Using RBMT Systems to Produce Bilingual Corpus for SMT

Xiaoguang Hu, Haifeng Wang, Hua Wu
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza
No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738, China
{huxiaoguang, wanghaifeng, wuhua}@rdc.toshiba.com.cn

Abstract

This paper proposes a method using the existing Rule-based Machine Translation (RBMT) system as a black box to produce synthetic bilingual corpus, which will be used as training data for the Statistical Machine Translation (SMT) system. We use the existing RBMT system to translate the monolingual corpus into synthetic bilingual corpus. With the synthetic bilingual corpus, we can build an SMT system even if there is no real bilingual corpus. In our experiments using BLEU as a metric, the system achieves a relative improvement of 11.7% over the best RBMT system that is used to produce the synthetic bilingual corpora. We also interpolate the model trained on a real bilingual corpus and the models trained on the synthetic bilingual corpora. The interpolated model achieves an absolute improvement of 0.0245 BLEU score (13.1% relative) as compared with the individual model trained on the real bilingual corpus.

1 Introduction

Within the Machine Translation (MT) field, by far the most dominant paradigm is SMT, but many existing commercial systems are rule-based. In this research, we are interested in answering the question of whether the existing RBMT systems could be helpful to the development of an SMT system. To find the answer, let us first consider the following facts:

- Existing RBMT systems are usually provided as a black box. To make use of such systems, the most convenient way might be working on the translation results directly.
- SMT methods rely on bilingual corpus. As a data driven method, SMT usually needs large bilingual corpus as the training data.

Based on the above facts, in this paper we propose a method using the existing RBMT system as a black box to produce a synthetic bilingual corpus[1], which will be used as the training data for the SMT system.

For a given language pair, the monolingual corpus is usually much larger than the real bilingual corpus. We use the existing RBMT system to translate the monolingual corpus into synthetic bilingual corpus. Then, even if there is no real bilingual corpus, we can train an SMT system with the monolingual corpus and the synthetic bilingual corpus. If there exist *n* available RBMT systems for the desired language pair, we use the *n* systems to produce *n* synthetic bilingual corpora, and *n* translation models are trained with the *n* corpora respectively. We name such a model the *synthetic model*. An interpolated translation model is built by linear interpolating the *n* synthetic models. In our experiments using BLEU (Papineni et al., 2002) as the metric, the interpolated synthetic model achieves a relative improvement of 11.7% over the best RBMT system that is used to produce the synthetic bilingual corpora.

[1] In this paper, to be distinguished from the real bilingual corpus, the bilingual corpus generated by the RBMT system is called a synthetic bilingual corpus.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 287–295, Prague, June 2007.