A Dependency-to-String Model for Chinese-Japanese SMT System

Hua Shan Yujie Zhang Lu Bai Te Luo
School of Computer and Information technology， Beijing Jiaotong University
{13120422, yjzhang, 13120379, 14120472}@bjtu.edu.cn

Abstract

This paper describes the Beijing Jiaotong University Chinese-Japanese machine translation system which participated in the 2st Workshop on Asian Translation (WAT2015). We exploit the syntactic and semantic knowledge encoded in dependency tree to build a dependency-to-string translation model for Chinese-Japanese statistical machine translation (SMT). Our system achieves a BLEU of 34.87 and a RIBES of 79.25 on the Chinese-Japanese translation task in the official evaluation.

1 Introduction

Motivated by representing the grammatical function of the constituents of a sentence or phrase，dependency grammar holds both syntactic and semantic knowledge．How to building translation model by exploiting the syntactic and semantic knowledge encoded in dependency tree has been now one of the most popular research topics in the recent years．

In dependency tree based models, researchers propose some tree decomposition methods or grammars to build translation model. These models can be classified into string-to-tree model, tree-to-tree model and tree-to-string model. Our system participated in WAT2015 (Nakazawa et al., 2015) adopts tree-to-string model. Particularly, we use the dependency-to-string translation method proposed by (Xie et al., 2011) in Chinese-Japanese translation task. This method proposes a novel tree decompose tion, which takes head-dependents relation (HDR) fragments as elementary structures of rule extraction. An HDR is a tree fragment composed of a head and all its dependents. In this method, the translation rules are expressed with the source side as generalized HDR fragments and the target sides as strings. The model takes substitution as the only operation and can specify reordering information directly into translation rules, thus requires no additional heuristics or reordering models as the previous works. And the model is more concise.

Section 2 describes dependency-to-string translation model in detail. Section 3 reports on our experiment results on a Chinese- SMT system. Section 4 concludes this paper.

2 Dependency-to-String Translation Model

In this paper, we describes the translation model in four aspects, dependency-to-string grammar, translation rule acquisition, the model and the decoding.

2.1 Dependency-to-String Grammar

A dependency structure for a sentence is a directed acyclic graph with words as nodes and modification relations as edges, each edge directing from a head to a dependent. Figure 1 (a) shows an example dependency structure of a Chinese sentence.

2010年 FIFA 世界杯 在 南非 成功 举行

Proceedings of the 2nd Workshop on Asian Translation (WAT2015), pages 82–86,
Kyoto, Japan, 16th October 2015.
2015 Copyright is held by the author(s).

2010 FIFA World Cup in South Africa successfully hold

Here are some properties of a HDR fragment :

1) head determines the syntactic category of HDR, and can often replace HDR;
2) head determines the semantic category of HDR; dependent gives semantic specification.

According to the above properties, we can represent the corresponding HDR fragment with head. The translation rules of dependency-to-string model can be classified into two categories:

-HDR rules, which represent the source side as generalized HDR fragments and the target sides as strings and act as both translation rules and reordering rules.

-H rules, which represent the source side as a word and the target side as words or strings and are used for translating words.

Figure 1 shows examples of the two translation rules. (b), (c) and (d) are three examples of HDR rules, and (d) is an example of H rules. In the figure, the nodes modified by "*" are head of HDR fragment. By the way, the three HDR rules describes translation ways of the same sentence pattern (that is, constituted by "noun phrase + preposition phrase + adverb + verb") and different contexts. Thereinto, rule (b) appoints its context completely, rule (c) restrains its context partially and rule (d) has no restraint for its context.

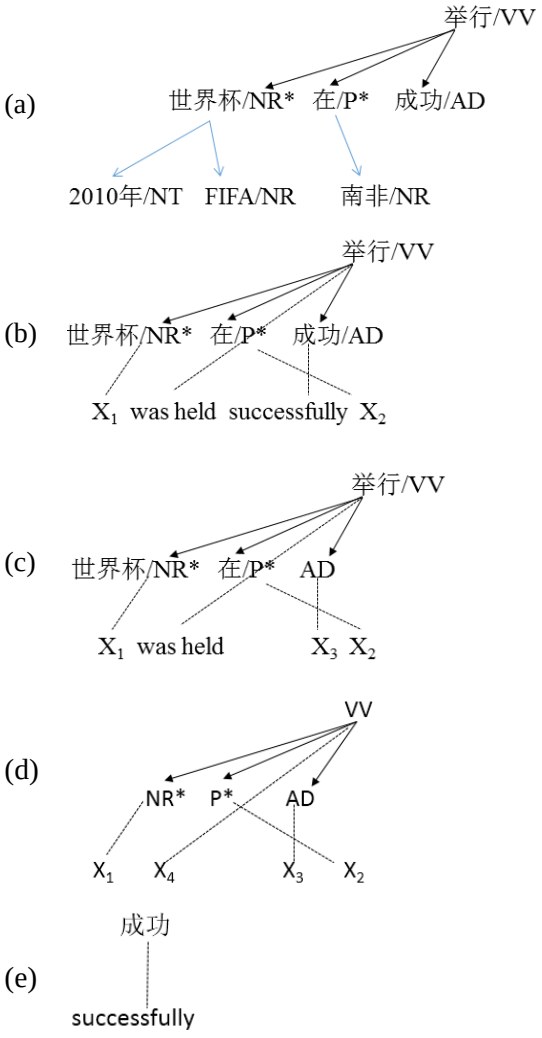

Figure 1: Examples of dependency structure (a), HDR rules (b), (c) , (d)and H rules (e).

2.2 Rule Acquisition

The rule acquisition of dependency-to-string model begins with a parallel corpus with word-aligned results, the source dependency structures and the target side sentence. We accomplish the rule automatic acquisition through the following three steps:

1) Tree annotation: annotate the necessary information on each node of depend ency trees for translation rule acquisition.
2) Acceptable HDR fragments identification: identify HDR fragments from the annotated trees for HDR rules generation.
3) HDR rules generation: generate a series of HDR rules according to the identified acceptable HDR fragments.

The following describes each of these in detail.

2.2.1 Tree Annotation and Acceptable HDR Fragments Identification

The tree annotation can be accomplished by a single postorder transversal of dependency tree *T*. For each node *n* of *T*, we annotated with head span *hsp(n)* and dependency span *dsp(n)* (Xie et al., 2011). During the recursive walk, we calculate *hsp(n)* according to alignment relation for each node *n* accessed. The *dsp(n)* can be obtained according to *hsp(n)* and dependency span of all dependents of *n*. After tree annotation, we can identify HDR fragments for HDR rules generation, according to head span and dependency span of each node.

2.2.2 HDR Rules and H Rules Generation

According to the identified acceptable HDR fragments, a series of lexicalized and unlexicalized HDR rules will be generated. This paper will not describe in detail about it and you can refer to (Xie et al., 2011).

H rules acquisition can be implemented as a sub procedure of HDR rules acquisition. Specifically, in the recursive walk of dependency tree, a H rule is generated according to alignment information for each node accessed.

2.3 Translation Model

Given the dependency-to-string grammar, for a given source language dependency tree *T*, it may generate more than one derivations *D* that convert a source dependency tree *T* into a target string e, thus producing varieties of candidate translations. To compare the candidate translations, we adopt a general log-linear model (Och and Ney, 2002) to define D as:

$$P(D) \propto \prod \phi_i (D)^{\lambda i} \qquad (1)$$

where ϕ_i (D) is feature function defined on derivation D and λ_i are the feature weights.

Our paper used seven features as follows:

1) translation probabilities: $P(t|s)$ and $P(s|t)$;
2) lexical translation probabilities: $P_{lex}(t|s)$ and $P_{lex}(s|t)$;
3) rule penalty: $exp(-1)$;
4) target word penalty: $exp(|e|)$;
5) language model : $P_{lm}(e)$;

2.4 Decoding

Our decoder is based on bottom up chart parsing algorithm that convert the input dependency structure into a target string. It finds the best derivation among all possible derivations *D*. Given a source dependency structure *T*, the decoder traverses each internal node *n* of *T* in post-order. And we process it as follows.

1) If *n* is a leaf node, it checks the rule set for matched translation rules H and uses the rules to generate candidate translation;
2) If *n* is a internal node, it enumerates all instances of the related sentence, clauses or phrases of the HDR fragment rooted at *n*, and checks the translation rule set for matched translation rules. If there is no matched rules, we construct a pseudo translation rule according to the word order of the HDR fragment in the source side;
3) Make use of Cube Pruning algorithm (Chiang, 2007; Huang and Chiang, 2007) to generate the candidate translation for the node *n*.

To balance the decoder's performance and speed, we use four constraints as follows:

1) Beam-threshold: we get the score threshold from the best score in the current stack multiplied by a fixed ratio. The candidate translations with a score worse than the score threshold will be discarded;
2) beam-limit: the maximum number of candidate translations in the beam;
3) rule-threshold: we get the rule score threshold from the best score multiplied by a fixed ratio in the rule table queue. The rules with a score worse than rule score threshold will be dis-

carded;

4) rule-limit: the maximum number of rules in the rule table queue.

For our experiments, we set the beam-threshold = 10^{-2}, beam-limit = 100, rule-threshold = 10^{-2} and rule-limit = 100.

3 Experiments

3.1 Data preparation

We use ASPEC[1] Chinese-Japanese paper excerpt corpus. The training data contains 672,315 sentence pairs, the development data contains 2090 sentence pairs, and the test data contains 2107 sentence pairs.

We employ the Stanford Word Segmenter[2] for Chinese word segmentation, with the standard of CTB. And we use JUMAN[3] for Japanese word segmentation.

After word segmentation, we use the Stanford Parser [4] for Chinese dependency parsing. The parser can create dependency tree for a Chinese sentence and provide the part-of-speech (POS) for each node and the dependency relation type for each edge. Meanwhile, the sentences with special symbols and the sentences whose dependency results contain cross phenomenon will be filtered out. As a result, about 590,000 sentence pairs were obtained for training data.

We apply SRI Language Modeling Toolkit[5] to train a 4-gram language model on the Japanese corpus preprocessed.

We obtain the word alignments by running GIZA++[6] on the corpus in both directions and applying "grow-diag-and" refinement.

We make use of MERT to tune the feature weights in order to maximize the system's BLEU[7] score on the development set.

[1] http://orchid.kuee.kyoto-u.ac.jp/ASPEC/
[2] http://nlp.stanford.edu/software/segmenter.shtml
[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN
[4] http://nlp.stanford.edu/software/lex-parser.shtml
[5] http://www.speech.sri.com/projects/srilm/
[6] http://www.statmt.org/moses/giza/GIZA++.html
[7] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v11b.pl

System	Rule #	BLEU	RIBES
Baseline	35M	34.25	78.94
Dep2str	8.8M	34.87	79.25

Table 1 The comparison results of the two systems

Then, we use dependency-to-string model described in Section 2 to build a Chinese-Japanese translation system. And use the BLUE score and RIBES score for evaluation.

3.2 Experiments and Evaluation Results

The Chinese-Japanese translation system (Dep2str) consists of three modules:

1) Rule extraction module: extract rules using the Chinese dependency tree, the Japanese sentence and alignment information of the training corpus.
2) Decoding module: decode the Chinese sentences for the n-best Japanese translations according to the model parameters that have been set.
3) Training module: train the translation model using minimum error rate to get the best parameters on the development data.

We then decode the test data using the system.

Table 1 shows the number of the extracted translation rules and the translation performance on the test data. Furthermore, we implemented a MOSES PBSMT system (Koehn et al., 2002) as the baseline for a comparison. In our experiments the value of the distortion limit of the baseline system is the default. The number of translation rules and translation performance of the baseline system are also showed in the table.

In terms of the number of translation rules, the number of the extracted translation rules in the baseline system is over 3 times more than that of dep2str system. We think that the lack of restrictions on syntactic structure resulted in this. In terms of translation performance, the BLEU score and RIBES score on the test data

achieved by dep2str system are higher than the baseline system by 0.62 and 0.31 respectively. These evaluation results illustrate that the translation system based on the dependency-to-string model is effective on the Chinese-Japanese translation task.

4 Conclusions

This paper describes the Beijing Jiaotong University Chinese-Japanese machine translation system participated in WAT2015. The system employs a dependency-to-string model, which takes the HDR fragments as elementary structures for the rule extraction and directly specifies the ordering information in translation rules, making the decoding algorithm simplified. The experiment results on the ASPEC data showed that the BLEU score and the RIBES score are increased by 0.62 and 0.31 respectively, compared with the phrase-based system.

At present, the accuracy of the Chinese dependency parsing is not very high, and our system's performance is affected by the accuracy. Meanwhile, we filtered out the sentences which could not be parsed by the dependency parser. This caused a decrease in the amount of training data by about 100 thousand sentence pairs. We think that the system's performance will be improved with Chinese dependency parsing with high accuracy.

References

Jun Xie，Haitao Mi，and Qun Liu．A novel dependency-to-string model for statistical machine translation．In Proceedings of the 201 1 Conference On Empirical Methods in Natumf Language Processing，pages 216—226，Edinburgh，Scotland，UK·，July 2011·Association for Computational Linguistics．

Nakazawa, Toshiaki and Mino, Hideya and Goto, Isao and Neubig, Graham and Kurohashi, Sadao and Sumita, Eiichiro. 2015. Overview of the 2nd Workshop on Asian Translation. In proceedings of the 2nd Workshop on Asian Translation(WAT2015), October 2015, Kyoto, Japan.

Franz Josef Och and Hermann Ney. 2002. Discriminative training and maximum entropy models for statistical machine translation. In Proceedings of 40th Annual Meeting of the Association for Computational Linguistics, pages 295–302, Philadelphia, Pennsylvania, USA, July.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran,Richard Zens, Chris Dyer, Ondrej Bojar, AlexandraConstantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions, pages 177-180.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In Proceedings of 225 ACL 2005, pages 263–270.

Liang Huang and David Chiang. 2007. Forest rescoring: Faster decoding with integrated language models. In Proceedings of ACL 2007, pages 144–151, Prague, Czech Republic, June.

Hideki Isozaki, Tsutomu Hirao, Kevin Duh, Kat suhito Sudoh, and Hajime Tsukada. 2010. Automatic evaluation of translation quality for distant language pairs. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 944-952.