# A Learning Approach to Improving Sentence-Level Machine Translation Evaluation

A Thesis presented

by

Alex Kulesza

to

Computer Science

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

April 6, 2004

# Abstract

The problem of evaluating machine translation (MT) systems is more challenging than it may first appear, as many diverse translations may often be considered equally correct. The task is even more difficult when practical circumstances require the evaluation to be done automatically, e.g., for incremental system development and error analysis. While several automatic metrics such as Bleu have been proposed and adopted for large-scale MT system discrimination, they all fail to achieve satisfactory levels of correlation with human judgements at the sentence level. Here, a new class of learning metrics based on support vector machines is proposed and shown to significantly improve upon current automatic evaluators, increasing performance halfway toward that achieved by independent human evaluators. Training only to classify translations as either machine- or human-produced avoids the myriad problems inherent in obtaining the desired human judgement targets, and is shown to nevertheless induce a strong correlation with those judgements. Future research includes exploring dependencies with respect to the feature set.

# Contents

# Chapter 1

# Introduction

Machine Translation (MT), the automated conversion of text from one natural language to another, has been a longstanding subject of interest and research. However, though in principle the goal of an MT system is clear, in reality a number of factors make the problem of evaluating success for such systems difficult. End users might apply MT toward a wide range of tasks from cross-language information retrieval to the full replacement of human translators, and each relies on a potentially unique set of characteristics from the translated output. Translating literature, for example, probably requires complete fluency and the ability to accurately convey extremely subtle ideas (depending even on the context of historically or thematically related works), while information-retrieval applications might only require a minimal "keyword" translation. Therefore, MT evaluation has a strong task-dependency. Additionally, the ambiguity and multiplicity of solutions even given strict qualitative criteria make the task of evaluation particularly difficult. Unlike problems such as automatic speech recognition, in which the goal is to exactly recognize a spoken utterance, MT is typically without a clear, objective target: "he walked the dog" and "he took the dog for a walk" may be equally good translations of a single foreign phrase, despite their syntactic differences.

The problem of MT evaluation has therefore necessarily received significant attention alongside the development of MT systems themselves. As it specifically aims to address the needs of humans who communicate with language, MT is most naturally and effectively evaluated through the manual efforts of such users, and extensive research has been conducted in order to describe, test, and taxonomize various methods and considerations involved in doing so [6, 7]. However, the human evaluation of MT systems, while providing the most direct and reliable assessment, has numerous drawbacks; primarily prohibitive are the time and expense involved in organizing and executing a user study. Subjects must be located, trained, presented with evaluation materials, and compensated; furthermore, to alleviate biases due to academic background or bilingual experience, a large number of subjects is usually preferable.

While manual methods may remain viable for isolated, large-scale evaluations, researchers should ideally benefit from the ability to quickly and accurately assess their own systems repeatedly as new ideas are implemented; in these situations, *automatic* methods have the potential to expedite the development of successful MT systems by greatly reducing the resources required for evaluation. In addition, new statistical MT systems have successfully incorporated training methods that directly optimize for the scores of automatic evaluation metrics [1]. Having reliable such metrics therefore implies that MT systems might be not only tested but trained very rapidly.

Of course, the primary problem with an automatic evaluation metric is the potential loss of accuracy involved in using a computer to perform a task that is most natural for humans; indeed, as described above, the precise goal of MT evaluation is not easily formalized. Furthermore, particularly when MT systems are trained directly using automatic metrics as a criteria, any weaknesses (e.g., situations in which scores are inappropriately inflated) are likely to be quickly exploited, rendering the metric useless. Therefore, a primary goal of any automatic MT evaluation metric should be a strong and consistent correlation with human judgements of the same outputs. Secondary goals may include the presentation of quantitative evaluations along one or more interpretable axes and efficiency with respect to computational or linguistic resources.

## 1.1 Automatic MT Evaluation Metrics

### 1.1.1 Definitions

A machine translation evaluation metric can be defined as follows: Assuming that an MT system takes foreign text $f$ and produces translated output text $e$, and that reference translations $e_1^*, e_2^*, \ldots, e_n^*$ of $f$, deemed to be "correct" (typically produced by human translators) are available, an evaluation metric is a function $F(e, e_1^*, e_2^*, \ldots, e_n^*)$. The task-dependency of evaluation implies that the output of the metric should not be binary; it should instead be presented along at least one continuous axis so that decisions of acceptability can be made with respect to particular applications. Here, the output of the metric is assumed to be a real number reflecting the correctness of $e$ in some monotone way. (Assume without loss of generality that the output has been transformed such that larger values always reflect a more positive evaluation.) The meaning of "correctness" will be more carefully stated below with respect to the goals of an automatic MT evaluation metric.

Note, however, that this definition is still not entirely general; evaluations might conceivably also rely on the source text $f$, for example. Such considerations are ignored here because $f$ is known to the MT system under evaluation; any dependency between $e$ and $f$ exploited by an automatic metric could thus be captured identically in the translation phase. In other words, if the metric were to rely on some automated notion of similarity between $e$ and $f$, a system designer could "cheat" by directly incorporating the very same measure into his or her system, thereby guaranteeing strong evaluations. Of course, if the measure is particularly reliable, this may not seem such a bad thing; furthermore, the relationship of $f$ with $e$ almost surely plays a large role in determining the success of a translation. However, given that humans are capable of evaluating MT outputs based on reference translations alone (and, in particular, without knowledge of the source language), and under the assumption that any automated measure is likely to have exploitable weaknesses of some kind, the source text $f$ is omitted here as a possible input to the automated evaluator in an attempt to reduce the possibility for developing pathological behavior. Moreover, the discovery of relationships between source and translated texts seems primarily the domain of MT itself, not MT evaluation.

Thus, it is assumed that only privileged information—information to which the evaluator has access and to which the system does not—is used for automatic evaluation, and that such information comes exclusively from reference translations. Again, human performance suggests that reference translations alone are sufficient for the task, but by no means implies that additional data would not provide further benefit. Other sources of information (alignments, logical representations of the concepts in the original text, etc.) might indeed be very useful, but do not seem in general to be widely available and so are not considered here.

### 1.1.2 Goals

In order to be useful, an automatic MT evaluation metric should provide some cost advantage while maintaining maximal "correctness" with respect to human evaluations. The former quality generally follows from the automated nature of the metric, though it is worth ensuring that the metric is efficiently computable and relies on human-produced data (such as reference translations) only when that data may be reused from evaluation to evaluation. More significantly, the latter quality is specified here as a strong and consistent correlation of metric outputs with quantitative assessments provided by human judges. To evaluate an evaluation metric, then, requires a corpus of translations, some of which are considered correct (reference translations), and the rest of which have been rated by humans with respect to the references (hypothesis translations).

The metric under consideration is first used to generate evaluations of the hypothesis translations. There then exist a variety of ways to compute correlations of these results with the numerical judgements of humans, depending on the goal of evaluation. Either a standard (Pearson) or rank (Spearman) correlation coefficient may be computed, the former reflecting a general-purpose linear relationship and the latter a reliability with respect to the ranking of various hypotheses. If evaluation is to serve the purpose of choosing the best MT system, rank correlation may be the best measure; on the other hand, quantifying the degree of separation between systems (or between development iterations of a single system) may require linear correlation. When evaluating automatic metrics here, both measures are presented; in general they seem to agree strongly.

A second important factor in computing correlations between automatic metrics and human judgements involves the *size* of the translations over which the evaluations have been conducted. Longer texts are likely to reduce the "noise" of evaluation (both human and automatic) and, when available, are adequate for the overall assessment of a system with respect to its peers. However, the correlation of a metric on long texts becomes increasingly less meaningful for purposes of error analysis; a metric that identifies when a system performs more successfully translating one novel than translating another will nevertheless generally fail to provide much insight into why this may be the case. On the other hand, a metric that can accurately gauge the quality of short-text translations may clearly illuminate the individual sentences or phrases with which the MT system has the most trouble, thereby leading to improvements in its performance. Furthermore, applications such as confidence estimation rely directly on local automatic assessments [8]. Since short-text evaluations can be averaged into evaluations of larger texts, a meaningful correlation for short texts is always preferable, though more difficult to achieve. Here, the standard textual unit used for evaluating automatic metrics is the sentence, in contrast to the larger texts often used to justify metrics in the past [3, 4]. This decision is primarily in response to conclusions drawn by two teams at the 2003 Johns Hopkins Workshop on Speech and Language Engineering, indicating a need for better sentence-level evaluation metrics [8, 9].

## 1.2 Previous Work

In recent years, a number of automatic evaluation metrics have been proposed and used for MT applications, including word error rate (WER), position-independent word error rate (PER) [2], Bleu score [3], NIST score [4], and the F-Measure [5]. Here, each is described briefly.

### 1.2.1 WER

WER, as applied to MT evaluation, is defined as the edit (or Levenshtein) distance between the hypothesis translation and a reference translation, where words are the units of insertion, deletion, and substitution. The minimal number of such operations required to transform the hypothesis into the reference forms the output of the WER metric. Therefore, the WER of hypothesis “he walked the dog” with respect to reference “he took the dog for a walk” is 4, as we can replace “walked” with “took” and then insert the three final words. WER is easily computed with a standard polynomial-time dynamic programming algorithm, and the negatively-correlated error value can be inverted to establish an evaluation scale with positive slope.

When multiple reference translations exist, the evaluation may be based on the minimum edit distance (most positive evaluation) between the hypothesis and any reference; this is sometimes referred to as multi-reference word error rate (mWER). WER has historically been used successfully as a general error metric for speech recognition applications.

### 1.2.2 PER

PER attempts to address the concern that WER might overly penalize re-orderings of words that are, in fact, acceptable. For example, “he went to the store” and “to the store he went” produce a WER of 4 but have arguably interchangeable meanings. PER, therefore, computes an edit distance in which the re-ordering of words is free; in the PER’s bag-of-words conception, error rates are at most equal to the WER. PER computation is simplified and can be achieved by removing all words in the shorter translation from the longer translation (if they appear) and returning the size of the remaining word-set. The previous example has a PER of 0. “He walked the dog” and “he took the dog for a walk” have a PER of 4; the similarity between the words “walk” and “walked” is not considered.

Again, when multiple references are available, the minimum distance with any reference may be used, and the error value may be inverted to obtain a positive evaluation.

### 1.2.3 BLEU score

Bleu was among the first metrics proposed specifically for the evaluation of MT systems [3]. The intuition applied in the development of Bleu suggests not only that hypotheses with many words appearing in references should receive positive evaluations (as with WER), but also that groups of consecutive words appearing in references should be further rewarded, since such groups may constitute matching "phrases" of some kind. Thus, Bleu is built on a set of statistics referred to as $n$-gram precisions. An $n$-gram is simply a list of $n$ consecutive words from a text; thus single words comprise unigrams, pairs of consecutive words bigrams, etc. The trigrams of "he walked the dog," for example, are "he walked the" and "walked the dog."

Measuring the precision of a hypothesis with respect to $n$-grams requires calculating the fraction of $n$-grams in the hypothesis exactly matching some $n$-gram from one of the references (only one hypothesis n-gram is allowed to match with any given reference n-gram). Precisions are computed for $n = 1, 2, 3, 4$ and labeled $p_1$ through $p_4$.

To calculate the Bleu score, then, precisions are simply combined via a geometric average, which tends to place emphasis on the typically much smaller $p_3$ and $p_4$ values:

$$\textsc{Bleu} = BP \cdot \exp\left(\sum_{n=1}^{4} \frac{\log p_n}{4}\right)$$

$BP$ is an exponential brevity penalty used to compensate for high precisions that may result when a hypothesis is very short. The details of computing $BP$ depend on the total size of the evaluation text [3].

For the example hypothesis "he walked the dog" and reference "he took the dog for a walk" the Bleu score is 0, since no trigrams or 4-grams match (precisions are 0). This fact illustrates that Bleu with a maximum $n$ of 4 is most effective when applied to translations of reasonable size (at least a sentence or more). Limiting $n$ to 2 for the purpose of illustration, however, the hypothesis unigrams (bold indicating a match in the reference) are "**he**," "walked," "**the**," and "**dog**." Thus, unigram precision is $\frac{3}{4}$. Bigrams are "he walked," "walked the," and "**the dog**," for a bigram precision of $\frac{1}{3}$. A geometric mean gives $\sqrt{\frac{3}{4} \cdot \frac{1}{3}} =$

$\frac{1}{2}$, and the BLEU score without brevity penalty is 0.5.

When first described, BLEU scores were shown to correlate highly with human judgements (a Pearson coefficient of 0.96) when averaged over large, 50-sentence texts from several human and machine translators. BLEU has also been used for directly training at least one state-of-the-art MT system [1].

### 1.2.4 NIST score

The NIST metric is based upon BLEU, but attempts to correct two potentially problematic behaviors [4]:

- Not all individual $n$-grams are equal; some may appear very commonly in a corpus and serve a general purpose ("he said that"), while others are more rare and, by implication, specific and important when they do arise ("striking blue fedora").
- A geometric mean tends to over-emphasize the small precisions for large $n$-grams.

In attempt to deal with the first point, the NIST score introduces an information weighting scheme based upon $n$-gram frequencies in a large corpus. (Often, in practice, the set of all reference translations is used.) If $k_n$ is the number of times a particular $n$-gram is seen in the corpus, and $k_{n-1}$ is the number of times the first $n-1$ words of the $n$-gram are seen, then the information weight for that $n$-gram is defined as $\log \frac{k_{n-1}}{k_n}$. The information weight therefore captures an estimate of the self-information of the $n$th word given the first $n-1$. Information weights are used to produce modified precisions $w_n$, calculated by summing the information weights of the matching $n$-grams and dividing by the total number of $n$-grams in the hypothesis.

In order to address the second point, the NIST score utilizes an arithmetic mean over the precisions; as a result, the much larger $p_1$ and $p_2$ values tend to dominate the computation. The NIST score also adds 5-grams to the computation, and incorporates a modified brevity penalty $BP'$:

$$NIST = BP' \cdot \sum_{n=1}^{5} \frac{w_n}{5}$$

Providing a convincing NIST score computation example is difficult due to the information weighting scheme; however, given such weights, the calculation is very straightforward, mirroring the example provided above for Bleu.

The NIST score was originally demonstrated to correlate slightly more strongly and consistently with human evaluations than Bleu when averaged over large texts, and is the metric used by the annual NIST MT system competition. However, the additional requirement of a large, representative corpus for generating information weights makes NIST potentially more sensitive to its input, and certainly requires greater care in implementation and set-up.

### 1.2.5 F-Measure

The F-Measure for MT evaluation was developed partly in response to the uninterpretability of automatic evaluation metrics such as Bleu or NIST, whose quantified outputs generally have little absolute meaning [5]. Attempting to combine the intuitions and successes of those metrics with a more general and standard framework for thinking about the problem of evaluation, the F-Measure incorporates not only precision but also its counterpart statistic, recall—the fraction of $n$-grams in the reference that also appear in the hypothesis.

Precision and recall can be defined in terms of sets; if the hypothesis and reference are sets $H$ and $R$ of constituent words, then (unigram) precision is $\frac{|H \cap R|}{|H|}$ and recall is $\frac{|H \cap R|}{|R|}$. However, to capture the notion that consecutive matching words might carry more importance than single words matching in isolation (handled in Bleu and NIST by the use of $n$-gram statistics), the F-Measure generalizes the concept of intersection by introducing "runs." A run is any set of consecutive words occurring in both the hypothesis and reference, and the maximum matching between hypothesis and reference is defined as the set $M^*$ of non-conflicting runs that maximizes

$$s(M) = \left( \sum_{m \in M} |m|^e \right)^{\frac{1}{e}}$$

for some exponent $e$, where $|m|$ refers to the number of words in run $m$. $|H \cap R|$ is replaced with $s(M^*)$ to give modified precision $\frac{s(M^*)}{|H|}$ and modified recall $\frac{s(M^*)}{|R|}$. The F-Measure is finally computed as the harmonic mean of the two, equivalent to:

$$\frac{2 \cdot s(M^*)}{|H| + |R|}$$

When $e = 1$, s(M) is equal to the number of words in common, and the F-Measure is merely the fraction of the total words (including both reference and hypothesis) that co-occur. When $e = 2$, the F-Measure can be approximated visually by the square root of the fraction of area covered by blocks whose diagonals reflect runs when the two translations are placed on two axes of a grid [5]. Interestingly, however, the greatest correlations with human judgements have been found when $e = 1$, suggesting that capturing the influence of successive matching words in this way is not generally useful to evaluation. Using $e = 1$ conveniently allows for a straightforward implementation, while $e = 2$ or higher requires a non-optimal greedy algorithm as the problem of locating $M$ most likely becomes NP-hard.

For example hypothesis “he walked the dog” and reference “he took the dog for a walk,” the maximum matching consists of runs “he” and “the dog”, so $s(M^*) = 3$ when $e = 1$ or $\sqrt{5}$ when $e = 2$. Since $|H| = 4$ and $|R| = 7$, the F-Measure is $\frac{6}{11}$ or $\frac{2\sqrt{5}}{11}$ for $e = 1$ and $e = 2$, respectively.

# Chapter 2

# Approach

## 2.1 Motivation

Experiments conducted at the 2003 Johns Hopkins Workshop on Speech and Language Engineering have suggested that the currently available automatic machine translation evaluation metrics provide insufficient correlation with human judgements when considered at the level of an individual sentence. Having collected pairs of human judgements (recorded on a scale from 1 to 5) for 633 hypothesis translations, each with respect to a single, randomly-chosen reference, the Confidence Estimation team reports that, although inter-judge correlation is surprisingly low, the correlation between human judges and automatic metrics is significantly lower [8]. Figure 2.1 illustrates the computed correlations for human judges and a series of automatic metrics over these 633 single-sentence hypotheses. (Human judgements are percentile-normalized to compensate for differing evaluation tendencies.)

Although, ideally, evaluation methodologies should be such that humans exhibit higher inter-judge correlation than seen here (and in similar studies, a sentiment shared by Turian et al. [5]), these results nevertheless clearly indicate that room for the improvement of automatic metrics is available at the sentence-level. Correlations of automatic evaluations averaged over larger texts were much higher, in agreement with previous results [3, 4].

Others have also noted poor automatic metric performance, particularly with respect to short translations. Turian et al. report, based on experiments rank-correlating the outputs

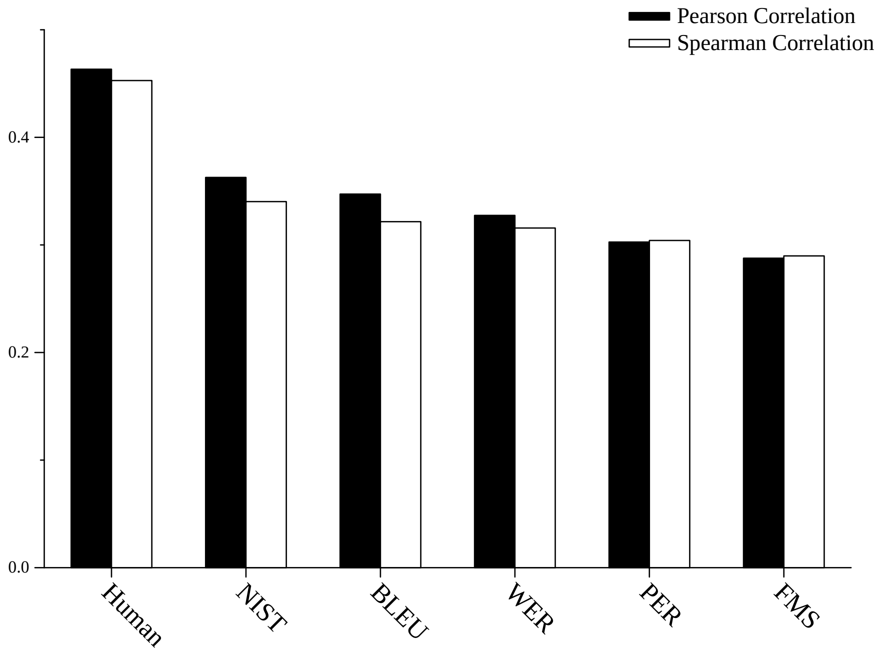

Figure 2.1: Pearson and Spearman correlations between normalized human judgements (captured on a scale from 1 to 5) and various human and automatic evaluators. Automatic metrics are calculated with respect to a set of four human references. The gap between human and automatic performance is statistically significant at 95%.

of Bleu, NIST, and the F-Measure with human judgements on translations from 1 to 100 sentences in length, that "even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgment [5]." The Syntax for Statistical Machine Translation team from the 2003 Workshop further notes in its final report that the Bleu metric, used for training and evaluation of the team's MT system, seems insensitive to syntactic changes on the sentence level that should be noticeable to human judges [9]. Such claims provide further anecdotal evidence for the poor correlations of automated evaluators with human judgements.

An automatic evaluation metric that can effectively determine the quality of a single-sentence translation thus has a variety of potential applications, among them finer-grained error analysis during the system design phase and post-translation confidence estimation using the automatic metric as a target for machine learning [8]. Additionally, if improved sentence-level accuracy indicates a lower "noise" level than that found with currently available metrics, we could expect to obtain generally more accurate evaluations from such a metric. Therefore, the goal of this thesis is the exploration and development of automatic MT error metrics that correlate more strongly and consistently with human judgements on a per-sentence basis.

## 2.2 Machine Learning

In attempting to create new automatic machine translation evaluation metrics that will address some of the recently observed outstanding difficulties, it is tempting to suggest the use of general-purpose machine-learning methods as a means of directly approximating human evaluations. Given a large training set of human-evaluated hypothesis translations and reference counterparts, it should be possible to directly learn human evaluation scores as a function of some feature set generated over the input. However, this approach is problematic for at least two reasons, both deriving fundamentally from the resource problem.

First, machine-learning methods universally adhere to a "more is better" principle with respect to the size of the training set. To successfully learn evaluation scores would require the initial development of a large set of human evaluations and, consequently, consume large amounts of time and money. If such a project were planned, it would need to involve a very carefully designed methodology in order to ensure that the evaluation data received is of the best possible quality and the greatest possible longevity. Given the extensive research history of human MT evaluation methods and practices, such a task might itself be very difficult or even prohibitive.

Second, however, even if such a gigantic resource could be created, it would be necessarily static, representing a fixed distribution of MT outputs on a fixed set of language pairs. With the constant evolution of MT systems, hypothesis translations judged in the project evaluations and subsequently used to generate metrics would eventually fail to adequately represent the population of MT outputs being considered. The training set would then no longer reflect the true distribution of hypothesis translations; under these conditions, machine-learning approaches are likely to fail.

It is crucial, then, that any machine-learning approach to automatic MT evaluation sustain the ability to be retrained and itself reevaluated with respect to modern and representative translation samples at regular intervals. In the limit one would like to retrain whenever a new group of MT systems (or a new iteration of possibilities for a single system under development) is to be evaluated, but if doing so requires the process of human evaluation, then the advantage of the automatic evaluator is neutralized.

## 2.3 An Alternate Training Criterion

To provide the necessary flexibility, as well as to alleviate the problems inherent in producing human evaluations of MT outputs, an alternate training criterion is proposed. Instead of attempting to perform a direct regression on human evaluations of hypothesis translations, a simpler classification problem might be more feasible: has a given hypothesis translation been produced by a machine or by a human? This question has the distinct advantage of being answerable with existing information—training sets can be as large as the corpora for which we have reference translations, since no human input is necessary to define the target classification. Furthermore, because human-produced reference translations can be fixed for a particular corpus, they remain a one-time startup cost and can be used repeatedly, regardless of any innovations in the design of MT systems. A metric predicated upon an ability to distinguish machine translations from human translations, therefore, can be retrained at will on new distributions of hypothesis MT outputs.

Of course, the classification criterion does not necessarily lend itself to the stated goal of improving correlation with human evaluators; in this sense it is less direct than the full regression approach. However, evidence suggests that, at least currently, MT outputs and human translations are easily distinguished by human evaluators, even when judgements are made without the knowledge that translations may be from multiple sources (see figure 2.2). Indeed, if MT outputs were of such quality that they could not be distinguished from human translations, evaluation would no longer be necessary; by definition such an MT system would be completely successful. Therefore, while less expressive than the regression criterion, the proposed classification is likely to establish broad performance groups in which humans reliably outperform machines, and therefore to encourage a macro-scale correlation with human judgements.

Furthermore, many machine-learning methods for classification do not simply produce a binary decision for each example, but in fact generate continuous outputs that can be interpreted as degrees of confidence in the final classification. Using such a method, the classification criterion actually induces a continuous-scale evaluator for which the output is some measure of "confidence" that a particular translation was produced by a human. If, in fact, the question of translation quality is closely related to the believability of a translation being produced by a human, then such a system might reasonably expect to achieve success as an MT evaluation metric. Determining the conditions under which such results are possible is a central question of this thesis.

Finally, a machine-learning approach carries the great advantage of flexibility. In addressing problems such as the lack of sensitivity to syntactic features or the desire to ensure that evaluators continue to distinguish modern systems as they improve, new features can be designed and incorporated into a retrained model. Machine-learning based automatic metrics, therefore, have the potential to form a large class of customizable, infinitely ad-

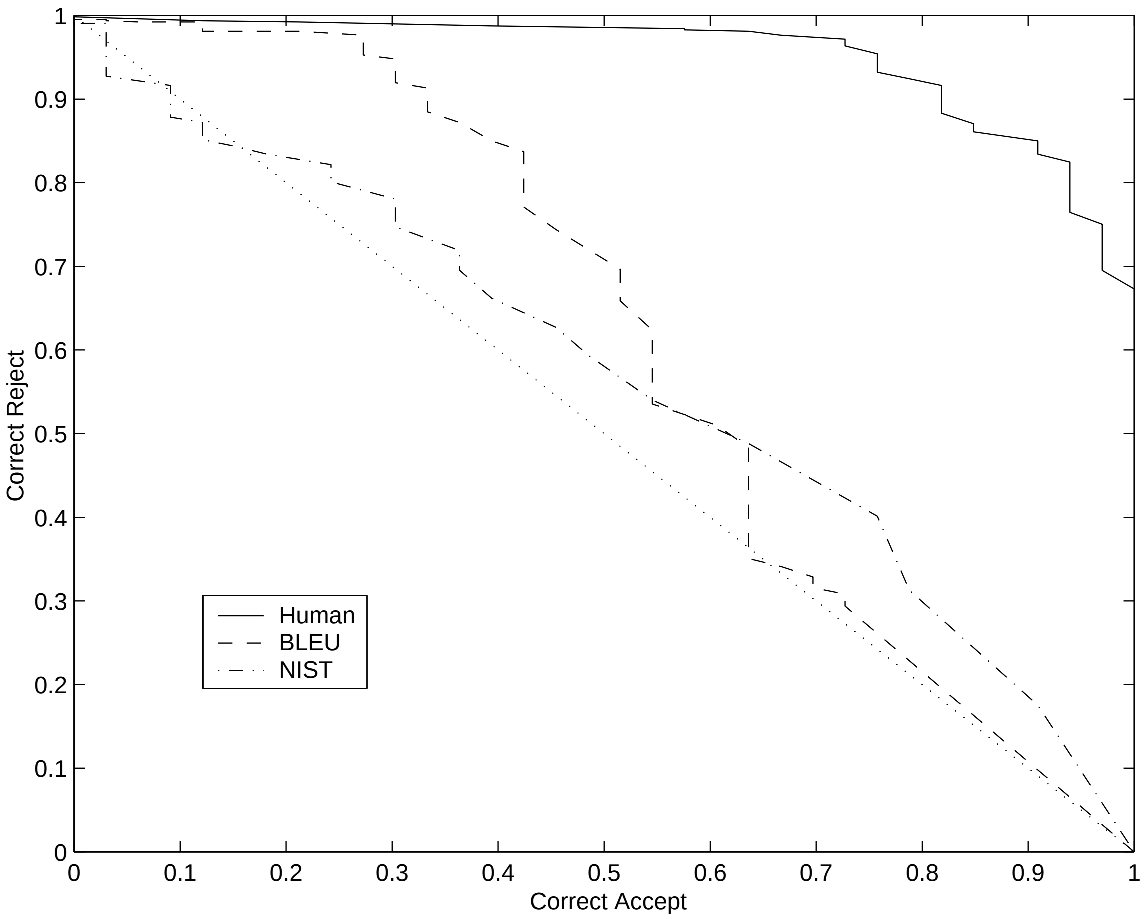

Figure 2.2: ROC curves reflecting the abilities of human evaluation scores, Bleu scores, and NIST scores to distinguish human translations from machine translations. The $x$-axis signifies the fraction of human translations correctly recognized as human, and the $y$-axis signifies the fraction of machine translations correctly recognized as machine. A curve is generated by considering all possible thresholds on a given evaluation metric's output; scores above the threshold are assumed to signify a human translation, and scores below the threshold are assumed machine-produced. The more a curve tends to the upper right, the greater the ability of the metric in question to distinguish human from machine translations. (The dotted diagonal is a baseline.) The human judgements come from Workshop experiments in which participants were told that all hypotheses were machine translations [8]; the data set actually consists of 633 machine translations and 70 human translations. Significantly, even without an awareness of the task, humans easily outperform standard automatic measures.

justable evaluation solutions. Though the continually changing nature of such a metric implies that its scores will hold little absolute meaning, the advantages with respect to detailed error analysis and adaptability might be significant.

# Chapter 3

# Implementation

## 3.1 Data Sets

Data for the training and testing of machine-learning based metrics are drawn from the test corpus and output lists of the alignment template statistical machine translation system that obtained the best results in the 2002 and 2003 DARPA MT evaluations and was a subject of research at the 2003 JHU Workshop [9]. The system is trained to produce Chinese to English translations of news articles that typically cover current political and economic events, and although the system consistently performs at high levels relative to other current MT systems, it is worth noting that the translations it produces are often poor, lacking in fluency despite being sometimes comprehensible upon extended consideration. A typical output (formatted for readability) is:

> As soldiers to fire on the spot, "the sudden shot and killed the Takhar and another from Hamas and another person was seriously injured, and there have been sent to a nearby hospital for emergency treatment."

While the machine translation system considered is believed to represent the current state-of-the-art in Chinese to English MT, the development of more successful systems in the future will likely have a large impact on the parameters meaningful to evaluation.

In order to apply general machine-learning methods, a series of examples must be drawn from the data, each consisting of an input and an output; general training algorithms can then be applied for learning to predict the output given the input. Recalling that the input to an evaluation metric is an $(n+1)$-tuple $(e, e_1^*, e_2^*, \ldots, e_n^*)$ containing a hypothesis and $n$ references, an input here consists of an English hypothesis translation $e$ based on a single Chinese source sentence, produced either by the above MT system and drawn randomly from a 100-best list or by a human reference translator, as well as three additional human reference translations $e_1^*, e_2^*, e_3^*$ of the same sentence. Care is taken to ensure that the hypothesis is never produced by the same translator as any of the references. The output used for the modified training criterion is a simple binary variable indicating whether the translation is produced by a machine or by a human. Due to the design of the classification problem, the output can is determined by the source of the hypothesis, needing no manual assistance. If direct regression were attempted, on the other hand, human evaluations of each hypothesis would be required.

In total, 21,144 examples (input/output pairs) have been extracted from the system, half of which contain machine-produced hypotheses and half of which contain human-produced hypotheses. These examples are split approximately 2:1 into training and validation sets of 14,120 and 7,024, respectively; an equal number of human and machine hypotheses are maintained in each set. The machine-learning layer is eventually optimized over the larger training set and then tested for classification accuracy on the validation set. The final assessment of any resulting metric, however, depends not on its classification abilities but on its continuous-output correlation with human judgements. For measuring this quality, a third data set is used—the test set—consisting of 633 hypothesis translations, all produced by the above MT system and evaluated by two independent human judges during an experiment at the 2003 JHU Workshop.

In the evaluation experiment, users familiar with natural-language research were asked to judge the quality of hypothesis translations with respect to single references produced by humans. Ratings were collected on a scale from 1–5, which was described to the users as in figure 3.1. To compensate for nevertheless differing interpretations of the evaluation scale (which are evident), ratings are percentile-normalized to the users reporting them. Each rating is assumed to carry a more positive evaluation than all lower ratings as well as

```
Please rate the quality of a given hypothesis translation with
respect to the reference on a scale from 1 to 5 as follows:

    Reference ex: bob walked the dog.

    1: Useless; captures absolutely none of the reference's meaning.
        ex: franklin is a doctor.
    2: Poor; contains a few key words, but little or no meaning.
        ex: dog banana walk.
    3: Mediocre; contains some meaning, but with serious errors.
        ex: the dog walked bob.
    4: Acceptable; captures most of the meaning with only small errors.
        ex: bob walk the dog.
    5: Human quality; captures all of the reference's meaning.
        ex: bob took the dog for a walk.
```

Figure 3.1: The five point evaluation scale used to collect human judgements for 633 machine-produced hypothesis translations.

half of the equal ratings given out by that user; for example, suppose users A and B rate hypotheses with the following frequencies:

| Rating | 1 | 2 | 3 | 4 | 5 | total |
|---|---|---|---|---|---|---|
| Frequency (user A) | 6 | 16 | 26 | 32 | 20 | 100 |
| Frequency (user B) | 18 | 36 | 30 | 16 | 0 | 100 |

Then a rating of 3 by user A is normalized to $(6 + 16 + \frac{26}{2})/100 = 35\%$, while a rating of 3 by user B is normalized to $(18 + 36 + \frac{30}{2})/100 = 69\%$. These normalizations produce results corresponding to the notion that a rating of 3 by user B is in reality much higher praise than the same rating given by an empirically kinder user A.

Rather than averaging the two independent ratings given to each hypothesis, the ratings are split randomly into two sets, each with a single rating for every hypothesis. One of these data sets is considered the "true" human judgement with which high correlation by a metric is taken to mean success, while the other is a "human" evaluation metric to which automatic metrics can be compared. (Note that, though generated in the same manner as the true human judgements, the human metric does not necessarily represent the best possible performance, since it is only based on a single rating for each hypothesis. A metric based on ratings averaged across many users would likely perform even better due to decreased noise.)

## 3.2 Features

Examples as described above cannot be fed directly into a machine-learning layer with any hope of success, however, simply because linguistic objects such as natural language translation hypotheses are too complex to allow for direct generalization based on any reasonable number of training examples. One way to effectively reduce the "dimensionality" of such objects is to manually preprocess them in order to extract a set of numeric features expected to be most salient for the classification task. Only the feature values, then, are passed as inputs to the machine-learning layer, which is not allowed to directly examine the various translations. The result is potentially improved generalization power at the cost of a strictly less informative input space.

Here, two major sets of features are considered; the differing results produced by each are discussed in the following chapter. The "basic" feature set consists of very simple statistics, many taken from the definitions of known automatic MT evaluation metrics like Bleu and WER. The basic set includes the following features:

- The minimum and maximum ratio of hypothesis length to reference length over the three references.
- Unmodified $n$-gram precisions for $n$ =1–5, calculated as described in section 1.2.3.
- Word error rate (WER), calculated as described in section 1.2.1.
- Position-independent word error rate (PER), calculated as described in section 1.2.2.

The basic feature set is designed primarily to determine if a machine-learning approach can make better use of the information already incorporated heuristically by current metrics.

The second, "extended" feature set includes all those features in the basic set, as well as additional features derived from statistical parses of the input hypothesis and reference translations using Collins's model 3 parser [12], with tagging by Ratnaparkhi's MXPOST tagger [13]. The probability score returned by the parser is itself used as a feature value, since human translations might be expected to have higher-probability parses than their machine-generated counterparts; it appears in the extended set both in raw, absolute form and as a ratio with the parse scores of the references. (Minimum and maximum such ratios over the three references are reported.) The remaining parse-based features are pairs of the form:

$$\left( \min_i \frac{c_e}{c_{e_i^*}},\ \max_i \frac{c_e}{c_{e_i^*}} \right),$$

where $c_t$ is the number of times a specific syntactic nonterminal occurs in translation $t$ according to the parser. Such features are computed for each of the following nonterminals: `ADJP`, `FRAG`, `NP`, `NPB`, `PP`, `SBAR`, and `VP`. For some sentences (particularly long, jumbled machine hypotheses), the parser fails entirely; in these situations, counts of 0 are reported for all nonterminals. All references, however, are parsed successfully, so division by zero does not occur. The extended feature set is designed to determine how syntactic information might aid a machine-learning based evaluation metric.

## 3.3 Learning

Many general methods exist for learning a classifier based on some set of labeled training instances; in order to create a continuous-scale evaluation metric from the simple classification criterion, however, a learning method that produces some form of "confidence" value must be utilized. In particular, constructs such as standard decision trees are not useful; a binary classification provides very limited information on the quality of a particular translation. Instead, support vector machines (SVMs) are used here due to the particularly nice conception of classification they provide, considering it as a problem of linear separation in some feature space.

SVMs operate like perceptrons, determining margin-maximizing linear separators between classes, but they allow for highly non-linear classification boundaries by performing linear classification in a potentially complex feature space [10]. The classification of a test example is determined by the side of the feature-space separator on which it falls. After training by optimizing a quadratic constraint problem, the SVM contains a separator defined by $x$ for which $\langle w, x \rangle = b$ for a vector $w$, a constant $b$, and an inner product $\langle \cdot, \cdot \rangle$ in the feature space (calculated indirectly with a kernel function). Thus, $sign(\langle w, x \rangle - b)$ determines the SVM's classification of an input $x$, in this case deciding whether a translation has been produced by a human or by a machine.

However, the separator is more than a simple classifier; it acts as an organization for the space of translations, defining the half-spaces of machine and human produced translations as well as a boundary subspace in which translations are equally human and machine-like. By computing not only the side of the boundary on which a particular example falls, but also the distance between the example and that boundary (removing the operator $sign$ in the expression above), a measure of "confidence" is obtained. Examples buried deep in the machine half-space (examples for which $\langle w, x \rangle - b$ is very positive or very negative, depending on the sign of the classifier) are perhaps very clearly machine produced, while those close to the boundary might have qualities seen more commonly in translations produced by humans, even if they still fall into the "machine" class. Those machine-produced examples that appear in the human half-space, on the other hand, have successfully fooled the classifier, and are likely to be of very high quality. A SVM can be used to define, therefore, a measure of confidence that is conceptually justified based on a simple classification criterion. It is worth noting that, because a SVM attempts to maximize the margin of its classifier (equivalent to minimizing $||w||$), irrelevant features should be appropriately ignored in the inner product and have little influence on a particular example's distance from the separator.

The Torch3 machine learning library implementation of SVMs for classification is used here in a slightly modified form [11]. Gaussian kernels are employed, and the quadratic constraint problem is optimized until the Karush-Kuhn-Tucker conditions are satisfied to within 0.01 error. Two parameters remain to be tuned: $C$, the trade off between margin maximization and error minimization, and $\sigma$, the standard deviation of the Gaussian kernel. The effects of adjusting these parameters are discussed in the following chapter. Once an SVM has been trained on the training set, its classification accuracy is measured on the validation set, and its correlation with human judgements is found by computing the Pearson and Spearman correlation coefficients between $\langle w, x \rangle - b$ and the true human evaluations for all examples $x$ in the test set.

# Chapter 4

# Results

The methods described produce an automatic machine translation evaluation metric that significantly outperforms current automatic metrics, as measured by the sentence level Pearson and Spearman correlation coefficients of the metric outputs with human judgements for the test set of 633 machine-produced hypotheses. Additionally, results demonstrate that given a particular feature set, optimizing for the modified training criterion in fact strongly encourages correlation with the true evaluations, implying that expensive human judgements are not required in general. However, different feature sets may produce metrics with widely differing performance characteristics; feature selection therefore seems to be a key to reliably producing machine-learning based evaluation metrics. The exact nature of the relationship between features and ultimate performance is left for future research.

Statistical significance measures over correlation coefficients given in the following sections are derived using Fisher's $z'$ transformation, which converts the skewed sampling distribution of Pearson correlation coefficients to a normal distribution from which the usual confidence intervals can be calculated.

## 4.1 Improved Correlation with Human Judgements

SVMs using Gaussian kernels define a class of models in which the specific model achieving best performance must be located; in order to choose values for the parameters $C$ and $\sigma$ (described in section 3.3) a grid search is employed. For each parameter, a discrete set of values is defined manually and laid along a unique axis. The resulting "grid" (two dimensional in this case) consists of all possible parameter value combinations, and for each combination a SVM is trained using the entire training set. The performance of each model is subsequently evaluated on the validation set by computing the percentage of hypotheses that are correctly classified as human or machine-produced—"classification accuracy"—and the best model is selected. If performance levels between adjacent models on the grid are significantly discontinuous, the grid may be refined and the process repeated until a satisfactory minimum is found.

Using the basic feature set, maximum validation performance is attained at SVM/kernel parameter values of $C = 50$ and $\sigma = 10$. The resulting overall classification accuracy is 64.4%, with 58.7% accuracy classifying human translations and 70.0% accuracy classifying machine translations. Classification accuracy, however, does not reflect the success of the metric; instead the desired characteristic is a strong correlation with human judgements. For the selected model, the Pearson correlation coefficient with respect to human evaluations over the test set is 0.38, and the Spearman correlation coefficient is 0.36.

Figure 4.1 shows the levels of correlation obtained by a single human judge (the human metric), the SVM metric using the optimal parameters, and various current automatic MT evaluation metrics. Raw data are presented in table 4.1. The SVM-based metric outperforms all other automatic evaluators at 95% significance, indicating that it is a more reliable sentence-level translation quality estimator. The SVM-based metric fails to reach the performance level of a human, but makes up approximately half of the gap between the best previously known automatic metrics and the human metric.

Note that in figure 4.1, the correlation results of the NIST score have not been included. Due to the dependence of the NIST metric on a corpus from which information weights can be generated, no single "true" NIST score exists. Perhaps as a result of such considerations (e.g., an inadequate corpus), the computed NIST scores exhibit a misleadingly low correlation with human judgements, and have been excluded for that reason. Previous work, however, suggests that the NIST metric should perform comparably to other metrics such as Bleu and FMS [5, 8].

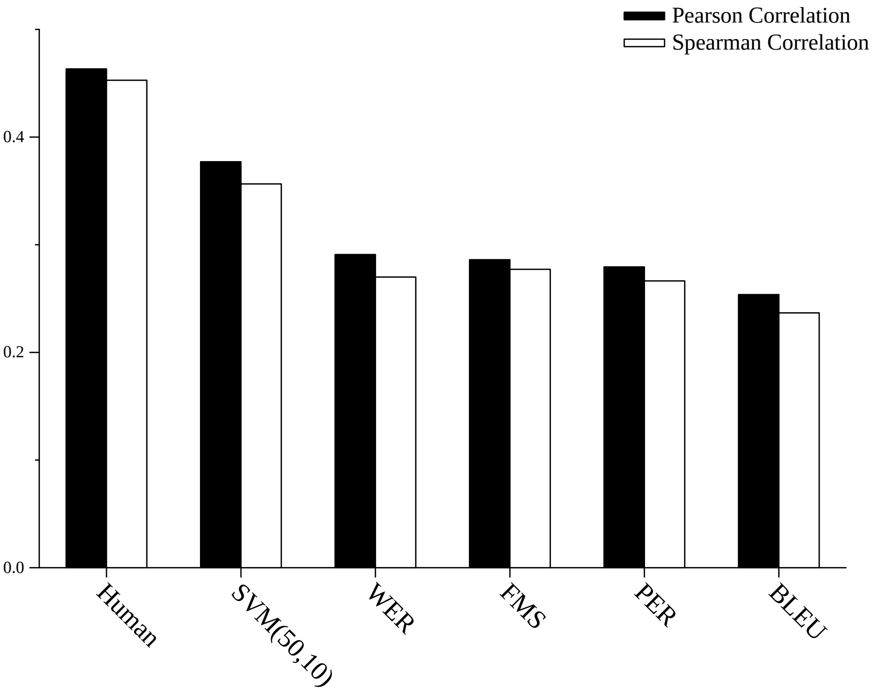

Figure 4.1: Pearson and Spearman correlation coefficients for the human metric, a machine-learned SVM metric, and four standard automatic metrics. The gap between the machine-learned metric and the other automatic metrics is statistically significant at 95%, as is the gap between the human and machine-learned metrics.

Also note that the correlations reported here are lower than those noted by Foster et al. and reproduced in figure 2.1, though the same 633 test examples and human judgements are used in both cases. The apparent discrepancy is due to the fact that the prior work considered metrics computed with respect to four reference translations, while here only three references are used so that the fourth may act as a human hypothesis. The meaningfulness of any metric naturally decreases as the example set of acceptable translations shrinks.

| Metric | Pearson Coefficient | Spearman Coefficient |
|---:|:---:|:---:|
| Human | 0.4633 | 0.4528 |
| SVM(50,10) | 0.3771 | 0.3563 |
| WER | 0.2909 | 0.270 |
| FMS | 0.2861 | 0.2772 |
| PER | 0.2794 | 0.2664 |
| Bleu | 0.2537 | 0.2367 |

Table 4.1: Data for figure 4.1

## 4.2 Success of the Modified Criterion

In addition to producing an automatic MT evaluation metric that outperforms current automatic metrics, however, results show that the classification criterion is also generally successful as an approximation to the goal of correlation with human judgements. For a modified criterion to be successful, procedures that optimize for that criterion must be shown to also optimize for the desired or true criterion; in this case, knowing that improving the classification accuracy of a SVM model tends to improve its correlation with human judgements should allow a metric designer to do without an expensive test set and tweak models to improve classification accuracy with faith that the desired effect will result. Figure 4.2 shows the strong empirical relationship observed between classification accuracy and human judgement correlation by plotting the two performance measures against each another for the entire grid search space of SVMs using the basic feature set. Significant positive correlation is apparent (Pearson coefficient of 0.855), indicating that tweaking SVM parameters to maximize classification accuracy tends to encourage correlation with human judgements as well.

In fact, this result justifies the somewhat blind model selection performed earlier. The chosen model exhibits not only the highest classification accuracy—observable using only the validation set, which is generated automatically—but also a human judgement correlation within 1.5% of the best achieved by any SVM using the basic feature set. This kind of ability to choose a successful metric without relying on resource-intensive human judgements is the key to the approach: once the reliability of the modified criterion is established, new metrics can be implemented, optimized, and applied to modern MT systems without requiring any additional data collection.

## 4.3 Sensitivity to Features

Upon performing similar experiments using the extended feature set, however, a striking characteristic emerges. As seen in figure 4.3, models trained on the larger feature set achieve a higher overall classification accuracy than those trained only on the basic features,

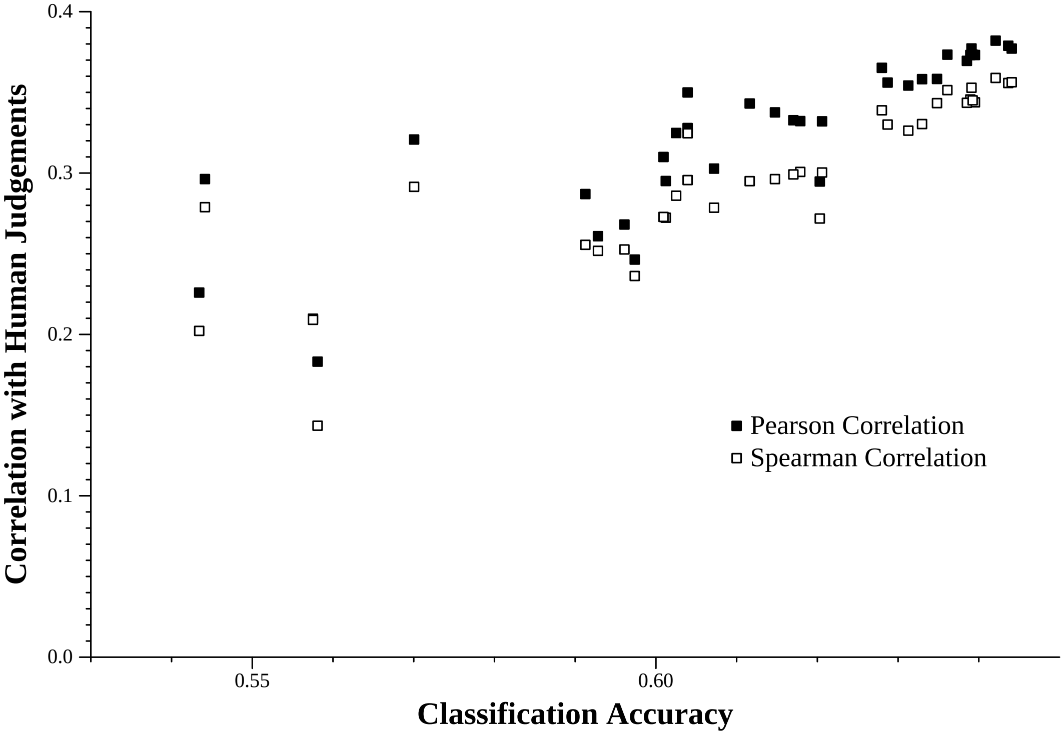

Figure 4.2: Automatic metric correlation with human judgements versus classification accuracy for a wide range of learning parameters. The results suggest that training for the simplified classification criterion induces the strong correlation necessary for a successful evaluation metric. The meta-correlation between classification accuracy and Pearson correlation results for SVM classifiers is 0.855, significant at 99%. Raw data are available in appendix A.

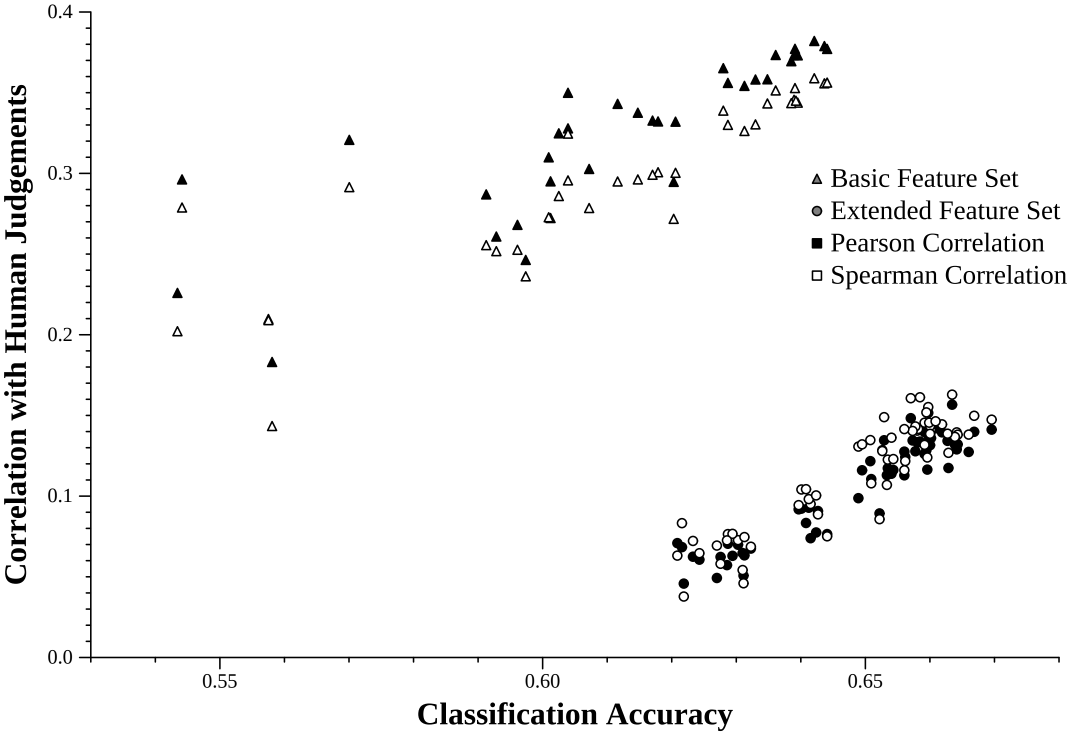

Figure 4.3: Automatic metric correlation with the human judgements versus classification accuracy for a wide range of learning parameters and two different feature sets. The results suggest that, though training for the classification criterion induces correlation in both cases, the specific feature set is a crucial factor in determining the success of the metric. Raw data are available in appendix A.

as should be expected; furthermore, the same positive correlation between classification accuracy and human judgement correlation is clearly apparent for these models as a group. However, the highest level of human judgement correlation reached is less than half that attained by the models trained on basic features; the seemingly linear relationship between the two performance measures is shifted toward high classification accuracy and low human judgement correlation when the extended feature set is employed. Therefore, while the modified criterion continues to encourage parameter optimization over a fixed set of features, the specific set of features in use can apparently have a large impact on the range of achievable performance. Parse-based features, in particular, have led the model to seriously degraded correlation with human judgements.

While the opaque nature of SVM classifiers makes it difficult to discern the exact cause of the problem, intuitively it seems possible that the additional features include information useful for classification and yet irrelevant to human evaluators. For example, it may be that the MT system, for whatever reason, produces outputs that parse with an inordinate number of `SBAR`s compared to the references—a fact that the training phase is likely to exploit—but that humans evaluate with no particular awareness of this phenomenon; perhaps `SBAR`s do not perceptually degrade the quality of the sentence. Such an explanation is particularly convincing for the parse-based features included in the extended set since parsing the disfluent outputs from the MT system can be difficult or impossible in many cases. The tagger and parser employed have been trained on the assumption that structure exists within their inputs; their job is merely to locate it. However, if such structure does not exist at all then the behavior of these models may be unpredictable, or worse, pathological. Indeed, the exploitable information might not come from the MT system actually producing many `SBAR`s, but from the tendency of the tagger/parser combination to produce many `SBAR`s when there is little actual structure to be deduced.

It is not currently clear how to address these issues. One suggestion might be that they do not need addressing at all—even if the results of the metric do not currently correlate with human judgements, there can be no harm in encouraging MT systems to produce outputs that appear more like human outputs; systems might reasonably reduce the number of `SBAR`s produced even if the perceptual effect for humans is negligible. Once these pathological behaviors with respect to the feature set are worked out, then, a metric could be retrained and would no longer exhibit the poor correlation seen here. It is also possible, however, that such a method would lead to much effort wasted on addressing issues that do not matter, or to circular system adjustments with which errors would merely move from one type to the next without being truly removed. In any case, feature selection is clearly an important part of applying machine learning methods to the problem of automatic MT evaluation. Though one successful feature set is identified here, current work aims to understand more generally the interactions between feature sets and correlation performance.

# Chapter 5

# Conclusions

Applying machine learning to the problem of machine translation evaluation is primarily difficult because the desired human judgement targets are too expensive to make available in large quantities, particularly as the population of MT outputs may be changing constantly. The solution proposed here is to approximate human judgements with a binary decision variable that simply reports whether a translation was produced by a human or by a machine, thereby eliminating the need for user data collection. Results indicate that this approach is very effective given certain feature sets, greatly improving sentence-level correlation between metric scores and human judgements over current automatic metrics, and that for a fixed set of input features, the classification criterion is strongly linked with such correlation. Thus, the method provides the ability to optimize for classification and obtain improved correlation even without access to a human evaluated test set.

## 5.1 Future Work

Other learning methods might be applied to the problem, however. Unsupervised learning could be used produce an organization of the data in which translations of a similar quality are placed near each other; such a map could then be translated to a continuous metric with only a few human evaluations as reference points. More generally, active learning methods might allow a variable cost/performance trade off, using as much evaluation data as is available while still taking advantage of a large unlabeled training set. These methods have the potential, however, to discover patterns in the data that are irrelevant to the task, and might end up grouping translations by some unknown property. Whether such a property drawn from the data would be any more or less informative than that of having been produced by a human or machine (effectively enforced here as the property of organization) is an open question.

Furthermore, the approach outlined here appears to have a strong dependence on the feature set with which it is provided, particularly as the training criteria is an approximation to the desired result. As discussed above, this effect may or may not present a serious problem, but it deserves a deeper examination. Current work involves training models on single features to see if features can be selected automatically or characterized in some way so as to assemble the best possible collection, as well as performing user studies to determine if the low correlation exhibited by the extended feature metric in fact produces noticeably worse real-world translation judgements. Another avenue for research might include training metrics over a larger range of system outputs, e.g., from many different MT systems, as perhaps greater variation in hypothesis translations would reduce exploitation of individual pathological behaviors by the training algorithm.

A final aspect of machine learning approaches to MT evaluation that deserves attention is the degree of customization that they potentially provide in addition to merely improved performance. While the focus here has been on improving correlation with human judgements, system designers might also find it useful to add features tuned specifically to issues of concern, thereby improving the metric's sensitivity to particular aspects of translations. For example, the Syntax for Statistical MT team at the 2003 Johns Hopkins workshop observed that Bleu was not particularly sensitive to syntactic improvements [9]; a learning-based metric, on the other hand, could be designed specifically to take syntax into account, thereby allowing finer-grained error analysis than is currently possible.

As statistical learning and otherwise data-driven methods in MT become more powerful, the potential for receiving valuable guidance from accurate, reliable, low-level automatic evaluation metrics is growing dramatically, and learning approaches to MT evaluation hold the promise of improved performance, as demonstrated here, and increased customization.

With the gains in flexibility and power over heuristic methods, however, comes a greater responsibility for understanding and controlling the metric's behavior, whether through the evaluation of many example hypothesis translations or the careful selection of features. Intelligent human input, then, remains paramount in ensuring that any automated evaluation metric is relevant and robust.

## 5.2 Acknowledgments

Many thanks to Stuart Shieber, George Foster, Simona Gandrabur, and Cyril Goutte, who gave me their time, guidance, and indispensable help throughout this process.

# Bibliography

[1] Franz Josef Och. “Minimum Error Rate Training for Statistical Machine Translation.” In “ACL 2003: Proc. of the 41st Annual Meeting of the Association for Computational Linguistics,” Japan, Sapporo, July 2003.

[2] C. Tillman, S. Vogel, H. Ney, H. Sawaf, and A. Zubiaga. “Accelerated DP-based Search for Statistical Translation.” In Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech ’97), pages 2667–2670, Rhodes, Greece, September 1997.

[3] Kishore A. Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “BLEU: a Method for Automatic Evaluation of Machine Translation.” Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY, September 2001.

[4] George Doddington. “Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics.” In Human Language Technology: Notebook Proceedings: 128-132, San Diego, 2002.

[5] Joseph P. Turian, Luke Shen, and I. Dan Melamed. “Evaluation of Machine Translation and its Evaluation.” In Proceedings of MT Summit IX; New Orleans, USA, 23-28 September 2003.

[6] Eduard Hovy, Margaret King, and Andrei Popescu-Belis. “Principles of Context-Based Machine Translation Evaluation.” Machine Translation, 16, pp. 1-33, 2002.

[7] Georges Van Slype. “Critical Methods for Evaluating the Quality of Machine Translation.” Prepared for the European Commission Directorate General Scientific and Technical Information and Information Management, Report BR-19142, Bureau Marcel van Dijk, 1979.

[8] George Foster, et el. “Confidence Estimation for Machine Translation.” Final Report, 2003 Johns Hopkins Workshop on Speech and Language Engineering, Baltimore, MD, 2003.

[9] Franz Josef Och, et el. “Syntax for Statistical Machine Translation.” Final Report, 2003 Johns Hopkins Workshop on Speech and Language Engineering, Baltimore, MD, 2003.

[10] Nello Cristianini. “Support Vector and Kernel Methods for Learning.” ICML Tutorial, Williamstown, MA, 2001.

[11] Ronan Collobert, Samy Bengio, and Johnny Marithoz. “Torch: a modular machine learning software library.” Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[12] Michael Collins. “Head-Driven Statistical Models for Natural Language Parsing.” PhD Dissertation, University of Pennsylvania, 1999.

[13] Adwait Ratnaparkhi. “A Maximum Entropy Part-Of-Speech Tagger.” In Proceedings of the Empirical Methods in Natural Language Processing Conference, University of Pennsylvania, 1996.

# Appendix A

# Data Tables

| | | σ | | | | |
|---|---|---|---|---|---|---|
| | | 10 | 25 | 50 | 75 | 100 |
| C | 5 | **a.** 0.6391<br>**b.** 0.3772<br>**c.** 0.3529 | **a.** 0.6392<br>**b.** 0.3734<br>**c.** 0.3450 | **a.** 0.6287<br>**b.** 0.3561<br>**c.** 0.3301 | **a.** 0.6206<br>**b.** 0.3320<br>**c.** 0.3003 | **a.** 0.6147<br>**b.** 0.3376<br>**c.** 0.2963 |
| | 10 | **a.** 0.6437<br>**b.** 0.3789<br>**c.** 0.3558 | **a.** 0.6395<br>**b.** 0.3732<br>**c.** 0.3438 | **a.** 0.6280<br>**b.** 0.3652<br>**c.** 0.3389 | **a.** 0.6179<br>**b.** 0.3322<br>**c.** 0.3007 | **a.** 0.6116<br>**b.** 0.3431<br>**c.** 0.2950 |
| | 25 | **a.** 0.6421<br>**b.** 0.3821<br>**c.** 0.3589 | **a.** 0.6390<br>**b.** 0.3731<br>**c.** 0.3456 | **a.** 0.6039<br>**b.** 0.3499<br>**c.** 0.3246 | **a.** 0.6170<br>**b.** 0.3327<br>**c.** 0.2992 | **a.** 0.6025<br>**b.** 0.3248<br>**c.** 0.2860 |
| | 50 | **a.** 0.6441<br>**b.** 0.3771<br>**c.** 0.3563 | **a.** 0.6385<br>**b.** 0.3695<br>**c.** 0.3435 | **a.** 0.5700<br>**b.** 0.3208<br>**c.** 0.2915 | **a.** 0.6072<br>**b.** 0.3027<br>**c.** 0.2785 | **a.** 0.6012<br>**b.** 0.2951<br>**c.** 0.2723 |
| | 75 | **a.** 0.6361<br>**b.** 0.3734<br>**c.** 0.3514 | **a.** 0.6330<br>**b.** 0.3582<br>**c.** 0.3304 | **a.** 0.5441<br>**b.** 0.2963<br>**c.** 0.2789 | **a.** 0.5928<br>**b.** 0.2609<br>**c.** 0.2518 | **a.** 0.6009<br>**b.** 0.3100<br>**c.** 0.2728 |
| | 100 | **a.** 0.6348<br>**b.** 0.3583<br>**c.** 0.3433 | **a.** 0.6313<br>**b.** 0.3542<br>**c.** 0.3263 | **a.** 0.5434<br>**b.** 0.2259<br>**c.** 0.2022 | **a.** 0.5974<br>**b.** 0.2464<br>**c.** 0.2362 | **a.** 0.5961<br>**b.** 0.2681<br>**c.** 0.2526 |
| | 150 | **a.** 0.6203<br>**b.** 0.2948<br>**c.** 0.2718 | **a.** 0.6039<br>**b.** 0.3279<br>**c.** 0.2957 | **a.** 0.5581<br>**b.** 0.1831<br>**c.** 0.1434 | **a.** 0.5575<br>**b.** 0.2098<br>**c.** 0.2090 | **a.** 0.5913<br>**b.** 0.2870<br>**c.** 0.2555 |

Table A.1: Classification accuracy (a), Pearson correlation with human judgements (b), and Spearman correlation (c) for each SVM grid point using the basic feature set.

| | | σ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 25 | 40 | 50 | 55 | 60 | 65 | 70 | 75 | 100 |
| C | 5 | **a.** 0.6219 | **a.** 0.6311 | **a.** 0.6323 | **a.** 0.6303 | **a.** 0.6287 | **a.** 0.6313 | **a.** 0.6294 | **a.** 0.6286 | **a.** 0.6270 |
| | | **b.** 0.0458 | **b.** 0.0508 | **b.** 0.0675 | **b.** 0.0699 | **b.** 0.0705 | **b.** 0.0633 | **b.** 0.0630 | **b.** 0.0572 | **b.** 0.0493 |
| | | **c.** 0.0378 | **c.** 0.0460 | **c.** 0.0686 | **c.** 0.0727 | **c.** 0.0765 | **c.** 0.0746 | **c.** 0.0766 | **c.** 0.0726 | **c.** 0.0694 |
| | 10 | **a.** 0.6209 | **a.** 0.6441 | **a.** 0.6427 | **a.** 0.6397 | **a.** 0.6412 | **a.** 0.6401 | **a.** 0.6408 | **a.** 0.6424 | **a.** 0.6415 |
| | | **b.** 0.0709 | **b.** 0.0765 | **b.** 0.0908 | **b.** 0.0918 | **b.** 0.0928 | **b.** 0.0922 | **b.** 0.0833 | **b.** 0.0775 | **b.** 0.0739 |
| | | **c.** 0.0631 | **c.** 0.0751 | **c.** 0.0887 | **c.** 0.0944 | **c.** 0.0981 | **c.** 0.1041 | **c.** 0.1043 | **c.** 0.1004 | **c.** 0.0951 |
| | 25 | **a.** 0.6310 | **a.** 0.6509 | **a.** 0.6560 | **a.** 0.6535 | **a.** 0.6543 | **a.** 0.6508 | **a.** 0.6495 | **a.** 0.6540 | **a.** 0.6489 |
| | | **b.** 0.0647 | **b.** 0.1106 | **b.** 0.1129 | **b.** 0.1173 | **b.** 0.1162 | **b.** 0.1217 | **b.** 0.1160 | **b.** 0.1139 | **b.** 0.0987 |
| | | **c.** 0.0542 | **c.** 0.1079 | **c.** 0.1160 | **c.** 0.1226 | **c.** 0.1230 | **c.** 0.1347 | **c.** 0.1321 | **c.** 0.1362 | **c.** 0.1307 |
| | 50 | **a.** 0.6276 | **a.** 0.6526 | **a.** 0.6596 | **a.** 0.6580 | **a.** 0.6600 | **a.** 0.6599 | **a.** 0.6582 | **a.** 0.6577 | **a.** 0.6585 |
| | | **b.** 0.0623 | **b.** 0.1284 | **b.** 0.1164 | **b.** 0.1334 | **b.** 0.1316 | **b.** 0.1365 | **b.** 0.1333 | **b.** 0.1278 | **b.** 0.1338 |
| | | **c.** 0.0581 | **c.** 0.1279 | **c.** 0.1239 | **c.** 0.1406 | **c.** 0.1386 | **c.** 0.1455 | **c.** 0.1416 | **c.** 0.1432 | **c.** 0.1613 |
| | 75 | **a.** 0.6243 | **a.** 0.6562 | **a.** 0.6592 | **a.** 0.6639 | **a.** 0.6609 | **a.** 0.6627 | **a.** 0.6573 | **a.** 0.6560 | **a.** 0.6595 |
| | | **b.** 0.0606 | **b.** 0.1242 | **b.** 0.1261 | **b.** 0.1320 | **b.** 0.1420 | **b.** 0.1342 | **b.** 0.1345 | **b.** 0.1276 | **b.** 0.1285 |
| | | **c.** 0.0646 | **c.** 0.1217 | **c.** 0.1318 | **c.** 0.1367 | **c.** 0.1464 | **c.** 0.1387 | **c.** 0.1404 | **c.** 0.1415 | **c.** 0.1519 |
| | 100 | **a.** 0.6233 | **a.** 0.6533 | **a.** 0.6642 | **a.** 0.6643 | **a.** 0.6696 | **a.** 0.6602 | **a.** 0.6619 | **a.** 0.6592 | **a.** 0.6570 |
| | | **b.** 0.0625 | **b.** 0.1130 | **b.** 0.1290 | **b.** 0.1321 | **b.** 0.1412 | **b.** 0.1359 | **b.** 0.1395 | **b.** 0.1396 | **b.** 0.1483 |
| | | **c.** 0.0722 | **c.** 0.1069 | **c.** 0.1395 | **c.** 0.1381 | **c.** 0.1474 | **c.** 0.1411 | **c.** 0.1445 | **c.** 0.1455 | **c.** 0.1607 |
| | 150 | **a.** 0.6216 | **a.** 0.6522 | **a.** 0.6629 | **a.** 0.6660 | **a.** 0.6634 | **a.** 0.6669 | **a.** 0.6600 | **a.** 0.6597 | **a.** 0.6529 |
| | | **b.** 0.0684 | **b.** 0.0893 | **b.** 0.1174 | **b.** 0.1274 | **b.** 0.1567 | **b.** 0.1399 | **b.** 0.1382 | **b.** 0.1516 | **b.** 0.1346 |
| | | **c.** 0.0832 | **c.** 0.0857 | **c.** 0.1268 | **c.** 0.1381 | **c.** 0.1629 | **c.** 0.1498 | **c.** 0.1435 | **c.** 0.1552 | **c.** 0.1489 |

Table A.2: Classification accuracy (a), Pearson correlation with human judgements (b), and Spearman correlation (c) for each SVM grid point using the extended feature set.