Syntax-Based Concept Alignment for Domain-Specific Machine Translation

Anonymous ACL submission

Abstract

1 Introduction

Grammar-based translation pipelines such as those based on Grammatical Framework (GF) (**?**) have been successfully employed in domain-specific Machine Translation (MT). What makes these systems well suited to the task is the fact that, when we constrain ourselves to a specific domain, where precision is most often more important than coverage, they can provide strong guarantees in terms of grammatical correctness.

Nevertheless, lexical exactness is, in this context, just as important as grammaticality. An important part of the design of the Controlled Natural Language (CNL) the grammar in such a system describes becomes, then, the creation of a translation lexicon. In many cases, this is done for the most part manually, resulting in a time consuming task requiring significant linguistic knowledge. When the grammar is designed based on a parallel corpus of example sentences, it is possible to automate part of this process by means of statistical word and phrase alignment techniques (**?**). None of them is, however, suitable for the common case in which only a limited number of example sentences is available.

In this paper, we propose an alternative approach to the automation of this task. While still being data-driven, our method is grammar-based and, as such, capable of extracting meaningful correspondences even from individual sentence pairs.

A further advantage of performing syntactic analysis is that we do not have to choose *a priori* whether to focus on the word or phrase level. Instead, we can simultaneously operate at different levels of abstraction, thus extracting both single- and multiword, even non-contiguous, correspondences.

For this reason, we refer to the task our system attempts to automate as *Concept Alignment* (CA). We call *concepts* the abstract units of translation, composed of any number of words, the system identifies, and represent them as *alignments*, i.e. pairs of semantically equivalent concrete expressions.

This paper is structured as follows. Section 2 starts by giving an overview of our approach to CA to then focus on our algorithm for extracting correspondences. It is concluded with a description of our method for converting the alignments obtained in this way to a GF translation lexicon. After that, Section 3 presents the results of our first evaluation of the system. Finally, Section 4 consists of a discussion of such results and some ideas for future work.

2 Methodology

2.1 Extracting concepts

2.1.1 Refinements of the basic algorithm

2.2 Generating grammar rules

3 Evaluation

4 Conclusions