Abstract

This paper describes the EBMT system of Kyoto University that participated in the OLYMPICS task at IWSLT 2012. When translating very different language pairs such as Chinese-English, it is very important to handle sentences in tree structures to overcome the difference. Many recent studies incorporate tree structures in some parts of translation process, but not all the way from model training (alignment) to decoding. Our system is a fully tree-based translation system where we use the Bayesian phrase alignment model on dependency trees and example-based translation. To improve the translation quality, we conduct some special processing for the IWSLT 2012 OLYMPICS task, including sub-sentence splitting, non-parallel sentence filtering, adoption of an optimized Chinese segmenter and rule-based decoding constraints.

1. Introduction

We consider that it is quite important to use linguistic information in the translation process when tackling very different language pairs such as Chinese-English and Japanese-English, and one of the most important pieces of information is sentence structure. Many recent studies incorporate some structural information into decoding, but rarely into alignment. In this paper, we adopt a fully tree-based translation framework based on dependency tree structures [1]. In the alignment step, we use Bayesian subtree alignment model based on dependency trees. Section 2 shows a brief description of the model. It is a kind of tree-based reordering model, and can capture non-local reorderings which sequential word-based models cannot often handle properly. In the translation step, we adopt an example-based machine translation (EBMT) system, handling examples which are discontinuous as a word sequence, but continuous structurally. It also considers similarities of neighboring nodes, which are useful for choosing suitable examples matching the context.

Figure 1 shows the overview of our EBMT system on Chinese-English translation. The translation example database is automatically constructed from a parallel training corpus by means of a Bayesian subtree alignment model. Note that both source and target sides of all the examples are stored in dependency tree structures. An input sentence is also parsed and transformed into a dependency structure. For all the sub-trees in the input dependency structure, matching examples are searched in the example database. This step is the most time consuming part, and we exploit a fast tree retrieval method [2]. There are many available examples for one sub-tree, and also, there are many possible sub-tree combinations. The best combination is detected by a log-linear decoding model with features described in Section 3.

In the example in Figure 1, five examples are used. They are combined and produce an output dependency tree. We call nodes surrounding those of the example, “bond” nodes. The bond nodes of one example are replaced by other examples, and thus examples can be combined.

We attended the IWSLT 2012 OLYMPICS task which is a Chinese-to-English text translation task. Based on the characteristic of this task, we conducted some special processing. We split sub-sentences and filtered non-parallel sentences to improve the quality of the supplied corpora. We adopted an optimized Chinese segmenter which can generate segmentation results that are much more similar to English to improve the alignment accuracy. To reduce the computational complexity, we adopted rule-based decoding constraints on the decoding. Details of the above special processing for this task are described in Section 4.

2. Bayesian Subtree Alignment Model based on Dependency Trees

Alignment accuracy is crucial for providing high quality corpus-based machine translation systems because translation knowledge is acquired from an aligned training corpus. For distant language pairs such as Chinese-English and Japanese-English, the word sequential models such as IBM models are quite inadequate (about 20% alignment error rate (AER)), and therefore it is important to improve the alignment accuracy itself. The differences between languages can be seen in Figure 2, which shows an example of Japanese-English. The word or phrase order is quite different for these languages. Another important point is that there are frequent many-to-one or many-to-many correspondences. For example, the Japanese noun phrase “受光素子” is composed of three words, whereas the corresponding English phrase consists of only one word “photodetector”, and the English func-
Input: 腕关节部的活检中确认了渗透到神经的肿瘤肿块

Output: in the biopsy of the wrist joint the tumor lump which permeated nerve were recognized

Figure 1: An example of Chinese-English translation.

Figure 2: Example of dependency trees and alignment of subtrees. The root of the tree is placed at the extreme left and words are placed from top to bottom.

Figure 3: Alignment results from bi-directional GIZA++. Black boxes depict the system output, while dark (Sure) and light (Possible) gray cells denote gold-standard alignments.

Even if the word order differs greatly between languages, phrase dependencies tend to hold between languages. This can be seen in Figure 2. Therefore, incorporating dependency analysis into the alignment model is useful for distant language pairs. We exploit Bayesian subtree alignment model based on dependency trees [3]. This model incorporates dependency relations of words into the alignment model and define the reorderings on the word dependency trees. Figure 2 shows an example of the dependency trees for Japanese and English.

3. Tree-based Translation

As a tree-based translation method, we adopt an example-based machine translation system [1]. In this section, we briefly introduce the translation procedure in our EBMT system.

3.1. Retrieval of Translation Examples

The input sentence is converted into the dependency structure as in the alignment step. Then, for each sub-tree, avail-
able translation examples are retrieved from the example database. Here the word “available” means that all the words in the focusing input sub-tree appear in the source tree of the example, and the dependency relations between the words are same. We use a fast, on-line tree retrieval technique [2] to get all the available examples from a large training corpus.

3.2. Selection of Translation Examples

We find the best combination of examples by tree-based log-linear model with features shown below:

- Size of examples
- Translation probability
- Root node of examples
- Parent node
- Child nodes
- Bond nodes
- NULL-aligned words
- Language model

Among the features, an important one is “Size of examples”. Translations that are composed of larger examples can achieve higher quality because translations inside the examples are stable.

3.3. Combination of Translation Examples

When combining examples, in most cases, bond nodes are available outside the examples, to which the adjoining example is attached. Using the bond information, we don’t need to consider word or phrase order. Bond information naturally solves the reordering problem. Figure 1 is an example of combining translation examples. The combination process starts from the example used for the root node of the input tree (the first one in Figure 1). Then the example for the child node of the sub-tree covered by the initial example is combined (the second and third examples). When combining the second example to the first one, “细胞” is used as bond node, and for the third example, “节” is used as bond node. The combination repeated until all the examples are combined into one target tree. Finally, the output sentence is generated from the tree structure.

Note that there are NULL-aligned nodes in the examples (the nodes which are not circled, such as ‘爷’, ‘部(part)’ and articles in English).

4. IWSLT 2012 OLYMPICS Task

In this section, we first briefly introduce the IWSLT 2012 OLYMPICS task. We then describe the special processing for this task including sub-sentence splitting, non-parallel sentence filtering, adoption of an optimized Chinese segmenter and rule-based decoding constraints. Finally we report the formal run evaluation results with discussion.

4.1. Task Description

The OLYMPICS task is carried out using parts of the HIT Olympic Trilingual Corpus (HIT) [4] and the Basic Travel Expression Corpus (BTEC) as an additional training corpus. The HIT corpus is a multilingual corpus that covers 5 domains (traveling, dining, sports, traffic and business) that are closely related to the Beijing 2008 Olympic Games. The HIT corpus contains around 52k sentences 2.8 million words in total. The BTEC corpus is a multilingual speech corpus containing tourism-related sentences. The BTEC corpus consists of 20k sentences including the evaluation data sets of previous IWSLT evaluation campaigns.

4.2. Sub-sentence Splitting

The corpora supplied for this task have a problem that there are many parallel sentences containing multiple sub-sentences. Since multiple sub-sentences in a single sentence decrease the parsing accuracy, splitting the sentences containing sub-sentences into individual sentences is necessary. Based on our observation, there are two different patterns in the HIT and BTEC corpus for this sub-sentences problem. In the HIT corpus, there are same number of punctuation marks (including comma, period, question mark and exclamation mark) in most parallel sentences with this problem, and can be split using these punctuation marks. Here is one example:

Zh: 我带了些矿泉水和茶，您喜欢喝什么?
      (I've brought some mineral water and some tea, which do you prefer?)

En: I've brought some mineral water and some tea. Which do you prefer?

In this example, Chinese sentence and English sentence have the same number of punctuation marks. Moreover, “我带了些矿泉水和茶” corresponds to “I've brought some mineral water and some tea” and “您喜欢喝什么” corresponds to “Which do you prefer”. Therefore, it can be split based on the punctuation.

In the BTEC corpus, most parallel sentences with this problem contain same number of EOS punctuation marks (i.e. period, question mark and exclamation mark), and can be split using EOS punctuation marks. Here is one example:

Zh: 非常感谢。你知道我不想赶不上它。
      (Thank you so much. You see I don’t want to miss it.)

En: Thank you so much. You see, I don’t want to miss it.

Therefore, we split the sub-sentences in the HIT and BTEC corpus based on the punctuation marks and EOS punctuation marks respectively.

4.3. Non-parallel Sentence Filtering

Another problem of the supplied corpora is that there are many non-parallel sentences in the HIT corpus. Here is one example:
Also, since Chinese and English may use punctuation (especially for the usage of comma) in different places of parallel sentences, the sub-sentence splitting method for the HIT corpus that we described in Section 4.2 can lead to non-parallel sentences. Here is one example:

Zh: 是的，这位女士要一杯曼哈顿酒，我要一杯马丁尼。
(Yes, this lady will have a Manhattan, and I’ll have a martini.)

En: What about you?

Zh: 我上牛津大学。
(I am studying at Oxford University.)

En: Yes, I think so. This lady will have a Manhattan and I’ll have a martini.

These non-parallel sentences can decrease the accuracy of alignment and translation performance. Therefore, we propose a filtering method to automatically filter the non-parallel sentences. Our proposed method is an extension of [5], which extracted parallel sentences from comparable corpora by treating it as a classification problem. We think non-parallel sentences filtering can also be solved by classification. We use the same features and classification model described in [5]. The dictionary we used is created from the lexical translation table obtained by running GIZA++ on the whole supplied corpora. We extract the best 5 translation equivalents having translation probability above 0.1 from the lexical translation table as our dictionary. For training data, we use 5,000 parallel sentences from the BTEC corpus, because of the good quality of the BTEC corpus. We create non-parallel sentences from the parallel sentences following the method described in [5]. We generate all the sentence pairs except the original parallel sentence pairs in the Cartesian product, and discard the pairs that do not fulfill the condition of a sentence ratio filter and a word-overlap filter. Then we randomly select 500 non-parallel sentences and add them to the training data. Test data is created using the same method by using another 5,000 parallel sentences from the BTEC corpus. Our data filtering method achieved high accuracy with precision of 97.10%, recall of 84.81% and F-score of 90.54% in the experiment.

We then applied the trained classifier to the HIT corpus for non-parallel sentence filtering and filtered around 1,000 sentence pairs. We conducted translation experiments to investigate the effect of non-parallel sentence filtering on translation quality. Preliminary experimental results showed that non-parallel sentence filtering has little effect on translation quality (only 0.02% BLEU score increased). We think the reason is that the classifier trained on the BTEC corpus does not work well on the HIT corpus because of the difference between these two corpora, thus some parallel sentences are also filtered in this process.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.1162</td>
</tr>
<tr>
<td>Optimized</td>
<td>0.1209</td>
</tr>
<tr>
<td>Optimized+Constrained</td>
<td>0.1271</td>
</tr>
</tbody>
</table>

Table 1: Results of preliminary translation experiments.

4.4. Optimized Chinese Segmenter

As there are no explicit word boundary markers in Chinese, word segmentation is considered as an important first step in machine translation. Research shows that optimal Chinese word segmentation for machine translation is dependent on the other language, therefore, a bilingual approach is necessary [6]. In this task, we adopted a Chinese segmenter optimized based on a bilingual perspective, which exploits common Chinese characters shared between Chinese and Japanese for Chinese word segmentation optimization [7]. The BLEU scores with and without Chinese segmenter optimization are given in Table 1, indicated as “Optimized” and “Baseline” respectively. Although the Chinese segmenter we used is optimized for Chinese-Japanese machine translation, it shows better translation performance compared to the Chinese segmenter without optimization. We think the reason is that the optimized segmentation results are much more similar to English in number, which can reduce the number of 1-to-n alignments and improve the alignment accuracy.

4.5. Rule-based Decoding Constraints

Translating long and complex sentences is a critical problem in machine translation, because it increases the computational complexity. Finch et al. [8] presented a simple yet efficient method to solve this problem. They split a sentence into smaller units based on part-of-speech (POS) tags and commas, and translate the split units separately. Following their method, we also split a sentence into smaller units during decoding. Our EBMT system tends to choose large examples. Since the development data of this task also has the sub-sentence problem (described in Section 4.3), our system may use examples across punctuation boundaries which can generate translations with unnatural word order. Therefore, we split a source sentence based on comma, period, question mark and exclamation mark for decoding. The BLEU score after constrained decoding is given in Table 1, indicated as “Optimized+Constrained”. The result shows that our method achieved better translation performance compared to unconstrained decoding.

4.6. Results

The official scores for the our EBMT system with respect to several of the automatic metrics used for the official evaluation are given in Table 2 (For rankings, please refer to [9]). The scores are low for this task. There are several reasons:

The major reason is the quality and quantity of the sup-
Table 2: The official results for the our EBMT system in terms of a variety of automatic evaluation metrics.

<table>
<thead>
<tr>
<th>Case/Punctuation</th>
<th>BLEU</th>
<th>METEOR</th>
<th>WER</th>
<th>PER</th>
<th>TER</th>
<th>GTM</th>
<th>NIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case and punc</td>
<td>0.1273</td>
<td>0.4628</td>
<td>0.7552</td>
<td>0.6398</td>
<td>71.1530</td>
<td>0.4591</td>
<td>4.1138</td>
</tr>
<tr>
<td>No case and no punc</td>
<td>0.1228</td>
<td>0.4137</td>
<td>0.8288</td>
<td>0.6860</td>
<td>79.7690</td>
<td>0.4301</td>
<td>4.3104</td>
</tr>
</tbody>
</table>

Although our EBMT system can generate adequate and fluent translations, we could not achieve satisfactory results in the run submission. Besides the difficulty of this task itself, our EBMT system suffers from the low accuracy of the Chinese parser. In the future, we aim to improve our system to achieve better translation quality even on limited small training data.

6. References


Conference of the European Association for Machine Translation (EAMT’12), Trento, Italy, May 2012.


