Abstract

This paper describes NTT SMT System 2008 presented at the patent translation task (PAT-MT) in NTCIR-7. For PAT-MT, we submitted our strong baseline system faithfully following a hierarchical phrase-based statistical machine translation [2]. The hierarchical phrase-based SMT is based on a synchronous-CFGs in which a paired source/target rules are synchronously applied starting from the initial symbol. The decoding is realized by a CYK-style bottom-up parsing on the source side with each derivation representing a translation candidate. We demonstrate the strong baseline for the PAT-MT English/Japanese translations.

Keywords: Statistical Machine Translation, Hierarchical Phrase-based SMT.

1 Introduction

We present NTT Statistical Machine Translation System 2008 for the patent translation task (PAT-MT) in NTCIR-7. Our system has been successfully demonstrated for the numbers of evaluation tasks, including NIST1, WMT[13] and IWSLT [10]. For PAT-MT Japanese/English translations, we employed a strong baseline system faithfully following a hierarchical phrase-based statistical machine translation [2]. Hierarchical phrase-based machine translation is formulated as a probabilistic synchronous context-free grammar (PSCFG) [1] in which string pairs are generated. The system uses a set of source terminal symbols $T_S$, a set of target terminal symbols $T_T$ and a set of non-terminal symbols $N$. Each production rule is realized as follows [2, 11]:

\[ X \rightarrow \langle \gamma, \alpha, \sim, w \rangle \]  

where $X \in N$, $\gamma \in [N \cup T_S]^*$ and $\alpha \in [N \cup T_T]^*$. $\gamma$ and $\alpha$ share the same number of non-terminals with each non-terminal mapped by $\sim$. $w \in \mathbb{R}$ is a real-valued weight associated with each rule. Starting from an initial non-terminal symbol, each non-terminal is recursively rewritten by the production rule's right hand side $\gamma$ and $\alpha$ associated with $\sim$.

Based on the synchronous-CFG formalism, we adopted the hierarchical phrase-based modeling by introducing some constraints to each production rule [2].

1. A single non-terminal category $X$ is used.

2. Each rule contains at most two non-terminals.

The set of production rules or grammar is automatically learned from word alignment annotated corpora. Specifically, given a bilingual data, we run GIZA++ [7] in two directions. Second, the word alignments are heuristically combined [8]. Finally, phrases are extracted that do not violate word alignment constraints [4]. At the same time, if there exists a phrase with potential embedded phrases, we treat the sub phrases as a non-terminal $X$ [2]. In order to eliminate the spuriously extracted grammar, we further restrict the form of production rules as follows:

3. Each rule contains at most five terminals in each of the source and target sides.

4. No adjacent non-terminals exist in the source side.

In addition to the automatically acquired rules, monotonic rules are added to reduce the data sparseness.

2 Hierarchical Phrase-based SMT

Hierarchical phrase-based SMT is formulated as a probabilistic synchronous context-free grammar (PSCFG) [1] in which string pairs are generated. The system uses a set of source terminal symbols $T_S$, a set of target terminal symbols $T_T$ and a set of non-terminal symbols $N$. Each production rule is realized as follows [2, 11]:

\[ X \rightarrow \langle \gamma, \alpha, \sim, w \rangle \]  

where $X \in N$, $\gamma \in [N \cup T_S]^*$ and $\alpha \in [N \cup T_T]^*$. $\gamma$ and $\alpha$ share the same number of non-terminals with each non-terminal mapped by $\sim$. $w \in \mathbb{R}$ is a real-valued weight associated with each rule. Starting from an initial non-terminal symbol, each non-terminal is recursively rewritten by the production rule's right hand side $\gamma$ and $\alpha$ associated with $\sim$.

Based on the synchronous-CFG formalism, we adopted the hierarchical phrase-based modeling by introducing some constraints to each production rule [2].

1. A single non-terminal category $X$ is used.

2. Each rule contains at most two non-terminals.

The set of production rules or grammar is automatically learned from word alignment annotated corpora. Specifically, given a bilingual data, we run GIZA++ [7] in two directions. Second, the word alignments are heuristically combined [8]. Finally, phrases are extracted that do not violate word alignment constraints [4]. At the same time, if there exists a phrase with potential embedded phrases, we treat the sub phrases as a non-terminal $X$ [2]. In order to eliminate the spuriously extracted grammar, we further restrict the form of production rules as follows:

3. Each rule contains at most five terminals in each of the source and target sides.

4. No adjacent non-terminals exist in the source side.

In addition to the automatically acquired rules, monotonic rules are added to reduce the data sparseness.
problem:

\[ X \rightarrow \langle X_1 \cdots X_n \rangle \]

where boxed indices denote one-to-one mapping of non-terminals between source and target sides.

Translation under PSCFG is regarded as the decoding problem which is cast as a parsing problem using the source side rules. Given a source sentence \( f \), we perform CKY-based parsing using the source yield of the productions rules. The best translation is generated from the target yield \( \hat{e}(D) \) of the best derivation \( D \) according to the weight \( w(D) \) [2].

\[
\hat{e} = \arg\max_{e : f(D) = f, \hat{e}(D) = e} w(D)
\]

The weight of a derivation \( w(D) \) is a \( \lambda \)-scaled linear combination of several (or many) feature functions \( \phi_i \) decomposed by rules \( r \) in \( D \):

\[
w(D) = \sum_{i} \sum_{r \in D} \lambda_i \phi_i(r)
\]

We employed a standard set of features, namely, relative count-based probabilities and lexical probabilities in two directions, various length penalties, and \( n \)-gram language models [2]. For an efficient intersection with \( n \)-gram language models, we introduce cube-pruning [3].

3 Evaluation

We exploited two sets of data for each direction. For the official baseline system, we used only a set of aligned sentence pairs, namely PSD-1. For the contrastive runs, we employed additional data: PSD-2 for additional production rules and PPD-1.2 for larger \( n \)-gram language models. We have also included English Web-1T 5-grams and Japanese Web-1T 7-grams.

All the corpora were case-preserved but normalized according to NFKC, an unicode standard for encoding normalization. The Japanese corpus was tokenized by mecab \(^3\). The English corpus was tokenized by an in-house developed tool following the tokenization standard described in English Web 1T data.

We found that the formal run test data and the tuning/training data come from different epoch with totally different notations for non-ascii letters in English, such as symbols used in equations. Therefore, we convert all the old-style symbol notations into new styles by reverse engineering the publicly available tools \(^4\).

Word alignment is annotated via an in-house spun tool which supports a variant of HMM alignment model [12] with various token factoring [13]. From the heuristically combined factored word alignment, hierarchical rules are extracted with each rule containing at most 5 terminals. The feature scaling factors are MERT tuned [6] using a combination of all the development data consisting of nearly 2,000 sentences with sentence length at most 40 words. The translation results in BLEU [9] are summarized in Table 1.\(^5\). The single-reference BLEU with 1,381 sentences (sBLEU) indicated that our system using only a small subset of data (an official run) resulted in better BLEU. However, the multiple-reference BLEU with 300 sentences (mBLEU) gain a small increase by employing all the data.

4 Conclusion

We presented our strong baseline system faithfully following hierarchical phrase-based machine translation [2]. The official results indicate that the performance is very competitive to the top ranked systems in terms of BLEU.

5 Acknowledgments

This work is partly supported by MEXT Grant-In-Aid for Scientific Research of Priority Areas: Cyber Infrastructure for the Information-Explosion Era.

References


\(^3\)http://mecab.sourceforge.net/

\(^4\)http://www.uspto.gov/web/offices/ac/ido/oeip/sgml/st32/redbook/grbv25x.html

\(^5\)Table 1 shows the single reference BLEU-S with 1,381 sentences and the double reference BLEU-m300-DE with 300 sentences.

<table>
<thead>
<tr>
<th></th>
<th>sBLEU</th>
<th>mBLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Japanese/English</td>
<td>27.20</td>
<td>35.93</td>
</tr>
<tr>
<td>+ Web 1T/PPD</td>
<td>26.88</td>
<td>36.05</td>
</tr>
<tr>
<td>English/Japanese</td>
<td>28.07</td>
<td></td>
</tr>
<tr>
<td>+ Web 1T/PPD</td>
<td>27.20</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Evaluation results by single-reference BLEU (sBLEU) and multiple-reference BLEU (mBLEU).


