Sentence-level MT evaluation without reference translations: Beyond language modeling

Michael Gamon, Anthony Aue, Martine Smets

Microsoft Research, Microsoft Customer Service and Support
mgamon|anthaue|msmets@microsoft.com

Abstract. In this paper we investigate the possibility of evaluating MT quality and fluency at the sentence level in the absence of reference translations. We measure the correlation between automatically-generated scores and human judgments, and we evaluate the performance of our system when used as a classifier for identifying highly dysfluent and ill-formed sentences. We show that we can substantially improve on the correlation between language model perplexity scores and human judgment by combining these perplexity scores with class probabilities from a machine-learned classifier. The classifier uses linguistic features and has been trained to distinguish human translations from machine translations. We show that this approach also performs well in identifying dysfluent sentences.

1. Introduction

It is widely recognized that the automatic evaluation of machine translation quality is crucial for at least two major tasks:

- comparisons of different systems on the same or similar data sets
- system-internal evaluations in the practical deployment of MT systems

There is a rapidly-growing body of literature on automatic metrics for the first scenario, starting with the BLEU metric (Papineni et al. 2002) and the NIST metric (Doddington 2002). Crucially, the BLEU metric and its relatives (e.g. Lin and Och 2004b, Babych and Hartley 2004, Soricut and Brill 2004) rely on one or more human reference translations for each machine-translated sentence to be evaluated. These metrics are all based on the idea that the more shared substrings the machine-translated sentence has with the human reference translation(s), the better the translation is. It has been shown that BLEU scores, despite their shortcomings, correlate surprisingly well with human judgment (Coughlin 2003). The BLEU metric and its relatives are typically computed for fixed multi-sentence test sets in order to track the performance of MT systems over time and to compare different MT systems with respect to these test sets.

Moving from multi-sentence evaluation to single-sentence and even word-level evaluation, Blatz et al. (2004) survey a number of approaches to the estimation of confidence. The training data for their experiments consist of machine translations and human reference translations. This work uses naive Bayes and multilayer perceptrons for classification.

For the system-internal evaluation of deployed MT systems, Quirk (2004) uses a small (350 sentence) corpus of machine translations that have been annotated for translation quality by human annotators. He represents translated sentences as feature vectors, training a classifier to emulate the human scoring. The features he uses include sentence perplexity according to a trigram language model (LM) of a training corpus, source sentence features such as length, translation features such as number and size of mappings, whether a translation comes from a learned mapping or from a dictionary, and tilings of source and translated sentences with respect to the training corpus. Quirk demonstrates that this approach can produce usable results for sentence-level confidence.

There are also several attempts at using machine-learned classifiers for the purpose of MT quality assessment both at the sentence level and for larger test sets. Corston-Oliver et al (2001) demonstrated that classifiers can distinguish quite

reliably at the sentence level between machine translations and human translations. Classification accuracy increases if linguistic features are added to purely perplexity-based features. Kulesza and Shieber (2004), in a similar approach, train a Support Vector Machine (SVM) classifier that is capable of reliably distinguishing machine translation output from human translations. They use a combination of features derived from n-gram precision, length, and word error rate with respect to human reference translations. Additionally, a confidence score produced by the classifier shows high correlation with human judgment.

To summarize, there are a number of automatic translation quality metrics both at the multi-sentence and single-sentence level. Most of these metrics (except Quirk 2004) require one or more human reference translations for each sentence to be evaluated. This is a reasonable requirement when the task is system comparison or tracking of system performance over time.

The use of MT systems in a production environment, however, requires the evaluation of massive amounts of machine-translated text. In practice, it is necessary to be able to assess the quality of all MT output in order to identify the particularly badly-translated sentences. In cases where MT is employed for the dissemination of large amounts of text (see e.g. Richardson 2004), two of the roles of the human translator are to identify systematic translation errors and to perform post-editing of low-quality MT output. In this scenario, a tool for detecting badly-translated sentences automatically and reliably is essential. It saves time and allows the translator to concentrate on the problematic sentences.

By definition, in MT for dissemination there is no human reference translation. The challenge, then, is to find a reasonable automatic evaluation metric for sentence-level translation quality that does not require human reference translations.

Translation quality involves both content and form. An ideal translation needs to capture the meaning of the source sentence and express it in a fluent target language sentence. In practice, automatic assessment of semantic adequacy is a much harder problem than evaluation of the fluency of a sentence. As we will discuss below, fluency assessment can serve as a proxy for overall translation quality as long as it correlates well enough with overall translation quality.

One readily-available solution for the evaluation of MT output fluency at the sentence level that does not require reference translations is the ngram language model (LM). LMs can be trained on a domain-specific corpus in the target language. A perplexity score can be calculated for each machine-translated sentence reflecting the degree to which the observed word sequence is "expected" compared to what has been observed in the training corpus. This approach has been used successfully to score output from different MT engines in multi-engine MT systems (Callison-Burch et al 2001, Akiba et al 2002, Nomoto 2003).

In this paper we attempt to improve on a sentence-level language model perplexity score by adding other sources of information about the fluency of the translation. The resources that our approach requires are:

i. a set of machine-translated sentences
ii. a corpus of target-language text from the same domain (but crucially not translations of the same source sentences that were used in (i))
iii. an automatic linguistic analysis system (parser)

By combining perplexity scores with scores provided by a classifier that is trained to distinguish machine-translated sentences from human translations, we are able to improve on the correlation between human judgments and perplexity scores alone. This classifier uses linguistic analysis features to complement the ngram-based language model. We also show that this system performs well when tasked with identifying the worst (i.e. most dysfluent) translations. All experiments were performed using an example-based machine translation system to translate technical documentation from English into French (Smets et al. 2003).

2. Experimental Setup

2.1. Data

We used a corpus of 1,566,265 French sentences from Microsoft technical documentation to train our language models.

Our training data for the SVM classifiers consists of 198,771 machine-translated sentences (English to French) from the Microsoft Product Support Services Knowledge Base (Richardson 2004), and 260,601 human-translated sentences from the same domain.

Language models and SVMs were then tested on a set of 500 held-out sentences which had been annotated by human annotators for both MT quality and fluency. The annotation consisted of separate scores on a scale of 1 to 4, where 1 means completely dysfluent or incomprehensible, and 4 means perfectly fluent or human-quality translation. For fluency annotation, the raters only took the target sentence into account. There were 6 raters for MT quality and 1 rater for fluency. The fluency rating was done independently of the MT quality evaluation to ensure that knowledge of the source sentence was not influencing the fluency evaluation.

The distribution of fluency and MT quality scores as assigned by the human evaluators to the test set are shown in Figure 1.

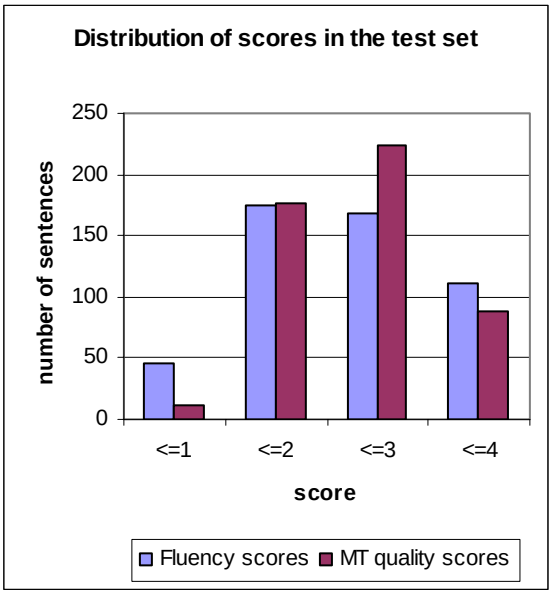

Figure 1: distribution of scores on the test set

2.2. Language Model

The French target language model was trained on all available data from the target domain. The data was preprocessed by converting all tokens to lower case, removing contractions (i.e. converting "d'" and "l'" to *de* and *le*, respectively), and removing punctuation. Then a 4-gram language model was built using interpolated Kneser-Ney smoothing. Kneser-Ney smoothing has been shown to outperform all other techniques for smoothing n-gram language models. (Kneser & Ney 1995, Goodman 2000). Per-sentence scores were computed by preprocessing the test data in the same manner as the training data, then computing cross-perplexity with the language model in the usual way:

$$P(s) = 2^{-\frac{1}{k}\sum_{i=1}^{k} \log p_{KN}(w_i|w_{i-1},w_{i-2},w_{i-3})}$$

Equation 1: Perplexity

In this formula, s is the sentence to be scored, consisting of tokens w_i through w_k , and p_{KN} is the conditional probability of the token given the three preceding tokens according to the language model.

2.3. SVM classifier

The SVM classifier was trained on the data described in section 2.1. The two values for the target feature are "1" for sentences produced by human translators and "0" for sentences produced by the MT system. The underlying assumption, as described in Kulesza and Shieber (2004) and Corston-Oliver et al. (2001), is that machine-translated output is known a priori to be of much worse quality than human translations. This classifier can then be used to evaluate a new sentence. If the new sentence is classified with high probability as a human translation, it is more similar to the human translations in the training set, and hence is likely to be of high fluency. If, on the other hand, the new sentence is classified as a machine-translated sentence, it is likely to have qualities similar to those only observed in machine-translated language; in other words it is likely to be less fluent. As Kulesza and Shieber (2004) point out, the class probability assigned by the classifier can serve as a quality score: the higher the probability that a sentence is human-translated, the better the quality of the sentence.

We used the SMO (Sequential Minimal Optimization) algorithm (Platt 1999) to train a linear SVM.

2.4. Feature vectors

For the purpose of training the SVM classifier, sentences were represented as vectors of binary

features. The features we used are based on linguistic analysis with the French NLPWin analysis system (Heidorn 2002). They fall into the following categories:

- trigrams of part of speech tags
- context-free grammar productions. Each syntactic node is represented with its label and the labels of all its daughters. For example, the feature "*NP::DETP:NOUN:PP*" indicates a noun phrase consisting of a determiner phrase, followed by a noun, followed by a prepositional phrase
- Semantic analysis features: Example: *+Def* indicating definiteness on a noun phrase
- Semantic features, part of speech, and semantic relationship to the parent node: *Def Noun PrepRel*, for example, indicates the presence of the definiteness feature on a nominal node that is in a prepositional relationship with its parent
- Semantic modification relations: *Verb Tsub Noun Tobj Pron*, for example, indicates a verb with a nominal logical subject and a pronominal logical object.

The use of this particular set of features is based on work in style classification (Gamon 2004), and is motivated by the desire to capture linguistic generalizations that go beyond surface ngram regularities. All features are extracted automatically. Parses are not vetted for quality or adequacy, and partial (non-spanning) parses are also exploited for feature extraction.

The total number of unique features extracted from the training data is 39254. In order to reduce the dimensionality of the feature vectors, we restricted the features to the top 2000 features according to the log likelihood ratio (Dunning 1993) of the feature with respect to the class label.

Training an SVM on the machine-translated and human-translated data described in section 2.1 produces a classifier that achieves 77.84% classification accuracy on the training set. If the data set is split 70/30 for training and testing, the result is similar with 77.59% accuracy. The baseline accuracy (choosing the most frequent target feature value) is 56.73%.

The SVM also produces a class probability that can be used as a score: the higher the probability that a sentence is in class "1" (i.e. human translations), the better its fluency, and vice versa.

3. Results

3.1. Correlation with human judgments

Reported correlation numbers are correlation coefficients as shown in Equation 2 where $\bar{x}$ and $\bar{y}$ are the sample mean.

$$Correl(X,Y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

Equation 2: correlation coefficient

3.1.1. Baseline correlations

In a first round of experiments, we determined the baseline correlations for BLEU scores, language model scores, and SVM scores with human judgments of fluency and MT quality. Since our approach is directed at gauging fluency of translation, we also need to establish that fluency indeed correlates well with general MT quality. The correlation between overall translation quality scores and fluency is 0.67, indicating that given the MT system and the data used in our experiments, fluency can be a reasonable approximation for overall translation quality.

BLEU achieves a relatively high correlation with overall translation quality (0.58) and a somewhat lower correlation with fluency (0.41). The LM by itself achieves a 0.34 correlation with fluency and a 0.29 correlation with overall translation quality. Note again that these correlation results are at the sentence level, hence are much lower than the correlations that are reported when comparing MT systems on a fixed test set (e.g. Soricut and Brill 2004). The higher correlation between BLEU and the human judgments is no surprise, since BLEU is able to take advantage of the human translation. Correlation of the SVM class probability scores with human judgments yields the worst results, at 0.09 with fluency and at 0.12 with MT quality.

Figure 2 shows the individual correlations with human judgments that the three different scoring methods achieve individually.

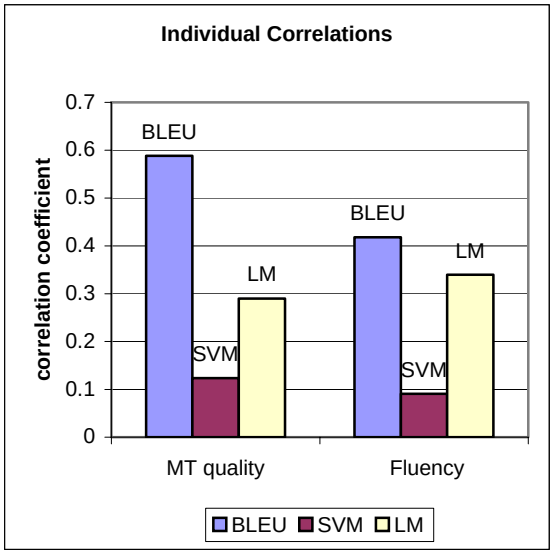

Figure 2: Individual correlations with human judgments

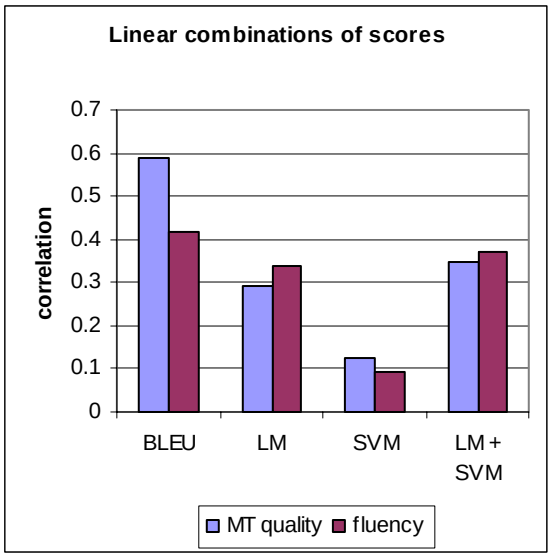

Figure 3: Linear combination of LM and SVM scores.

3.1.2. Combining the scores

The results in the previous sections clearly indicate that language model scores achieve a much higher correlation with human judgments than SVM class probability scores.

Given that these two scores are based on very different properties of the translation, namely surface string similarity with human-produced text and deeper linguistic analysis properties, the next natural step is to see if the scores can be combined in a way that maximizes correlation. In order to investigate this question, we use Powell's method (Powell 1964) to find a linear combination of scores from different approaches that maximizes the correlation with human judgments of fluency[1]. For this approach we split the data into a parameter-tuning set and a test set. We performed a 50/50 split on the 500-sentence data set, performing 2-fold cross validation on the two resulting subsets.

Figure 3 illustrates the correlations achieved by a linear combination of the scores. All numbers are based on averaging the results from 2-fold cross-validation, so the baseline results can differ slightly from those calculated on the whole test set. In the remainder of this paper, we will use the notation "+" in the figures to indicate a locally optimal linear combination of scores using Powell's method.

These results confirm that the information provided about fluency and MT quality by the two different metrics is at least to some degree additive, i.e. it is possible to improve on the individual scores by combining them into a single score.

In addition to the linear combination of scores, we also experimented with a number of other purely arithmetic combinations. Surprisingly, simple multiplication proves to be a remarkably effective way to combine LM perplexity score and SVM class probability score[2].

When these scores are combined by multiplication, correlation with fluency jumps from 0.37 for the linear combination of LM and SVM to 0.42. Correlation with overall translation quality also increases, from 0.35 to 0.42. Results are shown in Figure 4.

At this point, we do not understand why multiplying SVM scores with perplexity yields an increase in correlation compared to a linear combination. The effect is robust, though. It is statistically significant at the 95% level (using Fisher's z' transformation) and it was also observed in a second independent held-out set of 500 sentences that were manually annotated for fluency.

[1] We also experimented with optimizing correlation with the MT quality scores. This did not lead to significant improvements in that correlation, however.

[2] Technically, we have to multiply the LM score with (1-SVMscore) since the orientation of the two scores is reversed: the higher the SVM score, the more human-like is the translation. Higher perplexity scores, on the other hand, indicate lower fluency.

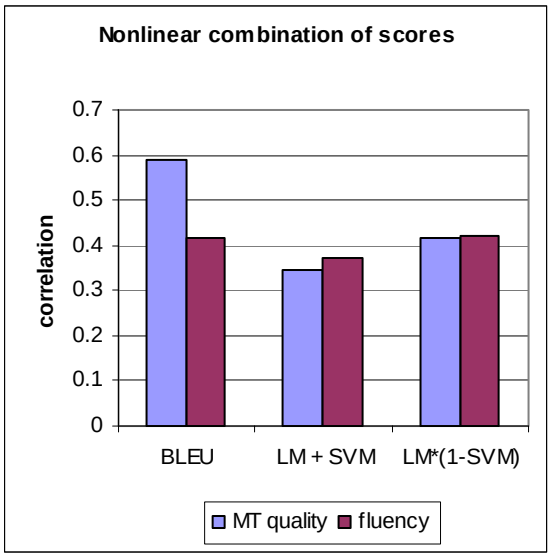

Figure 4: Nonlinear combination of SVM and LM scores

3.2. Classification: Identifying the worst translations

While we have shown that a particular combination of scores can achieve a relatively respectable correlation with human judgments, the ultimate goal of our experiments is to determine whether the developed metrics are practically effective or not. In fact, it is a generally-accepted assumption in the field that even the BLEU score correlation with human judgment at the sentence level is insufficient for evaluation.

The second evaluation scenario described in the introduction is focused on automatically detecting problem cases rather than assigning global evaluation scores or determining complete rankings of sentence quality. The practical use of the metrics in the dissemination scenario is to be able to rapidly detect the worst translations. Once they have been detected, the appropriate action can then be taken. For example, these particularly bad translations can be post-edited, manually re-translated, kept out of a translation memory system, etc.

For this more limited use of an automatic scoring mechanism, we need to assess whether the translations that score worst with our metrics correspond to bad translations according to human judgment. In order to evaluate the usefulness of our automatic metrics in this regard, we used the metrics as a classifier for identifying the worst-translated sentences. We evaluated this classifier on the human-annotated test data. The classification task was to identify low fluency and low MT quality scores, where „low" scores were either defined as scores <=1 or scores <= 2.

Classification was performed by assigning the worst n% of documents according to the scoring metric to the „Bad" class, and the rest of the documents to the „Good" class. In order to obtain precision/recall curves, we computed recall and precision for values of n in increments of 5%. For each of the intervals, we determined how many of the human-annotated low scores were correctly identified. Precision and recall curves for fluency and MT quality are shown in Figure 5 and Figure 6. Since there are only 11 sentences with an MT quality score of <= 1, we do not show the precision/recall results for that scenario. We tested four different metrics in this classification task: BLEU (as a point of comparison), the language model perplexity scores (LM), the SVM scores, and the multiplicative combination of language model and SVM scores (LM*(1-SVM)). As is apparent in the figures, the multiplicative combination of SVM and language model scores outperforms all other scores at most levels of recall. (This is true also for SVM scores, not shown for reasons of legibility). These results hold when classifying overall MT fluency as well as MT quality, indicating that for the MT system we used, fluency scores are indeed a very good indicator of overall MT quality. Whether this is the case in any given MT system, however, needs to be determined empirically. Systems that use string-based language models, which our system does not, may score well on fluency-based metrics, regardless of the semantic adequacy of a sentence.

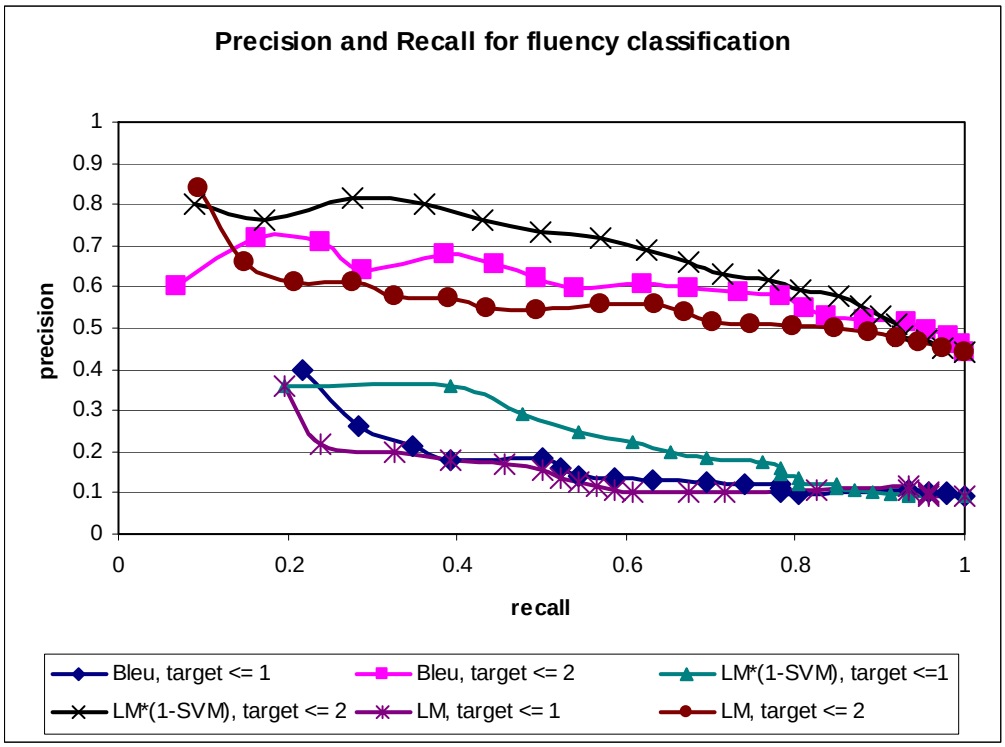

Figure 5: Precision and recall for fluency classification

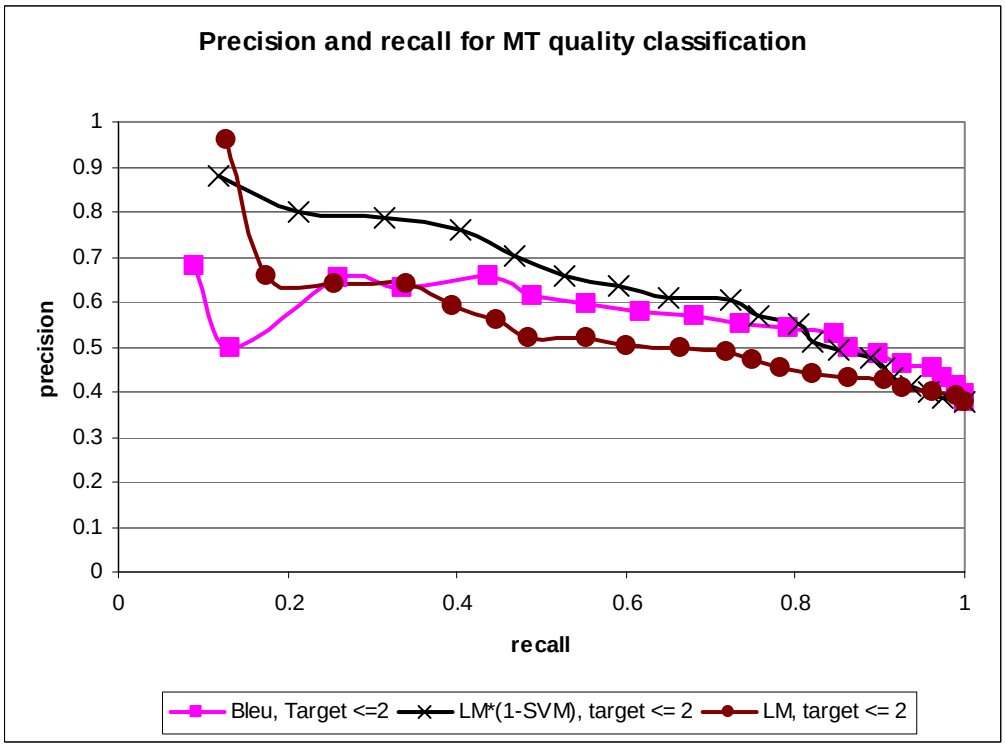

Figure 6: Precision and recall for MT quality classification

4. Conclusion

The question we set out to answer is whether it is possible to develop a sentence-level MT quality/fluency score that correlates reasonably well with human judgment and allows reliable identification of dysfluent translations, yet does not require reference translations. Both these requirements are imposed by practical considerations of MT for dissemination, where by definition reference translations are not available and where quality assessment at the sentence level is necessary.

We combined class probability scores from an SVM classifier trained to distinguish human translations from machine translations by using linguistic analysis features. We combined these SVM scores with standard language model perplexity scores.

We found (not too surprisingly) that none of the methods described in this paper achieve as high a correlation with human judgment on the sentence level as the BLEU score, which is computed with respect to a reference translation. However, we were able to substantially improve on the correlation of a baseline language model perplexity score by combining it with scores from an SVM classifier that had been trained on (non-aligned) human translations and machine translations in the relevant domain. Using linguistic analysis features, the SVM complements the surface-string-based perplexity scores of the language model.

Correlation with human judgments is a valid and established success criterion for automatic MT evaluation scores. For a deployed MT system, however, the task of automatically identifying the most dysfluent translations may be even more important than overall scoring. When formulated as a classification task for identification of the worst-translated sentences, our combined SVM and language model metric outperformed BLEU, and it also performed better than the language model and SVM individually.

In any given MT system, the usefulness of a fluency-based metric needs to be determined. In the example-based MT system we use, dysfluency is a very good indicator of poor overall MT quality. Instances of perfectly well-formed but semantically inadequate sentences are rare enough in the output of our system to make a fluency-based metric appropriate - an observation that may not hold true for string-based statistical systems. We believe, however, that the combination of classifier scores and language model scores is worth exploring in other systems as well, at least for the purposes of fluency ranking and candidate selection.

5. Acknowledgments

We wish to thank members of the Microsoft Research Natural Language Processing Group, in particular Chris Quirk, for helpful comments and discussions. We also thank the anonymous reviewers for their comments and suggestions.

6. References

AKIBA, Yasuhiro Taro WATANABE and Eiichiro SUMITA (2002): Using Language and Translation Models to Select the Best among Outputs from Multiple {MT} systems. Proceedings of COLING 2002, pp 8-14.

BLATZ, John, Eric FITZGERALD, George FOSTER, Simona GANDRABUR, Cyril GOUTTE, Alex KULESZA, Alberto SANCHIS, Nicola UEFFING (2004): Confidence Estimation for Machine Translation. In Proceedings of COLING 2004, pp. 315-321.

BABYCH, Bogdan and Anthony HARTLEY (2004): Extending the BLEU MT Evaluation Method with Frequency Weightings. Proceedings of ACL 2004, pp. 621-628.

CALLISON-BURCH, Chris and Raymond S. FLOURNOY (2001): A Program for Automatically Selecting the Best Output from Multiple Machine Translation Engines. Proceedings of MT Summit VIII, pp63-66.

CORSTON-OLIVER, Simon, Michael GAMON, and Chris BROCKETT (2001): A Machine Learning Approach to the Automatic Evaluation of Machine Translation. *Proceedings of ACL*, pp. 140-147.

COUGHLIN, Deborah (2003): Correlating Automated and Human Assessments of Machine Translation Quality. *Proceedings of MT Summit IX*, pp. 63-70.

DODDINGTON, George. (2002). Automatic Evaluation of Machine Translation Quality Using N-Gram Co-Occurrence Statistics. In Proceeding of the Second International Conference on Human Language Technology. San Diego, CA, pp. 138-145.

DUNNING, Ted (1993): Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics 19: 61-74.

GAMON, Michael (2004): Linguistic Correlates of Style: Authorship Classification with Deep Linguistic Analysis Features. Proceedings of COLING 2004, pp. 611- 617.

GOODMAN, Joshua (2000): A Bit of Progress in Language Modeling. Technical report, Microsoft Research.

HEIDORN, George E. (2002): Intelligent Writing Assistance. In R. Dale, H. Moisl, and H. Somers (eds.), A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text, Marcel Dekker, New York.

KNESER, Reinhard and Hermann NEY (1995): Improved backing-off for m-gram language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume 1, pp 181–184.

KULESZA, Alex and Stuart M. SHIEBER (2004): A Learning Approach to Improving Sentence-Level MT Evaluation. In Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation.

LIN, Chin-Yew and Franz OCH (2004a): ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. Proceedings of COLING 2004, pp. 501-507.

LIN, Chin-Yew and Franz OCH (2004b): Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics. Proceedings of ACL 2004, pp. 606-613.

NOMOTO, Tadashi (2003): Predictive Models of Performance in Multi-Engine Machine Translation. Proceedings of MT Summit IX, pp 269-276.

PAPINENI, Kishore A., Salim ROUKOS, Todd WARD and Wei-Jing ZHU (2002): BLEU: A Method for Automatic Evaluation of Machine Translation. Proceedings of ACL 2002, pp 311-318.

PLATT, John (1999): Fast training of SVMs using sequential minimal optimization. In: B. Schoelkopf, C. Burges and A. Smola (eds.) Advances in Kernel Methods: Support Vector Learning, MIT Press, Cambridge, MA, pp. 185-208.

POWELL, M. J. D. (1964): An efficient method for finding the minimum of a function of several variables without calculating derivates. Computer Journal 7.

QUIRK, Christopher (2004): Training a Sentence-Level Machine Translation Confidence Measure. In Proceedings of LREC 2004, pp 525-828.

RICHARDSON, Steve (2004): Machine Translation of Online Product Support Articles Using a Data-Driven MT System. Proceedings of AMTA 2004, pp.246-251.

SMETS, Martine, Michael GAMON, Jessie PINKHAM, Thomas REUTTER, and Martine PETTENARO (2003): High quality machine translation using a machine-learned sentence realization component. Proceedings of MT Summit IX, pp. 362-369.

SORICUT, Radu and Eric BRILL (2004): A Unified Framework for Automatic Evaluation using N-gram Co-Occurrence Statistics. Proceedings of ACL 2004, pp. 613-620.