On the Automatic Learning of Bilingual Resources: Some Relevant Factors for Machine Translation

Helena de M. Caseli, Maria das Graças V. Nunes, and Mikel L. Forcada

1 NILC – ICMC, University of São Paulo
CP 668P – 13.560-970 – São Carlos – SP – Brazil
helename,gracan@icmc.usp.br

2 Departament de Llenguatges i Sistemes Informàtics,
Universitat d’Alacant, E-03071 Alacant, Spain
mlf@ua.es

Abstract. In this paper we present experiments concerned with automatically learning bilingual resources for machine translation: bilingual dictionaries and transfer rules. The experiments were carried out with Brazilian Portuguese (pt), English (en) and Spanish (es) texts in two parallel corpora: pt-en and pt-es. They were designed to investigate the relevance of two factors in the induction process, namely: (1) the coverage of linguistic resources used when preprocessing the training corpora and (2) the maximum length threshold (for transfer rules) used in the induction process. From these experiments, it is possible to conclude that both factors have an influence in the automatic learning of bilingual resources.

Keywords: Machine translation, bilingual resources, automatic learning, parallel corpora.

1 Introduction

The ability to translate from one language to another has became not only a desirable but also a fundamental skill in the multilingual world every day accessible through the Internet. Due to working or social needs, there is a growing necessity of being able to get at least the gist of a piece of information in a different language. Unfortunately, this ability is not inherent.

The good news is that computers are getting more and more available to ease this task. This is one of the challenges of machine translation (MT) research and also of this paper. In particular, this paper is concerned not only with translating from one language to another but also with the automatic learning of bilingual resources useful for rule-based machine translation (RBMT) systems.

Traditionally, the bilingual resources used in RBMT systems are built by means of hard manual work. However, in the last years several methods have been proposed to automatically learn bilingual resources such as dictionaries [1,2,3,4] and transfer rules [5,6,7] from translation examples (parallel corpora).
A bilingual dictionary is a bilingual list of words and multiword units (possibly accompanied by morphological information) that are mutual translations. A transfer (or translation) rule, in turn, is a generalization of structural, syntactic or lexical correspondences found in the parallel sentences (translation examples).

In line with these initiatives, this paper presents experiments carried out to investigate the relevance of certain factors when automatically learning bilingual dictionaries and transfer rules following the methodology of the ReTraTos project.

Following this methodology, the bilingual dictionaries and the transfer rules are induced from automatically word-aligned (or lexically aligned) parallel corpora processed with morphological analysers and part-of-speech (PoS) taggers. The aspects under investigation in this paper are: (1) the coverage of linguistic resources used in preprocessing the training corpora and (2) the maximum length threshold used in the transfer rule induction process, that is, the number of source items that a transfer rule can contain. The evaluation of the translation quality was carried out with Brazilian Portuguese (pt), Spanish (es) and English (en) texts in two parallel corpora — pt–es and pt–en — by using the automatic metrics BLEU \[8\] and NIST \[9\].

Our interest in the above factors is motivated by previous results \[10\]. We now intend to investigate if a better coverage of the preprocessing linguistic resources can bring better MT results, and if the previous poor pt–en MT performance was due to the threshold length (too restrictive) used during the induction of the transfer rules. We think that the small length of source patterns was not sufficient to learn relevant syntactic divergences between those languages. To our knowledge, it is the first time that such a study is carried out for Brazilian Portuguese or other languages.

This paper is organized as follows. Section 2 presents related work on automatic induction of bilingual dictionaries and transfer rules. Section 3 describes briefly the induction methodology under study in this paper. Section 4 shows the experiments and their results, and section 5 ends this paper with some conclusions and proposals for future work.

2 Related Work

According to automatic evaluation metrics like BLEU \[8\] and NIST \[9\], the phrase-based statistical MT (SMT) systems such as \[11\] and \[12\] are the state-of-the-art in MT. Despite this fact, this paper is concerned with RBMT since the symbolic resources (dictionaries and rules) of RBMT suit better than the models of SMT to our research purpose: to know how translation is performed and what affects its performance.

This section presents briefly some of the methods proposed in the literature to automatically induce bilingual dictionaries and transfer rules. These methods can follow many different approaches, but usually they perform induction from a sentence-aligned parallel corpus.

\[http://www.nilc.icmc.usp.br_nilc/projects/retratos.htm\]