outperform the simple BLEU baseline by at least 5 correlation points[3]. On SMT data, the RTE system outperforms MT on 3 of the 4 corpora. On the NIST datasets, there is a draw: each system is able to model two of the datasets better. In sum, we find that MT, a regression model built on the component scores of several current MT metrics, already greatly outperforms an individual metric (BLEU), and provides a very strong baseline for RTE and RTE+MT.

Excitingly, the Stanford RTE system shows comparable or even higher accuracy than MT; apparently even without major changes, and without integration of the MT scores, it is sufficiently robust to deal with the partly ill-formed system translation. The

[3]Across datasets, the correlations are much higher for the NIST MT corpora. One definite reason for this is the coarser grading scale for the SMT datasets, which produces a high number of draws in the accuracy judgments.

	BLEU	MT	RTE	RTE+MT
NIST	57.7	64.8	62.7	67.5
SMT	33.0	38.6	43.0	45.1
NIST+SMT	40.4	46.2	48.7	51.6

Table 2: Experiment 2: Performance of MT evaluation systems on combined corpora (Spearman's ρ averaged over 10-fold cross validation)

dataset on which RTE performs worst, and is bested by MT by almost 2%, is the NIST Urdu-English dataset. This dataset is not only the smallest dataset, but also the one whose adequacy was rated the lowest of all datasets: The average was 3.3 on a seven-point scale, while all other datasets show an average adequacy rating of above 4. We interpret this as evidence that metrics like BLEU and TER are appropriate while system output is still poor, but become problematic once system output becomes reasonably good. Meanwhile, more advanced metrics such as RTE use better system translations to their advantage, allowing them to rate adequacy more accurately.

A final important result is that the hybrid system (RTE+MT) is able to consistently and substantially outperform both base systems. This indicates that the information in the RTE and MT features is complementary. Thus, like entailment, MT evaluation can benefit from a combination of shallow and deep semantic information.

3.2 Experiment 2: Combined corpora

Experiment 1 has provided a detailed comparison of the different models under different circumstances. However, it could be argued that the advantage of entailment-based evaluation is dependent on the ability to optimize the feature weights specifically for individual corpora and that these weights do not generalize across corpora. Also, for practical purposes it is desirable that evaluation systems can be applied to new corpora without the need for retraining.

Experiment 2 addresses these concerns by investigating the generalizability of learned MT evaluation models across datasets. We repeat the cross-validation method of Experiment 1 on the union of all NIST corpora, the union of all SMT corpora, and the union of all corpora. By and large, we see the same tendencies as in Experiment 1. The RTE system outperforms MT on the SMT and NIST+SMT dataset, but comes out worse on the NIST dataset.[4] The hybrid system again performs 2–3 points better than either of the two individual systems on all datasets, gaining a total of 10–12 correlation points over the BLEU baseline.

4 Resources and Availability

Entailment technology is much more heavyweight than traditional MT evaluation metrics. The Stanford RTE system builds on deep linguistic analysis in Stage 1 and

[4]If we remove the NIST08 Urdu corpus, the difference between the two systems is smaller, $\rho = 65.0$ (RTE) vs. $\rho = 65.6$ (MT). The greater profit that RTE can draw from the exclusion of Urdu supports our earlier interpretation that it is the comparatively bad performance of the Urdu MT that hampers RTE.

requires the computation of all pairwise lexical similarity scores between premise and hypothesis words in Stage 2. The average runtime per segment pair on the MT06 dataset is roughly 8 seconds on an AMD Opteron 2.6GHz processor, with a memory requirement of several GB.

However, we believe that comparatively simple changes can reduce runtime. These include sharing of processing between both entailment directions, caching of lexical similarity scores, and reduction to a smaller number of lexical resources. In addition, the independence between individual segment pairs makes scoring easily parallelizable.

We will make a runtime version of the entailment-based Stanford MT evaluation metric available shortly. Updated information can be found at `http://nlp.stanford.edu/software/mteval.shtml`.

5 Conclusion and Outlook

In this paper, we have proposed to replace the traditional, shallow approach to MT evaluation by a linguistically informed semantic approach. We have modeled the prediction of adequacy scores for system translations as an entailment task, predicting adequacy based on mutual entailment between translation candidate and reference translation. Our results indicate that even with only little changes, an existing system for recognizing textual entailment outperforms BLEU, and is able to match the performance of a committee of current MT evaluation metrics.

We believe our results show promise for textual entailment-based MT evaluation, for several reasons. First, we found entailment-based features can be combined with traditional MT evaluation metrics to obtain a hybrid "best of both worlds" system that consistently outperforms either individual approach. This allows for the incremental improvement of evaluation procedures.

Next, the entailment-based system delivers its worst predictions for the most ill-formed corpus. This is not surprising, since many entailment features are triggered by syntactic (mis-)matches, and the system is bound to go astray if the linguistic analysis fails. This problem can be addressed by making the core entailment detection more robust.

Finally, the entailment-based approach to MT evaluation provides a general architecture for integrating higher-level linguistic knowledge, such as larger-level semantic paraphrases. This allows entailment-based MT evaluation systems to keep up with the constantly improving quality of Machine Translation output, which is becoming increasingly difficult to measure with surface-oriented methods.

Probably the largest hurdle to the practical use of entailment technology for MT evaluation at the moment is its high resource requirements. We have outlined some possible improvements – but the high requirements have a practical advantage as well: they make it very difficult for systems to be built to "game" entailment-based scoring.

References

Callison-Burch, Chris, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*. Trento, Italy.

Dagan, Ido, Oren Glickman, Alfio Gliozzo, Efrat Marmorshtein, and Carlo Strapparava. 2006. Direct word sense matching for lexical substitution. In *Proceedings of ACL*. Sydney, Australia.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

de Marneffe, Marie-Catherine, Trond Grenager, Bill MacCartney, Daniel Cer, Daniel Ramage, Chloé Kiddon, and Christopher D. Manning. 2007. Aligning semantic graphs for textual inference and machine reading. In *Proceedings of the AAAI Spring Symposium*. Stanford, CA.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of HLT*, pages 128–132. San Diego, CA.

Harabagiu, Sanda and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of ACL*. Sydney, Australia.

Lin, Chin-Yew and Franz Josef Och. 2004. ORANGE: a method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of COLING*. Geneva, Switzerland.

MacCartney, Bill, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. 2006. Learning to recognize features of valid textual entailments. In *Proceedings of NAACL*.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*. Toulouse, France.

Snover, Matthew, 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.