Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation

Lucia Specia
Research Group in
Computational Linguistics,
University of Wolverhampton,
Wolverhampton, UK
l.specia@wlv.ac.uk

Jesús Giménez
TALP Research Center,
LSI Department,
Universitat Politècnica de Catalunya,
Barcelona, Spain
jgimenez@lsi.upc.edu

Abstract

We describe an effort to improve standard reference-based metrics for Machine Translation (MT) evaluation by enriching them with Confidence Estimation (CE) features and using a learning mechanism trained on human annotations. Reference-based MT evaluation metrics compare the system output against reference translations looking for overlaps at different levels (lexical, syntactic, and semantic). These metrics aim at comparing MT systems or analyzing the progress of a given system and are known to have reasonably good correlation with human judgments at the corpus level, but not at the segment level. CE metrics, on the other hand, target the system in use, providing a quality score to the end-user for each translated segment. They cannot rely on reference translations, and use instead information extracted from the input text, system output and possibly external corpora to train machine learning algorithms. These metrics correlate better with human judgments at the segment level. However, they are usually highly biased by difficulty level of the input segment, and therefore are less appropriate for comparing multiple systems translating the same input segments. We show that these two classes of metrics are complementary and can be combined to provide MT evaluation metrics that achieve higher correlation with human judgments at the segment level.

1 Introduction

Machine Translation (MT) evaluation metrics are essential for system development and system comparison. The most commonly used metrics like BLEU (Papineni et al., 2002) and NIST (Doddington, 2002) are based on reference translations to compute some form of overlap between n-grams in the MT system output and one or more human translations. More complex reference-based metrics replace or complement n-gram matching with alternative lexical features, such as lemma- or synonym-based alignment between the machine and human translations (Lavie and Agarwal, 2007), or sometimes use syntactic and semantic features, such as the matching of syntactic constituents, dependency relations or semantic roles (Liu and Gildea, 2005; Giménez and Màrquez, 2010b).

Although evaluation campaigns have shown that such metrics correlate reasonably well with human judgments at the corpus level, their correlation at the segment level (e.g. sentences) is usually much lower. While recent metrics like METEOR (Lavie and Agarwal, 2007) and TER (Snover et al., 2006a) try to overcome this limitation, segment-level scoring is still a limitation, particularly for the *de facto* metrics BLEU and NIST. Moreover, the scores given by existing metrics usually cannot be interpreted in absolute terms. For example, it is difficult to reason about a BLEU score of 0.35 in terms of translation quality, as such a score heavily depends on the corpus used for the evaluation (size, distribution of n-grams, etc.), the number of reference translations available and the type of MT system (rule-based, statistical, etc.), among other factors.

Confidence Estimation (CE) metrics, on the other hand, aim at providing a score to end-users of MT systems for each translated segment. End-users may include people using web-based systems to get the gist of text and professional translators using commercial systems to aid them to produce publishable quality translations. In our experimental setup, CE metrics estimate a quality indicator within a given numeric range, as opposed to binary "bad" / "good" judgments used in most previous work on Confidence Estimation. Therefore, an absolute numeric estimate is provided to the user, which can be directly interpreted according to a given task (translation post-editing, for example).

CE metrics cannot rely on reference translations, since unseen texts will usually be given to the system for translation. These metrics use features extracted given only the source and translation text, and optionally monolingual and bilingual corpora or information about the MT system used to produce the translations. Such features are given to a machine learning algorithm in order to learn a model to predict quality estimates for a certain language pair from data annotated with automatic scores (Blatz et al., 2004) or directly from data annotated with human scores (Quirk, 2004; Specia et al., 2009). CE metrics have been shown to correlate significantly better with human evaluation than standard metrics like BLEU and NIST (Specia et al., 2010b). Since by definition CE metrics are aimed at estimating the quality of a particular MT system for the translation of a given input segment, they are heavily dependent on features regarding the input segment, that is, features reflecting the difficulty of translating the source segment. Therefore, they are not very suitable for comparing multiple MT systems translating the same input segments.

While the two types of metrics have different objectives, we believe that the advantages of both can be exploited by combining them. We aim to enrich reference-based MT evaluation metrics by using the learning framework of CE metrics, as well as the reference-independent features used by such metrics, in order to improve the correlation of MT evaluation metrics with human judgments at the segment level, where a segment corresponds to a single sentence.

In general, the goal of automatic evaluation metrics is to approximate human judgments. We exploit a learning framework to directly estimate human scores, instead of estimating scores that correlate well with them. The type of human scores used for the annotation and to be predicted by the system will depend on the aspects of quality which are relevant for a given task. For example, one can annotate translations according to their post-editing needs, fluency or adequacy.

We have used recent developments in both types of metrics: the CE framework trained on human annotations, as proposed by Specia et al. (2009) and the evaluation metric combining a number of standard metrics like BLEU, NIST, METEOR and linguistic features, as proposed by Giménez and Màrquez (2010b). In the remaining of the paper we first refer to related work on MT evaluation and CE (Section 2), then give more details about the CE (Section 3) and evaluation (Section 4) metrics used, and report the experiments combining both (Section 5).

2 Related Work

While the combination of MT evaluation and CE metrics has not been attempted before, a number of previous efforts address related issues: using machine learning algorithms and human annotated data for MT evaluation, combining different MT evaluation metrics, using source-dependent features for MT evaluation and attempting to improve MT evaluation at the sentence-level.

The first attempt to tackle sentence-level MT evaluation as a learning problem was proposed by Corston-Oliver et al. (2001). A classifier is trained to distinguish between human translations (presumably good) and MT system translations (presumably bad) at the sentence level (*human-likeness classification*). Reference translations are used as examples of good translations, and machine translations as examples of bad translations. A number of language model and linguistic features are extracted based on the translations and/or references, including branching properties of the parser, function word density, etc. Similarly, Kulesza and Shieber (2004) and Gamon et al. (2005) use a number of reference-based features to predict human-likeness. While this approach has the advantage of not requiring human annotation, the predictions obtained have very low correlation with human judgments, which is an indication, as shown in (Albrecht and Hwa, 2007a), that high human-likeness does not necessarily imply good MT quality and vice-versa.

Albrecht and Hwa (2007a) use a regression algorithm with string-based and syntax-based features extracted from MT output, reference translations and target language corpus to improve sentence-level MT evaluation. Albrecht and Hwa (2007b; 2008) rely instead on *pseudo-references*, which are translations produced by other MT systems. The training is performed based on 1-5 human judgments for translation fluency and adequacy. This approach is the most closely related to ours, but it does not exploit source dependent and other CE features.

Padó et al. (2009) use a regression algorithm with features motivated by textual entailment between the translation and the reference sentences, along with lexical similarity and other linguistic features to predict pairwise preference judgments among MT hypotheses. Source-dependent or other CE features are not used.

Liu and Gildea (2007) exploit features that constrain the reference-based n-grams matchings according to the input segments. For example, they constrain the matching of words in the reference and MT output to those cases which are aligned to the same words in the source sentence. The features are combined using a learning framework trained to maximize the Pearson correlation of the combination of features with human judgments. Source features which are independent from the reference translations are not used.

To the best of our knowledge, the approach presented in this paper is the first to use a learning framework based on human annotation with an enriched feature set derived from the confidence estimation scenario.

3 Confidence Estimation Metrics

The CE framework used in this paper is similar to that proposed by Specia et al. (2009), with an alternative learning algorithm (Support Vector Machines (SVM) as opposed to Partial Least Squares (PLS)) and without explicit feature selection. The choice of the algorithm was motivated by practical reasons, since PLS requires more training steps for explicit feature selection, while SVM is able to weight features appropriately according to their relevance as part of the model learning process. We use the following implementation of SVM for regression in our experiments: epsilon-SVR algorithm with radial basis function kernel from the LIBSVM package (Chang and Lin, 2001), with the parameters γ, ϵ and *cost* optimized.

In order to perform the task of CE across different MT systems and language-pairs, Specia et al. (2009) define a number of shallow, language- and MT system-independent features, extracted from the input (source) sentences and their corresponding translation (target) sentences, and also monolingual and parallel corpora. The set of 74 features used in this paper, grouped here for space reasons, is the following:

- source & target sentence lengths and their ratios
- source & target sentence type/token ratio
- average source word length
- average number of occurrences of all target words within the target sentence
- source & target sentence 3-gram language model probabilities and perplexities obtained using large monolingual corpora
- target sentence 3-gram language model probability trained on a corpus of POS-tags of words
- percentage of 1 to 3-grams in the source sentence belonging to each frequency quartile of a large monolingual corpus
- alignment score (IBM Model 4) for source and target sentences and percentage of different types of word alignments, as given by GIZA++ (Och and Ney, 2003) using a large parallel corpus ($\sim$1.2 million sentences)
- average number of translations per source word in the sentence (as given by probabilistic dictionaries like IBM Model 1), unweighted or weighted by the (inverse) frequency of the words
- percentages of numbers, content- / non-content words in the source & target sentences
- number of mismatching opening/closing brackets and quotation marks in the target sentence
- percentages and number of mismatches of each of the following superficial constructions between the source and target sentences: brackets, punctuation symbols, numbers.

The datasets used to train the CE system and the process to annotate them are described in Section 5.

4 Reference-based Evaluation Metrics

For reference-based metrics, we rely on the repository of metrics available as part of the ASIYA Toolkit (Giménez and Màrquez, 2010a)[1]. This includes a rich set of n-gram-based metrics and metrics operating at different linguistic levels (lexical, syntactic and semantic). Linguistic metrics have been shown to produce more reliable system rankings than standard n-gram based metrics, especially when the systems under evaluation are of different natures (Giménez and Màrquez, 2007). They have also performed well in recent evaluation campaigns (Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010). Moreover, they have been shown to present a high degree of complementarity with lexical metrics. Some of the linguistic metrics suffer a substantial decrease as sentence-level quality predictors, mainly due to parsing errors. Therefore, better results are usually achieved by combining n-gram-based and linguistic metrics (Giménez and Màrquez, 2010b). A drawback of linguistic metrics is that they rely on automatic linguistic processors and are, therefore, language dependent and in general much slower to compute than n-gram based metrics.

For our experiments we have selected a representative set of 52 metrics. All these metrics are available for translations into English (datasets described in Section 5.2), however, only 28 of them are available for translations into Spanish (datasets described in Section 5.1). We denote by '$\dagger_e$' the metrics available for translations into English only. In the following, we provide a brief description of the metrics grouped according to the linguistic level at which they operate.

4.1 Lexical Similarity

BLEUs (Papineni et al., 2002) Smoothed cumulative 4-gram BLEU score as described by Lin and Och (2004b).

NIST (Doddington, 2002) Default cumulative 5-gram NIST score.

GTM (Melamed et al., 2003) Three variants of GTM taking different values of the e parameter ($e \in \{1, 2, 3\}$) weighting the importance of the matching length.

METEOR (Denkowski and Lavie, 2010) Four variants of METEOR 1.2:

- $\mathbf{METEOR_{ex}}$ → only exact matchings.
- $\mathbf{METEOR_{st}}$ → stem matching.
- $\mathbf{METEOR_{sy}}\dagger_e$ → synonym matching.
- $\mathbf{METEOR_{pa}}$ → paraphrase matching.

ROUGE (Lin and Och, 2004a). Four variants of ROUGE:

- $\mathbf{ROUGE}_{L}$ → longest common subsequence (LCS).
- $\mathbf{ROUGE}_{S\star}$ → skip bigrams with no max-gap-length.
- $\mathbf{ROUGE}_{SU\star}$ → skip bigrams with no max-gap-length, including unigrams.
- $\mathbf{ROUGE}_{W}$ → weighted longest common subsequence (WLCS) with weighting factor $w = 1.2$.

WER (Word Error Rate) (Nießen et al., 2000) We use $-$WER to make this into a precision metric.

PER (Position-independent Word Error Rate) (Tillmann et al., 1997) We use $-$PER.

TER (Translation Edit Rate) (Snover et al., 2006b) Four variants of $-$TER:

- **TER**→ default (i.e., no paraphrases).
- $\mathbf{TER_{base}}$ → base (i.e., no stemming, no synonymy, no paraphrases).
- $\mathbf{TER_{p}}\dagger_e$ → with phrase substitutions.
- $\mathbf{TER_{pA}}\dagger_e$ → tuned towards adequacy.

$\boldsymbol{O_l}$ (Lexical overlap) (Giménez and Màrquez, 2010b). This metric is a particular instance of a more general *Overlap* measure. System and reference translations are considered as unordered sets of linguistic elements with repetition. *Overlap* is then defined as the Jaccard index between the two sets, i.e., the cardinality of their intersection divided by the cardinality of their union. In the case of lexical overlap, linguistic elements are word forms. Several metrics based on computing overlap at other linguistic levels are listed in this section.

[1] http://www.lsi.upc.edu/~nlp/Asiya

4.2 Syntactic Similarity

On Shallow Parsing (SP)

SP-$O_p(\star)$ Average overlap between words belonging to the same part-of-speech.

SP-$O_c(\star)$ Average overlap between words belonging to chunks of the same type.

SP-NIST$_{l|p|c|iob}$ NIST score over sequences of: lemmas (l), parts of speech (p), base phrase chunks (c), and chunk labels (iob).

On Dependency Parsing (DP)$\dagger_e$

DP-HWC$_l$ Head-word chain matching (Liu and Gildea, 2005). Only chains up to length 4 are considered. We use three different variants according to the item type:

DP-HWC$_w$ word forms.

DP-HWC$_c$ grammatical categories.

DP-HWC$_r$ grammatical relations.

DP-$O_l(\star)$ Average lexical overlap between items according to their tree level.

DP-$O_c(\star)$ Average lexical overlap between terminal nodes according to their grammatical category.

DP-$O_r(\star)$ Average lexical overlap between items according to their grammatical relationship.

On Constituency Parsing (CP)

CP-$O_p(\star)$ Average overlap between words belonging to the same part-of-speech.

CP-$O_c(\star)$ Average overlap between words belonging to constituents of the same type.

CP-STM$_d$ Syntactic tree matching (Liu and Gildea, 2005). We use three different variants respectively considering subtrees up to depth 4, 5 and 6.

4.3 Semantic Similarity

On Named Entities (NE)$\dagger_e$

NE-$O_e(\star)$ Lexical overlap between NEs of the same type.

NE-$M_e(\star)$ Lexical matching between NEs of the same type. Matching differs from overlap in that it requires the matching of the full linguistic element, whereas overlap considers partial matchings as well.

On Semantic Roles (SR)$\dagger_e$

SR-$O_r(\star)$ Average lexical overlap between SRs of the same type.

SR-$M_r(\star)$ Average lexical matching between SRs of the same type.

SR-O_r Average role overlap, i.e., overlap between semantic roles independently from their lexical realization.

We also use a more restrictive variant of these metrics which requires SRs to be associated to the same verb: SR-$O_{rv}(\star)$, SR-$M_{rv}(\star)$ and SR-O_{rv}.

On Discourse Representations (DR)$\dagger_e$

DR-$O_r(\star)$ Average lexical overlap between DR structures of the same type.

DR-$O_{rp}(\star)$ Average overlap between part-of-speech tags associated to lexical items in DR structures of the same type.

DR-STM$_d$ This metric is analogous to the CP-STM metric, but applied to DR trees. We use three variants considering subtrees up to depth 4, 5 and 6.

4.4 Optimal Metric Combinations

We combine linguistic metrics using the ULC approach, i.e., taking their normalized arithmetic mean. Optimal metric combinations are determined by maximizing Pearson correlation with human assessments as described by Giménez and Màrquez (2010b). The optimal combinations found are shown in Table 1, where metrics for each dataset are sorted according to their individual correlation. For all datasets with translations into English, it was possible to find metric combinations that outperform any individual metric. These include lexical, syntactic and semantic metrics.

5 Combining Confidence Estimation and Reference-based Evaluation Metrics

We experiment with the following strategies to combine the Confidence Estimation (CE) and Reference-based Evaluation (RE) metrics:

CE+RE (SVM) Join all CE features and RE metrics together as features and train an SVM regressor based on human annotations.

CE+ULC (SVM) Join all CE features and and the optimal metric set suggested by ULC as features and train an SVM regressor based on human annotations.

We compare these strategies against the following alternatives:

CE (SVM) The CE framework on its own, trained on all CE features using an SVM regressor based on human annotations.

RE (SVM) All RE metrics as features to train an SVM regressor based on human annotations.

ULC (SVM) All ULC metrics as features to train an SVM regressor based on human annotations.

RE (linear) The linear combination of all RE metrics (their normalized average).

ULC (linear) The linear combination of the best RE metrics.

BLEU, NIST, METEOR and TER Standard MT evaluation metrics.

We experiment with these metrics on two types of datasets for different language pairs and text domains:

- Large sets of English→Spanish translations for Europarl data annotated by professional translators (Section 5.1), and
- Small sets of {German, Spanish, French}→English translations for news data annotated by volunteers as part of an evaluation campaign (Section 5.2).

We measure the performance of each metric/combination by its Pearson correlation with the scores given by human annotators. In what follows we give details about the two types of datasets and present the results of our experiments.

Dataset	Optimal Metric Set
de-en	ROUGE$_W$, CP-STM$_6$, DP-$O_r(\star)$, DR-STM$_6$
es-en	GTM$_3$, DR-STM$_4$, DP-HWC$_r$
fr-en	GTM$_3$, DP-$O_r(\star)$, CP-STM$_6$
en-es	GTM$_2$

Table 1: Optimal metric combinations using the ULC approach.

5.1 LSP English→Spanish Translations

Four datasets were produced in a controlled environment as part of a project with a Language Service Provider (LSP). Each dataset consist of 4,000 Spanish translations for English sentences taken from the Europarl development and test sets provided by WMT08 (Callison-Burch et al., 2008). The translations were produced by training four Statistical MT (SMT) systems on 1.2 million English-Spanish sentence pairs from the Europarl training corpus as also provided by WMT08: Matrax (Simard et al., 2005), Portage (Johnson et al., 2006), Sinuhe (Kääriäinen, 2009) and MMR (Maximum Margin Regression) (Saunders, 2008). In the following we anonymize these systems by arbitrarily naming them S1-S4.

The translations produced by each system were manually annotated by professional translators with 1-4 scores, which is a range commonly used by them to indicate the quality of translations with respect to the need for post-editing[2]:

- 1 = requires complete retranslation
- 2 = post editing quicker than retranslation
- 3 = little post editing needed
- 4 = fit for purpose

The resulting datasets consist of four sets of $4,000$ distinct {*source, translation, reference, human-score*} quadruples. The distribution of the human scores assigned varies from dataset to dataset. The average scores are: S1 = 2.835, S2 = 2.558, S3 = 2.508 and S4 = 1.338. More details about these datasets, along with the actual datasets for download, can be found in (Specia et al., 2010a).

Each dataset was randomly split into training (3,000) and test (900) using a uniform distribution. Identical samples were created for all datasets. The optimization of the SVM parameters was performed by cross-validation using five random subsamples of the training set (75% for validation training and 25% for validation test).

5.1.1 Results

Table 2 shows the results of our combination strategies compared against other metrics. The two combination strategies, particularly **CE+RE (SVM)**, consistently outperform all other metrics, especially

[2] A *Kappa* agreement (Cohen, 1960) of 0.65 was obtained.

those metrics which do not use machine learning and human annotations.

	S1	S2	S3	S4
CE+RE (SVM)	**0.608**	**0.591**	**0.582**	**0.540**
CE+ULC (SVM)	**0.597**	0.572	0.568	**0.540**
CE (SVM)	0.577	0.557	0.536	0.536
RE (SVM)	0.510	0.492	0.503	0.459
ULC (SVM)	0.417	0.399	0.442	0.414
RE (linear)	0.274	0.261	0.340	0.206
ULC (linear)	0.417	0.399	0.442	0.412
$\mathrm{METEOR}_{\mathbf{pa}}$	0.270	0.302	0.350	0.256
BLEUs	0.295	0.277	0.339	0.223
NIST	0.197	0.189	0.253	0.124
TER	0.193	0.171	0.267	0.144
Avg. Human	2.868	2.583	2.526	1.344

Table 2: Results of the experiments with the LSP datasets in terms of Pearson correlation with human scores at the sentence level. Figures in bold face represent the best results for a given dataset, where difference to the second best approach is statistically significant (paired t-test, $p < 0.05$). The average human scores on the test set is also given to provide and intuition on the overall quality of the translations (in [1-4]).

The gain in performance obtained by the combinations of CE and RE as compared to these metrics individually shows that they are indeed complementary. An interesting outcome is the difference in the performance of the linear combination of RE metrics (**RE (linear)**) against their combination using SVM trained on human annotation (**RE (SVM)**). The performance of the linear combination of RE metrics is considerably lower, close to the standard evaluation metrics. This may be partially due to the low quality of the resources used to produce the linguistic features for Spanish, but it shows that using learning framework is a more robust approach. The subset **ULC (linear)** is superior to all RE metrics together, which was expected since ULC is defined in terms of correlation with human judgments. There is no gain in using the learning framework **ULC (SVM)** in this case, since ULC is composed by a single metric.

It is worth noticing that the experiments with these four datasets constitute the ideal scenario for confidence estimation, since the machine learning algorithm is trained on translations from a single MT system at a time or, more specifically, given that features the input segments are not repeated within each dataset. This explains the considerably superior performance of the approaches using CE features. In what follows we present a scenario which is closer to that of MT evaluation for system comparison, where different MT systems are used to translate the same input segments.

5.2 WMT {Spanish, French, German}→English Translations

As an alternative type of dataset, we collected WMT09 (Callison-Burch et al., 2009) English translations of news texts from German (*de-en*), Spanish (*es-en*) and French (*fr-en*) produced by a number of MT systems, which had been annotated by humans according to post-editing needs:

- 1 (BAD) = the sentence is too bad to edit
- 2 (EDIT) = the sentence can be edited
- 3 (OK) = the sentence does not require editing

The actual systems producing the translations vary according to the language pair, and they include SMT as well as rule-based and hybrid systems: 21 *de-en* MT systems, 13 *es-en* MT systems, and 21 *fr-en* MT systems. In total, 100 different source sentences for each language pair were translated by one or more MT system and annotated by humans. Some translations were annotated more than once to check (inter- and intra-) annotator agreement. In those cases, the multiple human scores were averaged.

The number of distinct translations annotated for each system varies from 37 to 56, with most systems ranging between 40 and 50 annotated translations. Since these numbers are too small for training our learning framework, we put together translations produced by all MT systems for a given language pair. The resulting datasets consist of three sets of distinct {*source, translation, reference, human-score*} quadruples:

- 1,012 quadruples for *de-en*
- 645 quadruples for *es-en* and
- 974 quadruples for *fr-en*

The distribution of human scores also varies according to the dataset, but the average scores in the three datasets is very similar: *de-en* = 1.87, *es-en* = 1.82 and *fr-en* = 1.94.

Each dataset was randomly split into training (80%) and test (20%) using a uniform distribution. The SVM parameters were optimized by cross-validation using five random subsamples of the training set (75% for validation training and 25% for validation test).

It is worth noticing that these WMT datasets differ from the LSP datasets (Section 5.1) in many aspects. Mainly, they contain fewer {*source, translation, reference, human-score*} quadruples, even though translations from several MT systems were put together. Moreover, each dataset contains multiple translations produced by different MT systems for the same source sentence, and therefore the source sentence features are repeated many times. This is not an ideal scenario for CE, given that many features are extracted from the source sentence only, while others depend somehow on the source sentence (about 60% of the features). Finally, these datasets were annotated by volunteers who were not trained for the annotation task and were not necessarily fluent speakers of both languages. This is reflected in the low agreement between the annotators mentioned in the WMT09 report (Callison-Burch et al., 2009).

5.2.1 Results

Table 3 shows the results of our combination strategies compared against other metrics. The results for these datasets also show that combining CE and RE metrics is beneficial, outperforming other metrics. The use of the learning framework as opposed to linear combinations also consistently yields superior results. However, RE metrics clearly play a more important role with these datasets, since there are many source-dependent CE features which are the same across different translations. In particular, for the *de-en* dataset, RE features combined using SVM, i.e., **RE (SVM)**, performs as well as **CE+RE (SVM)**. The effectiveness of reference-based metrics is also shown by the relatively high scores obtained by their linear combination (**RE (linear)**) as well as individually, especially METEOR.

A concern with this dataset is the reliability of the human annotation. As we mentioned before, the annotation was not performed by trained translators, and therefore it is likely to be much less consistent than that of the LSP datasets. Additionally, while the scenario of the experiments with the WMT datasets is closer to that of MT evaluation, the fact that the datasets for different MT systems were put together makes CE features much less likely to contribute to the task. Ideally, a model for each MT system should be learned individually. The scores estimated for multiple translations for a given input segment (produced by different MT systems) could then be contrasted against each other for system comparison.

	de-en	es-en	fr-en
CE+RE (SVM)	**0.480**	0.334	**0.315**
CE+ULC (SVM)	0.437	**0.379**	0.238
CE (SVM)	0.356	0.319	0.210
RE (SVM)	**0.479**	0.292	0.306
ULC (SVM)	0.428	0.119	0.202
RE (linear)	0.348	0.142	0.227
ULC (linear)	0.418	0.213	0.216
METEOR$_{\mathbf{pa}}$	0.298	0.090	0.197
BLEUs	0.220	0.167	0.138
NIST	0.227	0.038	0.180
TER	0.239	0.129	0.213
Avg. Human	1.854	1.878	1.950

Table 3: Results of the experiments with the WMT-09 datasets in terms of Pearson correlation with human scores at the sentence level. Figures in bold face represent the best results for a given dataset, where difference to the second best approach is statistically significant (paired t-test, $p < 0.05$). The average human scores on the test set is also given to provide and intuition on the overall quality of the translations (in [1-3]).

6 Conclusions

We have presented an approach for MT evaluation in which recent metrics are enriched with features from confidence estimation and a learning mechanism based on human annotations. The proposed metric showed improved correlation with human judgments at the segment level with several datasets.

While the proposed approach requires human annotation to learn models to predict a quality score, previous work has shown that confidence estimation can achieve good performance with a reasonably small number of training examples.

Another drawback of our approach is the dependency of some of our metrics on linguistic resources. This poses a limitation on their applicability to other language pairs, as well as to their use in other tasks such as system optimization, since computing such metrics requires more time. In particular, for system optimization using standard methods, the use of CE features is also problematic, since the variations in the n-best list may not be large enough to be reflected by the features we use.

An important remark is that such approach can be much more flexible than standard evaluation metrics with respect to the aspect of translation quality under evaluation. BLEU and NIST for example are known to better reflect fluency aspects. The proposed approach allows estimating different aspects of quality, depending on features extracted and the way the human annotation is performed.

Acknowledgments

Jesús Giménez is funded by the Spanish Government (project OpenMT-2, TIN2009-14675-C03).

References

Joshua Albrecht and Rebecca Hwa. 2007a. A re-examination of machine learning approaches for sentence-level mt evaluation. In *45th Meeting of the Association for Computational Linguistics*, pages 880–887, Prague.

Joshua Albrecht and Rebecca Hwa. 2007b. Regression for sentence-level mt evaluation with pseudo references. In *45th Meeting of the Association for Computational Linguistics*, pages 296–303, Prague.

Joshua Albrecht and Rebecca Hwa. 2008. The role of pseudo references in MT evaluation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 187–190.

J. Blatz, E. Fitzgerald, G. Foster, S. Gandrabur, C. Goutte, A. Kulesza, A. Sanchis, and N. Ueffing. 2004. Confidence Estimation for Machine Translation. In *20th Coling*, pages 315–321, Geneva.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *3rd Workshop on Statistical Machine Translation*, pages 70–106, Columbus.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *4th Workshop on Statistical Machine Translation*, pages 1–28.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. 2010. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, April.

Simon Corston-Oliver, Michael Gamon, and Chris Brockett. 2001. A Machine Learning Approach to the Automatic Evaluation of Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147.

Michael Denkowski and Alon Lavie. 2010. Meteor-next and the meteor paraphrase tables: Improved evaluation support for five target languages. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 339–342, July.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. In *2nd Conference on Human Language Technology Research*, pages 138–145, San Diego.

Michael Gamon, Anthony Aue, and Martine Smets. 2005. Sentence-level MT evaluation without reference translations: beyond language modeling. In *10th Meeting of the European Association for Machine Translation*, Budapest.

Jesús Giménez and Lluís Màrquez. 2007. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 256–264.

Jesús Giménez and Lluís Màrquez. 2010a. ASIYA: An Open Toolkit for Automatic Machine Translation Evaluation. *To Appear in The Prague Bulletin of Mathematical Linguistics*.

Jesús Giménez and Lluís Màrquez. 2010b. Linguistic Features for Automatic MT Evaluation. *To Appear in Machine Translation*.

Howard Johnson, F. Sadat, George Foster, Roland Kuhn, Michael Simard, Eric Joanis, and S. Larkin. 2006. Portage with Smoothed Phrase Tables and Segment Choice Models. In *Workshop on Statistical Machine Translation*, pages 134–137, New York.

Matti Kääriäinen. 2009. Sinuhe – Statistical Machine Translation using a Globally Trained Conditional Exponential Family Translation Model. In *Conference on Empirical Methods in Natural Language Processing*, pages 1027–1036, Singapore.

Alex Kulesza and Stuart Shieber. 2004. A learning approach to improving sentence-level MT evaluation.

In *10th International Conference on Theoretical and Methodological Issues in Machine Translation*, Baltimore.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *2nd Workshop on Statistical Machine Translation*, pages 228–231, Prague.

Chin-Yew Lin and Franz Josef Och. 2004a. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chin-Yew Lin and Franz Josef Och. 2004b. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Ding Liu and Daniel Gildea. 2005. Syntactic Features for Evaluation of Machine Translation. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32.

Ding Liu and Daniel Gildea. 2007. Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. In *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 41–48.

I. Dan Melamed, Ryan Green, and Joseph P. Turian. 2003. Precision and Recall of Machine Translation. In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*.

Franz Josef Och and Herman Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Sebastian Padó, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Textual entailment features for machine translation evaluation. In *4th Workshop on Statistical Machine Translation*, pages 37–41, Athens.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Morristown.

Chris Quirk. 2004. Training a Sentence-Level Machine Translation Confidence Measure. In *4th Conference on Language Resources and Evaluation*, pages 825–828, Lisbon.

Craig Saunders. 2008. Application of Markov Approaches to Statistical Machine Translation. Technical report, SMART Project Deliverable 2.2.

Michel Simard, Nicola Cancedda, Bruno Cavestro, Marc Dymetman, Eric Gaussier, Cyril Goutte, Kenji Yamada, Philippe Langlais, and Arne Mauser. 2005. Translating with Non-contiguous Phrases. In *Conference on Empirical Methods in Natural Language*, pages 755–762, Vancouver.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006a. A Study of Translation Edit Rate with Targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the America*, pages 223–231, Cambridge, MA.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006b. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231.

Lucia Specia, Marco Turchi, Nicola Cancedda, Marc Dymetman, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *13th Conference of the European Association for Machine Translation*, pages 28–37, Barcelona.

Lucia Specia, Nicola Cancedda, and Marc Dymetman. 2010a. A dataset for assessing machine translation evaluation metrics. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Lucia Specia, Dhwaj Raj, and Marco Turchi. 2010b. Machine translation evaluation versus quality estimation. *Machine Translation*, pages 1–12.

Christoph Tillmann, Stefan Vogel, Hermann Ney, A. Zubiaga, and H. Sawaf. 1997. Accelerated DP based Search for Statistical Translation. In *Proceedings of European Conference on Speech Communication and Technology*.