Automatic Cross-Lingual Similarization of Dependency Grammars for Tree-based Machine Translation

Wenbin Jiang [1] and **Wen Zhang** [1] and **Jinan Xu** [2] and **Rangjia Cai** [3]
[1]Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences, China
[2]School of Computer and Information Technology, Beijing Jiaotong University, China
[3]Research Center of Tibetan Information, Qinghai Normal University, China
jiangwenbin@ict.ac.cn

Abstract

Structural isomorphism between languages benefits the performance of cross-lingual applications. We propose an automatic algorithm for cross-lingual similarization of dependency grammars, which automatically learns grammars with high cross-lingual similarity. The algorithm similarizes the annotation styles of the dependency grammars for two languages in the level of classification decisions, and gradually improves the cross-lingual similarity without losing linguistic knowledge resorting to iterative cross-lingual cooperative learning. The dependency grammars given by cross-lingual similarization have much higher cross-lingual similarity while maintaining non-triviality. As applications, the cross-lingually similarized grammars significantly improve the performance of dependency tree-based machine translation.

1 Introduction

Due to the inherent syntactic regularity of each language and the discrepancy between annotation guidelines of linguists, there is not necessarily structural isomorphism between grammars of different languages. For many cross-lingual scenarios such as information retrieval and machine translation, relationships between linguistic units are expected to be (at least roughly) consistent across languages (Hwa et al., 2002; Smith and Eisner, 2009). For cross-lingual applications, syntactic structures with high cross-lingual similarity facilitates knowledge extraction, feature representation and classification decision. The structural isomorphism between languages, therefore, is an important aspect for the performance of cross-lingual applications such as machine translation.

To achieve effective cross-lingual similarization for two grammars in different languages, an adequate algorithm should both improve the cross-lingual similarity between two grammars and maintain the non-triviality of each grammar, where non-triviality indicates that the resulted grammars should not give flat or single-branched outputs. Different from constituency structures, dependency structures are lexicalized without specialized hierarchical structures. Such concise structures depict the syntactic or semantic relationships between words, and thus have advantage on many cross-lingual scenarios. It is worth to perform cross-lingual similarization for dependency grammars, but the special property of dependency grammars makes it hard to directly adopt the conventional structure transformation methods resorting to hand-crafted rules or templates.

Both graph-based models (McDonald et al., 2005) and transition-based models (Nivre et al., 2006) factorize dependency parsing into fundamental classification decisions, that is, the relationships between words or the actions applied to current states. We assume that cross-lingual similarization can also be factorized into fundamental classification decisions, and propose an automatic cross-lingual similarization algorithm for dependency grammars according to this assumption. The algorithm conducts cross-lingual similarization on the level of classification decisions

Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 501–510,
Austin, Texas, November 1-5, 2016.

with simple blending operations rather than on the level of syntactic structures with complicated hand-crafted rules or templates, and adopts iterative cross-lingual collaborative learning to gradually improve the cross-lingual similarity while maintaining the non-triviality of grammars.

We design an evaluation metric for the cross-lingual similarity of dependency grammars, which calculates the consistency degree of dependency relationships across languages. We also propose an effective method to measure the *real* performance of the cross-lingually similarized grammars based on the transfer learning methodology (Pan and Yang, 2010). We validate the method on the dependency grammar induction of Chinese and English, where significant increment of cross-lingual similarity is achieved without losing non-triviality of the grammars. As applications, the cross-lingually similarized grammars gain significant performance improvement for the dependency tree-based machine translation by simply replacing the parser of the translator.

2 Graph-based Dependency Parsing

Dependency parsing aims to link each word to its arguments so as to form a directed graph spanning the whole sentence. Normally the directed graph is restricted to a dependency tree where each word depends on exactly one parent, and all words find their parents. Given a sentence as a sequence n words:

$$x = x_1\ x_2\ ..\ x_n$$

dependency parsing finds a dependency tree y, where $(i, j) \in y$ is an edge from the head word x_i to the modifier word x_j. The root $r \in x$ in the tree y has no head word, and each of the other words, $j(j \in x \textbf{ and } j \neq r)$, depends on a head word $i(i \in x \textbf{ and } i \neq j)$.

Following the edge-based factorization method (Eisner, 1996), the score of a dependency tree can be factorized into the dependency edges in the tree. The graph-based method (McDonald et al., 2005) factorizes the score of the tree as the sum of the scores of all its edges, and the score of an edge is defined as the inner product of the feature vector and the weight vector. Given a sentence x, the parsing procedure searches for the candidate dependency tree with the maximum score:

$$\begin{aligned} y(x) &= \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \mathbf{S}(y) \\ &= \underset{y \in \mathbf{GEN}(x)}{\operatorname{argmax}} \sum_{(i,j) \in y} \mathbf{S}(i, j) \end{aligned} \quad (1)$$

Here, the function $\mathbf{GEN}$ indicates the enumeration of candidate trees. The MIRA algorithm (Crammer et al., 2003) is used to train the parameter vector. A bottom-up dynamic programming algorithm is designed for projective parsing which gives projective parsing trees, and the Chu-Liu-Edmonds algorithm for non-projective parsing which gives non-projective parsing trees.

3 Cross-Lingual Similarization

Since structural analysis can be factorized into fundamental classification decisions, we assume that the adjustment of the analysis results can be factorized into the adjustment of the fundamental decisions. The classification decision for graph-based dependency parsing is to classify the dependency relationship between each pair of words, and we hope it works well to conduct cross-lingual similarization on the level of dependency relationship classification. In this work, we investigate the automatic cross-lingual similarization for dependency grammars on the level of fundamental classification decisions, to avoid the difficulty of using hand-crafted transformation rules or templates.

In this section, we first introduce the evaluation metric for cross-lingual similarity, then describe the automatic cross-lingual similarization algorithm, and finally give a method to measure the real performance of the cross-lingually similarized grammars.

3.1 Evaluation of Cross-Lingual Similarity

The cross-lingual similarity between two dependency structures can be automatically evaluated. Dependency parsing is conducted on sentences, so we take bilingual sentence pairs as the objects for evaluation. The calculation of cross-lingual similarity needs the lexical alignment information between two languages, which can be obtained by manual annotation or unsupervised algorithms.

Given a bilingual sentence pair x_α and x_β, their dependency structures y_α and y_β, and the word

alignment probabilities $\mathcal{A}$, the cross-lingual similarity can be calculated as below:

$$d(y_\alpha, y_\beta) = \frac{\sum_{(i,j)\in y_\alpha} \sum_{(i',j')\in y_\beta} \mathcal{A}_{i,i'}\mathcal{A}_{j,j'}}{\sum_{(i,j)\in y_\alpha} \sum_{i',j'\in x_\beta} \mathcal{A}_{i,i'}\mathcal{A}_{j,j'}} \quad (2)$$

The bracketed word pair indicates a dependency edge. The evaluation metric is a real number between 0 and 1, indicating the degree of cross-lingual consistency between two dependency structures. For the cross-lingual similarity between bilingual paragraphs, we simply define it as the average over the similarity between each sentence pairs.

3.2 Factorized Cooperative Similarization

The fundamental decisions for graph-based dependency parsing are to evaluate the candidate dependency edges. The cross-lingual similarization for fundamental decisions can be defined as some kinds of blending calculation on two evaluation scores, of which the one is directly given by the grammar of the current language (current grammar), and the other is bilingually projected from the grammar of the reference language (reference grammar).

For the words i and j in the sentence x_α in the current language, their evaluated score given by the current grammar is $\mathbf{S}_\alpha(i,j)$, which can be calculated according to formula 1. The score bilingually projected from the reference grammar, $\mathbf{S}^\beta(i,j)$, can be obtained according to the translation sentence x_β in the reference language and the word alignment between two sentences:

$$\mathbf{S}^\beta(i,j) = \sum_{i',j'\in x_\beta} \mathcal{A}_{i,i'}\mathcal{A}_{j,j'}\mathbf{S}_\beta(i',j') \quad (3)$$

where i' and j' are the corresponding words of i and j in the reference sentence x_β, $\mathcal{A}_{i,j}$ indicates the probability that i aligns to j, and $\mathbf{S}_\beta(i',j')$ is the evaluated score of the candidate edge (i',j') given by the reference grammar.

Given the two evaluated scores, we simply adopt the linear weighted summation:

$$\mathbf{S}^\beta_\alpha(i,j) = (1-\lambda)\mathbf{S}_\alpha(i,j) + \lambda\mathbf{S}^\beta(i,j) \quad (4)$$

where λ is the relative weight to control the degree of cross-lingual similarization, indicating to which degree we consider the decisions of the reference grammar when adjusting the decisions of the current grammar. We have to choose a value for λ to achieve an appropriate speed for effective cross-lingual similarization, in order to obtain similarized grammars with high cross-lingual similarity while maintaining the non-triviality of the grammars.

In the re-evaluated full-connected graph, the decoding algorithm searches for the cross-lingually similarized dependency structures, which are used to re-train the dependency grammars. Based on the cross-lingual similarization strategy, iterative cooperative learning simultaneously similarizes the sentences in the current and reference languages, and gradually improves the cross-lingual similarity between two grammars while maintaining the non-triviality of each monolingual grammar. The whole training algorithm is shown in Algorithm 1. To reduce the computation complexity, we choose the same λ for the cross-lingual similarization for both the current and the reference grammars. Another hyper-parameter for the iterative cooperative learning algorithm is the maximum training iteration, which can be determined according to the performance on the development sets.

3.3 Evaluation of Similarized Grammars

The *real* performance of a cross-lingually similarized grammar is hard to directly measured. The accuracy on the standard testing sets no longer reflects the actrual accuracy, since cross-lingual similarization leads to grammars with annotation styles different from those of the original treebanks. We adopt the transfer learning strategy to automatically adapt the divergence between different annotation styles, and design a *transfer classifier* to transform the dependency regularities from one annotation style to another.

The training procedure of the transfer classifier is analogous to the training of a normal classifier except for the features. The transfer classifier adopts guiding features where a guiding signal is attached to the tail of each normal feature. The guiding signal is the dependency path between the pair of words in the source annotations, as shown in Figure 2. Thus, the transfer classifier learns the statistical regularity of the transformation from the annotations of the cross-lingually similarized grammar to the annotations of the original treebank. Figure 1 shows

Algorithm 1 Cooperative cross-lingual similarization.

function BISIMILARIZE($\mathbf{G}_\alpha$, $\mathbf{G}_\beta$, λ, $\mathbf{C}$) ▷ $\mathbf{C}$ includes a set of sentence pairs (x_α, x_β)
  repeat
    $\mathbf{T}_\alpha, \mathbf{T}_\beta \leftarrow$ BIANNOTATE($\mathbf{G}_\alpha, \mathbf{G}_\beta, \lambda, \mathbf{C}$) ▷ it invokes BIPARSE to parse each (x_α, x_β)
    $\mathbf{G}_\alpha \leftarrow$ GRAMMARTRAIN($\mathbf{T}_\alpha$)
    $\mathbf{G}_\beta \leftarrow$ GRAMMARTRAIN($\mathbf{T}_\beta$)
  until SIMILARITY($\mathbf{G}_\alpha, \mathbf{G}_\beta$) converges ▷ according to formula 2, averaged across $\mathbf{C}$
  return $\mathbf{G}_\alpha$, $\mathbf{G}_\beta$
function BIPARSE($\mathbf{G}_\alpha$, $\mathbf{G}_\beta$, λ, x_α, x_β, $\mathcal{A}$)
  $y_\alpha \leftarrow \text{argmax}_y(1-\lambda)\mathbf{S}_\alpha(y) + \lambda\mathbf{S}^\beta(y)$ ▷ according to formula 4
  $y_\beta \leftarrow \text{argmax}_y(1-\lambda)\mathbf{S}_\beta(y) + \lambda\mathbf{S}^\alpha(y)$
  return y_α, y_β

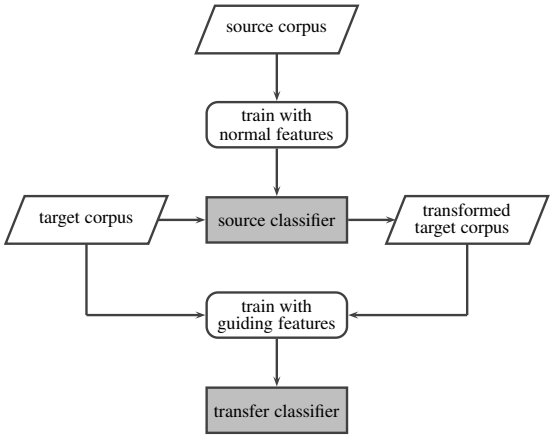
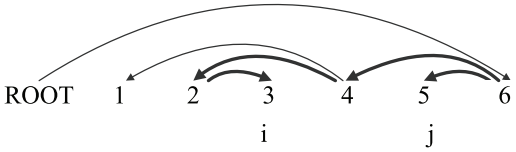


Figure 1: The training procedure of the transfer classifier.

the training pipeline for the transfer classifier, where source corpus and target corpus indicate the cross-lingually similarized treebank and the manually annotated treebank, respectively.

In decoding, a sentence is first parsed by the cross-lingually similarized grammar, and then parsed by the transfer classifier with the result of the similarized grammar as guiding signals to obtain the final parsing results. The improvement achieved by the transfer classifier against a normal classifier trained only on the original treebank reflects the promotion effect of the cross-lingually similarized grammar. The accuracy of the transfer classifier, therefore, *roughly* indicates the real performance of the cross-lingually similarized grammar.



path(i =3, j = 5) = up-up-up-down

Figure 2: The guiding signal for dependency parsing, where $path(i, j)$ denotes the dependency path between i and j. In this example, j is a son of the great-grandfather of i.

4 Tree-based Machine Translation

Syntax-based machine translation investigates the hierarchical structures of natural languages, including formal structures (Chiang, 2005), constituency structures (Galley et al., 2006; Liu et al., 2006; Huang et al., 2006; Mi et al., 2008) and dependency structures (Lin, 2004; Quirk et al., 2005; Ding and Palmer, 2005; Xiong et al., 2007; Shen et al., 2008; Xie et al., 2011), so the performance is restricted to the quality and suitability of the parsers. Since the trees for training follow an annotation style not necessarily isomorphic to that of the target language, it would be not appropriate for syntax-based translation to directly use the parsers trained on the original treebanks. The cross-lingually similarized grammars, although performing poorly on a standard testing set, may be well suitable for syntax-based machine translation. In this work, we use the cross-lingually similarized dependency grammars in dependency tree-to-string machine translation (Xie et al., 2011), a state-of-the-art translation model resorting to dependency trees on the source side.

Treebank	Train	Develop	Test
CTB	1-270 400-931 1001-1151	301-325	271-300
WSJ	02-21	22	23

Table 1: Data partitioning for CTB and WSJ, in unit of section.

5 Experiments and Analysis

We first introduce the dependency parsing itself, then describe the cross-lingual similarization, and finally show the application of cross-lingually similarized grammars in tree-based machine translation. For convenience of description, a grammar trained by the conventional dependency model is named as *original grammar*, a grammar after cross-lingual similarization is named as *similarized grammar*, and the transferred version for a similarized grammar is named as *adapted grammar*.

5.1 Dependency Parsing

We take Chinese dependency parsing as a case study, and experiment on Penn Chinese Treebank (CTB) (Xue et al., 2005). The dependency structures are extracted from the original constituency trees according to the head-selection rules (Yamada and Matsumoto, 2003). The partitioning of the dataset is listed in the Table 1, where we also give the partitioning of Wall Street Journal (WSJ) (Marcus et al., 1993) used to train the English grammar. The evaluation metric for dependency parsing is unlabeled accuracy, indicating the proportion of the words correctly finding their parents. The MIRA algorithm is used to train the classifiers.

Figure 3 gives the performance curves on the development set with two searching modes, *projective* searching and *non-projective* searching. The curves show that the non-projective searching mode fall behind of the projective one, this is because the dependency structures extracted from constituency trees are projective, and the projective search mode implies appropriate constraints on the searching space. Therefore, we use the projective searching mode for the evaluation of the original grammar. Table 2 lists the performance of the original grammar on the CTB testing set.

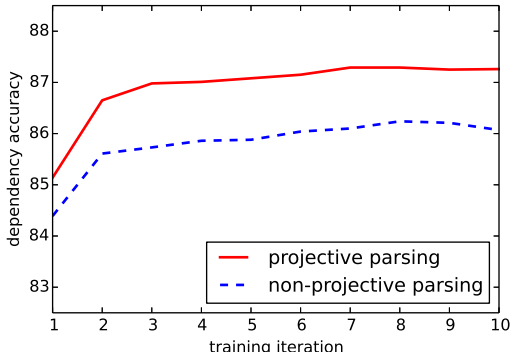


Figure 3: The developing curves of Chinese dependency parsing with both projective and non-projective searching modes.

5.2 Cross-Lingual Similarization

The experiments of cross-lingual similarization are conducted between Chinese and English, with FBIS Chinese-English dataset as bilingual corpus. The Chinese sentences are segmented into words with the character classification model (Ng and Low, 2004), which is trained by MIRA on CTB. The word sequences of both languages are labeled with part-of-speech tags with the maximum entropy hidden markov model (Ratnaparkhi and Adwait, 1996), which is reimplemented with MIRA and trained on CTB and WSJ. The word alignment information is obtained by summing and normalizing the 10 best candidate word alignment results of GIZA++ (Och and Ney, 2003).

The upmost configuration for cross-lingual similarization is the searching mode. On the Chinese side, both projective and non-projective modes can be adopted. For English, there is an additional *fixed* mode besides the previous two. In the fixed mode, the English dependency grammar remains unchanged during the whole learning procedure. The fixed mode on the English side means a degenerated version of cross-lingual similarization, where only the Chinese grammars are revolved during training. The combination of the searching modes for both languages results in a total of 6 kinds of searching configurations. For each configuration, the learning algorithm for cross-lingual similarization has two hyper-parameters, the coefficient λ and maximum iteration for iterative learning, which should be tuned first.

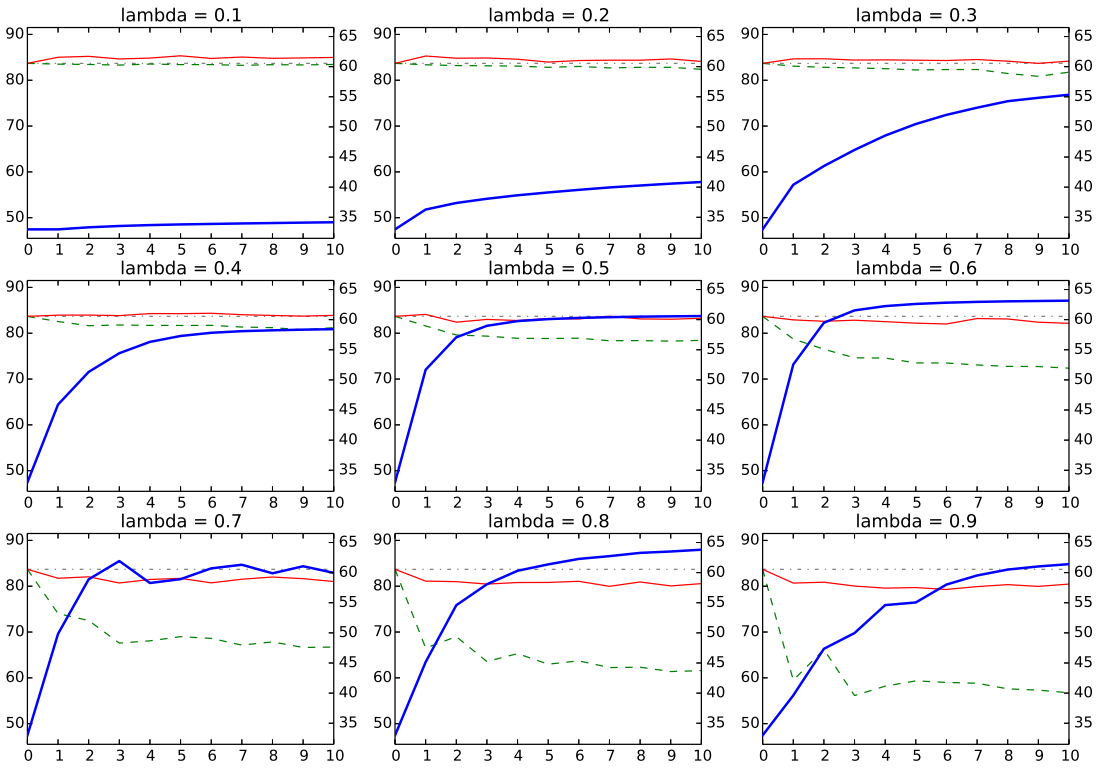


Figure 4: The developing curves of cross-lingual similarization with projective searching on both languages. X-axis: training iteration; Left Y-axis: parsing accuracy; Right Y-axis: cross-lingual similarity. Thin dash-dotted line (gray): accuracy of the baseline grammar; Thin dashed line (green): direct accuracy of cross-lingually similarized grammars; Thin solid line (red): adaptive accuracy of cross-lingually similarized grammars; Thick sold line (blue): the cross-lingual similarity of grammars.

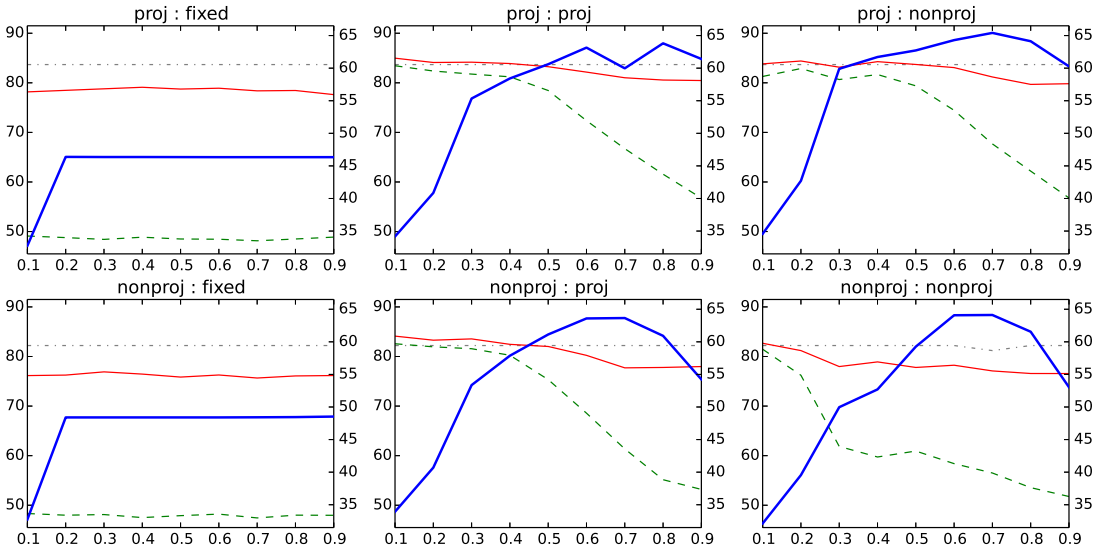


Figure 5: The developing curves of cross-lingual similarization with all searching configurations. X-axis: coefficient λ; Left Y-axis: parsing accuracy; Right Y-axis: cross-lingual similarity. The lines indicate the same as in Figure 4.

5.2.1 Determination of Hyper-Parameters

We select a subset of 40,000 sentence pairs out of the FBIS dataset, and use it as the smaller bilingual corpus to tune the hyper-parameters. For the coefficient λ we try from 0.1 to 0.9 with step 0.1; and for the iterative learning we simply set the maximum iteration as 10. The developing procedure results in a series of grammars. For the configuration with projective searching modes on both sides, a total of 90 pairs of Chinese and English grammars are generated. We use three indicators to validate each similarized grammar generated in the developing procedure, including the performance on the similarized grammar itself (direct accuracy), the performance of the corresponding adapted grammar (adaptive accuracy), and the cross-lingual similarity between the similarized grammar and its English counterpart. Figure 4 shows the developing curves for the configuration with projective searching on both sides. With the fixed maximum iteration 10, we draw the developing curves for the other searching configurations with respect to the weight coefficient, as shown in Figure 5.

We find that the optimal performance is also achieved at 0.6 in most situations. In all configurations, the training procedures increase the cross-lingual similarity of grammars. Along with the increment of cross-lingual similarity, the direct accuracy of the similarized grammars on the development set decreases, but the adaptive accuracy given by the corresponding adapted grammars approach to that of the original grammars. Note that the projective searching mode is adopted for the evaluation of the adapted grammar.

5.2.2 Selection of Searching Modes

With the hyper-parameters given by the developing procedures, cross-lingual similarization is conducted on the whole FBIS dataset. All the searching mode configurations are tried and 6 pairs of grammars are generated. For each of the 6 Chinese dependency grammars, we also give the three indicators as described before. Table 2 shows that, cross-lingual similarization results in grammars with much higher cross-lingual similarity, and the adaptive accuracies given by the adapted grammars approach to those of the original grammars. It indicates that the proposed algorithm improve the cross-lingual similarity without losing syntactic knowledge.

To determine the best searching mode for tree-based machine translation, we use the Chinese-English FBIS dataset as the small-scale bilingual corpus. A 4-gram language model is trained on the Xinhua portion of the Gigaword corpus with the SRILM toolkit (Stolcke and Andreas, 2002). For the analysis given by non-projective similarized grammars, The *projective transformation* should be conducted in order to produce projective dependency structures for rule extraction and translation decoding. In details, the projective transformation first traverses the non-projective dependency structures just as they are projective, then adjusts the order of the nodes according to the traversed word sequences. We take NIST MT Evaluation testing set 2002 (NIST 02) for developing , and use the case-sensitive BLEU (Papineni et al., 2002) to measure the translation accuracy.

The last column of Table 2 shows the performance of the grammars on machine translation. The cross-lingually similarized grammars corresponding to the configurations with projective searching for Chinese always improve the translation performance, while non-projective grammars always hurt the performance. It probably can be attributed to the low performance of non-projective parsing as well as the inappropriateness of the simple projective transformation method. In the final application in machine translation, we adopted the similarized grammar corresponding to the configuration with projective searching on the source side and non-projective searching on the target side.

5.3 Improving Tree-based Translation

Our large-scale bilingual corpus for machine translation consists of 1.5M sentence pairs from LDC data, including LDC2002E18, LDC2003E07, LDC2003E14, Hansards portion of LDC2004T07, LDC2004T08 and LDC2005T06. The source sentences are parsed by the original grammar and the selected cross-lingually similarized grammar. The alignments are obtained by running GIZA++ on the corpus in both directions and applying grow-diag-and refinement (Koehn et al., 2003). The English language model is trained on the Xinhua portion of the Gigaword corpus with the SRILM toolkit (Stol-

Grammar	Similarity (%)	Dep. P (%)	Ada. P (%)	BLEU-4 (%)
baseline	34.2	84.5	84.5	24.6
proj : fixed	46.3	54.1	82.3	25.8 $_{(+1.2)}$
proj : proj	63.2	72.2	84.6	26.1 $_{(+1.5)}$
proj : nonproj	64.3	74.6	84.7	26.2 $_{(+1.6)}$
nonproj : fixed	48.4	56.1	82.6	20.1 $_{(-4.5)}$
nonproj : proj	63.6	71.4	84.4	22.9 $_{(-1.7)}$
nonproj : nonproj	64.1	73.9	84.9	20.7 $_{(-3.9)}$

Table 2: The performance of cross-lingually similarized Chinese dependency grammars with different configurations.

System	NIST 04	NIST 05
(Liu et al., 2006)	34.55	31.94
(Chiang, 2007)	35.29	33.22
(Xie et al., 2011)	35.82	33.62
Original Grammar	35.44	33.08
Similarized Grammar	**36.78**	**35.12**

Table 3: The performance of the cross-lingually similarized grammar on dependency tree-based translation, compared with related work.

cke and Andreas, 2002). We use NIST 02 as the development set, and NIST 04 and NIST 05 as the testing sets. The quality of translations is evaluated by the case insensitive NIST BLEU-4 metric.

Table 3 shows the performance of the cross-lingually similarized grammar on dependency tree-based translation, compared with previous work (Xie et al., 2011). We also give the performance of constituency tree-based translation (Liu et al., 2006) and formal syntax-based translation (Chiang, 2007). The original grammar performs slightly worse than the previous work in dependency tree-based translation, this can ascribed to the difference between the implementation of the original grammar and the dependency parser used in the previous work. However, the similarized grammar achieves very significant improvement based on the original grammar, and also significant surpass the previous work. Note that there is no other modification on the translation model besides the replacement of the source parser.

From the perspective of performance improvement, tree-to-tree translation would be a better scenario to verify the effectiveness of cross-lingual similarization. This is because tree-to-tree translation suffers from more serious non-isomorphism between the source and the target syntax structures, and our method for cross-lingual similarization can simultaneously similarize both the source and the target grammars. For dependency-based translation, however, there are no available tree-to-tree models for us to verify this assumption. In the future, we want to propose a specific tree-to-tree translation method to better utilize the isomorphism between cross-lingually similarized grammars.

6 Related Work

There are some work devoted to adjusting the syntactic structures according to bilingual constraints to improve constituency tree-based translation (Huang and Knight, 2006; Ambati and Lavie, 2008; Wang et al., 2010; Burkett and Klein, 2012; Liu et al., 2012). These efforts concentrated on constituency structures, adopted hand-crafted transformation templates or rules, and learnt the operation sequences of structure transformation on the bilingual corpora. Such methods are hard to be directly applied to dependency structures due to the great discrepancy between constituency and dependency grammars. There are also work on automatically adjusting the syntactic structures for machine translation resorting to self-training (Morishita et al., 2015), where the parsed trees for self-training are selected according to translation performance. Our work focuses on the automatic cross-lingual similarization of dependency grammars, and learnt grammars with higher cross-lingual similarity while maintaining the non-triviality of the grammars.

There are substantial efforts that have been made in recent years towards harmonizing syntactic representations across languages. This includes the HamleDT project (Zeman et al., 2012; Zeman et al., 2014), as well as the Universal Dependencies initiative (Petrov et al., 2012; McDonald et al., 2013).

Our work aims to automatically harmonize the dependency representations resorting to bilingual correspondence, thus can be grouped into the building strategies for harmonized or universal dependencies. These existing annotated treebanks would also permit interesting control experiments, both for the measurement of similarity and for parsing.

7 Conclusion and Future Work

We propose an automatic cross-lingual similarization algorithm for dependency grammars, design an automatic evaluation metric to measure the cross-lingual similarity between grammars, and use the similarized grammars to improve dependency tree-based machine translation. Experiments show the efficacy of this method. The cross-lingual similarization in this paper is still *soft similarization*, it is worth to investigate the *hard similarization*, where the syntactic structures are totally isomorphic between two languages. Of course, in such syntactic structures, the syntactic nodes should be super-node, that is, a graph containing one or more basic syntactic nodes. Hard similarization could be more suitable for cross-lingual applications, and we leave this aspect for future research.

Acknowledgments

The authors are supported by National Natural Science Foundation of China (Contract 61379086 and 61370130). Jiang is also supported by Open-end Fund of the Platform of Research Database and Information Standard of China (No. qhkj2015-01). We sincerely thank the anonymous reviewers for their insightful comments.

References

Vamshi Ambati and Alon Lavie. 2008. Improving syntax driven translation models by re-structuring divergent and non-isomorphic parse tree structures. In *Proceedings of Student Research Workshop of AMTA*.

David Burkett and Dan Klein. 2012. Transforming trees to improve syntactic convergence. In *Proceedings of EMNLP-CNLL*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the ACL*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, pages 201–228.

Koby Crammer, Ofer Dekel, Shai Shalev-Shwartz, and Yoram Singer. 2003. Online passive aggressive algorithms. In *Proceedings of NIPS*.

Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the ACL*.

Jason M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of COLING*, pages 340–345.

Michel Galley, Jonathan Graehl, Kevin Knight, and Daniel Marcu. 2006. Scalable inference and training of context-rich syntactic translation models. In *Proceedings of the COLING-ACL*.

Bryant Huang and Kevin Knight. 2006. Relabeling syntax trees to improve syntax-based machine translation quality. In *Proceedings of NAACL*.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Statistical syntax-directed translation with extended domain of locality. In *Proceedings of the AMTA*.

Rebecca Hwa, Philip Resnik, Amy Weinberg, and Okan Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the ACL*.

Philipp Koehn, Franz Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL*.

Dekang Lin. 2004. A path-based transfer model for machine translation. In *Proceedings of the COLING*.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the ACL*.

Shujie Liu, Chi-Ho Li, Mu Li, and Ming Zhou. 2012. Re-training monolingual parser bilingually for syntactic smt. In *Proceedings of EMNLP-CNLL*.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. In *Computational Linguistics*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of ACL*, pages 91–98.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundagez, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Leez. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*.

Haitao Mi, Liang Huang, and Qun Liu. 2008. Forest-based translation. In *Proceedings of the ACL*.

Makoto Morishita, Koichi Akabe, Yuto Hatakoshi, Graham Neubig, Koichiro Yoshino, and Satoshi Nakamura. 2015. Parser self-training for syntax-based machine translation. In *Proceedings of IWSLT*.

Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? In *Proceedings of EMNLP*.

Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Labeled pseudoprojective dependency parsing with support vector machines. In *Proceedings of CoNLL*, pages 221–225.

Franz Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE TKDE*.

Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the ACL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. *Proceedings of LREC*.

Chris Quirk, Arul Menezes, and Colin Cherry. 2005. Dependency treelet translation: Syntactically informed phrasal smt. In *Proceedings of the ACL*.

Ratnaparkhi and Adwait. 1996. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2008. A new string-to-dependency machine translation algorithm with a target dependency language model. In *Proceedings of ACL*.

David Smith and Jason Eisner. 2009. Parser adaptation and projection with quasi-synchronous grammar features. In *Proceedings of EMNLP*.

Stolcke and Andreas. 2002. Srilm - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 311–318.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Re-structuring, re-labeling, and re-alignment for syntax-based machine translation. *Computational Linguistics*.

Jun Xie, Haitao Mi, and Qun Liu. 2011. A novel dependency-to-string model for statistical machine translation. In *Proceedings of EMNLP*.

Deyi Xiong, Qun Liu, and Shouxun Lin. 2007. A dependency treelet string correspondence model for statistical machine translation. In *Proceedings of Workshop on SMT*.

Nianwen Xue, Fei Xia, Fu-Dong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. In *Natural Language Engineering*.

H Yamada and Y Matsumoto. 2003. Statistical dependency analysis using support vector machines. In *Proceedings of IWPT*.

Daniel Zeman, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Jan Hajič, and Zdeněk Žabokrtský. 2012. Hamledt: To parse or not to parse?

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. Hamledt: Harmonized multi-language dependency treebank. *Language Resources & Evaluation*.