Training a Sentence-Level Machine Translation Confidence Measure

Christopher B. Quirk

Microsoft Research
One Microsoft Way
Redmond, WA 90852 USA
chrisq@microsoft.com

Abstract

We present a supervised method for training a sentence level confidence measure on translation output using a human-annotated corpus. We evaluate a variety of machine learning methods. The resultant measure, while trained on a very small dataset, correlates well with human judgments, and proves to be effective on one task based evaluation. Although the experiments have only been run on one MT system, we believe the nature of the features gathered are general enough that the approach will also work well on other systems.

1. Introduction

Machine Translation (MT) technology is maturing and becoming more pervasive, yet the quality of output is still not ideal: a large portion of MT output is not fluent, fails to preserve meaning, or is otherwise undesirable. Nonetheless current MT systems are useful for a variety of scenarios, including quick translations for browsing, aiding human translators by making a first pass at translation, and in cross-lingual transformation retrieval.

To maximize the utility of existing technology, it is useful to have a confidence measure: some automatic estimate of how effectively a sentence was translated. Such a measure could be used to highlight suspect sentences in a translated document, to present only helpful translations to human translators, to select the best output from multiple MT systems, or to re-rank n-best lists of translations. Discriminative models, such as those presented in this paper, are also applicable to a wide variety of MT systems. As described below, these methods can be applied to non-statistical MT systems just as easily as statistical MT systems.

2. Related Work

Several distinct traditions are evident in the MT confidence estimation literature. There are hand trained measures that gauge the quality of a particular translation (Bernth, 1999), useful in several applications such as selection of the best translation from a set of MT systems or re-ranking n-best translation lists. Unfortunately this requires time-consuming and recurring hand-tuning for any given domain. More recently, statistical measures for estimating confidence on the word and phrase level have appeared (Gandrabur and Foster, 2003; Ueffing et al., 2003). Although this removes the expensive tuning process, the resultant measure is less generally applicable, since a single word translated poorly is not necessarily indicative of overall translation quality, and it is difficult to integrate word-level confidence scores into MT systems when most operate on the sentence level.

Finally, a recent workshop (Blatz et al., 2003) has investigated training sentence-level confidence measures using a variety of fuzzy match scores. While this sounds promising given the recent successes of automated evaluations, we've found a strong correlation between fuzzy-match scores and human quality judgments only on a very high level (Coughlin, 2003). Training against a loose approximation of your desired target feature severely limits the quality of results. Although they dismiss the possibility that a measure can be trained effectively on a small dataset, we find that training on even a small human-tagged dataset outperforms a large automatically-tagged dataset, at least on our particular feature set.

3. Methodology

An ideal confidence estimate would approximate the quality score assigned by a human. Therefore we applied supervised machine learning algorithms to human quality judgments in an attempt to learn a correspondence between features emitted by the translation system and a human quality score.

3.1. MT system and feature set

We first instrumented the hybrid machine translation system MSR-MT (Menezes and Richardson, 2001) to gather a variety of features. At a high level, MSR-MT is a syntactically-informed example based system, trained on domain specific translation resources from parallel sentence aligned bilingual corpora. MSR-MT uses parsers for both source and target

language which produce *logical form* (LF) representations (a predicate argument structure representation). In the training phase, both sides of the parallel corpus are parsed, the LF structures are aligned, and aligned LF chunks of varying sizes are stored in a translation database. At translation time, the system first parses the input sentence into an LF. The translation database is consulted for matching LF chunks, a subset of these chunks is selected, and the target sides of these chunks are merged into a single target logical form. Either hand-crafted or machine learned generation systems are then used to generated a target string from this target LF.

To prepare for training a confidence measure, we instrumented this system to produce a large number of features upon translating a sentence. The first main category of features models characteristics of the source sentence and how difficult it is to parse. We measure the perplexity of the sentence according to a trigram language model: sentences less like our training set are probably more difficult to translate. Since the translation process depends heavily on a parsed LF, we include whether the parser was able to find a spanning parse as a boolean feature. Also the size of the input sentence is somewhat indicative of translation difficulty: shorter sentences have fewer ambiguities and hence tend to be easier to translate. We can also gather similar features on the target side: sentence length, and perplexity according to a trigram language model.

We also gathered features about the translation process itself. While MSR-MT isn't a purely statistical MT system (and hence has no native translation probability estimates), we can measure translation effectiveness in a variety of ways. Since this system depends heavily on learned transfer mappings, we gathered information about the number and average size of the learned mappings—larger mappings generally lead to contextually better translations. When these mappings are unavailable, we combine information from learned dictionaries and human dictionaries to prevent untranslated words from appearing in the target side. Therefore we also emitted counts and percentages of words translated by each information source, both as a whole and subcategorized by type (it's much worse to let a preposition fall through untranslated than a noun, for instance). We also included ratios of the monolingual features: target length over source length, and target perplexity over source perplexity.

Finally, given that MSR-MT is trained on a large bilingual corpus, we expect translation quality to be tightly linked with how well the sentence material is covered in the training corpus. Therefore we compute the minimal tiling of the source sentence and the MT output using substrings of the training corpus; for any word not in the training corpus, we create an implicit one word tile.[1] We include both the average tiling size of the source and the target as well as their difference and ratio.

3.2. Data

The particular version of MSR-MT used in these experiments was a system trained on 351,026 sentence pairs of Spanish-English technical data, such as product manuals from Microsoft products, technical support, and other technical documentation.

We used a held-out, unseen set of 500 sentence pairs for confidence training and test data.[2] These 500 sentences were then translated with the instrumented MSR-MT as described above, saving aside the features gathered from translating each sentence. The MT output and the reference translation were given to human annotators, who were asked to grade the sentence on the following scale of 1 to 4:

- 4 = **Ideal**: Not necessarily a perfect translation, but grammatically correct, and with all information accurately transferred.
- 3 = **Acceptable**: Not perfect (stylistically or grammatically odd), but definitely comprehensible, *and* with accurate transfer of all important information.
- 2 = **Possibly Acceptable**: Possibly comprehensible (given enough context and/or time to work it out); some information transferred accurately.
- 1 = **Unacceptable**: Absolutely not comprehensible and/or little or no information transferred accurately.

Each sentence was judged by six different human annotators. Using a large number of annotators improves our confidence in the gold standard judgment, though it does make gauging inter-annotator agreement difficult. One way to estimate agreement is to first compute a mean score for each sentence, then look at the difference between each individual judgment and the mean for that sentence. This suggests agreement is actually quite good: 71% of the 3,000 judgments are within 1/2 point of the mean, and 95% are within 1 point of the mean.

We split the 500 sentences into 350 sentences for training and 150 sentences for test. For the purposes of

[1] This can be computed efficiently by first building a suffix tree (Gusfield, 1997) of each side of the training corpus to find all possible tilings efficiently, then using a simple dynamic programming solution to find the optimal tiling.

[2] Although it would be preferable to have a larger supervised set, producing human judgments of MT output is slow and expensive; such a small set had to suffice for these experiments.

Dataset	Linear Regression	SVM	Perceptron
350 sentences, WER $<$ 5th percentile	0.74	0.72	0.71
350 sentences, WER $<$ 30th percentile	0.76	0.73	0.74
15,000 sentences, WER $<$ 5th percentile	0.78		0.70
15,000 sentences, WER $<$ 30th percentile	0.66		0.61
350 sentences, human $\geq$ 3.0	0.83	0.82	0.77

Table 1: IROC of each method trained on various tagged datasets

training and testing binary classifiers, we considered those sentences with a mean human judgment greater than or equal to 3.0 as high quality.

3.3. Automatically evaluated data

This manually-annotated dataset is quite small by today's standards. As a comparison, in (Blatz et al., 2003) n-best lists ranging in size from 101 to 16,384 from approximately 5,000 sentences were used, which means that between 500,000 and 80,000,000 distinct training points were available. Yet this plentiful data source lacks human annotations; automatic judgments were the only reasonable course of action for such a large set of data. To gauge the importance of human-annotated judgments, we translated and gathered features on an additional unseen set of 15,000 sentence pairs. In (Blatz et al., 2003), it was found that WER[3] had the best correlation with human judgments among the sentence-level fuzzy match scores. Therefore we computed the WER between each translation output and the reference translation and produced binary classification judgments based on two cutoffs: WER less than the 5th and 30th percentile.

3.4. Training methods

We trained several different classifiers on the given data. Again for the sake of comparison with (Blatz et al., 2003), we attempted to use perceptrons with differing numbers of hidden layers. We also trained standard implementations of support vector machines and decision trees. In addition, we applied a variation on linear regression: a fast method unlikely to overfit on a small training set if the feature set is also small. Since several of the features were constant, collinear, or nearly collinear, training a linear regression on the full feature set lead to a degenerate solution. Instead we used a greedy feature selection algorithm for the linear regression: choose the single feature with the greatest correlation coefficient against the target feature, and then incrementally add the feature that maximizes the correlation coefficient until no increase in correlation occurs.

[3]Here, WER is defined as string edit distance normalized by the length of the string edit alignment.

4. Evaluation

4.1. Relative evaluation

First we used ROC curves to evaluate the performance of each method trained on each tagged corpus: the 350 sentence human-tagged corpus as well as various subsets of the 15,000 sentence automatically-tagged corpus. An ROC plots the sensitivity (the true positive rate) against the specificity (the true negative rate) to and thus provides an effective way to objectively compare binary classifiers. To produce the ROC curve for a particular method trained on a particular corpus, we computed the confidence score for each sentence in the test output, computed sensitivity and specificity for each possible cutoff point, and plotted those points on a graph.

When plotted in this manner, an ideal classifier will have a curve that stays maximally close to the upper-left corner, where a random classifier will stay on the line from (0,0) to (1,1). The integral of the curve, or IROC, provides a single numeric evaluation for overall performance: an ideal classifier will have an IROC of 1, where a random classifier will have an IROC of 0.5.

Table 1 is a comparison of the IROC values for each training set and method. Importantly, all classifiers trained on just 350 sentences of human-tagged data outperformed those trained on as much as 15,000 sentences of automatically-tagged data. Ablation suggests that more data is not the answer: performance improves marginally with the addition of orders of magnitude more data. The training method also makes a significant impact. We were surprised to find that a simple linear regression is so effective, although our bias toward selecting a few highly-correlated features is much less prone to overfitting even on a small dataset. Unfortunately we were unable to provide results for SVMs trained on the 15,000 sentence corpora; training on these sets was simply too slow (more than 1 week CPU time). Including up to ten hidden units in perceptron training had no significant impact on the IROC, hence those numbers were omitted.

4.2. Task based evaluation

In addition to providing a relative ranking of the performance of each of these methods, the evaluation of the previous section provides a benchmark of how the confidence measures would perform on the task of identifying high-quality translations. In the interest of showing a clear and objective gain from using the confidence measure, we've selected a secondary task-based evaluation: selecting the best translation from a pair of systems.

We took the same test set used above and retranslated using the Systran machine translation system. The Systran output was then evaluated by six human annotators using the methodology described in §3.2, and the resultant scores were combined into a single mean human score.

One way to use the confidence score as a translation selection mechanism is to pick a cutoff point for the confidence estimate, and use only those sentences from MSR-MT when the confidence was above that cutoff, falling back to the other translation system otherwise. Ideally this cutoff point would be selected so as to maximize human score. Therefore we split the 150 testing sentences into a subset of 50 for parameter selection and 100 for evaluation. We evaluated translation selection by confidence score in addition to an upper-bound oracle selection (where the best scoring translation is always selected) and random selection (to provide a baseline) on those 100 sentences. The confidence estimate used here is the linear regression trained on 350 human-tagged sentences.

Method	*Human score*
Systran alone	2.11
MSR-MT alone	2.51
Confidence estimate selection	2.55
Oracle selection	2.65
Random selection	2.31

Table 2: Translation selection by confidence score

Table 2 demonstrates that the confidence estimate can be used to boost the overall score, albeit by only a small amount. Yet this is a promising result for several reasons. First, the confidence estimate was only provided for MSR-MT translations, though we're selecting from two systems; better results could presumably be found by also modeling information about the Systran output. Also note that Systran performs worse than MSR-MT on average, yet the confidence metric succeeds in selecting a set that outperforms MSR-MT alone.

5. Conclusion

As noted in previous work (Blatz et al., 2003), a massive amount of tagged data is necessary to train discriminative models with large feature sets. Our results demonstrate that usable models can be constructed from small yet indicative feature sets on a very small human-tagged dataset, and that even such a small human-tagged dataset is preferable to a large automatically-tagged dataset.

Most importantly, though, we must explore new feature sets. Given that we're using discriminative models, we're free to use features that incorporate any amount of external resources. This provides an opportunity to explore a wide variety of statistical MT models even on non-statistical MT systems.

6. Acknowledgments

We would like to thank John Platt and Max Chickering for the use of their SVM and Decision Tree tools. Without the careful human evaluations conducted by the Butler Hill group and coordinated by Mo Corston-Oliver, this research could not have been conducted. Finally, insightful conversations with Joshua Goodman regarding automatic means of augmenting human data sets and Bob Moore regarding evaluation techniques are gratefully appreciated.

7. References

Bernth, Arendse, 1999. A confidence index for machine translation. *Proceedings of TMI-99*.

Blatz, John, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing, 2003. Confidence estimation for statistical machine translation. *Johns Hopkins Summer Workshop Final Report*.

Coughlin, Deborah, 2003. Correlating automated and human asessments of machine translation quality. *Proceedings of the MT Summit IX*.

Gandrabur, Simona and George Foster, 2003. Confidence estimation for text prediction. *Proceedings of the Conference on Natural Language Learning (CoNLL 2003)*.

Gusfield, Dan, 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Menezes, Arul and Stephen D. Richardson, 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. *Proceedings of the Workshop on Data-driven Machine Translation*, 39.

Ueffing, Nicola, Klaus Macherey, and Hermann Ney, 2003. Confidence meausres for statistical machine translation. *Proceedings of the MT Summit IX*.