Integrating Meaning into Quality Evaluation of Machine Translation

Osman Başkaya[1], **Eray Yıldız**[1], **Doruk Tunaoğlu**[1], **M. Tolga Eren**[1], **and A. Seza Doğruöz**[2]

[1]Huawei Turkey Research and Development Center, Istanbul, Turkey
{osbaskaya,doruktuna,tolgaeren}@gmail.com, eray.yildiz@huawei.com

[2]Independent Researcher
a.s.dogruoz@gmail.com

Abstract

Machine translation (MT) quality is evaluated through comparisons between MT outputs and the human translations (HT). Traditionally, this evaluation relies on form related features (e.g. lexicon and syntax) and ignores the transfer of meaning reflected in HT outputs. Instead, we evaluate the quality of MT outputs through meaning related features (e.g. polarity, subjectivity) with two experiments. In the first experiment, the meaning related features are compared to human rankings individually. In the second experiment, combinations of meaning related features and other quality metrics are utilized to predict the same human rankings. The results of our experiments confirm the benefit of these features in predicting human evaluation of translation quality in addition to traditional metrics which focus mainly on form.

1 Introduction

Machine translation (MT) systems translate large chunks of data automatically across languages. Although these systems may achieve high level accuracies using form related features (e.g. lexical and syntactic), they often fail to carry over the meaning embracing the form. Example (1) highlights the meaning difference between a Human Translation (HT) and an MT output for the same source sentence:

Example (1)

HT: "*Your feet's too big.*"[1]
MT: "*Your feet is too great.*"[2]

Although the form is often preserved, MT outputs may sound "strange" or "different" in comparison to HT ones due to the loss of meaning. Therefore, human translators generally enrich the text with the appropriate tone, style and sentiments during translation. Current quality evaluation metrics like BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007) are based on form related features and do not directly consider the transfer of meaning (e.g. sentiment and style) in MT. Some of these metrics check for synonyms and paraphrases but this approach is still limited to the coverage of the corresponding pair tables. In other words, these metrics do not explicitly evaluate the transfer of meaning in MT. Our main goals are:

- to find out whether the transfer of meaning related features (e.g. sentiment and style) in MT influences the human judgment of translation quality.
- to compare meaning and form related features for quality evaluation of MT.
- to measure whether meaning and form related features can be combined to improve the performance of existing MT quality evaluation metrics.

[1] WMT'15 Finnish to English test set, reference translation, segment id:440

[2] WMT'15 Finnish to English test set, translated by system: UoS.4059, segment id:440

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 210–219,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

By using publicly available parallel corpora (Tenth Workshop on Statistical Machine Translation (WMT15)), we achieve our goals with two experiments described in Section 5. Our results indicate that combining meaning related features with form related ones approximates to the human judged rankings better than the BLEU metric. These combined features also improve the performance of other MT quality evaluation metrics by 0.5-2 percentage points.

2 Related Work

So far, MT studies have focused mostly on features related to form (e.g., lexical and syntactic features) for the automatic evaluation of MT quality (e.g., BLEU (Papineni et al., 2002) and METEOR (Lavie and Agarwal, 2007)). BLEU metric is based on n-gram matching of the HT and MT texts and used widely in the MT community to evaluate the MT quality. METEOR employs both word matching scores and the linguistic information (e.g., synonyms and stemming) in contrast to BLEU. Following studies have evaluated MT quality with various features: POS tags (Dahlmeier et al., 2011), morphemes (Tantuğ et al., 2008), sentence structure (Li et al., 2012), named entities (Buck, 2012), semantic textual similarity (Castillo and Estrella, 2012), paraphrasing (Snover et al., 2006), semantic roles (Lo and Wu, 2011) and language models (Stanojevic and Simaan, 2014). Recently, Yu et al. (2015) proposed another metric (i.e. DPMFComb) which is a combination of a syntax-based metric and some other evaluation metrics in Asiya[3]. At WMT15, DPMFComb obtained the best results at the metrics task for system-level evaluation of translation into English tasks.

Although previous methods require human reference translation, recent methods (e.g. *quality estimation metrics*), aim to eliminate the necessity of human translation. These methods apply Machine Learning (ML) techniques using lexical (e.g. average source/target token length), syntactic (e.g. ratio of percentage of POS tags in the source/target sentences), and statistical features (e.g. source/target sentence LM probability, word alignment probabilities, etc.) (Stymne et al., 2014; Langlois, 2015; Shah et al., 2015). Interested reader may also benefit from the survey on MT evaluation metrics by Han and Wong (2016).

Src. Lng.	Domain	# of Sent's	# of Jdg's
Czech	News Texts	2496	20224
Finnish	News Texts	1744	10757
French	News Forum	2136	12189
German	News Texts	1989	12880
Russian	News Texts	2407	14924
Total		10772	70974

Table 1: WMT15 Test Data Statistics grouped by source languages. The domain of source text, the number of sentences and the number of human judgments are presented for each source language.

Chen and Zhu (2014) explore sentiment consistency between MT and HT texts to improve the MT quality by incorporating sentiment related features (e.g. subjectivity, polarity, intensity and negation). By using these features in their MT system, they improved the BLEU score by 1.1 point on NIST Chinese-to-English translation dataset[4]. Mohammad et al. (2015) also investigate the sentiment consistency between MT and HT texts with a different motivation. They improve sentiment analysis performance for Arabic by translating available resources (e.g., sentiment lexicon, sentiment annotated data) from English to Arabic. Although sentiment analysis of English translations of Arabic texts obtain competitive results to current state-of-the-art Arabic sentiment analysis systems, they did not evaluate the MT output quality.

There are also studies using MT systems to enrich labeled data for sentiment analysis by translating between languages and leveraging sentiment scores (Wan, 2009; Demirtaş and Pechenizkiy, 2013; Hiroshi et al., 2004). However, none of these studies employ meaning related features to evaluate the MT quality.

3 Data Description

3.1 2015 Workshop on Statistical Machine Translation

We utilized WMT15[5] parallel corpora (Bojar et al., 2015) which include several tasks (e.g., standard news translation task, a metrics task, a tuning task, a task for run-time estimation of machine translation quality, and an automatic post-editing task). 24 institutions participated in the translation task with a total of 68 machine translation systems. The WMT15 data includes:

[3] `http://asiya-faust.cs.upc.edu/`

[4] `http://www.nist.gov/speech/tests/mt`

[5] `http://www.statmt.org/wmt15/`

	Human Translation (HT) Output	Machine Translation (MT) Output
1.	*Adam, you see badly what you are looking at.*	*Adam, you see what you look at.*
2.	*Of course I don't hate you.*	*Of course I hate you.*
3.	*This is business news*	*This is supposed to be of business news*
4.	*The views of Chinese towards white people is similar!*	*The Chinese think like white people!*

Table 2: Examples of MT Errors in WMT15 Dataset

- Source sentences
- Reference human translations (HT)
- Machine translations (MT)
- Human judgments (e.g. from 1 (*best*) to 5 (*worst*)) for each MT text.

The data is available for five language pairs: Czech (ces)-English, French (fre)-English, German (deu)-English, Finnish (fin)-English, and Russian (rus)-English. Domains of the test data are the same for all languages except for French. The test data for the French-English language pair was fetched from a news discussion forum instead of news texts. Table 1 shows the statistics for the test data. The target language is English for all source languages. The domain of source text, the number of sentences and the number of human judgments are presented. All data was based on the news text corpora except for French-English pair.

In order to evaluate the quality of each MT system, Bojar et al. (2015) conducted a human evaluation using Appraise[6] (Federmann, 2012) which is an open source toolkit (similar to Amazon Mechanical Turk[7]). Each segment consists of a source sentence in the original language (e.g. Czech), its corresponding human translation (English), and 5 anonymous MT system translations (English).

To make the task more consistent and to increase the number of data points, the organizers treated almost identical system translations as one. Even though exactly 5 translations are presented to each judge in a segment, there may be more than 5 MT systems that are ranked. Judges rate the segments from 1 (best) to 5 (worst) by the quality of translated sentences (allowing ties).

In total, there were 29,007 segments, each of which would have produced at least 10 individual system comparisons (e.g., A>B, B>C, A=C, C>B, etc.). To map these individual comparisons to system scores, the organizers used TrueSkill [8] (Herbrich et al., 2006), a Bayesian skill ranking algorithm (similar to Elo used in Chess (Elo, 1978)) and fed these individual bilateral comparisons to TrueSkill. A score is produced for each participated system. In this study, we utilized the HT texts, MT system translations and human judgments in our experiments.

3.2 Features

Table 2 provides examples of MT errors in comparison to HT. All example translations (MT vs. HT texts) are selected from the WMT15 dataset based on the lowest (5) rankings by human judges. Although translations overlap at the word level, they convey quite different meanings. In example (1), the word *'badly'* has disappeared in MT output and led to a loss of information. In example (2), a negated sentence is translated as an affirmative sentence by the MT system. Example (3) illustrates how the MT system generates a more speculative sentence than HT sentence. The pair in example (4) differs in terms of formality between MT vs HT output. MT evaluation metrics may attribute high scores for these pairs since they mainly focus on lexical and syntactic matching. However, as our examples demonstrate, meaning could easily be lost if we rely only on form related MT system evaluation metrics.

To investigate the consistency between MT and HT texts for sentiment and stylistic features, we make use of *sentiment polarity, subjectivity, connotation, negation, speculation, readability* and *formality* to measure how these features influence the quality of translation with respect to human rankings.

[6] https://github.com/cfedermann/Appraise

[7] https://www.mturk.com/mturk/

[8] http://research.microsoft.com/en-us/projects/trueskill/

Sentiment Polarity indicates whether the designated sentence has an affirmative or negative sentiment. To measure the impact of this feature, we use *Vader*, a rule based sentiment analysis tool (Hutto and Gilbert, 2014). It utilizes grammatical and syntactical rules. In the experiments performed by Hutto and Gilbert (2014), *Vader* outperforms several competing sentiment analysis approaches.

Additionally, we trained a machine learning (ML) based sentiment analyzer using a deep learning approach described by Yildiz et al. (2016). Their architecture is a Convolutional Neural Network (CNN) which takes pre-trained word vectors[9] as input and applies interleaved convolution and pooling operations. The top layer in the network is Softmax layer which computes the probability of assigning a class (positive, negative).

We adopted this architecture and trained a network using Stanford Twitter Sentiment Corpus[10]. The training set contains 1.6 million tweets automatically labeled as positive or negative from various domains while the test set is labeled manually. This ML based sentiment analyzer achieves 90.1% accuracy and outperforms the SVM classifier reported by Go et al. (2009).

Subjectivity indicates whether a text expresses an opinion. In order to compute the subjectivity scores, we trained our architecture using the *sentiment polarity and subjectivity dataset*[11] (Pang and Lee, 2004) which includes 5000 subjective and 5000 objective sentences. We applied 10-fold cross validation to the data and obtained 91.50% average accuracy.

Connotation indicates cultural or emotional association carried by words that appear in sentences (Feng et al., 2013). In contrast to the sentiment polarity, connotation polarity indicates subtle shades of sentiment beyond denotative or surface meaning of text. The words which do not express sentiment can carry a positive or negative connotation.

For instance, "life" and "home" are considered neutral with regard to the sentiment analysis. However, they convey a positive connotation (Carpuat, 2015). We use the connotation polarity of each word in a sentence to compute connotation score using a normalized version of the formulation proposed by Carpuat (2015). The connotation polarities of the words are obtained by looking up a lexicon which is constructed by Feng et al. (2013). We used the following formula to compute the connotation score:

$$CS = \frac{\#positive - \#negative}{\#total} \quad (1)$$

where *CS* is the connotation score, *#positive* indicates the number of the words with positive connotation and *#negative* indicates the number of the words with negative connotation.

This formula assigns a continuous value between 1 and -1 to the sentence as a connotation score. The values close to 1 indicate that a given sentence carries a positive connotation while the values close to -1 indicate a negative connotation.

Negation turns an affirmative statement into a negative one. We also detect the effects of negation feature in our experiments. Konstantinova et al. (2012) present a freely available dataset which contains 400 reviews (50 each from 8 domains such as movies and consumer products) annotated by linguists for negation and speculation. We train our deep learning model with these datasets and obtain 96.65% accuracy for negation.

Speculation is used to express levels of certainty. We obtain 95.55% accuracy using the same dataset and method for negation.

Readability measures the ease of reading and comprehending a text (Dale and Chall, 1948). For readability measurement we use Flesch reading-ease test in which higher scores indicate that the text is easier to read. The Flesch readability score (Kincaid et al., 1975) is calculated using the sentence length and the number of syllables per word as presented in the formula below.

$$Flesch = 206.835 - 1.015\frac{A}{B} - 84.6\frac{C}{A} \quad (2)$$

where *A* is the number of words, *B* is the number of sentences and *C* is the number of syllables in a given text.

In addition to the rule based readability measurement, we use an ML based readability metric "*simplicity*" as described by Vajjala and Meurers (2016). They extract various syntactic, psycholinguistic and lexical features from text and

[9] https://code.google.com/archive/p/word2vec/

[10] http://help.sentiment140.com/for-students

[11] http://www.cs.cornell.edu/people/pabo/movie-review-data/rotten_imdb.tar.gz

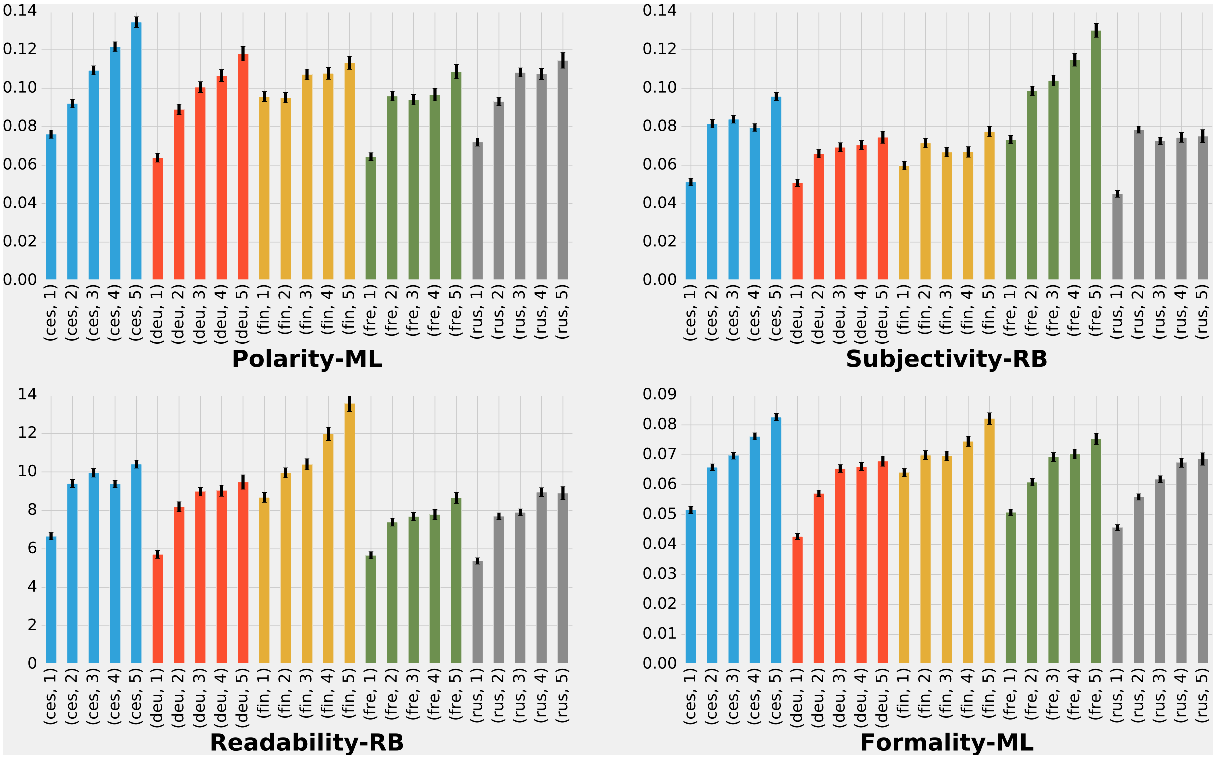

Figure 1: Means of absolute differences between the feature scores of MT and HT outputs. x-axis denotes the language and human rank pairs. Error bars indicate 95% confidence intervals.

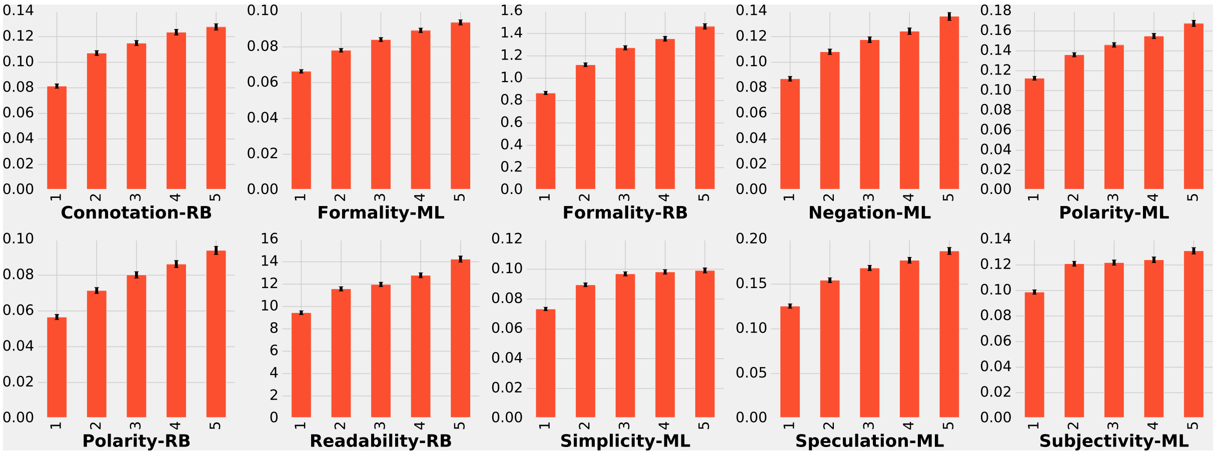

Figure 2: Means of absolute differences between the features scores of MT outputs and the corresponding HT outputs for all features. x-axis denotes the human rankings. Error bars indicate 95% confidence intervals.

train a pair-wise classifier using them. The training data is a sentence-aligned corpus constructed from news articles and Wikipedia pages and their simplified versions. The method correctly classifies the simplified and complex sentences in terms of their reading level with an accuracy of over 80%.

Formality Heylighen and Dewaele (1999) state formality as the most important dimension of variation between styles. They define the formality score as a function of POS tag frequencies. The formality score is given in Equation 3 where NF is the frequency of nouns, AdjF is the adjective frequency, PF is the preposition frequency, ArtF is the article frequency, PrpF is the proper noun frequency, VF is the verb frequency, AdvF is the adverb frequency and IF is the interjection frequency.

$$F = \frac{NF + AdjF + PF + ArtF}{2} - \frac{PrpF + VF + AdvF + IF}{2} + 50 \quad (3)$$

Additionally, we use an ML based formality score obtained by training the mentioned architecture on the dataset introduced by Pavlick and Tetreault (2016). We have observed 80.71% accuracy through 10-fold cross validation.

All metrics were normalized between (0, 1) except the Readability and Formality. Since these two metrics are formula-based, we avoided interfering with their original scales.

4 Method

In the WMT15 task, the language pairs are divided into two groups depending on whether English is the source or the target language. We only utilize the pairs where English is the target language due to the richness in resources. For each feature, MT texts are ranked using the following approach:

1. Compute the score for HT text (e.g., 0.65).
2. Compute the scores for MT texts (e.g. A=.79, B=.25, C=.20, D=.95, E=.30).
3. Compute the absolute difference between MT scores and the HT score (e.g., $\hat{A} = .14$, $\hat{B} = .40$, $\hat{C} = .45$, $\hat{D} = .30$, $\hat{E} = .35$).
4. Rank the systems according to these differences where a smaller value corresponds to a better ranking (e.g. 1=A, 2=D, 3=E, 4=B, 5=C).

Figure 1 shows absolute differences for four features with respect to language and human rankings of MT system output. For instance, *Subjectivity-RB* feature captures the differences between ranks when the source language is French but cannot achieve the same performance for Finnish and Russian translations. Moreover, *Readability-RuleBased (RB)* and *Formality-MachineLearning (ML)* seem to perform well for all languages whereas *Polarity-ML* falls short for French.

Figure 2 illustrates the trend for all features in which absolute score difference between MT system outputs and HT text is low for high rankings (e.g., 1) and high for the low (e.g., 5) ones. Therefore, high ranked translations preserve the meaning better than the low ranked ones. Note that both figures are descriptive and do not correspond to an objective evaluation directly.

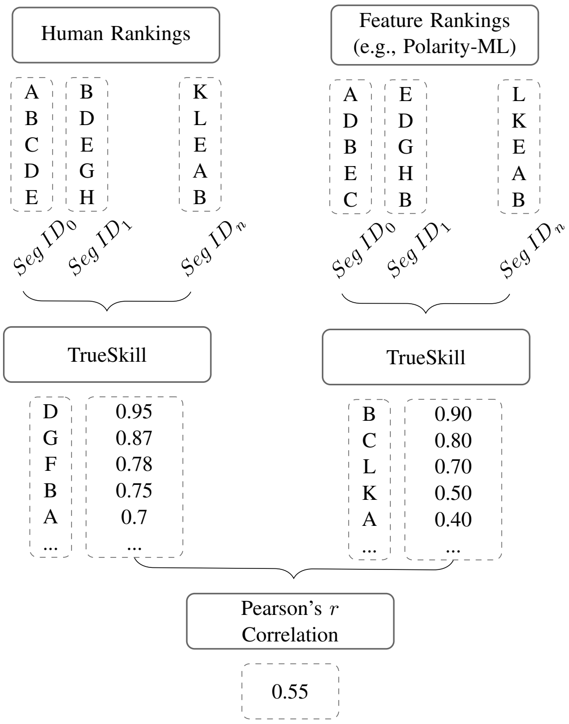

Figure 3: Steps to map human rankings and feature rankings (for example, *Polarity-ML*) to system-wide scores. A to L denotes individual rankings and SegID$_i$ denotes the i^{th} segment in the WMT15 test set.

5 Experiments

5.1 Experiment #1: Impact of Individual Features on Translation Quality

This experiment investigates the correlation between each feature and MT translation quality evaluated by rankings of human judges. Using the rankings described in Section 4, we followed the "System-Based Evaluation Methodology" by Stanojević et al. (2015). After obtaining the rankings for each feature as described in the previous section, we used TrueSkill to map segment rankings to system-wide scores (see Figure 3). Next, we compared TrueSkill scores obtained per feature and human judgments with Pearson's r correlation using the scripts provided by the WMT15 Metrics Task[12]. As stated in (Stanojević et al., 2015) the script performs bootstrap resampling of 1000 samples while calculating the correlation scores and the 95% confidence intervals.

[12] http://www.statmt.org/wmt15/metrics-task/wmt15-metrics-results.tgz

	All	Ces	Deu	Fin	Fre	Rus
Connotation-RB	74.2 ± 2.2	**87.6 ± 0.8**	86.1 ± 2.0	41.7 ± 3.5	87.8 ± 1.8	67.9 ± 2.6
Formality-ML	74.9 ± 2.1	62.7 ± 1.1	85.3 ± 1.9	**85.9 ± 2.0**	80.3 ± 2.2	60.3 ± 3.1
Formality-RB	**80.7 ± 1.7**	67.9 ± 1.1	**92.5 ± 1.5**	68.0 ± 2.8	**97.4 ± 0.9**	**78.0 ± 2.4**
Negation-ML	48.8 ± 2.7	42.2 ± 1.4	61.2 ± 3.0	33.4 ± 3.5	78.5 ± 2.1	28.8 ± 3.6
Polarity-ML	67.1 ± 2.3	54.0 ± 1.2	78.2 ± 2.2	65.6 ± 2.7	76.1 ± 2.4	61.7 ± 2.8
Polarity-RB	78.6 ± 2.1	79.2 ± 1.0	88.6 ± 1.7	75.6 ± 2.5	81.4 ± 2.4	67.9 ± 2.7
Readability-RB	76.7 ± 2.0	66.4 ± 1.2	79.8 ± 2.2	79.7 ± 2.1	85.3 ± 2.0	72.4 ± 2.6
Simplicity-ML	41.5 ± 2.8	17.6 ± 1.4	54.9 ± 3.0	18.4 ± 3.7	77.5 ± 2.5	39.1 ± 3.5
Speculation-ML	62.2 ± 2.4	41.4 ± 1.3	68.3 ± 2.7	63.6 ± 2.9	86.2 ± 2.1	51.7 ± 3.2
Subjectivity-ML	61.1 ± 2.6	56.3 ± 1.3	66.4 ± 2.9	42.5 ± 3.2	75.9 ± 2.6	64.3 ± 2.9
BLEU	91.6 ± 1.4	95.8 ± 0.6	86.5 ± 2.0	92.9 ± 1.4	97.5 ± 0.9	85.1 ± 2.2
DPMFComb	96.2 ± 0.9	96.0 ± 0.5	97.0 ± 0.9	95.1 ± 1.2	98.0 ± 0.8	95.0 ± 1.1
Meteor	94.9 ± 1.0	94.8 ± 0.5	95.5 ± 1.0	96.3 ± 1.0	95.1 ± 1.2	92.7 ± 1.4
Random-Baseline	0.0 ± 2.9	-28.4 ± 1.5	47.6 ± 3.2	-65.9 ± 2.8	-3.6 ± 3.6	50.4 ± 3.4

Table 3: Pearson's r correlation between Trueskill scores of a metric and human judgments with the corresponding 95% confidence intervals are shown. Each row represent either a meaning related feature (top) or a selected metric from WMT15 (bottom). ML stands for machine learning and RB stands for rule based method.

We have used three metrics from WMT15 Metrics Task for comparison, BLEU and METEOR and DPMFComb. DPMFComb was selected since it was the best system in overall score in system-based evaluation of WMT15 Metrics Shared Task and the best performing evaluation metric for three out of five languages.

Results Table 3 shows all the Pearson's correlations. Overall, *Formality-RB* obtains the highest correlation score (80.7%) among all features. However, DPMFComb (96.2%), BLEU and METEOR are better than the rest. The exceptions are German for BLEU and French for METEOR. For German, *Formality-RB* (92.5%) outperforms BLEU (86.5%). For French, *Formality-RB* (97.4%) beats the METEOR score (95.1%). In addition, Rule-Based (RB) systems perform better than Machine Learning (ML) ones. For example, *Formality-RB* and *Polarity-RB* outperform *Formality-ML* and *Polarity-ML* respectively.

Meaning related features outperform *Random* baseline as expected. The random baseline is computed by assigning random ranks (1-5) to each translation in each segment. We assigned uniformly random ranks to all sentences without considering the language. Although its performance may vary per language, its overall performance is 0.0 (± 2.9).

5.2 Experiment #2: Impact of Combined Features on Translation Quality

As discussed in Section 4, our approach is fundamentally different than MT evaluation metrics such as BLEU. Results of our first experiment indicated strong correlations between quality scores of the features and human rankings. Therefore, we also investigate whether we can predict human rankings of MT translated text by combining these features since they capture different aspects of translation.

In contrast with Experiment 1, this experiment focuses on training systems that combine several features to predict human rankings. As input, BLEU, METEOR and DPMFComb metrics are utilized in combination with the feature scores. We experimented with several classifiers from RankLib[13] to train the ensemble systems and opted to utilize a Random Forest (Liaw and Wiener, 2002) based approach which produced the best 5-fold cross validation score.

First, we obtained scores and rankings for each translation using the Random Forest classifiers for the following combinations:

1. All meaning related features
2. All meaning related features + BLEU

[13] https://sourceforge.net/p/lemur/wiki/RankLib/

	All	Ces	Deu	Fin	Fre	Rus
ALL+DPMFComb	**96.8** ± **0.8**	95.9 ± 0.4	**97.5** ± **0.8**	**97.9** ± **0.8**	**98.6** ± **0.6**	94.3 ± 1.3
DPMFComb	96.2 ± 0.9	**96.0** ± **0.5**	97.0 ± 0.9	95.1 ± 1.2	98.0 ± 0.8	**95.0** ± **1.1**
ALL+Meteor	95.8 ± 0.9	95.4 ± 0.5	96.6 ± 1.0	97.7 ± 0.8	98.0 ± 0.8	91.2 ± 1.6
Meteor	94.9 ± 1.0	94.8 ± 0.5	95.5 ± 1.0	96.3 ± 1.0	95.1 ± 1.2	92.7 ± 1.4
ALL+BLEU	93.5 ± 1.2	93.4 ± 0.6	92.3 ± 1.5	96.8 ± 0.9	97.6 ± 0.9	87.3 ± 1.9
ALL	92.0 ± 1.2	87.5 ± 0.7	93.4 ± 1.3	94.5 ± 1.2	97.8 ± 0.8	86.8 ± 1.8
BLEU	91.6 ± 1.4	95.8 ± 0.6	86.5 ± 2.0	92.9 ± 1.4	97.5 ± 0.9	85.1 ± 2.2

Table 4: Pearson's r correlation between Trueskill scores of a metric and human judgments with the corresponding 95% confidence intervals are shown. ALL represents the combination of all meaning related features.

3. All meaning related features + Meteor
4. All meaning related features + DPMFComb

Then, we calculated the Trueskill scores for translation systems and finally fed them into WMT15 scrips to obtain Pearson's r correlation similar to the first experiment.

Results Combined meaning related features outperform the BLEU score (Table 4). Even though the margin is relatively low, it is a promising indication. Moreover, combining them with a metric increases the performance of the metric: 1.9pp for BLEU, 0.9pp for METEOR and 0.6pp for DPMFComb. In other words, these features can utilize some meaning or style related information which is not captured by the conventional MT evaluation metrics.

6 Discussion & Conclusion

In this paper, we investigate how meaning related features influence the automatic evaluation of MT systems. Our experiments prove the additional benefit of these features in predicting human evaluation of translation quality. More specifically, we find that:

- MT systems that are ranked higher by human judges preserve the meaning (features such as polarity, formality and readability) better than the low ranked ones.
- Rankings of MT output generated according to meaning based features correlate highly with human rankings on translation quality (See Figure 2).
- When meaning related features are combined with form related lexical features, human evaluation of MT system quality can be predicted with a higher accuracy. (See Table 4).

Extracting meaning related features from text and using form related features for MT evaluation have been studied separately. However, integrating meaning related features into MT quality evaluation can capture the meaning preservation from source to target languages. Our experiments prove that this integrated approach achieves a only slightly better performance than the form based metrics (e.g. BLEU). Moreover, our experiments indicate that the meaning related features can boost the performance of BLEU, METEOR and DPMFComb metrics without even specific optimization. Therefore, our method of integrating meaning related features to MT systems with ranking components can also improve the performances of other metrics instead of only relying on form based features.

Commonly used evaluation metrics (e.g. BLEU and METEOR) require a reference human translation to assess the quality of MT. We also use human translation as a reference since most meaning related feature extraction tools are only available for English and limited for other languages. Although there are studies assessing the quality of MT systems without human translation, meaning related features are still not integrated to MT systems yet. As new tools for other languages become available, we plan to extend our work to implement MT quality estimation for these languages as well. As future work, we will investigate the ways to develop more ”human-like” MT systems by employing these meaning related and stylistic features in the training of MT systems or in post-processing steps such as parameter tuning.

Acknowledgments

This work is partially funded by 3140951 numbered TUBITAK-TEYDEB grant.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 1–46, Lisbon, Portugal, September. Association for Computational Linguistics.

Christian Buck. 2012. Black box features for the wmt 2012 quality estimation shared task. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 91–95. Association for Computational Linguistics.

Marine Carpuat. 2015. Connotation in translation. In *6th Workshop On Computational Approaches to Subjectivity, Sentiment and Social Media Analysis WASSA 2015*, page 9.

Julio Castillo and Paula Estrella. 2012. Semantic textual similarity for mt evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 52–58. Association for Computational Linguistics.

Boxing Chen and Xiaodan Zhu. 2014. Bilingual sentiment consistency for statistical machine translation. In *EACL*, pages 607–615.

Daniel Dahlmeier, Chang Liu, and Hwee Tou Ng. 2011. Tesla at wmt 2011: Translation evaluation and tunable metric. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 78–84. Association for Computational Linguistics.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Erkin Demirtaş and Mykola Pechenizkiy. 2013. Cross-lingual polarity detection with machine translation. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 9. ACM.

Arpad E. Elo. 1978. *The rating of chessplayers, past and present*. Arco Pub.

Christian Federmann. 2012. Appraise: an open-source toolkit for manual evaluation of mt output. *The Prague Bulletin of Mathematical Linguistics*, 98:25–35.

Song Feng, Jun Seok Kang, Polina Kuznetsova, and Yejin Choi. 2013. Connotation lexicon: A dash of sentiment beneath the surface meaning. In *ACL*, pages 1774–1784.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1:12.

Aaron Li-Feng Han and Derek Fai Wong. 2016. Machine translation evaluation: A survey. *arXiv preprint arXiv:1605.04515.*

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, pages 569–576.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center Leo Apostel, Vrije Universiteit Brüssel.*

Kanayama Hiroshi, Nasukawa Tetsuya, and Watanabe Hideo. 2004. Deeper sentiment analysis using machine translation technology. In *Proceedings of the 20th international conference on Computational Linguistics*, page 494. Association for Computational Linguistics.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International AAAI Conference on Weblogs and Social Media.*

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document.

Natalia Konstantinova, Sheila CM De Sousa, Noa P. Cruz Díaz, Manuel J. Maña López, Maite Taboada, and Ruslan Mitkov. 2012. A review corpus annotated for negation, speculation and their scope. In *LREC*, pages 3190–3195.

David Langlois. 2015. Loria system for the wmt15 quality estimation shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 323–329, Lisbon, Portugal, September. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231. Association for Computational Linguistics.

Liangyou Li, Zhengxian Gong, and Guodong Zhou. 2012. Phrase-based evaluation for machine translation. In *Proceedings of COLING 2012: Posters*, pages 663–672, Mumbai, India, December. The COLING 2012 Organizing Committee.

Andy Liaw and Matthew Wiener. 2002. Classification and regression by randomforest. *R News*, 2(3):18–22.

Chi-kiu Lo and Dekai Wu. 2011. Meant: an inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility via semantic frames. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 220–229. Association for Computational Linguistics.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2015. How translation alters sentiment. *Journal of Artificial Intelligence Research*, 54:1–20.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

Kashif Shah, Varvara Logacheva, Gustavo Paetzold, Frédéric Blain, Daniel Beck, Fethi Bougares, and Lucia Specia. 2015. Shef-nn: Translation quality estimation with neural networks. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 342–347, Lisbon, Portugal, September. Association for Computational Linguistics.

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 association for machine translation in the Americas*, pages 223–231.

Miloš Stanojevic and Khalil Simaan. 2014. Beer: Better evaluation as ranking. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 414–419.

Miloš Stanojević, Amir Kamran, Philipp Koehn, and Ondřej Bojar. 2015. Results of the wmt15 metrics shared task. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 256–273, Lisbon, Portugal, September. Association for Computational Linguistics.

Sara Stymne, Jörg Tiedemann, and Joakim Nivre. 2014. Estimating word alignment quality for smt reordering tasks. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 275–286, Baltimore, Maryland, USA, June. Association for Computational Linguistics.

A. Cüneyd Tantuğ, Kemal Oflazer, and Ilknur Durgar El-Kahlout. 2008. Bleu+: a tool for fine-grained bleu computation.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 235–243. Association for Computational Linguistics.

Eray Yildiz, Caglar Tirkaz, H. Bahadir Sahin, Mustafa Tolga Eren, and Omer Ozan Sonmez. 2016. A morphology-aware network for morphological disambiguation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hui Yu, Qingsong Ma, Xiaofeng Wu, and Qun Liu. 2015. CASICT-DCU Participation in WMT2015 Metrics Task. *EMNLP 2015*, page 417.