Zh-Ja Patent MT based on Syntactic Pre-ordering

Abstract
This paper presents our Chinese-to-Japanese patent machine translation system for WAT 2015 (Group ID: ntt) that uses syntactic pre-ordering over Chinese dependency structures. A head word and its modifier words are reordered by hand-written rules or a learning-to-rank model. Our system outperforms baseline phrase-based machine translations and competes with baseline tree-to-string machine translations.

1 Introduction
Patent documents, which well-structured written documents that describe the technical details of inventions, are expected to have almost no semantic ambiguities caused by indirect or rhetorical expressions. Due to this aspect, patent documents are good candidates for literal translation, which most machine translation (MT) approaches aim to do.

One technical challenge for patent machine translation is the complex syntactic structure of patent documents, which typically have long sentences that complicate MT reordering, especially for the word order in distant languages. Chinese and Japanese have similar word order in noun modifiers but different subject-verb-object order, requiring long distance reordering in translation. In this year's WAT evaluation campaign (Nakazawa et al., 2015), we tackle long distance reordering by syntactic pre-ordering based on Chinese dependency structures (Sudoh et al., 2014) in a Chinese-to-Japanese patent translation task.

Our system basically consists of three components: Chinese syntactic analysis (word segmentation, part-of-speech (POS) tagging, and dependency parsing) adapted to patent documents; dependency-based syntactic pre-ordering; and a standard phrase-based statistical MT.

This paper describes our system's details and discusses our evaluation results.

2 System Overview
Figure 1 shows a brief workflow of our Chinese-to-Japanese MT system. Its basic architecture is standard with syntactic pre-ordering. Input sentences are first applied to word segmentation and POS tagging, parsed into dependency trees, reordered using pre-ordering rules or a pre-ordering model, and finally translated into Japanese by a phrase-based statistical MT.

3 Chinese Syntactic Analysis: Word Segmentation, Part-of-Speech Tagging, and Dependency Parsing
Word segmentation and POS tagging are solved jointly (Suzuki et al., 2012) for better Chinese word segmentation based on the POS tag sequences. The dependency parser produces un-

<table>
<thead>
<tr>
<th>PBMT</th>
<th>Data-driven</th>
<th>Rule-based</th>
<th>T2S</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIBES</td>
<td>0.781</td>
<td>0.812</td>
<td>0.822</td>
</tr>
<tr>
<td>Human</td>
<td>n/a</td>
<td>8.00</td>
<td>16.25</td>
</tr>
</tbody>
</table>

~3-4% positive effect in RIBES
~2% positive effect in BLEU