Using Translation Paraphrases from Trilingual Corpora to Improve Phrase-Based Statistical Machine Translation: A Preliminary Report

Francisco Guzmán Herrera Leonardo Garrido Luna
Instituto Tecnológico de Monterrey
Centro de Sistemas Inteligentes
Monterrey, México
guzmanhe@gmail.com *leonardo.garrido@itesm.mx*

Abstract

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation. Nevertheless, Statistical Machine Translation effectiveness is limited to situations where large amounts of training data are available. Therefore, the broader the coverage of a SMT system is, the better the chances to get a reasonable output are. In this paper we propose a method to improve quality of translations of a phrase-based Machine Translation system by extending phrase-tables with the use of translation paraphrases learned from a third language. Our experiments were done translating from Spanish to English pivoting through French.

1. Introduction

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation [4]. There have been several attempts to improve the performance of such systems. Non-syntactic phrase-based translation systems[9] certainly outperform word-based systems[21]. Nevertheless, Statistical Machine Translation (SMT) effectiveness is limited to situations where large amounts of data are available.

Such a condition, limits the performance of SMT systems over"low density" language pairs [5]. Scarce training data, often leads to a low coverage problem, that is, a low amount of learned translations for a language pair. In this paper we will discuss a method for expanding learned translations by means of a third language, so coverage is augmented and translation quality incremented.

This paper is organized as follows: In Sec. 2, we give an outline of the related work being done in phrase-based SMT. In Sec. 3 we describe the coverage problem and how extending phrase-tables we can tackle this problem. In Sec. 4, we describe thoroughly the translation paraphrases we used in our experiments. In Sec. 5, we explain the methodology followed throughout our experimentation and in Sec. 6 we discuss the results. In Sec. 7, we discuss our results and propose further improvements to our system.

978-0-7695-3124-3/08 $25.00 © 2008 IEEE
DOI 10.1109/MICAI.2007.34