

Translation Paraphrases in Phrase-Based Machine Translation

Francisco Guzmán and Leonardo Garrido

Center for Intelligent Systems
ITESM Campus Monterrey, Mexico
guzmanhe@gmail.com leonardo.garrido@itesm.mx

Abstract. In this paper we present an analysis of a phrase-based machine translation methodology that integrates paraphrases obtained from an intermediary language (French) for translations between Spanish and English. The purpose of the research presented in this document is to find out how much extra information (i.e. improvements in translation quality) can be found when using Translation Paraphrases (TPs). In this document we present an extensive statistical analysis to support conclusions.

1 Introduction

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation [1]. There have been several attempts to improve the performance of such systems. Non-syntactic phrase-based translation systems[2] certainly outperform word-based systems[3].

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 [4]. Scarce training data, often leads to a low coverage problem, that is, a low amount of learned translations for a language pair.

There are several efforts trying to improve translation quality of SMT systems. Many state-of-the-art systems involve the introduction of syntactic information to phrase-based machine translations [5,6,7,8,9].

On the other hand, we find several efforts which do not use syntactic information. One main topic of discussion is the usage of paraphrases. For example Callison [4] improves translation quality by giving alternatives to broaden coverage of a phrase-based machine translation system through the use of paraphrases. They use paraphrases in cases when a phrase is not found in their phrase-tables. Other effort is conducted by Guzman and Garrido [10] who obtain what they call “translation paraphrases” from pivoting through an intermediary language.

In this paper we analyze their methodology to assess whether the inclusion of Translation Paraphrases (TP) in a SMT system are useful to improve translation quality, in comparison to systems that do not include such features.