Harmonizing word alignments and syntactic structures for extracting phrasal translation equivalents

Dun Deng, Nianwen Xue and Shiman Guo
Computer Science Department
Brandeis University
415 South Street, Waltham, MA 02453
ddeng@brandeis.edu, xuen@brandeis.edu, shim@brandeis.edu

Abstract

Accurate identification of phrasal translation equivalents is critical to both phrase-based and syntax-based machine translation systems. We show that the extraction of many phrasal translation equivalents is made impossible by word alignments done without taking syntactic structures into consideration. To address the problem, we propose a new annotation scheme where word alignment and the alignment of non-terminal nodes (i.e., phrases) are done simultaneously to avoid conflicts between word alignments and syntactic structures. Relying on this new alignment approach, we construct a Hierarchically Aligned Chinese-English Parallel Treebank (HACEPT), and show that all phrasal translation equivalents can be automatically extracted based on the phrase alignments in HACEPT.

1 Introduction

During the past two decades since the emergence of the statistical paradigm of Machine Translation (MT) (Brown et al., 1993), the field of Statistical Machine Translation (SMT) has attained consensus on the need for structural mappings between languages in MT. Accurately identifying structural mappings (i.e., phrasal translation equivalents) is critical to the performance of both phrase-based systems (Koehn, Och, and Marcu, 2003; Och and Ney, 2004) and syntax-based systems (Chiang, 2005; Chiang, 2007; Galley et al., 2004). The fact is that phrasal translation equivalents are identified based on word alignments, so how word alignments are done directly affects the identification of phrasal translation equivalents. As reported by (Zhu, Li, and Xiao, 2015), even one spurious word alignment can prevent some desirable phrasal translation equivalents from being extracted. The unfortunate fact is that spurious word alignments abound in current word-aligned parallel texts used for extracting phrasal translation equivalents. This is because the word alignments in these parallel texts, whether they are induced in an unsupervised manner such as that described by (Och and Ney, 2003) or manually annotated based on existing word alignment standards such as (Li, Ge, and Strassel, 2009) and (Melamed, 1998), are generally done as an independent task without taking syntactic structures into consideration. As a result, conflicts between word alignments and syntactic structures are inevitable, and when such a conflict arises, the extraction of desirable phrasal translation equivalents will be impossible.

To address this shortcoming, we designed a hierarchical alignment scheme in which word-level alignment (namely alignment of terminal nodes) and phrase-level alignments (namely alignment of non-terminals) are done simultaneously in a coordinated manner so that conflicts between word alignments and syntactic structures are avoided. Based on this alignment scheme, we constructed a Hierarchically Aligned Chinese-English Parallel Treebank (HACEPT) which currently has 9,897 sentence pairs. We show that this hierarchically aligned corpus provides a new way to extract hierarchical translation rules and can be used as a training corpus to learn this type of alignments.

The rest of the paper is organized like this: Section 2 shows how phrasal translation equivalents can be

Proceedings of SSST-9, Ninth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 1–9,
Denver, Colorado, June 4, 2015.

made impossible to extract by word alignments done without considering syntactic structures. To avoid the problem, Section 3 introduces our new alignment scheme and how HACEPT is constructed using the scheme. Section 4 shows how hierarchical translation rules can be extracted from the phrase alignments in HACEPT. We also provide statistics about two important aspects of the rules, namely the distributions of terminal and non-terminal nodes in the rules and the number of terminal nodes contained in a single rule. Section 5 discusses some work in the literature that are related to what is discussed in this paper. Section 6 concludes the paper and points out furture work to do.

2 Spurious word alignments impede the extraction of phrase pairs

Spurious word alignments arise in any word alignment practice where the alignment is done as an independent task without taking sysntactic structures into consideration, regardless of whether the alignment is automatically generated by utilizing a word aligner such as the GIZA++ toolkit (Och and Ney, 2003) or manually annotated using alignment standards such as (Li, Ge, and Strassel, 2009) and (Melamed, 1998). (Zhu, Li, and Xiao, 2015) has described how a spurious word alignment in automatically generated word alignments prevents some phrasal translation equivalents from being extracted. In this section, we will show how spurious word alignments in human annotated word alignments make the extraction of phrasal translation equivalents impossible.

Consider the following example quoted from (Li, Ge, and Strassel, 2009), where the relevant word alignment in each sentence/phrase pair is highlighted by underlining. Note that the word alignments are done without taking syntactic structures into consideration, as can be told from the fact that all the underlined aligned multi-word strings do not correspond to a constituent in a Penn TreeBank (Marcus, Santorini, and Marcinkiewicz, 1993) or Chinese TreeBank (Xue et al., 2005) parse tree.

1a. He is visiting Beijing <> 他 正 访问 北京

1b. He has gone to Beijing <> 他 去 北京 了

1c. to quickly and efficiently solve the problem <> 迅速 有效地 解决 问题

1d. Results can be obtained by doing experiments <> 做 实验 可以 得出 结果

1e. We fully agree with the Chinese position that there is only one China in the world <> 我们 完全 同意 中方 的 立场, 世界 上 只有 一 个 中国

Just like the spurious word alignment discussed by (Zhu, Li, and Xiao, 2015), the underlined word alignment in each of the sentence/phrase pair above makes it impossible to extract at least one diserable phrasal translation equivalent. For each of the sentence/phrase pair in (1), (2) lists the phrasal translation equivalents that cannot be extracted due to the word alignment done in that pair:

2a. visiting Beijing <> 访问 北京

2b. gone to Beijing <> 去 北京

2c. solve the problem <> 解决 问题

2d. doing experiments <> 做 实验

2e. the Chinese position <> 中方 的 立场

The reason why the phrasal translation equivalents in (2) cannot be extracted is because a word in a phrase on one side is aligned to a word that is not part of the corresponding phrase on the other side. Take (2c) for instance. The Chinese verb 解决/solve in the phrase 解决 问题 is aligned to both *solve* and *to* in (1c), which is not part of the phrase *solve the problem*. As a result, the phrase pair in (2c) cannot be obtained.

It is not desirable that legitimate phrase pairs such as those in (2) cannot be extracted. To fix the problem, Section 3 proposes a new alignment scheme.

3 Hierarchical alignment and the creation of HACEPT

Hierarchical alignment is a new alignment scheme where both terminal nodes (words) and non-terminal nodes (linguistic phrases) between parallel parse trees are aligned in a coordinated way so that conflicts in the form of redundancies and incompatabilities between word alignments and syntatic structures are avoided. We use this scheme to construct HACEPT with the goal of providing the field of MT with

2

a resource that has human annotated tree-structured mappings for MT training purposes.

The word alignment done in HACEPT differs from the common practice of word alignment in the field (Melamed, 1998; Li, Ge, and Strassel, 2009) in that the requirement that every word in a sentence pair needs to be word-aligned is relaxed. On the word level, we only align words that have an equivalent in terms of lexical meaning and grammatical function. For those words that do not have a translation counterpart, we leave them unaligned at word level and instead the appropriate phrases in which they appear. This strategy makes sure that both redundancies and incompatibilities between word alignments and syntactic structures are avoided. In addition, artificial ambiguities are also eliminated. These points will be illustrated in the discussion of the concrete example in Figure 1 below.

We take the Chinese-English portion of the Parallel Aligned Treebank (PAT) described in (Li et al., 2012) for annotation. Our data have three batches: one batch is weblogs, one batch is postings from online discussion forums and one batch is news wire. The English sentences in the data set are annotated based on the original Penn TreeBank (PTB) annotation stylebook (Bies et al., 1995) as well as its extensions (Warner et al., 2004), while the Chinese sentences in the data set are annotated based on the Chinese TreeBank (CTB) annotation guidelines (Xue and Xia, 1998) and its extensions ZhangandXue12. The PAT has no phrase alignments and the word alignments in it are done under the requirement that all the words in a sentence should be aligned.

Next we discuss our annotation procedure in detail. Our annotators are presented with sentence pairs that come with parallel parse trees. The task of the annotator is to decide, first on the word level and then on the phrase level, if a word or phrase needs to be aligned at all, and if so, to which word or phrase it should be aligned. The decisions about word alignment and phrase alignment are not independent, and must obey well-formedness constraints as outlined in (Tinsley et al., 2007):

a. A non-terminal node can only be aligned once.
b. if Node n_c is aligned to Node n_e, then the descendants of n_c can only be aligned to descendants of n_e.
c. if Node n_c is aligned to Node n_e, then the ancestors of n_c can only be aligned to ancestors of n_e.

This means that once a word alignment is in place, it puts constraints on phrase alignments. A pair of non-terminal nodes (n_c, n_e) cannot be aligned if a word that is a descendant of n_c is aligned to a word that is not a descendant of n_e on the word level.

Let us use the concrete example in Figure 1 to illustrate the annotation process, which is guided by a set of detailed annotation guidelines. On the word level, only those words that are connected with a dashed line are aligned since they have equivalents. Note that the Chinese prenominal modification marker 的 and the existential verb 有/have, and the English determiner *the*, the relative pronoun *who*, the preposition *of*, the expletive subject *it*, the copular verb *is*, the infinitive marker *to* and the conjunction word *both* are all left unaligned on the word level. Aligning these words will generate artificial ambiguous cases and create both redundancies and incompatibilities between word alignments and parse trees.

For instance, if 的 is to be word-aligned, it could be glued to the preceding verb 喋喋不休 and the whole string will be aligned to *harp*. Note that 喋喋不休 and *harp* are both unambiguous and form a one-to-one correspondence. With the word alignment between 喋喋不休 的 and *harp*, we make the unambiguous *harp* correspond to both 喋喋不休 and 喋喋不休 的 (and possibly more strings), thus creating a spurious ambiguity. Also note that the string 喋喋不休 的 does not form a constituent in the Chinese parse tree, so the word alignment is incompatible with the syntactic structure of the sentence. By leaving 的 unaligned, we avoid both the spurious ambiguity and the incompatibility.

As for redundancies, consider the English determiner *the*, which has no translation counterpart in the Chinese sentence. If *the* is to be word-aligned, it could be attached to the noun *people* and the whole string *the people* will be aligned to 人们. This will create a redundancy, since the English parse tree already groups *the* and *people* together to form an NP, and therefore there is no need to repeat this information on the word level by attaching *the* to *people*, especially when the word alignment also generates a spurious ambiguity for 人们, which unambiguously

3

means *people* but is aligned to *the people*.

With word alignments in place, next the annotator needs to perform phrase alignments. Note that word alignments place restrictions on phrase alignments. For instance, VP_{c0} cannot be a possible alignment for VP_{e1}, because 通常, a descendant of VP_{c0}, is aligned to *often*, which is not a descendant of VP_{e1}. For a phrase that does have a possible alignment, the annotator needs to decide whether the possible phrase alignment can be actually made. This is a challenging task since, for a given phrase, there usually are more than one candidate from which a single alignment needs to be picked. For instance, for the English ADJP, there are in total two possible phrase alignments, namely VP_{c6}, and VP_{c7}, both of which obey the well-formedness constraints. Since a non-terminal node is not allowed to be aligned to multiple non-terminal nodes on the other side, the annotator needs to choose one among all the candidates. This highlights the point that the alignment of non-terminal nodes cannot be deterministically inferred from the alignment of terminal nodes. This is especially true given our approach where some terminal nodes are left unaligned on the word level. For instance, the reason why VP_{c7} is a possible alignment for ADJP is because the word 有 is left unaligned. If 有 were aligned with, say, *is*, VP_{c7} could not be aligned with ADJP since *is* is not a descendant of ADJP and aligning the two nodes will violate Constraint *b*.

While Constraints *b* and *c* can be enforced automatically given the word alignments, the decisions regarding the alignment of non-terminal nodes which satisfy Constraint *a* are based on linguistic considerations. One key consideration is to determine which non-terminal nodes encapsulate the grammatical relations signaled by the unaligned words so that the alignment of the non-terminal nodes will effectively capture the unaligned words in their syntactic context. When identifying non-terminal nodes to align, we follow two seemingly conflicting general principles:

- Phrase alignment should not sever key dependencies involving the grammatical relation signaled by an unaligned word.
- Phrase alignment should be minimal, in the sense that the phrase alignment should contain only the elements involved in the grammatical relation, and nothing more.

The first principle ensures that the grammatical relation is properly encapsulated in the aligned non-terminal nodes. For example in Figure 1, if we attach the English preposition *on* to *tolls* and aligning them to 通行费, we would fail to capture the lexical dependency between *harp* and *on*. Aligning VP_{c2} with VP_{e2} captures the dependency.

The first principle in and of itself is insufficient to produce desired alignment. Taken to the extreme, it can be trivially satisfied by aligning the two root nodes of the sentence pair. We also need the alignment to be minimal, in the sense that aligned non-terminal nodes should contain only the elements involved in the grammatical relation, and nothing more. These two requirements used in conjunction ensure that a unique phrase alignment can be found for each unaligned word. The phrase alignments in Figure 1 which are indicated by blue dotted lines, all satisfy these two principles.

Following the principles and the procedure introduced above, we constructued HACEPT,[1] which has 9,897 sentence pairs. In the next section, we show how the alignments in HACEPT can help to extract translation rules.

4 Extracting hierarchical translation rules in HACEPT

Hierarchical translation rules can be automatically extracted from the phrase alignments in HACEPT. Take a pair of aligned non-terminal nodes (n_c, n_e), a translation rule based on the alignment between n_c and n_e can be extracted like this: Check each of the immediate daughter nodes of both n_c and n_e. For any of the daughter nodes that is aligned, stop looking down into the node and keep the phrase category label of the node as a variable the rule. For each daughter node that is not aligned, recursively traverse its children until either an aligned node is found, in which case its phrase category label will be kept as a variable in the rule, or a terminal node is

[1] As of the writing of this paper, we are in the process of doing adjudication on the double annotation done to create HACEPT. We look forward to finishing adjudication soon and releasing the resouce to the public.

4

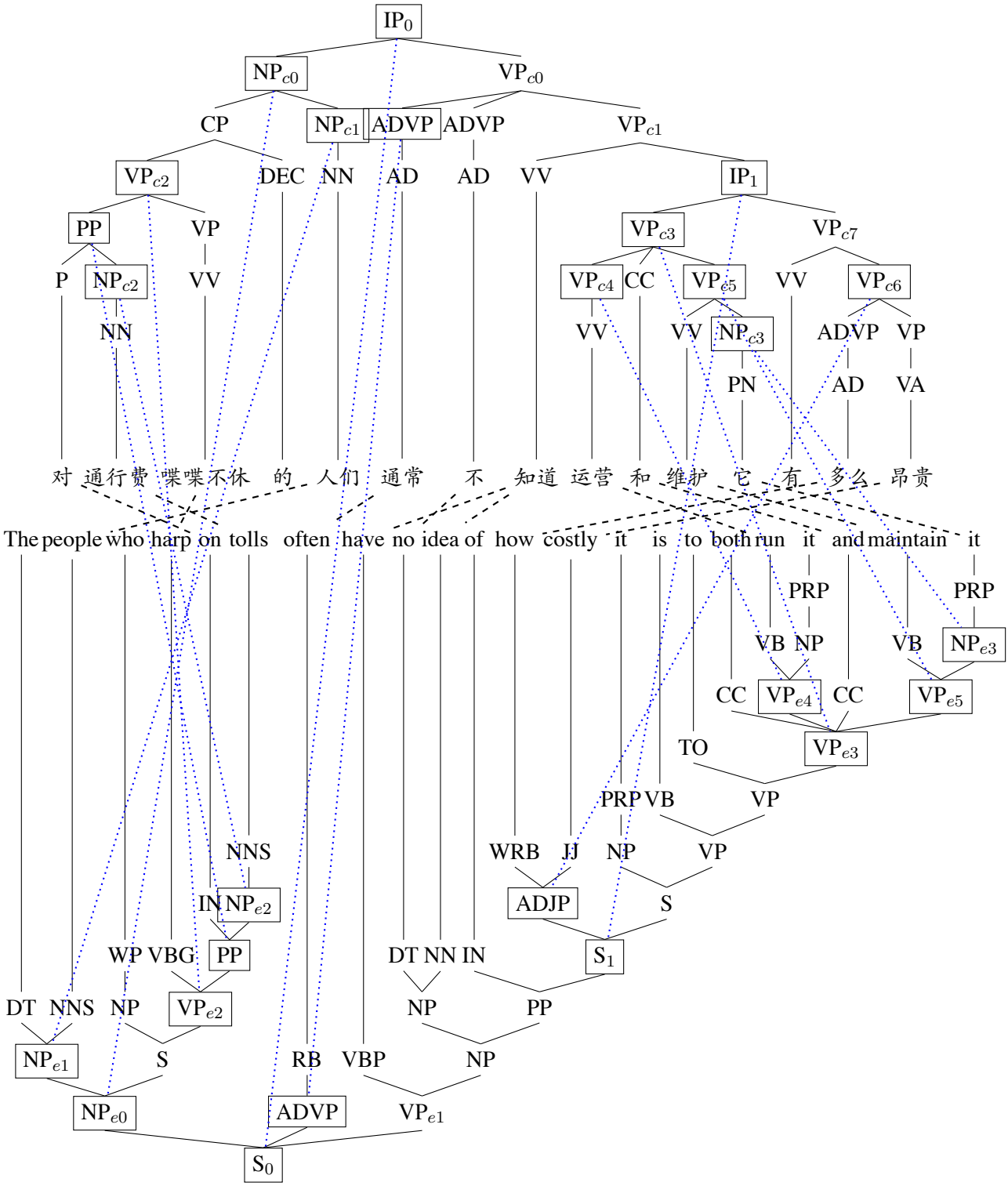

Figure 1: A hierarchically aligned sentence pair

5

reached, in which case the word is included as part of the translation rule.

To illustrate the rule extraction process specified above, let us take the phrase alignment between NP_{c0} and NP_{e0} in Figure 1 for instance. The search starts top-down from the two root nodes. On the Chinese side, NP_{c0} has two immediate daughter nodes: CP and NP_{c1}. NP_{c1} is aligned, so we stop looking inside the node and just keep the phrase category label of the node as part of the rule. CP is not aligned, so we keep checking its two immediate daughter nodes: VP_{c2} and DEC. VP_{c2} is aligned and will not be further checked. DEC is not aligned and dominates the terminal node 的, which will be kept in the rule. Since DEC is the last node inside NP_{c0} and a terminal node is reached, the search on the Chinese side ends. The same procedure will simultaneously take place on the English side, and when the search is done, we will get the translation rule in (3) below:

(3) $NP_{c0} \Leftrightarrow NP_{e0}$:

VP_{c2} 的 NP_{c1} <> NP_{e1} who VP_{e2}

Note that the rule contains both terminals (的 and *who*) and non-terminals represented by phrase category labels.

The rule in (3) illustrates one type of rule, namely the rules containing both terminal and non-terminal nodes. There are also rules with only terminal nodes and rules with only non-terminal nodes. Figure 1 has quite a few examples for the former and an example is given below:

(4) $NP_{c2} \Leftrightarrow NP_{e2}$:

通行费 <> tolls

The rule above contains only terminals. Figure 1 does not contain an example for rules with only non-terminals, but such rules do exist and here is a common example:

(5) IP $\Leftrightarrow$ S:

NP_{subj} VP_{pred} <> NP_{subj} VP_{pred}

The rule above illustrates parallel sentences whose subjects and predicates are both aligned.

Table 1 provides the statistics of the distribution of the three types of rules in HACEPT.

Rule types	No.	Percentage
with only terminals	52379	50.46
with only non-terminals	2621	2.53
with both	48796	47.01
Total	103796	100

Table 1: Rule distribution

Given the importance of hierarchical translation rules for MT, a natural question to ask about the hierarchical translation rules extracted from HACEPT is this: are these rules usable? The most crucial factor deciding the usability of a rule is its length in terms of the number of terminal nodes it contains. If a rule contains too many terminal nodes, it cannot be easily used for MT purposes. Table 2 provides the statistics about the number of terminal node (TN) in the extracted rules:

TN	Rule	Percentage	Cumulative
0	6974	6.72	6.72
1	4017	3.87	10.59
2	30829	29.70	40.29
3	18780	17.09	58.38
4	12897	12.43	70.81
5	9387	9.04	79.85
6	6079	5.86	85.71
7	4404	4.24	89.95
More than 7	10429	10.05	1

Table 2: Rule length

As shown by the table, 89.95 percent of the rules contain 7 or less than 7 terminal nodes. There are still 10 percent of the rules that contain more than 7 terminal nodes.

One primary reason that increases the number of terminal nodes in a rule is how the parse trees are designed. To be specific, some parts of the parse trees are designed to be flat, presumably for the sake of increasing treebank annotation throughput, but this makes some otherwise legitimate phrase alignments inaccessible unless we change the underlying parse trees. When a phrase alignment cannot be made, some terminal nodes will be left out to appear in the rule. This is illustrated by Figure 1.

On the Chinese side, there is a node, namely VP_{c0}, which dominates the predicate part of the sentence.

On the English side, the predicate part of the English sentence is split into ADVP and VP_{e1}, and there is no single node dominating these two nodes. As a result, VP_{c0} has no phrase alignment. Suppose a node VP_{e0} is created that includes ADVP and VP_{e1} as its immediate daughters, VP_{c0} and VP_{e0} could be aligned. (7) below is the rule based on the alignment between the two sentences in Figure 1, and (8) is the rule based on the alignment between the two sentences if a node is created for the predicate of the English sentence and aligned to VP_{c0}.

(6) $IP_0 \Leftrightarrow S_0$

NP_{c0} ADVP 不 知道 IP_1 <> NP_{e0} ADVP have no idea of S_1

(7) $IP_0 \Leftrightarrow S_0$

NP_{c0} VP_{c0} <> NP_{e0} VP_{e0}

($VP_{e0} \Rightarrow$ ADVP VP_{e1})

Note that the rule in (6) has 6 terminal nodes in total whereas the rule in (7) has none. This is a good example to illustrate the fact that a flat structure makes some legitimate phrase alignment impossible and as a result increases the number of terminal nodes in a rule.

There is another place in Figure 1 that has the same problem. Note that the Chinese VP_{c0} has three immediate daughter nodes: ADVP, ADVP, and VP_{c1}. This structure is flat and can become deeper if an intermediate node is created to dominate the second ADVP and VP_{c1}. This node can then combine with the first ADVP to form VP_{c0}. Note that this intermediate node will serve as the phrase alignment of VP_{e1}, which cannot be unaligned in the figure. With the phrase alignment between VP_{e1} and the hypothetical intermediate node (call it VP_{c9}), the number of terminal nodes in (6) will be reduced to zero even without the creation of VP_{e0} in (7). The new rule looks like this:

(8) $IP_0 \Leftrightarrow S_0$

NP_{c0} ADVP VP_{c9} <> NP_{e0} ADVP VP_{e1}

($VP_{c9} \Rightarrow$ ADVP VP_{c1})

In the near future, we plan to binarize the flat structures as illustrated above to create some intermediate nodes, which can be aligned and reduce the number of terminal nodes in existing rules.

5 Related work

To address the problem caused by spurious word alignments, there has been research done to improve word alignment quality by incorporating syntactic information into word alignments (May and Knight, 2007; Fossum, Knight, and Abney, 2008). Another research direction has been explored to conduct syntactic alignment between parse trees (Tinsley et al., 2007; Pauls et al., 2010; Sun, Zhang, and Tan, 2010b; Sun, Zhang, and Tan, 2010a), and implements syntactic rule extraction based on syntactic alignment instead of word alignment. Our work reported in Section 3 can be viewed as a combination of these two lines of research.

There has also been reasearch done to automatically obtain phrasal translation equivalents (Ambati and Lavie, 2008; Hanneman, Burroughs, and Lavie, 2011; Lavie, Parlikar, and Ambati, 2008; Zhu, Li, and Xiao, 2015). This line of research is different from our work in two respects:

First, word alignment as the foundation of phrase-pair extraction is done differently in the two approaches. Automatic extraction of phrase pairs uses automatically generated word alignments, where there are lots of spurious word alignments, which, as pointed out by (Zhu, Li, and Xiao, 2015), are harmful to rule extraction and affect translation quality. By contrast, HACEPT is free of spurious word alignments. As already mentioned in Section 3, all the word alignments in HACEPT are compatible with the syntactic structures and will not block any legitimate phrase alignment.

Second, phrase alignment is inferred from word alignment in automatic approaches. As reported by (Ambati and Lavie, 2008), in places where language-particular function words such as English auxiliary verbs that exist in one language but not the other are involved, there are usually more than one candiate in the language that has the function words for a phrase in the language that does not have a counterpart of the function words. Automatic inference cannot always make the right decision in such situations. We

7

have strict standards for choosing the correct phrase alignment in such cases and as a result, HACEPT can function as a training corpus for automatic approaches.

6 Conclusion

In this paper, we report a resource we have constructed with a novel alignment scheme. The corpus contains both word and phrase alignments and can help extract hierarchical translation rules and train syntax-based MT models. The next step is, of course, to do MT experiments with this resource to see if it indeed helps to improve system performance.

Acknowledgments

This work is supported by the IBM subcontract No. 4913014934 under DARPA Prime Contract No. 0011-12-C-0015 entitled "Broad Operational Language Translation". We would like to thank Libin Shen and Salim Roukos for their inspiration and discussion during early stages of the project, Abe Ittycheriah and Niyu Ge for their help with setting up the data, Loretta Bandera for developing and maintaining the annotation tool, and two anonymous reviewers for their helpful comments. We are grateful for the hard work of our annotators Hui Gao, Tseming Wang and Lingya Zhou. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor or any of the people mentioned above.

References

Ambati, Vamshi and Alon Lavie. 2008. Improving syntax driven translation models by re-structuring divergent and non-isomorphic parse tree structures. In *Proceedings of AMTA-2008 Student Research Workshop*, pages 235--244.

Bies, Ann, Mark Ferguson, Karen Katz, Robert MacIntyre, Victoria Tredinnick, Grace Kim, Mary Ann Marcinkiewicz, and Britta Schasberger. 1995. Bracketing guidelines for treebank ii style penn treebank project. Technical report, University of Pennsylvania.

Brown, Peter F., Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263--311.

Chiang, David. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263--270. Association for Computational Linguistics.

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

Fang, Licheng and Chengqing Zong. 2008. An efficient approach to rule redundancy reduction in hierarchical phrase-based translation. In *Proceedings of NLP-KE '08 International Conference on Natural Language Processing and Knowledge Engineering*, pages 1--6.

Fossum, V., K. Knight, and S. Abney. 2008. Using syntax to improve word alignment precision for syntax-based machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 44--52.

Galley, Michel, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In Susan Dumais, Daniel Marcu, and Salim Roukos, editors, *HLT-NAACL 2004: Main Proceedings*, pages 273--280.

Hanneman, Greg, Michelle Burroughs, and Alon Lavie. 2011. A general-purpose rule extractor for scfg-based machine translation. In *Proceedings of SSST-5, Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 135--144.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 48--54.

Lavie, Alon, Alok Parlikar, and Vamshi Ambati. 2008. Syntax-driven learning of sub-sentential translation equivalents and translation rules from parsed parallel corpora. In *Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 87--95.

Li, Xuansong, Niyu Ge, and Stephanie Strassel. 2009. Tagging guidelines for chinese-english word alignment. Technical report, Linguistic Data Consortium.

Li, Xuansong, Stephanie Strassel, Stephen Grimes, Safa Ismael, Mohamed Maamouri, Ann Bies, and Nianwen Xue. 2012. Parallel aligned treebanks at ldc: New challenges interfacing existing infrastructures. In *Proceedings of LREC-2012*, Istanbul, Turkey.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313--330.

May, J. and K. Knight. 2007. Syntactic re-alignment models for machine translation. In *Proceedings of the*

8

2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP--CoNLL), pages 360--68.

Melamed, I. Dan. 1998. Annotation style guide for the blinker project. Technical report, University of Pennsylvania.

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

Och, Franz Josef and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417--449.

Pauls, A., D. Klein, D. Chiang, and K. Knight. 2010. Unsupervised syntactic alignment with inversion transduction grammars. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT--NAACL)*, pages 118--26.

Sun, J., M. Zhang, and C.L. Tan. 2010a. Discriminative induction of sub-tree alignment using limited labeled data. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 1047--55.

Sun, J., M. Zhang, and C.L. Tan. 2010b. Exploring syntactic structural features for sub-tree alignment using bilingual tree kernels. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 306--15.

Tinsley, John, Ventsislav Zhechev, Mary Hearne, and Andy Way. 2007. Robust language pair-independent subtree alignment. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.

Warner, Colin, Ann Bies, Christine Brisson, and Justin Mott. 2004. Addendum to the penn treebank ii style bracketing guidelines: Biomedical treebank annotation. Technical report, University of Pennsylvania.

Xue, Nianwen and Fei Xia. 1998. The bracketing guidelines for penn chinese treebank project. Technical report, University of Pennsylvania.

Xue, Nianwen, Fei Xia, Fudong Chiou, and Martha Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural Language Engineering*, 11(2):207--238.

Zhu, Jingbo, Qiang Li, and Tong Xiao. 2015. Improving syntactic rule extraction through deleting spurious links with translation span alignment. *Natural Language Engineering*, pages 1--23.

9