

Discriminative Reranking for Machine Translation

Libin Shen

Supervisors: **Aravind K. Joshi** ^{*}, **Anoop Sarkar** [†]

Department of Computer and Information Science

University of Pennsylvania

libin@linc.cis.upenn.edu

Abstract

We propose to use discriminative machine learning algorithms to rerank the output of a baseline machine translation (MT) system. A baseline system first provides a set of N-best candidate translations. Then we extract various features from those candidates, and use these features to rerank the N-best list via discriminative machine learning algorithms.

1 Introduction

The proposed work here for the academic year 2003-04 comes from my participation in the JHU 03 Workshop. The work is also part of a larger project of using discriminative machine learning in natural language processing. Discriminative reranking for machine translation is an important application of our research project.

1.1 Statistical Machine Translation

After the IBM Models (Brown et al., 1993) was proposed in 1993, various generative models have been proposed for Statistical Machine Translation (SMT) in the last ten years. In the IBM Models, the source-channel formalism, which was previously used in speech recognition, was applied to machine translation. Machine translation was reduced to a decoding problem. The training was to estimate two sets of parameters, the language model and the translation model. The language model is used to compute the generative probability of a translation, and the translation model is used to compute the generative probability of a source sentence given its translation.

The IBM Models are strictly word-based models, so they did not take advantage of the linguistic knowledge dominating the encoding and decoding of natural languages. Due to the lack of syntactic knowledge, the IBM models cannot handle many translation patterns, such as long-distance movement, relative distortion, questions, etc.

^{*} Dept. of Computer and Information Science, Univ. of Pennsylvania, joshi@linc.cis.upenn.edu

[†] School of Computing Science, Simon Fraser University, anoop@cs.sfu.ca

Thereafter, many MT models were proposed as extensions of the IBM models. These systems used more linguistically knowledge, so as to decrease the decoding perplexity and increase the translation quality.

Wang and Waibel (1998) proposed an alignment model based on shallow model structures. Since their translation model reordered phrases directly, it achieved higher accuracy for translation between languages with different word orders. In (Och et al., 1999), a two-level alignment model was employed to utilize shallow phrase structures. Alignment between templates was used to handle phrase reordering, and word alignments within a template was used to handle phrase to phrase translation.

However, phrase level alignment cannot handle long distance reordering effectively. Parse trees have also been used in alignment models. Yamada and Knight (2001) proposed a tree to string model for alignment. Parse trees of the source language were used in the alignment model. Gildea (2003) proposed a tree to tree alignment model in which subtree cloning was used to handle more reordering in parse trees.

The source channel model has been successfully used in speech recognition, but the performance of source channel based MT systems is still unsatisfactory. The most important reason is that the underlying independence assumptions of translation models are too strong. For machine translation, we need a large context and a rich feature set. However, it is rather difficult to incorporate various features in a large context via generative models.

1.2 Reranking Techniques

Parsing is another field of natural language processing in which generative models have been widely used. In the recent year, the so-call reranking techniques, especially discriminative reranking, have resulted in significant improvements in parsing.

The reranking approach is as follows. First, a baseline parser generates N-best results. Then we extract features from these N-best parses, and use these features to rerank the N-best list. Compared with generative models, the reranking approach searches for the global optimum, if the quality of N-best results is good enough. It also allows the use of local and global features of various kinds, which are unavailable in generative models.

Various learning algorithms have been employed in parse reranking, such Boosting (Collins, 2000), Perceptron (Collins and Duffy, 2002), Support Vector Machines (Shen et al., 2003) and Log-linear models (Charniak, 2000; Collins, 2000). The reranking technique gives rise to 13.5% error deduction in labeled recall/precision over the previous best generative parsing systems.

2 Discriminative Reranking for MT

Inspired by the works in parse reranking, we plan to explore the application of discriminative reranking to machine translation. We hope to improve the performance of MT systems by exploiting the advantages of the reranking techniques.

The procedure of MT reranking is similar to parse reranking. We first use a baseline MT system to generate N-best translations. Then, we analyze the structures of the source sentence and its translation by POS tagging, parsing, and derivation tree extraction. We then extract features from these linguistic structures. Finally, we rerank the N-best translations with respect to these features.

2.1 Baseline MT system

A baseline MT system is expected to use a simple language model and a simple translation model. For example an baseline model uses a phrase based alignment model instead of a tree based alignment model. The baseline MT system also provides word to word alignments between the source sentence and its translation, which will be used in our feature extraction. Since the baseline is simple in principle, it is relatively easier to implement and runs much faster than systems using complicated encoding and decoding models.

Another important issue is the structure of the N-best training dataset. In the JHU Workshop 03, the SMT team used a development dataset ¹ of 993 Chinese sentences, each of which has 1000-best English translations. However, this dataset restricts the use of features. For example, we cannot use any individual lexical items as features due to the risk of overfitting.

In our approach, the training dataset contains a large number of Chinese sentences each of which has about a small N-best candidates. In this way, to rerank the N-best lists is more like a classification problem instead of a regression problem because of the size of the N-best list. This way is more suitable for the nature of machine translation, considering the lack of reliable measures of the quality of a translated sentence. The score of a translation is unreliable, so the regression problem suffers too much noise. But it is not difficult to find a set of good translations and a set of bad translations. Classification between good and bad translations is much more stable.

2.2 Features Extraction

The nature of the new data set allows us to search a much larger space of the features, i.e. arbitrary fragments of deep syntactic structures, with much less risk of overfitting. Due to the characteristic of the dataset and the feature space, it is very natural for us to use various discriminative machine learning algorithms, such as

¹This development dataset is used to train the reranking algorithm.

Boosting, Perceptron, and SVMs, which have been successfully used in the parse reranking task.

For machine translation, two classes of features can be used to handle language models and alignment models respectively. For the first class, we may use features extracted from translations, such as N-gram POS tags, NP chunks, segments of parse trees, and LTAG elementary trees. For the second class, we plan to use pairs of aligned syntactic structures (or fragments of these structures) as features.

2.3 Discriminative Reranking for MT

We also plan to explore variants of these algorithms to make them more suitable for the MT reranking task. For example, in the JHU workshop 03 we proposed the Multi-Bias Perceptron algorithm (SMT Team, 2003) to rerank machine translated sentences. This algorithm is adapted from the traditional Perceptron algorithm, and it trains a unique bias only for the English translations corresponding to the same Chinese sentence. The new algorithm is more suitable to MT reranking than the traditional Perceptron algorithm.

2.4 Preliminary Experiments

In the summer workshop, we designed two preliminary experiments of using individual structures as features. The training data for our perceptron algorithm contains 993 Chinese sentences. Each Chinese sentence has 1000 best translations generated by Och's system (Och et al., 1999). We list the 1000 best translations with respect to delta BLEU scores. Then the top 30% of the translations are used as positive samples, and the bottom 30% of the translations are used as negative samples. The test data had 878 Chinese sentences. Each one has four reference translations. The BLEU score on the test set is used for evaluation.

In our first experiment, we use aligned POS-tag sequences as features. For each pair of the aligned templates, we first replace all the words with POS-tags. Then the POS-tag sequence on the English side is used as a feature. Furthermore, pairs of the corresponding sequences are also used as a features. Thus, each translation is represented by a vector defined on these POS-tag based features. The value of a feature is 1 if the corresponding POS-tag sequence or pair of sequences appears in a samples. Otherwise, the value of the feature is set to 0. We say the feature is active in this sample if its value is 1.

The feature space of the POS-tag features is about 30,000. However, the feature space is very sparse. There are about 31 active features for each sample on average. We use the same training and test data set as reported in (SMT Team, 2003). The BLEU score on the training set is 34.2%, and the BLEU score on the test set is 30.9%.

Our next experiment uses fragments of translation parse trees as features. For each translation, we first parse it with the Collins' parser. Then we use subtrees in the parse tree as features for that translation. We use the following subtrees in our experiment.

- Rules used in derivation, which is equal to (parent node, all child nodes) structure.
- (parent node, two adjacent child nodes)
- (parent node, three adjacent child nodes)

By parent node we mean the constituent tag of the parent nodes. Similar for the child nodes.

The feature space of the subtree features is about 65,000. There are about 100 active features for each sample on average. The BLEU score on the training set is 30.3%, and the BLEU score on the test set is 30.5%.

Although the performance of the reranking result of the perceptron algorithm is not convincing, this approach is still attractive. Since the training data contains only 997 Chinese features, we cannot use some useful features such as lexicalized structures. The result of using only the POS-tags features is already as good as the well optimized baseline system.

We need to further research in adapting discriminative algorithm to MT reranking. The Multi-Bias perceptron algorithm is our first step in this direction. On the other hand, we will also work on a data set of different style, i.e. 100,000 Chinese sentences with 20 translations for each one, so that discriminative machine learning algorithms become more useful.

2.5 Advantages of Discriminative Reranking

Compared with previous work in SMT, the approach of discriminative reranking has the following advantages.

Discriminative ranking enables us to use global features which are unavailable for the baseline system. Second, we can use features of various kinds and need not worry about fine grained smoothing issue.

Furthermore, the reranking approach helps to decrease the decoding complexity. The underlying generative system uses simple language models and alignment models. For example, we can use the template based alignment as the baseline system. Then we extract syntactic structures from the source and target sentences, and we can get the alignment on deep syntactic structures. Thus, we decrease the computational complexity of tree alignment.

Finally, our statistical machine learning approach is theoretically well founded, and has been shown effective in many NLP tasks. Compared with some generative

models for machine translation, the reranking system is relatively easy to implement since the decoding on complex structures is avoided.

3 Research Approach and Equipments

We will build a training dataset of 100,000 Chinese sentences. Each sentence will have 10 good translations and 10 bad translations selected from a 1000 best list generated by a Chinese-English MT system that uses simple language models and alignment models, for example phrase based models. Then, we extract syntactic structures from a word aligned Chinese sentence and its English translation, and use those individual structures as the features for that translation. Discriminative machine learning algorithms, such as variants of Perceptron, Winnow and Boosting, will be applied to the reranking of the machine translated sentences.

As to the dataset, we plan to use the Chinese-English data used in **NIST 2003 MT Evaluation**. So we need access to that dataset.

For the baseline MT system, we hope to use some existing systems, for example, Franz Och's template based MT system that we have used in the JHU 03 workshop.

References

- P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.
- E. Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of NAACL 2000*.
- Michael Collins and Nigel Duffy. 2002. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of ACL 2002*.
- Michael Collins. 2000. Discriminative reranking for natural language parsing. In *Proceedings of the 7th International Conference on Machine Learning*.
- D. Gildea. 2003. Loosely tree-based alignment for machine translation. In *ACL 2003*.
- F. Och, C. Tillmann, and H. Ney. 1999. Improved alignment models for statistical machine. In *EMNLP-WVLC 1999*.
- L. Shen, A. Sarkar, and A. K. Joshi. 2003. Using LTAG based features in parse reranking. In *Proc. of EMNLP 2003*.
- SMT Team. 2003. Final report. JHU Summer Workshop 2003.
- Y. Wang and A. Waibel. 1998. Modeling with structures in statistical machine translation. In *COLING-ACL 1998*.
- K. Yamada and K. Knight. 2001. A syntax-based statistical translation model. In *ACL 2001*.