Influence of Part-of-Speech and Phrasal Category Universal Tag-set in Tree-to-Tree Translation Models

Francisco Oliveira, Derek F. Wong, Lidia S. Chao, Liang Tian, Liangye He
Department of Computer and Information Science,
University of Macau, Macao SAR, China
{olifran, derekfw, lidiasc}@umac.mo,
{tianliang0123, wutianshui0515}@gmail.com

Abstract

Tree-to-tree Statistical Machine Translation models require the use of syntactic tree structures of both the source and target side in learning rules to guide the translation process. In order to accomplish the task, available treebanks for different languages are used as the main resources to collect necessary information to handle the translation task. However, since each treebank has its own defined tags, a barrier is inherently created in highlighting alignment relationships at different syntactic levels for different tag-sets. Moreover, these models are typically over constrained. This paper presents a unified tag-set for all languages at Part-of-Speech and Phrasal Category level in tree-to-tree models. Different experiments are conducted to study for its feasibility, efficiency, and translation quality.

1 Introduction

The study of Statistical Machine Translation (SMT) (Lopez, 2008) relying on syntactic information has received wide attention in recent years. In particular, syntactic information is being integrated either on the source or target or both side(s) in training translation models for handling the translation task. In *hierarchical* models (Chiang, 2007) that consider syntactic information (Zollmann and Venugopal, 2006), the input sentence is analyzed and translated by synchronous context free grammars (SCFG) hierarchically with extra linguistic information. In *string-to-tree* SMT models (Galley et al., 2004; Zhang et al., 2011), the output of the translation always follows a grammatical syntax of the target language. In *tree-to-string* SMT models (Liu et al., 2006; Wu et al., 2010), source side syntax is used to generate the translation output. Finally, by considering the syntax of both the source and target languages, *tree-to-tree* SMT models (Zhang et al., 2008; Liu et al., 2009) tend to be the best among the previous models. Basically, all of these models require two extra components: (1) *syntax parsers* (He et al., 2012; Petrov et al., 2006) in obtaining annotated syntax trees for training the models, and (2) *monolingual treebanks* (a detailed list can be found in Petrov et al. (2012)) for training the parsers. Currently, many of them are publicly available through Internet, institutions and data consortiums.

Independently from the method used, although there are many treebanks available, they typically have their own tag-set defined for different languages, ranging from tens to hundreds of tags, which is hard to conduct the research in a multilingual environment. As a consequence, Petrov et al. (2012) developed a universal Part-of-Speech (POS) tag-set for twenty five different languages. However, at phrasal level, disagreements between the languages remain undefined.

This paper presents a study of the application of universal tag-set from POS to phrasal category level in tree-to-tree translation models. In the POS tag level, we basically used the universal tag-set proposed by Petrov et al. (2012) in mapping original tags into universal ones. In order to fulfill the missing relationships at phrasal category level, a mapping work of phrasal tags for Chinese (Zh), English (En), French (Fr), German (De), and Portuguese (Pt) is presented. The main objective is to partially relax syntactic constraints imposed to the original models by having more generalizations in the unified tag-set proposed. With fewer tags defined between languages, fewer syntax rules will be extracted during the training phase, which reduces the computation load, possible rule ambiguities, and increases the translation efficiency. Although we only focus on five languages, extensions to other languages are possible.

International Joint Conference on Natural Language Processing, pages 907–911,
Nagoya, Japan, 14-18 October 2013.

Tag	Chinese	English	French	German	Portuguese
CNP	CLP, NP, QP, UCP	NP, NAC, NX, WHNP, QP	NP	CNP, MPN, NM, NP	np
CVP	VP, VCD, VCP, VNV, VPT, VRD, VSB	VP	VN, VP, VPpart, VPinf	CVP, VP, VZ	x, vp
CAJP	ADJP	ADJP, WHADJP	AP	AA, AP, CAP, MTA	ap, adjp
CAVP	ADVP, DNP, DP, LCP	ADVP, WHADVP, PRT	AdP	AVP, CAVP	advp
CPP	PP	PP, WHPP	PP	CAC, CPP, PP	pp
CS	FRAG, IP	S, SBAR, SBARQ, SINV, SQ, PRN, FRAG, RRC	ROOT, SENT, Ssub, Sint, Srel	CS, PSEUDO, S	fcl, icl, acl, cu, sq
CCONJP	CP	CONJP	*No mapping tag*	*No mapping tag*	*No mapping tag*
CCOP	*No mapping tag*	UCP	CCOP	CCP, CO	*No mapping tag*
CX	LST, PRN	X, INTJ, LST	*No mapping tag*	CH, CVZ, DL, ISU, QL	*No mapping tag*

Table 1: Mappings from original Phrasal Category to Universal tags

This paper is organized as follows. Section 2 gives the mapping details from POS and phrasal category level tags into universal ones. Section 3 presents the application of universal tags in tree-to-tree models. Section 4 details the experiment results conducted. Section 5 introduces related work followed by a conclusion.

2 Universal Tag-set

A two level universal tag-set is defined in the annotation of syntactic trees for different languages. In the first level, a universal POS tag-set (Petrov et al., 2012) is converted for all leave nodes. It consists of twelve different tags, including: *NOUN* (noun), *VERB* (verb), *ADJ* (adjective), *ADV* (adverb), *PRON* (pronoun), *DET* (determiner and article), *ADP* (preposition and postposition), *NUM* (numeral), *CONJ* (conjunction), *PRT* (particle), "." (punctuation marks) and *X* (others). However, some tags proposed in their original work are not considered at this stage. For example, the original tag *NP* in English, which is supposed to be converted into *NOUN* at POS level, is only changed to *CNP* at the phrasal category stage for better differentiating its actual meaning at tree level.

In phrasal category level, nine universal tags are defined for higher level nodes: *CNP* (noun phrase), *CVP* (verb phrase), *CAJP* (adjective phrase), *CAVP* (adverb phrase), *CPP* (preposition phrase), *CS* (sentence/sub-sentence), *CCONJP* (conjunction phrase), *CCOP* (coordinated phrase), and *CX* (others). Corresponding mappings at a phrasal category level for Zh, En, Fr, De, and Pt language are listed in Table 1.

The proposed conversion is carefully designed by studying the actual meaning of the original tags based on previously published work. Although it is common to find out disagreements between tag-sets across different languages due to their inherent characteristics, the objective of this paper is to unify different tags which are used in most of the treebanks at clause level.

3 Rule Extraction Process

The rule extraction process for tree-to-tree models based on universal tag-set is similar to hierarchical phrase-based model (Chiang, 2007), which considers SCFG rules for handling the translation task. The main difference is that rules where there are syntactic labels for non-terminals are extracted. Given a word aligned sentence tree pair $T(f_1^J)$ and $T(e_1^I)$, each rule in the model is a three tuple consisting of variables $ST(f_{j_1}^{j_2})$, $ST(e_{i_1}^{i_2})$, and $\widetilde{A}$ respectively. $ST(f_{j_1}^{j_2})$ is a sub-tree covering the interval span $[j_1, j_2]$ of $T(f_1^J)$; similarly, $ST(e_{i_1}^{i_2})$ denotes the target sub-tree covering the interval span $[i_1, i_2]$ of $T(e_1^I)$; and $\widetilde{A}$ is the alignment between terminals and leaf non-terminals of the two trees, such that $\forall(j, i) \in \widetilde{A} : j_1 \leq j \leq j_2 \leftrightarrow i_1 \leq i \leq i_2$ holds.

The extraction process starts with standard phrase extraction, and for all the phrases found, a rule is created for each instance. Based on this initial rule set, the rest of all possible rules are iden-

tified based on a simple criterion: these phrases should be subsumed by larger pairs in this set. As an example, if there is another rule $<ST(\gamma)$ ||| $ST(\alpha)$ ||| $A'>$ such that the pair (γ, α) includes another sub-phrase ($f_{j_1'}^{j_2'}$, $e_{i_1'}^{i_2'}$), i.e. γ=$\gamma_1 f_{j_1'}^{j_2'} \gamma_2$ and α=$\alpha_1 e_{i_1'}^{i_2'} \alpha_2$, then a new rule $<ST(\gamma_1 X \gamma_2)$ ||| $ST(\alpha_1 X \alpha_2)$ ||| $\hat{A}>$ will be created, where $\hat{A}$ contains alignment information for all the terminals and non-terminals. As syntax information is provided for both sides, for each pair, it must have a node in both trees which subsumes the corresponding string. In other words, non-terminal label checks to their related syntax nodes are necessary in assigning correct tags to all non-terminals in the rules.

[*NP*][*NP*] 在 巴黎 [*VRD*][*VBD*] 。 [*IP*] |||
[*NP*][*NP*] [*VRD*][*VBD*] *in Paris* . [*S*] |||
0-0 1-2 2-3 3-1 4-4 (1)

As an example, in rule (1), the top node of the source tree is [*IP*], the top node of the target tree is [*S*], and both trees have five children. Alignment information between terminals and non-terminals is associated by their numerical positions. It might appear cases in which the source and target node have different tags assigned due to language divergences. As an example, in order to have a valid substitution of [*VRD*][*VBD*], it requires to have a rule in which the source has a *VRD* tag and the target has a *VBD* tag. Thus, for all non-terminals except the top node, it consists of the source and target tag.

Once all the rules are learned from the entire corpus, probability scores are calculated, which are used in the decoding stage. In addition, glue rules are added in allowing combinations of partial translation fragments monotonically.

The proposed mapping from the original into universal tag-set is advantageous in two aspects. Firstly, in some sense, after the conversion is performed, some rules become more generalized and relaxed compared to the original model. As an example, in Chinese tag-set, as verb phrase related tags (*VP*, *VCD*, *VCP*, *VNV*, *VPT*, *VRD*, *VSB*) are all grouped into *CVP*, more coverage in the selection of rules is expected. In particular, suppose that in the original tag-set, "想 一 想" (think) is tagged as *VCD* (verb compounds), while in universal tag-set, it is tagged as *CVP*. In this case, it is obvious that the phrase "想 一 想" (think) can only be associated to rules with *VCD* but not to verb phrases (*VP*), which limits its usage. As a consequence, a wider coverage of rules is available during the decoding process.

Secondly, since many similar tags in the original tag-set are grouped as only one universal tag, many rules will be merged together, resulting in a smaller size compared to the original model.

4 Experiments

The training environment is executed in a server equipped with a Xeon processor at 2.9GHz, with 192G physical memory. All the experiments are carried out in Moses toolkit (Koehn et al., 2007). Different language pairs are considered in the experiments, including Fr-En, De-En, Zh-En, and Zh-Pt. The bilingual data we used for Fr-En and De-En are extracted from Europarl Parliament (version 7), while Zh-En and Zh-Pt parallel information are extracted from online web-sites. All sentences are parsed by Berkeley parser (Petrov et al., 2006) and word-aligned by using GIZA++ based on five iterations of IBM model 1, three for IBM models 3 and 4, and five for HMM alignment (Och and Ney, 2003). We used a 5-gram language model for all the languages based on the SRILM toolkit (Stolcke, 2002).

Different test sets are considered, including: news-test (NT) data (2009, 2010, 2011) for Fr-En and De-En, which are extracted from the international workshop of SMT (WMT) held annually by the ACL's special interest group for MT; test data for Zh-En and Zh-Pt are extracted from online web pages.

We limited the length of the sentences to be less than fifty, and all of them should be valid aligned parse trees for all the training and testing data. For Chinese, a segmentation model (Zhang et al., 2003) is used for detecting word boundaries.

Table 2 shows the translation quality measured in terms of BLEU metric (Papineni et al., 2002) with the original (Ori.) and universal (Uni.) tag-set. When Chinese is considered as the source, results are lower than the ones targeted for European languages, probably affected by the corpus selection, size of the corpus, parsing success rate, non-standard linguistic phenomena (Wong et al., 2012), etc. In particular, we observed that the parsing accuracy (either on the original or universal tag-set) for Chinese language is lower compared

	Fr-En		De-En	
	Ori.	Uni.	Ori.	Uni.
NT 2009	11.57	11.59*	9.64	9.66
NT 2010	10.81	10.84*	10.48	10.55*
NT 2011	12.12	12.15	9.43	9.44
	Zh-En		Zh-Pt	
	Ori.	Uni.	Ori.	Uni.
Test Data	4.79	4.85	3.87	3.88

Table 2: Translation quality comparison

Language Pair	System	VmPeak (KB)	Rule Size
Fr-En	Ori.	1,002,040	1,223,261
	Uni.	982,208	1,190,177
De-En	Ori.	761,108	926,907
	Uni.	745,724	887,317
Zh-En	Ori.	826,032	853,315
	Uni.	812,144	832,099
Zh-Pt	Ori.	686,932	813,405
	Uni.	682,308	804,356

Table 3: Memory usage and rule size

with other languages, which possibly led to poorer alignment relationships at tree level. However, there is an improvement for all the language pairs with different test sets considered by comparing with the baseline approach. Moreover, we measured the improvements over the baseline based on the significant test method proposed by Koehn (2004). The results that are significantly better than the baseline at $p = 0.05$ are shown by *. For NT 2010, the results are totally significant, while others' significance rate is better at a range between 97% and 99.4%.

Table 3 measures the average peak virtual memory (VmPeak) usage, and the actual number of rules generated. It is concluded that there is a decrease of 2% in terms of the peak virtual memory compared to the baseline, and a decrease of 1% to 4% in terms or distinct rules.

In short, although the improvement in terms of the translation quality is not high, it significantly reduces not only the rule table size but also memory requirements, which is very beneficial when larger data are considered.

5 Related Work

Some of earlier work focused in describing alignment relationships in dependency tree-to-tree structures based on synchronous tree mapping grammars (Eisner, 2003), and synchronous dependency insertion grammars (Ding and Palmer, 2005). However, their work is targeted on dependency grammars, which is simpler than CFG equivalent formalisms (Fox, 2002). Other studies reported the use of syntactic information from conventional bilingual parsed trees. Zhang et al. (2008) proposed a tree sequence alignment model for bilingual trees. Liu et al. (2009) considered packed forests instead of 1-best trees for the whole translation process. Although both methods tend to increase rule coverage and to relax the over-constrained problem, they require tailored and sophisticated decoders. Zhai et al. (2011) considered the addition of bilingual phrases and binarization of parse trees to deal with the problems.

In this work, we proposed the substitution of original tags into universal ones, which has a higher level of abstraction in partially increasing the rule coverage while reducing the size of the rule table. Moreover, our approach does not require big changes in tree-to-tree models for accomplishing the translation task.

6 Conclusion

This paper presents the application of universal tag-set defined at the POS and the phrasal category level to tree-to-tree models. A phrasal category tag-set is defined for Chinese, English, French, German, and Portuguese. With the universal tag-set, learned rules become more generalized and compact. Moreover, this could partially relax the over-constrained disadvantage of traditional tree-to-tree models. Based on the experiment results, better accuracy is obtained compared with the baseline (without tag conversion) and better efficiency due to the reduced number of rules in the proposed method. In the future, we intend to further evaluate the proposed strategy for more languages, with proper universal tags defined, and to study their actual relationships in the learned rules in deducing new strategies to further reduce the rule table size.

Acknowledgments

This work is supported by the Research Committee of University of Macau and Science and Technology Development Fund of Macau under the grants UL019B/09-Y3/EEE/LYP01/FST and 057/2009/A2, respectively.

References

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*, 33(2):201–228.

Yuan Ding and Martha Palmer. 2005. Machine translation using probabilistic synchronous dependency insertion grammars. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 541–548.

Jason Eisner. 2003. Learning non-isomorphic tree mappings for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2*, pages 205–208.

Heidi J. Fox. 2002. Phrasal cohesion and statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 304–311.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proceedings of the North American Chapter of the ACL: Human Language Technologies*, pages 273–280.

Liangye He, Derek F. Wong, and Lidia S. Chao. 2012. Adapting multilingual parsing models to sinica treebank. In *Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 211–215.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 177–180.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string alignment template for statistical machine translation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 609–616.

Yang Liu, Yajuan Lü, and Qun Liu. 2009. Improving tree-to-tree translation with packed forests. In *Proceedings of the 4th International Joint Conference on Natural Language Processing*, pages 558–566.

Adam Lopez. 2008. Statistical machine translation. *ACM Computing Survey*, 40(3):1–49.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*, pages 433–440.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2089–2096.

Andreas Stolcke. 2002. Srilm-an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Fai Wong, Francisco Oliveira, and Yiping Li. 2012. Hybrid machine aided translation system based on constraint synchronous grammar and translation corresponding tree. *Journal of Computers*, 7(2):309–316.

Xianchao Wu, Takuya Matsuzaki, and Jun'ichi Tsujii. 2010. Improve syntax-based translation using deep syntactic structures. *Machine Translation*, 24(2):141–157.

Feifei Zhai, Jiajun Zhang, Yu Zhou, and Zong Chengqing. 2011. Simple but effective approaches to improving tree-to-tree model. In *Proceedings of MT Summit XIII*, pages 261–268.

Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. Hhmm-based chinese lexical analyzer ictclas. In *Proceedings of the second SIGHAN workshop on Chinese language processing - Volume 17*, pages 184–187.

Min Zhang, Hongfei Jiang, AiTi Aw, Haizhou Li, Chew Lim Tan, and Sheng Li. 2008. A tree sequence alignment-based tree-to-tree translation model. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 559–567.

Jiajun Zhang, Feifei Zhai, and Chengqing Zong. 2011. Augmenting string-to-tree translation models with fuzzy use of source-side syntax. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 204–215.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 138–141.