Textual Entailment Features for Machine Translation Evaluation

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

Abstract

We present two regression models for the prediction of pairwise preference judgments among MT hypotheses. Both models are based on feature sets that are motivated by *textual entailment* and incorporate lexical similarity as well as local syntactic features and specific semantic phenomena. One model predicts absolute scores; the other one direct pairwise judgments. We find that both models are competitive with regression models built over the scores of established MT evaluation metrics. Further data analysis clarifies the complementary behavior of the two feature sets.

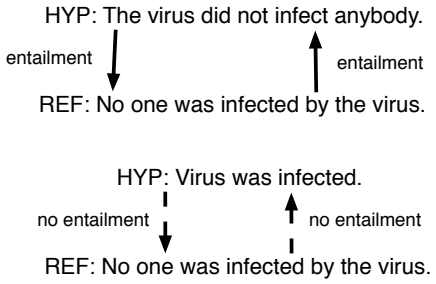

Figure 1: Entailment status between an MT system hypothesis and a reference translation for good translations (above) and bad translations (below).

1 Introduction

Automatic metrics to assess the quality of machine translations have been a major enabler in improving the performance of MT systems, leading to many varied approaches to develop such metrics. Initially, most metrics judged the quality of MT hypotheses by *token sequence match* (cf. BLEU (Papineni et al., 2002), NIST (Doddington, 2002). These measures rate systems hypotheses by measuring the overlap in surface word sequences shared between hypothesis and reference translation.

With improvements in the state-of-the-art in machine translation, the effectiveness of purely surface-oriented measures has been questioned (see e.g., Callison-Burch et al. (2006)). In response, metrics have been proposed that attempt to integrate more linguistic information into the matching process to distinguish linguistically licensed from unwanted variation (Giménez and Màrquez, 2008). However, there is little agreement on what types of knowledge are helpful: Some suggestions concentrate on *lexical* information, e.g., by the integration of word similarity information as in Meteor (Banerjee and Lavie, 2005) or MaxSim (Chan and Ng, 2008). Other proposals use *structural* information such as dependency edges (Owczarzak et al., 2007).

In this paper, we investigate an MT evaluation metric that is inspired by the similarity between this task and the *textual entailment* task (Dagan et al., 2005), which suggests that the quality of an MT hypothesis should be predictable by a *combination* of lexical and structural features that model the matches and mismatches between system output and reference translation. We use supervised regression models to combine these features and analyze feature weights to obtain further insights into the usefulness of different feature types.

2 Textual Entailment for MT Evaluation

2.1 Textual Entailment vs. MT Evaluation

Textual entailment (TE) was introduced by Dagan et al. (2005) as a concept that corresponds more closely to "common sense" reasoning than classical, categorical entailment. Textual entailment 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.*

Information about the presence or absence of entailment between two sentences has been found to be beneficial for a range of NLP tasks such as Word Sense Disambiguation or Question Answering (Dagan et al., 2006; Harabagiu and Hickl, 2006). Our intuition is that this idea can also be fruitful in MT Evaluation, as illustrated in Figure 1. Very good MT output should entail the reference translation. In contrast, missing hypothesis material breaks forward entailment; additional material breaks backward entailment; and for bad translations, entailment fails in both directions.

Work on the recognition of textual entailment (RTE) has consistently found that the integration of more syntactic and semantic knowledge can yield gains over

*This paper is based on work funded by the Defense Advanced Research Projects Agency through IBM. The content does not necessarily reflect the views of the U.S. Government, and no official endorsement should be inferred..

Proceedings of the 4th EACL Workshop on Statistical Machine Translation , pages 37–41,
Athens, Greece, 30 March – 31 March 2009.

surface-based methods, provided that the linguistic analysis was sufficiently robust. Thus, for RTE, "deep" matching outperforms surface matching. The reason is that linguistic representation makes it considerably easier to distinguish admissible variation (i.e., paraphrase) from true, meaning-changing divergence. Admissible variation may be lexical (synonymy), structural (word and phrase placement), or both (diathesis alternations).

The working hypothesis of this paper is that the benefits of deeper analysis carry over to MT evaluation. More specifically, we test whether the features that allow good performance on the RTE task can also predict human judgments for MT output. Analogously to RTE, these features should help us to differentiate meaning preserving translation variants from bad translations.

Nevertheless, there are also substantial differences between TE and MT evaluation. Crucially, TE assumes the premise and hypothesis to be well-formed sentences, which is not true in MT evaluation. Thus, a possible criticism to the use of TE methods is that the features could become unreliable for ill-formed MT output. However, there is a second difference between the tasks that works to our advantage. Due to its strict compositional nature, TE requires an accurate semantic analysis of all sentence parts, since, for example, one misanalysed negation or counterfactual embedding can invert the entailment status (MacCartney and Manning, 2008). In contrast, human MT judgments behave more additively: failure of a translation with respect to a single semantic dimension (e.g., polarity or tense) degrades its quality, but usually not crucially so. We therefore expect that even noisy entailment features can be predictive in MT evaluation.

2.2 Entailment-based prediction of MT quality

Regression-based prediction. Experiences from the annotation of MT quality judgments show that human raters have difficulty in consistently assigning absolute scores to MT system output, due to the number of ways in which MT output can deviate. Thus, the human annotation for the WMT 2008 dataset was collected in the form of *binary pairwise preferences* that are considerably easier to make (Callison-Burch et al., 2008). This section presents two models for the prediction of pairwise preferences.

The first model (ABS) is a regularized linear regression model over entailment-motivated features (see below) that predicts an absolute score for each reference-hypothesis pair. Pairwise preferences are created simply by comparing the absolute predicted scores. This model is more general, since it can also be used where absolute score predictions are desirable; furthermore, the model is efficient with a runtime linear in the number of systems and corpus size. On the downside, this model is not optimized for the prediction of pairwise judgments.

The second model we consider is a regularized logistic regression model (PAIR) that is directly optimized to predict a weighted binary preference for each hypothesis pair. This model is less efficient since its runtime is quadratic in the number of systems. On the other hand, it can be trained on more reliable pairwise preference judgments. In a second step, we combine the individual decisions to compute the highest-likelihood total ordering of hypotheses. The construction of an optimal ordering from weighted pairwise preferences is an NP-hard problem (via reduction of CYCLIC-ORDERING; Barzilay and Elhadad, 2002), but a greedy search yields a close approximation (Cohen et al., 1999).

Alignment score(3)	Unaligned material (10)
Adjuncts (7)	Apposition (2)
Modality (5)	Factives (8)
Polarity (5)	Quantors (4)
Tense (2)	Dates (6)
Root (2)	Semantic Relations (4)
Semantic relatedness (7)	Structural Match (5)
Compatibility of locations and entities (4)	

Table 1: Entailment feature groups provided by the Stanford RTE system, with number of features

Both models can be used to predict system-level scores from sentence-level scores. Again, we have two method for doing this. The basic method (BASIC) predicts the quality of each system directly as the percentage of sentences for which its output was rated best among all systems. However, we noticed that the manual rankings for the WMT 2007 dataset show a tie for best system for almost 30% of sentences. BASIC is systematically unable to account for these ties. We therefore implemented a "tie-aware" prediction method (WITHTIES) that uses the same sentence-level output as BASIC, but computes system-level quality differently, as the percentage of sentences where the system's hypothesis was scored *better or at most ε worse than the best system*, for some global "tie interval" ε.

Features. We use the Stanford RTE system (MacCartney et al., 2006) to generate a set of entailment features (RTE) for each pair of MT hypothesis and reference translation. Features are generated in both directions to avoid biases towards short or long translations. The Stanford RTE system uses a three-stage architecture. It (a) constructs a robust, dependency-based linguistic analysis of the two sentences; (b) identifies the best alignment between the two dependency graphs given similarity scores from a range of lexical resources, using a Markov Chain Monte Carlo sampling strategy; and (c) computes roughly 75 features over the aligned pair of dependency graphs. The different feature groups are shown in Table 1. A small number features are real-valued, measuring different quality aspects of the alignment. The other features are binary, indicating matches and mismatches of different types (e.g., alignment between predicates embedded under compatible or incompatible modals, respectively).

To judge to what extent the entailment-based model delivers improvements that cannot be obtained with established methods, we also experiment with a feature set

formed from a set of established MT evaluation metrics (TRADMT). We combine different parametrization of (smoothed) BLEU (Papineni et al., 2002), NIST (Doddington, 2002), and TER (Snover et al., 2006), to give a total of roughly 100 features. Finally, we consider a combination of both feature sets (COMB).

3 Experimental Evaluation

Setup. To assess and compare the performance of our models, we use corpora that were created by past instances of the WMT workshop. We optimize the feature weights for the ABS models on the WMT 2006 and 2007 absolute score annotations, and correspondingly for the PAIR models on the WMT 2007 absolute score and ranking annotations. All models are evaluated on WMT 2008 to compare against the published results.

Finally, we need to set the tie interval ε. Since we did not want to optimize ε, we simply assumed that the percentage of ties observed on WMT 2007 generalizes to test sets such as the 2008 dataset. We set ε so that there are ties for first place on 30% of the sentences, with good practical success (see below).

Results. Table 2 shows our results. The first results column (Cons) shows consistency, i.e., accuracy in predicting human pairwise preference judgments. Note that the performance of a random baseline is not at 50%, but substantially lower. This is due to (a) the presence of contradictions and ties in the human judgments, which cannot be predicted; and (b) WMT's requirement to compute a total ordering of all translations for a given sentence (rather than independent binary judgments), which introduces transitivity constraints. See Callison-Burch et al. (2008) for details. Among our models, PAIR shows a somewhat better consistency than ABS, as can be expected from a model directly optimized on pairwise judgments. Across feature sets, COMB works best with a consistency of 0.53, competitive with published WMT 2008 results.

The two final columns (BASIC and WITHTIES) show Spearman's ρ for the correlation between human judgments and the two types of system-level predictions.

For BASIC system-level predictions, we find that PAIR performs considerably worse than ABS, by a margin of up to $\rho = 0.1$. Recall that the system-level analysis considers only the top-ranked hypotheses; apparently, a model optimized on pairwise judgments has a harder time choosing the best among the top-ranked hypotheses. This interpretation is supported by the large benefit that PAIR derives from explicit tie modeling. ABS gains as well, although not as much, so that the correlation of the tie-aware predictions is similar for ABS and PAIR.

Comparing different feature sets, BASIC show a similar pattern to the consistency figures. There is no clear winner between RTE and TRADMT. The performance of TRADMT is considerably better than the performance of BLEU and TER in the WMT 2008 evaluation, where $\rho \leq 0.55$. RTE is able to match the performance of an ensemble of state-of-the-art metrics, which validates our hope that linguistically motivated entailment features are sufficiently robust to make a positive contribution in MT evaluation. Furthermore, the two individual feature sets are outperformed by the combined feature set COMB. We interpret this as support for our regression-based combination approach.

Model	Feature set	Cons (Acc.)	BASIC (ρ)	WITHTIES (ρ)
ABS	TRADMT	0.50	0.74	0.74
ABS	RTE	0.51	0.72	**0.78**
ABS	COMB	0.51	0.74	0.74
PAIR	TRADMT	0.52	0.63	0.73
PAIR	RTE	0.51	0.66	**0.77**
PAIR	COMB	0.53	0.70	**0.77**
WMT 2008 (worst)		0.44	0.37	
WMT 2008 (best)		0.56	0.83	

Table 2: Evaluation on the WMT 2008 dataset for our regression models, compared to results from WMT 2008

Moving to WITHTIES, we see the best results from the RTE model which improves by $\Delta\rho = 0.06$ for ABS and $\Delta\rho = 0.11$ for PAIR. There is less improvement for the other feature sets, in particular COMB. We submitted the two overall best models, ABS-RTE and PAIR-RTE with tie-aware prediction, to the WMT 2009 challenge.

Data Analysis. We analyzed at the models' predictions to gain a better understanding of the differences in the behavior of TRADMT-based and RTE-based models. As a first step, we computed consistency numbers for the set of "top" translations (hypotheses that were ranked highest for a given reference) and for the set of "bottom" translations (hypotheses that were ranked worst for a given reference). We found small but consistent differences between the models: RTE performs about 1.5 percent better on the top hypotheses than on the bottom translations. We found the inverse effect for the TRADMT model, which performs 2 points *worse* on the top hypotheses than on the bottom hypotheses. Revisiting our initial concern that the entailment features are too noisy for very bad translations, this finding indicates some ungrammaticality-induced degradation for the entailment features, but not much. Conversely, these numbers also provide support for our initial hypothesis that surface-based features are good at detecting very deviant translations, but can have trouble dealing with legitimate linguistic variation.

Next, we analyzed the average size of the score differences between the best and second-best hypotheses for correct and incorrect predictions. We found that the RTE-based model predicted on average almost twice the difference for correct predictions ($\Delta = 0.30$) than for incorrect predictions ($\Delta = 0.16$), while the difference was considerably smaller for the TRADMT-based model ($\Delta = 0.17$ for correct vs. $\Delta = 0.13$ for incorrect). We believe it is this better discrimination on the top hypothe-

Segment	TRADMT	RTE	COMB	Gold
REF: Scottish NHS boards need to improve criminal records checks for employees outside Europe, a watchdog has said. HYP: The Scottish health ministry should improve the controls on extra-community employees to check whether they have criminal precedents, said the monitoring committee. [1357, lium-systran]	Rank: 3	Rank: **1**	Rank: 2	Rank: 1
REF: Arguments, bullying and fights between the pupils have extended to the relations between their parents. HYP: Disputes, chicane and fights between the pupils transposed in relations between the parents. [686, rbmt4]	Rank: **5**	Rank: 2	Rank: 4	Rank: 5

Table 3: Examples of reference translations and MT output from the WMT 2008 French-English News dataset. Rank judgments are out of five (smaller is better).

ses that explains the increased benefit the RTE-based model obtains from tie-aware predictions: if the best hypothesis is wrong, chances are much better than for the TRADMT-based model that counting the second-best hypothesis as "best" is correct. Unfortunately, this property is not shared by COMB to the same degree, and it does not improve as much as RTE.

Table 3 illustrates the difference between RTE and TRADMT. In the first example, RTE makes a more accurate prediction than TRADMT. The human rater's favorite translation deviates considerably from the reference translation in lexical choice, syntactic structure, and word order, for which it is punished by TRADMT. In contrast, RTE determines correctly that the propositional content of the reference is almost completely preserved. The prediction of COMB is between the two extremes. The second example shows a sentence where RTE provides a worse prediction. This sentence was rated as bad by the judge, presumably due to the inappropriate translation of the main verb. This problem, together with the reformulation of the subject, leads TRADMT to correctly predict a low score (rank 5/5). RTE's deeper analysis comes up with a high score (rank 2/5), based on the existing semantic overlap. The combined model is closer to the truth, predicting rank 4.

Feature Weights. Finally, we assessed the importance of the different entailment feature groups in the RTE model.[1] Since the presence of correlated features makes the weights difficult to interpret, we restrict ourselves to two general observations.

First, we find high weights not only for the score of the alignment between hypothesis and reference, but also for a number of syntacto-semantic match and mismatch features. This means that we do get an additional benefit from the presence of these features. For example, features with a negative effect include dropping adjuncts, unaligned root nodes, incompatible modality between the main clauses, person and location mismatches (as opposed to general mismatches) and wrongly handled passives. Conversely, some factors that increase the prediction are good alignment, matching embeddings under factive verbs, and matches between appositions.

[1]The feature weights are similar for the COMB model.

Second, we find clear differences in the usefulness of feature groups between MT evaluation and the RTE task. Some of them, in particular structural features, can be linked to the generally lower grammaticality of MT hypotheses. A case in point is a feature that fires for mismatches between dependents of predicates and which is too unreliable on the SMT data. Other differences simply reflect that the two tasks have different profiles, as sketched in Section 2.1. RTE exhibits high feature weights for quantifier and polarity features, both of which have the potential to influence entailment decisions, but are relatively unimportant for MT evaluation, at least at the current state of the art.

4 Conclusion

In this paper, we have investigated an approach to MT evaluation that is inspired by the similarity between this task and textual entailment. Our two models – one predicting absolute scores and one predicting pairwise preference judgments – use entailment features to predict the quality of MT hypotheses, thus replacing surface matching with syntacto-semantic matching. Both models perform similarly, showing sufficient robustness and coverage to attain comparable performance to a committee of established MT evaluation metrics.

We have described two refinements: (1) combining the features into a superior joint model; and (2) adding a confidence interval around the best hypothesis to model ties for first place. Both strategies improve correlation; however, unfortunately the benefits do not currently combine. Our feature weight analysis indicates that syntacto-semantic features do play an important role in score prediction in the RTE model. We plan to assess the additional benefit of the full entailment feature set against the TRADMT feature set extended by a proper lexical similarity metric, such as METEOR.

The computation of entailment features is more heavyweight than traditional MT evaluation metrics. We found the speed (about 6 s per hypothesis on a current PC) to be sufficient for easily judging the quality of datasets of the size conventionally used for MT evaluation. However, this may still be too expensive as part of an MT model that directly optimizes some performance measure, e.g., minimum error rate training (Och, 2003).

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*, pages 65–72, Ann Arbor, MI.

R. Barzilay and N. Elhadad. 2002. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, 17:35–55.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proceedings of EACL*, pages 249–256, Trento, Italy.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 70–106, Columbus, OH.

Yee Seng Chan and Hwee Tou Ng. 2008. MAXSIM: A maximum similarity metric for machine translation evaluation. In *Proceedings of ACL-08: HLT*, pages 55–62, Columbus, Ohio.

William W. Cohen, Robert E. Schapire, and Yoram Singer. 1999. Learning to order things. *Journal of Artificial Intelligence Research*, 10:243–270.

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

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

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

Jesús Giménez and Lluís Màrquez. 2008. A smorgasbord of features for automatic MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 195–198, Columbus, Ohio.

Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of ACL*, pages 905–912, Sydney, Australia.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of Coling*, pages 521–528, Manchester, UK.

Bill MacCartney, 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*, pages 41–48, New York City, NY.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of ACL*, pages 160–167, Sapporo, Japan.

Karolina Owczarzak, Josef van Genabith, and Andy Way. 2007. Dependency-based automatic evaluation for machine translation. In *Proceedings of the NAACL-HLT / AMTA Workshop on Syntax and Structure in Statistical Translation*, pages 80–87, Rochester, NY.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of ACL*, pages 311–318, Philadelphia, PA.

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