Shared Task Proposal: Syntactic Paraphrase Ranking

Michael White
Department of Linguistics
The Ohio State University
Columbus, OH 43210 USA
mwhite@ling.ohio-state.edu

Abstract

We describe a new shared task on syntactic paraphrase ranking that is intended to run in conjunction with the main surface realization shared task. Taking advantage of the human judgments collected to evaluate the surface realizations produced by competing systems, the task is to automatically rank these realizations—viewed as syntactic paraphrases—in a way that agrees with the human judgments as often as possible. The task is designed to appeal to developers of surface realization systems as well as machine translation evaluation metrics: for surface realization systems, the task sidesteps the thorny issue of converting inputs to a common representation; for MT evaluation metrics, the task provides a challenging framework for advancing automatic evaluation, as many of the paraphrases are expected to be of high quality, differing only in subtle syntactic choices.

1 Introduction

For the first surface realization shared task, the organizers considered running a follow-on task for evaluating automatic evaluation metrics—along the lines of similar meta-evaluations carried out for machine translation in recent years—though it was deferred for lack of time. For the second surface realization shared task, we propose to generalize this metrics meta-evaluation task to also usefully encompass realization ranking, where the various realizations generated for a given input in the main task are viewed as syntactic paraphrases of the original corpus sentence. The syntactic paraphrasing shared task comprises three tracks, described in the next section; in each case, the task is to automatically reproduce the relative preference judgments gathered during the human evaluation of the surface realization main task. As explained further below, developers of realization systems that can generate and optionally rank multiple outputs for a given input will be encouraged to participate in the task, which will test the system's ability to produce acceptable paraphrases and/or to rank competing realizations.

The objectives of the shared task are as follows:

broaden participation We expect developers of automatic quality metrics in the MT community to be interested in the proposed task, which is anticipated to be both more focused (with lexical choice largely excluded) and more challenging than in the MT case, given the generally high level of quality in realization results: as realization quality increases, the metrics' task becomes more difficult, since the paraphrases of a given sentence often involve subtle differences between acceptable and unacceptable variation. In an earlier study of the utility of automatic metrics with Penn Treebank (PTB) surface realization data (Espinosa et al., 2010), we observed moderate correlations between the most popular metrics and human judgments, though lower than the levels seen with MT data.

promote reuse of human judgments The task is intended to test the effectiveness of realization ranking models in a way that reuses human judgments, making it possible to carry out re-

INLG 2012 Proceedings of the 7th International Natural Language Generation Conference, pages 150–153,
Utica, May 2012.

Track	Reference Sentence	PTB Gold	PTB Auto
Realization Ranking	N	Y	N
Hybrid	Y	Y	N
Metrics Meta-Eval	Y	N	Y

Table 1: Additional inputs for the three realization tracks

producible system comparisons.

mitigate input conversion issues Realizer evaluations have typically focused on single-best outputs, where the depth and specificity of system inputs has a large impact on quality, making comparative evaluation difficult. While the surface realization shared task seeks to address this issue by developing common ground input representations, to date it has proved to be difficult to adapt existing systems to work with these inputs. By focusing on ranking paraphrases that are distinct from the reference sentence, the proposed task may provide a way to mitigate these issues, as discussed below.

2 Three Tracks: From Realization Ranking to Metrics Meta-Evaluation

We propose three tracks for the task, going from pure realization ranking to metrics meta-evaluation, with a hybrid case in the middle. For all three tracks, the input is a set of pairs of syntactic paraphrases (distinct from the reference sentence), and the output is the preferred member of each pair, where the goal is to match the human judgments of relative preference. The tracks differ in the additional inputs that systems may use in determining which member of each pair is preferred (see Table 1). In the realization ranking track, the task is to rank order the paraphrases for a given sentence, *without* having access to the reference sentence, using a realization ranking model. To do so, each system is allowed to use its own "native" inputs derived from the Penn Treebank and PTB-based resources. To the extent that a system's statistical ranking model can be used to assign a score to any possible realization, the ranking task can be accomplished by simply ranking the realizations by model score. As such, following this strategy, the task is one of **analysis by synthesis**.

For non-statistical realizers, or ones that cannot assign a score to any possible realization, there is an alternative strategy available, namely to **automatically approximate HTER**. Snover et al. (2006) demonstrate that the human-targeted translation edit rate (HTER) represents a reliable and easily interpretable method of evaluating MT output. With this method, a human annotator produces a targeted reference sentence which is as close as possible to the MT hypothesis while being fully acceptable; from the targeted reference, the TER score then represents a normalized post-edit score, which has been shown to correlate with human ratings at least as well as more complex competing metrics. As Madnani (2010) points out, generated paraphrases of the reference sentence can be used to approximate HTER scoring, as the closest acceptable paraphrase of a reference sentence should correspond to the version of the MT hypothesis with minimal changes to make it acceptable. Indeed, in the limit, it should be possible to use a system that can enumerate all and only the acceptable paraphrases of a reference sentence to fully implement HTER scoring.

Naturally, it is possible to combine the analysis-by-synthesis and approximating HTER strategies. One particularly simple way to do so is to (1) use an n-best list of realizations with normalized scores, (2) find the realization with the minimum TER score for each paraphrase to rank, then (3) combine the realizer's model score with the TER score, e.g. just by subtraction (weights for the combination could also be optimized using machine learning).

Regarding the issue of whether fair comparisons can be made when each system is allowed to use its own PTB-derived "native" input, note that it is unclear whether using shallow, specific inputs is necessarily advantageous for ranking a range of possible realizations, all distinct from the reference sentence: in the limit, a realizer input that completely specifies the reference sentence (and no other variants) is of no help at all, as in this case the approximating HTER strategy reduces to just doing TER scoring against the reference sentence.

Turning now to the metrics meta-evaluation track, here the the task is to rank order a set of realizations for a given sentence, starting with the reference sentence and *nothing else*. In principle, it should be possible to use any MT metric for this task off-the-

shelf. It should also be possible for realization systems to participate in this track, if they can be paired with a parser that produces inputs for the realizer, or a parser whose outputs can be converted to realizer inputs. To do so, strategies employed in the realization ranking track can be combined with ones that make use of the reference sentence.

Finally, between these two tracks is a hybrid track, where one is allowed to substitute automatic parses with gold parses. This track can be viewed as providing a way to estimate an upper bound on approaches that pay attention to how well a sentence expresses an intended meaning, while also arguably representing the most sensible way to automatically evaluate outputs in a data-to-text setting, where intended meanings can be reliably represented.

3 Pilot Experiments

In this section, we present two pilot experiments intended to demonstrate the feasibility of the task. The experiments use the human judgments collected in Espinosa et al.'s (2010) study, which consist of adequacy and fluency ratings from two judges for a variety of realizations for PTB Section 00. The realizations in the corpus were generated using several OpenCCG realization ranking models (White and Rajkumar, 2009) and using the XLE symbolic realizer with subsequent n-gram ranking (paraphrases involving WordNet substitutions were excluded). For comparison purposes, three well-known metrics (BLEU, METEOR and TER) were tested, along with three OpenCCG ranking models: (I) a generative baseline model, incorporating three n-gram models as well as Hockenmaier's (2003) generative model; (II) a model additionally incorporating a slew of discriminative features, extending White & Rajkumar's model with dependency ordering features; and (III) a model adding one additional feature for minimizing dependency length. Note that Models II and III are very similar, usually yielding the same single-best output, though occasionally differing in important ways; by contrast, both models represent a substantial refinement of Model I.

The two experiments investigate different strategies for approaching the hybrid task. The first experiment investigates the approximating-HTER strategy (with an analysis-by-synthesis component) using a 20-best list. For simplicity, edit rate (edit distance normalized by the number of words in the reference sentence) was used to find the realization in the 20-best list that was closest to the paraphrase to be ranked. The score for the paraphrase was then calculated by normalizing the realizer model score for the closest realization (linearly interpolating using the min and max scores across all 20-best lists), subtracting the edit rate, and adding in the metric score, for each of BLEU, METEOR and TER.[1] Since edit rate is less reliable than TER, as it overly penalizes phrasal shifts, the metric score was used alone in cases where the edit rate exceeded 0.5.

The results of the first experiment appear in Table 2. Human judgments were combined by averaging the summed adequacy and fluency ratings from each judge. Excluding exact match realizations, 2838 pairs of realizations with distinct combined scores (from approximately 250 sentences) were used to judge ranking accuracy. Here, BLEU substantially outperforms METEOR and TER, and combining Models I-III with BLEU does not yield significant differences in ranking accuracy. Note, however, that using TER scores rather than edit rate, and optimizing the way the model scores are combined with the TER score and BLEU score, could perhaps yield significant improvements. With METEOR and TER, combining the model score, edit rate and metric score in the simplest way does yield highly significant improvements. With the METEOR combination, Model II achieves a highly significant improvement over Model I, though in other cases, only trends are observed across models.

The second experiment investigates the analysis-by-synthesis strategy more directly. Here, the realizer's search was guided to reproduce each paraphrase where possible, with model scores then calculated where an exact match could be achieved. The results appear in Table 3 for 474 pairs with differing combined human judgments. The first column shows the ranking accuracy using the model scores by themselves; the subsequent columns compare the accuracy using BLEU, METEOR and TER against using the model score added to the metric score. Here we see from the first column that Model II substantially outperforms Model I, showing the

[1] TER scores were inverted for consistency.

	BLEU	Model+BLEU	METEOR	Model+METEOR	TER	Model+TER
Model I	71.2	70.2	58.6	65.4 (***)	59.7	68.7 (***)
Model II	-	70.8	-	66.7 (***† † †)	-	69.4 (***†)
Model III	-	**71.3** (†)	-	**67.1** (***)	-	**69.9** (***)

Table 2: Pairwise accuracy percentage on reproducing human judgments of relative adequacy plus fluency of syntactic paraphrases, using **n-best realizations** from three OpenCCG ranking models and minimum edit rate in combination with MT metrics (significance: * for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$ in comparison to MT metric, using McNemar's test; similarly for number of daggers in comparison to model in previous row)

	Model	BLEU	Model+BLEU	METEOR	Model+METEOR	TER	Model+TER
Model I	62.2	67.7	**73.0** (***)	49.2	65.4 (***)	50.6	73.8 (***)
Model II	**67.1** († † †)	-	72.2 (***)	-	68.6 (***††)	-	74.9 (***)
Model III	66.2	-	72.6 (***)	-	**68.8** (***)	-	**75.1** (***)

Table 3: Pairwise accuracy percentage on reproducing human judgments of relative adequacy plus fluency of syntactic paraphrases, using **exact targeted realizations** from three OpenCCG ranking models and minimum edit rate in combination with MT metrics (significance: * for $p < 0.1$, ** for $p < 0.05$, *** for $p < 0.01$ in comparison to MT metric, using McNemar's test; similarly for number of daggers in comparison to model in previous row)

ability of the ranking task to discriminate among models of varying sophistication, though the model differences are largely washed out when the model scores are combined with metric scores. In the subsequent columns, we see that METEOR and TER are only performing at chance (50%) on these particular ranking cases, while adding the model scores and metric scores does much better, with Model III plus TER performing the best overall, as might have been expected. Even with BLEU, which performs decently on its own, adding in the model scores achieves substantial (and highly significant) gains.

4 Task Organization

The proposed syntactic paraphrase ranking task is intended to be run as a straightforward extension of the main surface realization shared task. For development and training purposes, the human judgments collected for the first surface realization shared task will be made available; the data from Espinosa et al.'s study is already publicly available as well. For test data, the human judgments collected for evaluation during the second surface realization shared task will be used. Ideally enough systems will enter the main task to enable many pairwise comparisons per sentence, and enough judges can be employed to allow majority preferences to be used as the gold standard. As baselines for the metrics meta-eval and hybrid tracks, the BLEU, NIST, METEOR and TER metrics will be run by the organizers. Time permitting, a baseline system that works with n-best realization scores will also be made available, so that any developer of a realization system that can produce n-best outputs can easily participate.

Acknowledgments

This work was supported in part by NSF grant no. IIS-1143635. Thanks go to the anonymous reviewers for helpful comments and discussion.

References

Dominic Espinosa, Rajakrishnan Rajkumar, Michael White, and Shoshana Berleant. 2010. Further meta-evaluation of broad-coverage surface realization. In *Proc. of EMNLP-10*, pages 564–574.

Julia Hockenmaier. 2003. *Data and models for statistical parsing with Combinatory Categorial Grammar*. Ph.D. thesis, University of Edinburgh.

Nitin Madnani. 2010. *The Circle of Meaning: From Translation to Paraphrasing and Back*. Ph.D. thesis, University of Maryland.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proc. of AMTA-06*, pages 223–231.

Michael White and Rajakrishnan Rajkumar. 2009. Perceptron reranking for CCG realization. In *Proc. of EMNLP-09*, pages 410–419.