RECURSIVE NEURAL NETWORK BASED WORD TOPOLOGY MODEL FOR HIERARCHICAL PHRASE-BASED SPEECH TRANSLATION

Shixiang Lu, Wei Wei, Xiaoyin Fu, Bo Xu

Institute of Automation, Chinese Academy of Sciences, Beijing, China
{shixiang.lu,wei.wei.media,xiaoyin.fu,xubo}@ia.ac.cn

ABSTRACT

Recursive word topology structure is commonly found in natural language sentences, and discovering this structure can help us to not only identify the units that a sentence contains but also how they interact to form a whole. In this paper, we explore a novel recursive neural network (RNN) based word topology model (WordTM) for hierarchical phrase-based (HPB) speech translation, which captures the topological structure of the words on the source side in a syntactically and semantically meaningful order. Experiments show that our WordTM significantly outperforms the state-of-the-art soft syntactic constraints.

Index Terms— recursive neural network, word topology model, hierarchical phrase-based speech translation

1. INTRODUCTION

Hierarchical phrase-based (HPB) translation model [1, 2] with synchronous context free grammar has provided many attractive benefits in expressing translation knowledge, and effectively maintained the strengths of the phrase-based [3] translation model. However, it only uses the hierarchical rules that span any string of words in the source side input sentence, and does not normally use any syntactic information derived from linguistic knowledge or treebank data [4].

Consider the translation pair in Fig. 1 with the listed hierarchical rules in Fig. 2 and parts of the derivation trees in Fig. 3. Although the correct and incorrect derivations can be generated from the hierarchical rules, their topological structures of the words are quite different on the source side, and the word topology structure under the correct rule derivation seems more reasonable and meets the syntactic constraints. This difference in word topology structure is useful to distinguish the correct translation assumptions but has not been considered by current HPB translation system yet.

The max-margin based recursive neural network (RNN) [5, 6] has successfully parsed natural language words based on deep learned semantic transformations of their original features, and outperforms state-of-the-art approaches. It discovers the hierarchical word structure in natural language using a recursive paradigm, which is a kind of common phenomenon in HPB translation, and we believe this is also helpful to capture the words topology structure and distinguish the correct translation assumptions for HPB speech translation.

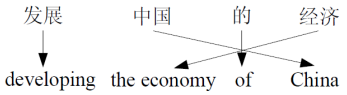

Fig. 1. Example of Chinese-English translation pair.

rule id	hierarchical rule
r_1	$X \rightarrow \langle$发展 X_1, developing $X_1\rangle$
r_2	$X \rightarrow \langle X_1$ 的经济, the economy of $X_1\rangle$
r_3	$X \rightarrow \langle$中国, China$\rangle$
r_4	$X \rightarrow \langle X_1$ 的 X_2, X_2 of $X_1\rangle$
r_5	$X \rightarrow \langle$经济, the economy$\rangle$

Fig. 2. Examples of hierarchical rules extracted from corpus.

In this paper, we adapt and extend the max-margin based RNN into HPB translation with *force decoding* and *converting tree*, and propose a RNN based word topology model from the source side to capture the topological structure of the words and solve derivation problem for our HPB speech translation SimuTalk[1] system. We map the words into the vector semantic representations, and then merge the words into phrases in a syntactically and semantically meaningful order for capturing the word topology structure on the source side. As show in Fig. 3, we assume that the word topology structure under the correct rule derivation tree, which generates correct translation result, are more agree with syntactically and semantically meaningful word merging order.

Similar researches have been done to introduce linguistic commitment for HPB translation. [4] and [7] introduced soft syntactic constraints based on parses of the source language

This work was supported by 863 program in China (No. 2011AA01A207).

[1] http://s2s.ia.ac.cn/speechtrans/

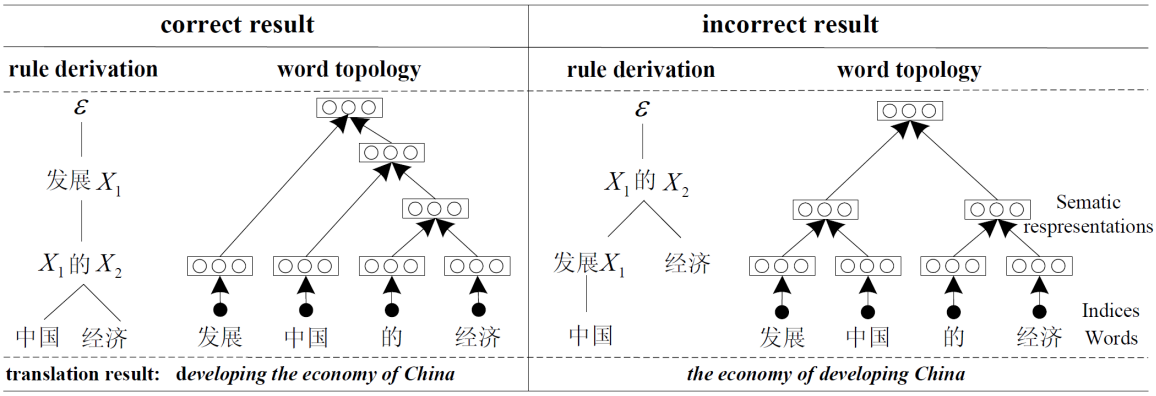

Fig. 3. The derivation trees which lead to correct and incorrect translation results on the source side.

to reward hierarchical rules that were respected with syntactic categories, while [8] constrained the application of hierarchical rules to respect in the target language syntax. Compared with them, our approach is simple and effective, which captures not only syntactic but also semantic commitment.

2. RECURRENT NEURAL NETWORK BASED WORD TOPOLOGY MODEL (WORDTM)

2.1. Model Description

Our WordTM is used to capture the topological structure of the words in the source side sentence using a recursive paradigm, as show in Fig. 3. Using the word semantic representations as input, our WordTM computes (i) a score that is higher when neighboring words should be merged into a phrase, (ii) a new semantic representation for this phrase, and (iii) class label of the phrase types, such as NP or VP. WordTM is trained so that the score is high when neighboring word/phrase has the same class label. After words with the same label are merged into phrase, neighboring phrases are merged to form a longer phrase or a full sentence recursively in a syntactically and semantically order. These merging decisions implicitly define a tree structure in which each node has associated with it the RNN outputs (i)-(iii).

Semantic Representation. We map words to a vector representation in a similar ways as [9] using the Chinese corpus. These word representations are stored in a word embedding matrix $L \in \mathbb{R}^{n\times|V|}$, where $|V|$ is the vocabulary size and n is the semantic space dimensionality. The operation to retrieve the ith word's semantic representation can be seen as a projection layer where we use a binary vector e_k which is zero in all positions except at the kth index,

$$x_i = Le_k \in \mathbb{R}^n \tag{1}$$

Merging Decision. Given two word vector semantic representations, the goal of our WordTM is twofold using the merging decision recursively. Firstly, we computes a new vector representation of the phrase which would combine the two word vector representations. The merging decision is defined as triples ($p \rightarrow c_1c_2$). As in Fig. 4, each such triplet denotes that a parent node p has two children nodes and each children node c_k can be either an input node x_i or a nonterminal node y_j in the tree. We restrict the RNN to two layers, where output layer has the same dimensionality as each input vector in input layer. The RNN computes the potential parent representation for these possible child nodes as

$$p = sigmoid(W[c_1; c_2] + b) \tag{2}$$

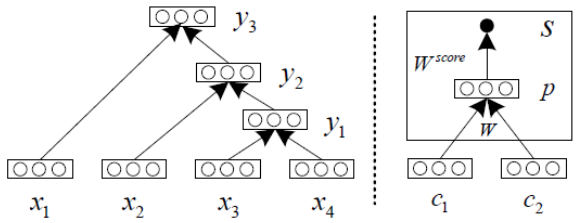

Fig. 4. An example binary tree with a simple RNN. The same weight matrix is replicated and used to compute all non-leaf node representations.

Secondly, we scores how likely this is a correct phrase. We compute a local score of this merging decesion using a simple inner product with a row vector $W^{score} \in \mathbb{R}^{1\times n}$ as

$$s = W^{score}p \tag{3}$$

Class Label. We leverage distributed representation p by adding to each RNN parent node (after removing the scoring layer) a softmax layer to predict class labels, such as NP or VP,

$$label_p = softmax(W^{label}p) \tag{4}$$

Max-Margin Learning. We formulate a global, regularized risk objective in a max-margin framework [5, 6] for parameter learning. Let the training data consist of (sentence,

tree) pairs: (x_i, y_i). We denote the set of all possible binary merging trees that can be constructed from an input sentence as $A(x_i)$. We want to maximize the following objective

$$J = \sum_i s(x_i, y_i) - \max_{y \in A(x_i)} (s(x_i, y) + \Delta(y, y_i)) \quad (5)$$

where the structure loss Δ penalizes trees more when they deviate from the incorrect tree.

For different trees of the source-side current translation sentence, we compute the total score of each tree as the sum of scores of each collapsing merging decision

$$s(x_i, y_i) = \sum_{d \in T(y_i)} s_d(c_1, c_2) \quad (6)$$

This sum score is our *word topology feature* which is used to judge the source-side word topology structure. The trees with higher scores are more agree with syntactically and semantically meaningful order.

A span is a pair of indices which indicate the left and right most leaf nodes under a node in the tree. Let $T(y_i)$ denote the set of spans coming from all nonterminal nodes of the tree. We choose as our loss function a penalization of incorrect spans and add a penalization term λ[2] to each incorrect decision

$$\Delta(y, y_i) = \sum_{d \in T(y)} \lambda 1 d \notin T(y_i) \quad (7)$$

2.2. Parameter Estimation

Our objective J is not differentiable due to the hinge loss. Therefore, following [5, 6], we will generalize gradient ascent via the sub-gradient method which computes a gradient-like direction called the sub-gradient. For any of our parameters (W, W^{score}, W^{label}), such as W, the gradient becomes:

$$\frac{\partial J}{\partial W} = \sum_i \frac{\partial s(x_i, y_i)}{\partial W} - \frac{\partial s(x_i, y_{max})}{\partial W} \quad (8)$$

For each source sentence in HPB translation, finding the highest scoring binary merging tree quickly and accurately (including the penalization) can be achieved with the following two steps: *force decoding* and *converting tree*.

Force Decoding. First, following our previous work [10], we introduce a force decoding to obtain the rule derivation trees by taking the translation pairs from the bilingual corpus. In this way, the source and target sentences are considered to be the input and output sentences for the translation task. The rule derivation trees on both sides are synchronously constructed from bottom to top as a CKY parser. We use the same features as in [2], except the LM feature, as all of the translation assumptions in force decoding are the same with the target sentence in the bilingual corpus. Thus, during decoding, each node in the derivation trees is generated only if the translation sentence fragment is exactly matched with the target side. When it has reached the top of the derivation tree, the derivation tree of hierarchical rules can be obtained during trace back, and we only trace the top-7 rule derivation trees.

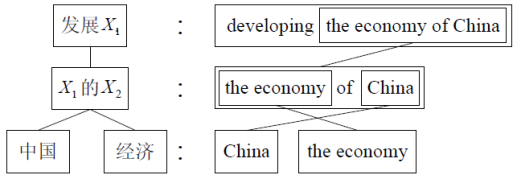

Fig. 5. Example of derivation tree with the hierarchical rules in Fig. 2.

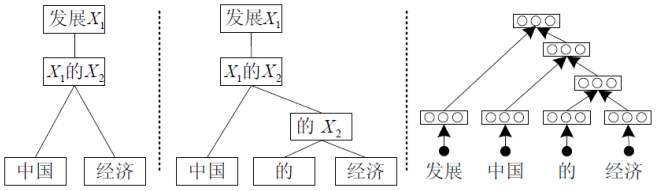

Fig. 6. Example of converting the derivation tree in Figure 5 into a binary tree for WordTM.

Fig. 5 illustrates the force decoding processes that obtain one of the rule derivations with hierarchical rules for the sentence pair in Fig. 1. The nonterminal symbols on the target side of hierarchical rules have been replaced by the translation assumptions. The derivation tree on the left is generated when the translated sub strings on the target side.

Converting Tree. Second, we would further convert the rule derivation tree on the source side into a binary tree for WordTM, then get the detail training samples for WordTM. Note that the grammar (merging decision) in RNN is not context free and only consists of the glue rule $X \rightarrow XX$. So this step is necessary for our WordTM learning.

Fig. 6 shows how to convert the rule derivation tree into a binary tree for WordTM. In our experiments, we define three following operations for converting tree,

1. For glue rule S, we consider it has the same operation with the merging decision.
2. For the hierarchical rule which has one nonterminal, we also consider it is same to the merging decision, such as the rule r_1 in Figure 2. "发展 X_1" is merge by "发展" and "X_1 (中国 的 经济)".
3. For the hierarchical rule which has two nonterminals, such as the rule r_2, we consider it is consisted by two merging decisions. We merge "的" (terminal) and "X_2 (经济)" (the right nonterminal) into "的 X_2 (的 经济)" firstly， and then merge "的 X_2 " and "X_1 (中国)" into "X_1 的 X_2 (中国 的 经济)". Note that there is another hierarchical rule which has two nonterminals, such as "与 X_1 有 X_2" [2]. We first merge "与 X_1" and "有 X_2" respectively, and then merge "与 X_1 有 X_2".

[2]As in Socher et al. (2011), a value of 0.05 was used for λ.

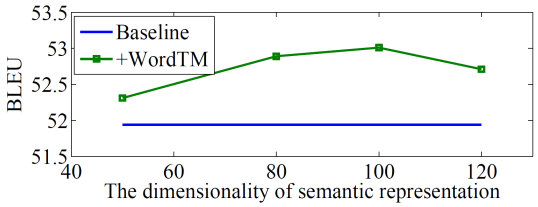

Fig. 7. Results on the development set with different dimensionality of semantic representation.

3. EXPERIMENTS AND RESULTS

3.1. Experimental Setup

We conduct experiments on the following Chinese-English speech translation task. Our development set is IWSLT 2005 test set (506 sentences), and our test set is IWSLT 2007 test set (489 sentences). The bilingual corpus is the Chinese-English part of Basic Traveling Expression corpus and China-Japan-Korea corpus, which contains 0.387M sentence pairs with 3.5/3.82M Chinese/English words. The LM training corpus is from the English side of the bilingual data.

The word semantic representation embedding matrix is trained on Chinese side of the bilingual corpus. Our vocabulary consists of all words that occurred more than twice in the corpus, remaining words are mapped to an unknown word.

The baseline system is following the same constrains as in [2]. In the contrast experiments, we add our *word topology feature*, the sum of scores of each collapsing merging decision in different source-side rule derivation tree of the current translation sentence, and this feature is combined into a standard log-linear model [11].

The LM are trained using the SRILM toolkit with modified Kneser-Ney smoothing. We perform pairwise ranking optimization (PRO) [12] to tune feature weights on the development set, and evaluate the translation quality using case-insensitive BLEU-4 score [13].

3.2. Experimental Results

Fig. 7 shows that different dimensionality of semantic representation affects the performance of WordTM, and 100-dimensional semantic representation gets the best performance. We use 100 dimensionality semantic representation for the next experiments.

Table 1 presents the main results on the test set. For the contrast experiments, we introduce source-side soft syntactic constraints (*coarse and fine-gained features*) [4] based on parses of the source language to reward hierarchical rules. The results show that adding word topology feature (+WordTM) is useful to distinguish the correct translation assumptions on the source side and significantly better than the baseline features, with an increase of 0.91 BLEU points. Compared with the state-of-the-art soft syntactic constraints ("+syntax (coarse)" and "+syntax (fine)"), our WordTM is more simple (without syntactic parses) and effective (with an increase of 0.36/0.30 BLEU points over coarse/fine-gained features).

Features	Dev	Test
Baseline	51.94	40.62
+syntax (coarse)	52.58	41.17*
+syntax (fine)	52.65	41.23*
+WordTM (100 dimensionality)	53.01	41.53**

Table 1. Results on the test set, and the improvements are statistically significant by the bootstrap resampling [14]. *: significantly better than the baseline ($p < 0.05$), **: significantly better than "soft syntactic constraints" ($p < 0.01$).

Features	Dev	Test
Baseline	50.84	42.32
+syntax (coarse)	51.41	42.75*
+syntax (fine)	51.45	42.81*
+WordTM (100 dimensionality)	51.96	43.29**

Table 2. Results with BLEU scores on the large data set. The meaning of "*" and "**" are similar to Table 1.

3.3. Large Data Set

We also conduct experiments on a larger bilingual corpus, which is partly used in our SimuTalk system. The larger bilingual corpus are collected from web data (such as, the bilingual subtitles in shooter, and the example bilingual sentence pairs in Jinshan, Baidu and Youdao Dictionary), which contains 4.3M parallel sentences pairs with 53/55M Chinese/English words, and they are most relevant to the spoken language domain. Similarly, the LM training corpus is from English side of the bilingual data, and the word semantic representation embedding matrix is trained on Chinese side of the bilingual corpus with 100 dimensionality.

The final BLEU score results are shown in Table 2. In the scenario with a large data set, our WordTM still significantly outperforms the state-of-the-art soft syntactic constraints.

4. CONCLUSIONS

In this paper, we adapt and extend the max-margin based RNN into HPB translation with force decoding and converting tree, and propose a RNN based word topology model for our HPB speech translation SimuTalk system, which successfully captures the topological structure of the words on the source side in a syntactically and semantically meaningful order. Experiments show that our WordTM significantly outperforms the state-of-the-art soft syntactic constraints for HPB translation. In the future, we will extend our RNN based word topology model into the target side, and captures both source- and target-side words topology structure, as to further improve the performance of our HPB speech translation.

5. REFERENCES

[1] D. Chiang, "hierarchical phrase-based model for statistical machine translation," in *Proceedings of ACL*, 2005, pp. 263–270.

[2] D. Chiang, "Hierarchical phrase-based translation," *Computational Linguistics*, vol. 33, no. 2, pp. 201–228, 2007.

[3] P. Koehn, F. J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of NAACL*, 2003, pp. 48–54.

[4] D. Chiang, Y. Marton, and P. Resnik, "Online large-margin training of syntactic and structural translation features," in *Proceedings of EMNLP*, 2008, pp. 224–233.

[5] R. Socher, C. D. Manning, and A. Y. Ng, "Learning continuous phrase representations and syntactic parsing with recursive neural networks," in *Proceedings of NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2010.

[6] R. Socher, C. C. Lin, A. Y. Ng, and C. D. Manning, "Parsing natural scenes and natural language with recursive neural networks," in *Proceedings of ICML*, 2011.

[7] Y. Marton and P. Resnik, "Soft syntactic constraints for hierarchical phrased-based translation," in *Proceedings of ACL*, 2008, pp. 1003–1011.

[8] A. Zollmann and A. Venugopal, "Syntax augmented machine translation via chart parsing," in *Proceedings of WMT*, 2006, pp. 138–141.

[9] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *Proceedings of ICML*, 2008, pp. 160–167.

[10] X. Fu, W. Wei, L. Fan, S. Lu, and B.Xu, "Nesting hierarchical phrase-based model for speech-to-speech translation," in *Proceedings of ISCSLP*, 2012, pp. 368–372.

[11] F. J. Och and H. Ney, "Discriminative training and maximum entropy models for statistical machine translation," in *Proceedings of ACL*, 2002, pp. 295–302.

[12] M. Hopkins and J. May, "Tuning as ranking," in *Proceedings of EMNLP*, 2011, pp. 1352–1362.

[13] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of ACL*, 2002, pp. 311–318.

[14] P. Koehn, "Statistical significance tests from achine translation evaluation," in *Proceedings of ACL*, 2004, pp. 1352–1362.