A Tool for the Acquisition of Japanese–English Machine Translation Rules Using Inductive Learning Techniques

Hussein Almuallim

Info. and Computer Science Dept.
King Fahd University of
Petroleum and Minerals
Dhahran 31261, Saudi Arabia

Yasuhiro Akiba   Takefumi Yamazaki
Akio Yokoo   Shigeo Kaneda

NTT Network Information System Labs.
1-2356, Take, Yokosuka-shi
Kanagawa-ken 238-03
Japan

Abstract

This work addresses the problem of constructing translation rules for ALT-J/E—a knowledge-based Japanese–English translation system developed at NTT. We introduce the system ATRACT, *which is a semi-automatic knowledge acquisition tool designed to facilitate the construction of the desired translation rules through the use of inductive machine learning techniques. Rather than building rules by hand from scratch, a user of* ATRACT *can obtain good candidate rules by providing the system with a collection of examples of Japanese sentences along with their English translations. This learning task is characterized by two factors: (i) It involves exploiting a huge amount of semantic information as background knowledge. (ii) Training examples are "ambiguous". Currently, two learning methods are available in* ATRACT. *Experiments show that these methods lead to rules that are very close to those composed manually by human experts given only a reasonable number of examples. These results suggest that* ATRACT *will significantly contribute to reducing the cost and improving the quality of ALT-J/E translation rules.*

AI Topic: Machine learning, Machine translation, Natural language, Knowledge acquisition.
Domain Area: Machine translation.
Language/Tool: LISP, C, Sun Sparcstation 2.
Status: Initial version implemented. Graphical interface under implementation.
Effort: Approximately 1.5 person-year.
Impact: ATRACT, our learning-based system, will significantly contribute to reducing the cost and improving the quality of the translation rules in the ALT-J/E Japanese–English translation system.

1 Introduction

A critical issue in AI research is to overcome the knowledge acquisition bottleneck in knowledge-based systems. As a knowledge base is expanded, adding more knowledge and fixing previous erroneous knowledge become increasingly costly. Moreover, maintaining the integrity of large knowledge bases has proven to be a very challenging task.

ALT-J/E, which is an experimental Japanese-English translation system developed at Nippon Telegraph and Telephone Corporation (NTT), is one example of a knowledge-based system in which such "symptoms" are being experienced. One major component of this system is its huge collection of *translation rules.* Each of these rules associates a Japanese sentence pattern with an appropriate English pattern. To translate a Japanese sentence into English, ALT-J/E looks for the rule whose Japanese pattern matches the sentence best, and then uses the English pattern of that rule for translation.

So far, ALT-J/E translation rules have been composed manually by extensively trained human experts. To qualify for this job, an expert must not only master both English and Japanese, but also be very familiar with various components of the system. Each time the rules are expanded or altered, the new set of rules must then be "debugged" using a collection of test cases. Usually, several iterations are needed to arrive at translation rules of acceptable quality.

Creating new translation rules as well as refining existing ones have proven to be extremely difficult and time-consuming because these tasks require considering a huge space of possible combinations (rules in ALT-J/E are expressed in terms of as much as

3000 "semantic categories"). Furthermore, it is usually troublesome for an expert to modify rules created by another expert (or even by the same expert after some period of time) because it is hard to figure out the reasoning behind the particular choices made in a given rule. The high costs involved make the manual creation of ALT-J/E's translation rules impractical. Indeed, in spite of the vast amount of resources spent on building the current rules of ALT-J/E, faults in these rules are still detected from time to time, making system maintenance a continuous requirement.

In this paper, we present ATRACT—an interactive knowledge acquisition tool designed to aid the developers of the ALT-J/E system in composing the desired translation rules. ATRACT (which stands for **A**utomatic **T**ranslation **R**ules **AC**quisition **T**ool) employs inductive machine learning techniques to learn translation rules from examples of Japanese sentences and their English translations. Given a collection of such examples, the system proposes one or more candidate translation rules. If these rules are not satisfactory, the user can provide more examples of cases where these rules are inaccurate and let the system learn new rules that can handle these cases. Alternatively, the user can pick up the best rule and modify it by hand when such modification is obvious.

With ATRACT, the user is relieved from exploring the huge space of alternatives she/he has to consider when constructing translation rules manually from scratch—a job which only extensively trained experts can perform. The task is now turned into a search for some reasonable rule(s) that explains the given training examples, where the search is handled automatically by a learning algorithm. This not only saves the user's time, but also makes it unnecessary for the user to be an expert of the ALT-J/E system. Moreover, this approach significantly reduces the "subjectivity" of the rules since the intervention of human experts is minimized. This is particularly important because the immense number of translation rules (currently over 10,000) requires employing a team of experts over an extended period of time.

ATRACT provides the user with a menu of multiple learning techniques to choose from. Several factors, however, distinguish our learning task from other conventional learning problems, and thus, limit what learning algorithms can be included. In particular, learning in our domain involves the effective utilization of a large amount of semantic knowledge, and the ability to learn from "ambiguous" training examples. Currently, two selected inductive machine learning algorithms are available in ATRACT. This paper reports experimental results which show that the rules learned by these algorithms are very close to the rules manually composed by human experts. In most cases, given a reasonable number of training examples, our system is able to find rules that are more than 90% accurate when compared to the manually composed rules. With these encouraging results, ATRACT is currently being augmented with a user friendly graphical interface to allow its use in the actual development of the ALT-J/E system.

The rest of this document is organized as follows. We begin in Section 2 by a brief overview of the ALT-J/E Japanese-English translation system. We then outline in Section 3 the manual procedure for acquiring ALT-J/E's translation rules and list some of the shortcomings of that procedure. In Section 4, we propose an alternative procedure using the ATRACT system. We then describe in Section 5 the inductive learning approaches used within ATRACT, followed by an experimental evaluation of these approaches in Section 6. Finally, conclusion remarks are stated in Section 7.

2 ALT-J/E: A brief overview

ALT-J/E, the **A**utomatic **L**anguage **T**ranslator: **J**apanese to **E**nglish, is one of the most advanced and well-recognized systems for translating Japanese to English. It is the largest such system in terms of the amount of knowledge it comprises. In this work, we are concerned with the following components of the ALT-J/E system:

1. The Semantic Hierarchy,
2. The Semantic Dictionary, and
3. The Translation Rules.

We briefly describe each of these components below. For more details about the ALT-J/E system, we refer the reader to [2, 3, 4].

As shown in Figure 1, the **Semantic Hierarchy** is a sort of concept thesaurus represented as a tree structure in which each node is called a *semantic category*, or a *category* for simplicity. Edges in this structure represent "is-a" relations among the categories. For example, "Agents" and "People" (see Figure 1) are both categories. The edge between these two categories indicates that any instance of "People" is also an instance of "Agents". The current version of ALT-J/E's Semantic Hierarchy is 12 levels deep and has about 3000 nodes. The **Semantic Dictionary** maps