Chapter 4
Machine Learning Approaches for Bangla Statistical Machine Translation

Maxim Roy
Simon Fraser University, Canada

ABSTRACT

Machine Translation (MT) from Bangla to English has recently become a priority task for the Bangla Natural Language Processing (NLP) community. Statistical Machine Translation (SMT) systems require a significant amount of bilingual data between language pairs to achieve significant translation accuracy. However, being a low-density language, such resources are not available in Bangla. In this chapter, the authors discuss how machine learning approaches can help to improve translation quality within as SMT system without requiring a huge increase in resources. They provide a novel semi-supervised learning and active learning framework for SMT, which utilizes both labeled and unlabeled data. The authors discuss sentence selection strategies in detail and perform detailed experimental evaluations on the sentence selection methods. In semi-supervised settings, reversed model approach outperformed all other approaches for Bangla-English SMT, and in active learning setting, geometric 4-gram and geometric phrase sentence selection strategies proved most useful based on BLEU score results over baseline approaches. Overall, in this chapter, the authors demonstrate that for low-density language like Bangla, these machine-learning approaches can improve translation quality.

INTRODUCTION

Machine Translation (MT) from Bangla to English has recently become a priority task for the Bangla Natural Language Processing (NLP) community. MT is a hard problem because of the highly complex, irregular and diverse nature of natural language. MT refers to computerized systems that utilize software to translate text from one natural language into another with or without human assistance. It is impossible to accurately model all the linguistic rules and relationships that shape the translation process, and therefore MT has to make decisions based on incomplete data.

DOI: 10.4018/978-1-4666-3970-6.ch004

In order to handle this incomplete data, a principled approach is to use statistical methods to make optimum decisions given incomplete data. Statistical Machine Translation (SMT) uses a probabilistic framework to automatically translate text from one language to another. Using the co-occurrence counts of words and phrases from the bilingual parallel corpora where sentences are aligned with their translation, SMT learns the translation of words and phrases. From the initial word-based translation models, research on SMT has seen dramatic improvement. At the end of the last decade the use of context in the translation model, which is known as a phrase-based SMT approach, led to a clear improvement in translation quality.

In SMT massive amounts of parallel text in the source and target language are required to achieve high quality translation. However, there are a large number of languages that are considered "low-density," either because the population speaking the language is not very large, or if insufficient amounts of bilingual text are available involving that language. Bangla is one such language. Bangla, one of the more important Indo-Iranian languages, is the sixth-most popular in the world and spoken by a population that now exceeds 250 million. Geographical Bangla-speaking population percentages are as follows: Bangladesh (over 95%), and the Indian States of Andaman and Nicobar Islands (26%), Assam (28%), Tripura (67%), and West Bengal (85%). The global total includes those who are now in diaspora in Canada, Malawi, Nepal, Pakistan, Saudi Arabia, Singapore, United Arab Emirates, United Kingdom, and United States. Although being among the top ten most widely spoken languages around the world, the Bangla language still lacks significant research in the area of NLP specifically in SMT.

SMT systems require a significant amount of bilingual data between language pairs to achieve significant translation accuracy. However, being a low-density language, such resources are not available in Bangla. In this chapter we discuss how machine learning approaches can help to improve translation quality within as SMT system without requiring a huge increase in resources.

We provide a novel semi-supervised learning and active learning framework for SMT, which utilizes both, labeled and unlabeled data. We propose two semi-supervised learning techniques for sentence selection within a Bangla-English phrase-based SMT System. We also propose several effective active learning techniques for sentence selection from a pool of untranslated sentences, for which we ask human experts to provide translations. We perform detailed experimental evaluations on the sentence selection methods and demonstrate that these sentence selection techniques can help to improve translation quality in SMT.

Overall, in this chapter we demonstrate that for low-density language like Bangla, these machine-learning approaches can improve translation quality.

BACKGROUND

Semi-Supervised Learning

Semi-supervised learning refers to the use of both labeled and unlabeled data for training. Semi-supervised learning techniques can be applied to SMT when a large amount of bilingual parallel data is not available for language pairs. Sarkar, Haffari, and Ueffing (2007) explore the use of semi-supervised model adaptation methods for the effective use of monolingual data from the source language in order to improve translation accuracy.

Self-training is a commonly used technique for semi-supervised learning. In self-training a classifier is first trained with a small amount of labeled data. The classifier is then used to classify the unlabeled data. Typically, the most confident unlabeled points, together with their predicted labels, are added to the training set. The classifier is retrained and the procedure repeated. Note