Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System

Chung-hye Han1, Benoit Lavoie2, Martha Palmer1, Owen Rambow3, Richard Kittredge2, Tanya Korelsky2, Nari Kim4, and Myunghee Kim2

1 Dept. of Computer and Information Sciences/IRCS
Univ. of Pennsylvania, Philadelphia, PA 19104, USA
{chunghye, mpalmer}@linc.cis.upenn.edu
2 CoGenTex, Inc., Ithaca, NY 14850-1589, USA
{benoit, richard, tanya, myunghee}@cogentex.com
3 ATT Labs-Research, B233, Florham Park, NJ 07932, USA
rambow@research.att.com
4 Konan Technology, Inc., Seoul 135-090, Korea
nari@konantech.co.kr

Abstract. This paper describes an approach for handling structural divergences and recovering dropped arguments in an implemented Korean to English machine translation system. The approach relies on canonical predicate-argument structures (or dependency structures), which provide a suitable pivot representation for the handling of structural divergences and the recovery of dropped arguments. It can also be converted to and from the interface representations of many off-the-shelf parsers and generators.

1 Introduction

This paper describes an approach for handling structural divergences ([1,3,4,8]) and recovering dropped arguments for Korean to English translation. Given that the two languages are very different from each other in structure, many challenging problems arise, demanding sophisticated linguistic modeling. The basic elements of our approach include:

- Transfer rules based on syntactic lexico-structural transfer ([8]);
- Conversion rules using a Korean predicate-argument lexicon for converting parsed syntactic structures produced by an off-the-shelf Korean parser ([12]) to the syntactic structures used for transfer;

* The work reported in this paper was supported by contract DAAD 17-99-C-0008 awarded by the Army Research Lab to CoGenTex, Inc., with the University of Pennsylvania as a subcontractor and NSF Grant - VerbNet, IIS 98-00658. Owen Rambow’s contribution to this paper was made when he was with CoGenTex, Inc. and Nari Kim’s contribution was made when she was a visiting researcher at IRCS, UPenn.
– Generation rules using an English realization lexicon for recovering dropped arguments after transfer.

The current implementation and processing of the transfer, conversion and generation rules is done uniformly, using a syntactic lexico-structural based framework ([5]). Declarative transformation specifications indicate how the lexemes and their relevant syntactic structures (essentially, their syntactic projection along with syntactic/semantic features) are mapped from one level to another. A similar approach was used in previous work for English to Arabic and English to French translations ([6,7]).

The corpus for this project is a set of Korean/English parallel texts that consist of battle scenario message traffic and military language training manual. These contain information on typical military events such as troop movement, intelligence gathering, and equipment supplies, among others. Each half has roughly 50,000 word tokens, and 6000 sentences.

This paper is structured as follows. In section 2, we introduce some linguistic issues that pose problems for Korean/English MT. In section 3, we present a brief overview of the implemented system. Section 4 presents the linguistic knowledge bases used for conversion, transfer and argument recovery. We conclude with sections 5 and 6 with a brief comparison to different approaches in other MT systems (e.g., LCS and CCLINC) and a discussion of future work. Although our system handles transfer in both Korean-to-English and English-to-Korean directions, in this paper we mainly concentrate on the Korean-to-English direction for the sake of exposition.

2 Some Linguistic Issues in Korean/English Machine Translation

While English canonically has rigid subject-verb-object (SVO) order, Korean is a verb-final language with free word order. For instance, ditransitive sentences in English have ‘subject-verb-indirect object-direct object’ order, as shown in the target sentence in Table 1. The corresponding Korean sentence can have ‘direct object-indirect object-subject-verb’ order, as shown in the source sentence in Table 1. In our system, the grammatical functions of argument NPs are identified by the use of Yoon’s Korean parser and conversion rules using the predicate-argument lexicon.

Unlike English, argument NPs can be deleted in Korean. For instance, in the source sentence in Table 2 which is a conditional sentence, the subject NP in the if-clause has been deleted and the subject NP and the object NP in the main clause have been deleted. Ideally, all the missing arguments should be identified in the output as in the target sentence in Table 2. With the addition of a discourse component, the references of the missing arguments can be restored.

1 Korean examples in this paper are romanized for convenience.