Enriching Statistical Translation Models using a Domain-independent Multilingual Lexical Knowledge Base

Miguel García, Jesús Giménez and Lluís Màrquez

TALP Research Center, LSI Department
Universitat Politècnica de Catalunya
Jordi Girona Salgado 1–3, E-08034, Barcelona
{tgarcia,jgimenez,lluism}@lsi.upc.edu

Abstract. This paper presents a method for improving phrase-based Statistical Machine Translation systems by enriching the original translation model with information derived from a multilingual lexical knowledge base. The method proposed exploits the Multilingual Central Repository (a group of linked WordNets from different languages), as a domain-independent knowledge database, to provide translation models with new possible translations for a large set of lexical tokens. Translation probabilities for these tokens are estimated using a set of simple heuristics based on WordNet topology and local context. During decoding, these probabilities are softly integrated so they can interact with other statistical models. We have applied this type of domain-independent translation modeling to several translation tasks obtaining a moderate but significant improvement in translation quality consistently according to a number of standard automatic evaluation metrics. This improvement is especially remarkable when we move to a very different domain, such as the translation of Biblical texts.

1 Introduction

One of the main criticisms against empirical methods in general, and Statistical Machine Translation (SMT) in particular, is their strong domain dependence. Since parameters are estimated from a corpus in a specific domain, the performance of the system on a different domain is often much worse. This flaw of statistical and machine learning approaches is well known and has been largely described in the NLP literature, for a variety of tasks, e.g., parsing [1], word sense disambiguation [2], and semantic role labeling [3].

In the case of SMT, domain dependence has very negative effects in translation quality. For instance, in the 2007 edition of the ACL MT workshop (WMT07), an extensive comparative study between in-domain and out-of-domain performance of MT systems built for several European languages was conducted [4]. Results showed a significant difference in MT quality between the two domains for all statistical systems, consistently according to a number of automatic evaluation metrics. In contrast, the differences reported in the case of rule-based or hybrid MT systems were less significant or inexistent, and even in some cases the performance of such systems out of the domain was higher than in the corpus domain. The reason is that, while these systems are