Augmenting String-to-Tree and Tree-to-String Translation with Non-Syntactic Phrases

Matthias Huck and **Hieu Hoang** and **Philipp Koehn**
School of Informatics
University of Edinburgh
10 Crichton Street
Edinburgh EH8 9AB, UK
{mhuck,hhoang,pkoehn}@inf.ed.ac.uk

Abstract

We present an effective technique to easily augment GHKM-style syntax-based machine translation systems (Galley et al., 2006) with phrase pairs that do not comply with any syntactic well-formedness constraints. Non-syntactic phrase pairs are distinguished from syntactic ones in order to avoid harming effects. We apply our technique in state-of-the-art string-to-tree and tree-to-string setups. For tree-to-string translation, we furthermore investigate novel approaches for translating with source-syntax GHKM rules in association with input tree constraints and input tree features.

1 Introduction

Syntax-based statistical machine translation systems utilize linguistic information that is obtained by parsing the training data. In *tree-to-string* translation, source-side syntactic tree annotation is employed, while *string-to-tree* translation exploits target-side syntax. The syntactic parse tree annotation constrains phrase extraction to syntactically well-formed phrase pairs: spans of syntactic phrases must match constituents in the parse tree. Standard phrase-based and hierarchical phrase-based statistical machine translation systems, in contrast, allow all phrase pairs that are consistent with the word alignment (Koehn et al., 2003; Chiang, 2005).

A restriction of the phrase inventory to syntactically well-formed phrase pairs entails that possibly valuable information from the training data remains disregarded. While we would expect phrase pairs that are not linguistically motivated to be less reliable, discarding them altogether might be an overly harsh decision. The quality of an inventory of syntactic phrases depends heavily on the tree annotation scheme and the quality of the syntactic parses of the training data. Phrase pairs that do not span constituents in the tree annotation obtained from syntactic parses can provide reasonable alternative segmentations or alternative translation options which prove to be valuable to the decoder.

In this work, we augment the phrase inventories of string-to-tree and tree-to-string translation systems with phrase pairs that are not induced in the syntax-based extraction. We extract continuous phrases that are consistent with the word alignment, without enforcing any constraints with respect to syntactic tree annotation. Non-syntactic phrases are added as rules to the baseline syntactic grammar with a fill-up technique. New rules are only added if their right-hand side does not exist yet. We extend the glue grammar with a special glue rule to allow for application of non-syntactic phrases during decoding. A feature in the log-linear model combination serves to distinguish non-syntactic phrases from syntactic ones. During decoding, the decoder can draw on both syntactic and non-syntactic phrase table entries and produce derivations which resort to both types of phrases. Such derivations yield hypotheses that make use of the alternative segmentations and translation options provided through non-syntactic phrases. The search space is more diverse, and in some cases all hypotheses from purely syntax-based derivations score worse than a translation that applies one or more non-syntactic phrases. We empirically demonstrate that this technique can lead to substantial gains in translation quality.

Our syntactic translation models conform to the GHKM syntax approach as proposed by **G**alley, **H**opkins, **K**night, and **M**arcu (Galley et al., 2004) with composed rules as in (Galley et al., 2006) and (DeNeefe et al., 2007). State-of-the-art GHKM string-to-tree systems have recently shown very competitive performance in public

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 486–498,
Baltimore, Maryland USA, June 26–27, 2014.

evaluation campaigns (Nadejde et al., 2013; Bojar et al., 2013). We apply the GHKM approach not only in a string-to-tree setting as in previous work, but employ it to build tree-to-string systems as well. We conduct tree-to-string translation with text input and additionally adopt translation with tree input and input tree constraints as suggested for hierarchical translation by Hoang and Koehn (2010). We also implement translation with tree input and feature-driven soft tree matching. The effect of augmenting the systems with non-syntactic phrases is evaluated for all variants.

2 Outline

The remainder of the paper is structured as follows: We review some of the basics of syntax-based translation in the next section (Section 3) and sketch the characteristics of our GHKM string-to-tree and tree-to-string translation frameworks.

In Section 4, we describe our technique to augment GHKM-style syntax-based systems with phrase pairs that do not comply with any syntactic well-formedness constraints.

Section 5 contains the empirical part of the paper. We first describe our experimental setup (5.1), followed by a presentation of the translation results (5.2). We also include a few translation examples (5.3) in order to illustrate the differences between the syntax-based baseline systems and the setups augmented with non-syntactic phrases. The empirical part is concluded with a brief discussion (5.4).

In the final part of the paper (Section 6), we give a survey of previous work that has dealt with problems related to overly restrictive syntactic grammars for statistical machine translation, inadequate syntactic parses, and insufficient coverage of syntactic phrase inventories. A broad spectrum of diverse methods has been proposed in the literature, many of which are quite dissimilar from ours but nevertheless related. We conclude the paper in Section 7.

3 Syntax-based Translation

In syntax-based translation, a probabilistic synchronous context-free grammar (SCFG) is induced from bilingual training corpora. The parallel training data is word-aligned and annotated with syntactic parses on either target side (string-to-tree), source side (tree-to-string), or both (tree-to-tree). A syntactic phrase extraction procedure extracts rules which are consistent with the word-alignment and conform with certain syntactic validity constraints.

Extracted rules are of the form $A, B \rightarrow \langle \alpha, \beta, \sim \rangle$. The right-hand side of the rule $\langle \alpha, \beta \rangle$ is a bilingual phrase pair that may contain non-terminal symbols, i.e. $\alpha \in (V_F \cup N_F)^+$ and $\beta \in (V_E \cup N_E)^+$, where V_F and V_E denote the source and target terminal vocabulary, and N_F and N_E denote the source and target non-terminal vocabulary, respectively. The non-terminals on the source side and on the target side of rules are linked in a one-to-one correspondence. The $\sim$ relation defines this one-to-one correspondence. The left-hand side of the rule is a pair of source and target non-terminals, $A \in N_F$ and $B \in N_E$.

Decoding is typically carried out with a parsing-based algorithm, in our case a customized version of CYK+ (Chappelier and Rajman, 1998). The parsing algorithm is extended to handle translation candidates and to incorporate language model scores via cube pruning (Chiang, 2007).

3.1 GHKM String-to-Tree Translation

In GHKM string-to-tree translation (Galley et al., 2004; Galley et al., 2006; DeNeefe et al., 2007), rules are extracted from training instances which consist of a source sentence, a target sentence along with its constituent parse tree, and a word alignment matrix. This tuple is interpreted as a directed graph (the *alignment graph*), with edges pointing away from the root of the tree, and word alignment links being edges as well. A set of nodes (the *frontier set*) is determined that contains only nodes with non-overlapping closure of their spans.[1] By computing *frontier graph fragments*—fragments of the alignment graph such that their root and all sinks are in the frontier set—the GHKM extractor is able to induce a minimal set of rules which explain the training instance. The internal tree structure can be discarded to obtain flat SCFG rules. Minimal rules can be assembled to build larger *composed rules*.

Non-terminals on target sides of string-to-tree rules are syntactified. The target non-terminal vocabulary of the SCFG contains the set of labels of the frontier nodes, which is in turn a subset

[1] The *span* of a node in the alignment graph is defined as the set of source-side words that are reachable from this node. The *closure* of a span is the smallest interval of source sentence positions that covers the span.

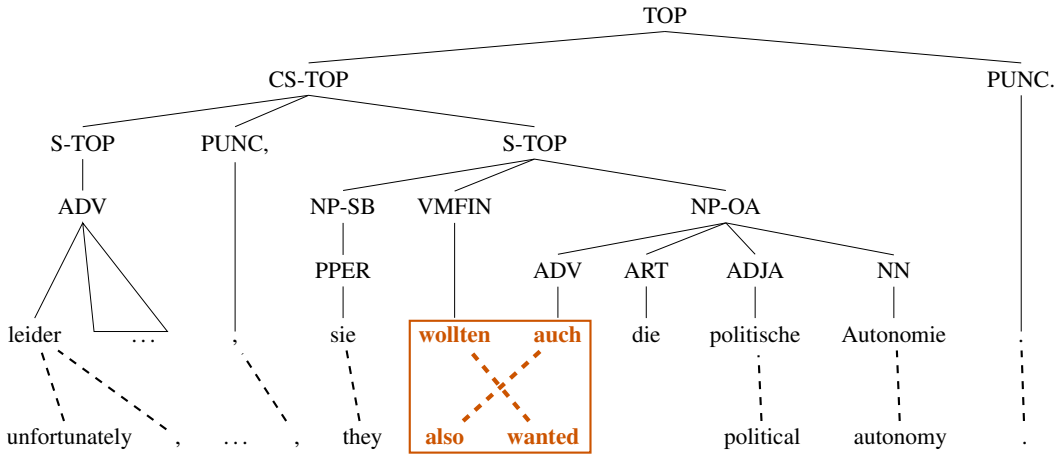


Figure 1: Word-aligned training sentence pair with target-side syntactic annotation.

of (or equal to) the set of constituent labels in the parse tree. It furthermore contains an initial non-terminal symbol Q. Source sides of the rules are not decorated with syntactic annotation. The source non-terminal vocabulary contains a single generic non-terminal symbol X.

In addition to the extracted grammar, the translation system makes use of a special *glue grammar* with an *initial rule*, *glue rules*, a *final rule*, and *top rules*. The glue rules provide a fall back method to just monotonically concatenate partial derivations during decoding. As we add tokens which mark the sentence start ("<s>") and the sentence end ("</s>"), the rules in the glue grammar are of the following form:

Initial rule:

$$X, Q \rightarrow \langle \text{<s>}\ X^{\sim 0}, \text{<s>}\ Q^{\sim 0} \rangle$$

Glue rules:

$$X, Q \rightarrow \langle X^{\sim 0} X^{\sim 1}, Q^{\sim 0} B^{\sim 1} \rangle$$

for all $B \in N_E$

Final rule:

$$X, Q \rightarrow \langle X^{\sim 0}\ \text{</s>}, Q^{\sim 0}\ \text{</s>} \rangle$$

Top rules:

$$X, Q \rightarrow \langle \text{<s>}\ X^{\sim 0}\ \text{</s>}, \text{<s>}\ B^{\sim 0}\ \text{</s>} \rangle$$

for all $B \in N_E$

3.2 GHKM Tree-to-String Translation

The described techniques for GHKM string-to-tree translation can be adjusted for tree-to-string translation in a straightforward manner. Rules are extracted from training instances which consist of a source sentence along with its constituent parse tree, a target sentence, and a word alignment matrix. We omit the details.

For GHKM tree-to-string translation, we investigate three decoding variants:

Tree-to-string translation with text input. The decoder can construct any source-side syntactic analysis that the grammar permits, very similar to string-to-tree translation.

Tree-to-string translation with tree input and input tree constraints. Syntactic annotation over the input data is provided to the decoder. The source-side syntactic non-terminals of a tree-to-string translation rule need to match the constituent span in the input sentence, otherwise the rule cannot be applied. This variant follows the method that was suggested for hierarchical translation by Hoang and Koehn (2010).

Tree-to-string translation with tree input and input tree features. Syntactic annotation over the input data is provided to the decoder. No hard matching constraints are imposed, but the decoder is informed about matches and mismatches of the syntactic annotation in the rules and in the input tree. It takes them into account for the score computation.

4 Non-Syntactic Phrases for GHKM Translation

The syntactic constraints in GHKM extraction can unfortunately prevent useful phrase pairs from being included in the phrase inventory. Consider the example in Figure 1: the highlighted phrase pair $\langle \textit{also wanted}, \textit{wollten auch} \rangle$ cannot be extracted from this training instance for string-to-tree translation.

In the standard phrase-based approach, in contrast, all continuous phrases that are consistent with the word alignment are extracted (Och et al., 1999; Och, 2002). The set of continuous bilingual phrases $\mathscr{BP}(f_1^J, e_1^I, A)$, given a training instance comprising a source sentence f_1^J, a target sentence e_1^I, and a word alignment $A \subseteq \{1,...,I\} \times \{1,...,J\}$, is defined as follows:

$$\mathscr{BP}(f_1^J, e_1^I, A) =$$
$$\Big\{ \langle f_{j_1}^{j_2}, e_{i_1}^{i_2} \rangle : \exists (i,j) \in A : i_1 \leq i \leq i_2 \wedge j_1 \leq j \leq j_2$$
$$\wedge \forall (i,j) \in A : i_1 \leq i \leq i_2 \leftrightarrow j_1 \leq j \leq j_2 \Big\}$$

Consistency for continuous phrases is based upon merely two constraints in this definition: (1.) At least one source and target position within the phrase must be aligned, and (2.) words from inside the source phrase may only be aligned to words from inside the target phrase and vice versa. The highlighted phrase pair from the example does not violate these constraints.

In order to augment our GHKM syntax-based systems with non-syntactic phrases, we obey the following procedure:

- The set $\mathscr{BP}$ is extracted from all training instances, and phrase translation probabilities are computed separately from those in the syntactic phrase inventory.
- Non-syntactic phrases are converted to rules by providing a special left-hand side non-terminal X.
- A phrase table fill-up method is applied to enhance the syntactic phrase inventory with entries from the non-syntactic phrase inventory. Non-syntactic rules are only added to the final grammar if no syntactic rule with the same (source *and* target) right-hand side is present. This method is inspired by previous work in domain adaptation (Bisazza et al., 2011).
- The glue grammar is extended with a new glue rule

  $$X, Q \rightarrow \langle X^{\sim 0} X^{\sim 1}, Q^{\sim 0} X^{\sim 1} \rangle$$

  that enables the system to make use of non-syntactic rules in decoding.
- A binary feature is added to the log-linear model (Och and Ney, 2002) to distinguish non-syntactic rules from syntactic ones, and to be able to assign a tuned weight to the non-syntactic part of the grammar.

5 Empirical Evaluation

We evaluate the effect of augmenting GHKM syntax-based translation systems—both string-to-tree and tree-to-string—with non-syntactic phrase pairs on the English→German language pair using the standard newstest sets of the Workshop on Statistical Machine Translation (WMT) for testing.[2] The experiments are conducted with the open-source *Moses* implementations of GHKM rule extraction (Williams and Koehn, 2012) and decoding with CYK+ parsing and cube pruning (Hoang et al., 2009).

5.1 Experimental Setup

We work with an English–German parallel training corpus of around 4.5 M sentence pairs (after corpus cleaning). The parallel data originates from three different sources which have been eligible for the constrained track of the ACL 2014 Ninth Workshop on Statistical Machine Translation shared translation task: Europarl (Koehn, 2005), News Commentary, and the Common Crawl corpus as provided on the WMT website. Word alignments are created by aligning the data in both directions with MGIZA++ (Gao and Vogel, 2008) and symmetrizing the two trained alignments (Och and Ney, 2003; Koehn et al., 2003). For string-to-tree translation, we parse the German target side with BitPar (Schmid, 2004).[3] For tree-to-string translation, we parse the English source side of the parallel data with the English Berkeley Parser (Petrov et al., 2006).

When extracting syntactic phrases, we impose several restrictions for composed rules, in particular a maximum number of twenty tree nodes per rule, a maximum depth of five, and a maximum size of five. We discard rules with non-terminals on their right-hand side if they are singletons in the training data.

Only the 100 best translation options per distinct source side with respect to the weighted phrase-level model scores are loaded by the decoder. The decoder is configured with a maximum chart span of 25 and a rule limit of 100.

A standard set of models is used in the baselines, comprising phrase translation probabilities and lexical translation probabilities in both direc-

[2] http://www.statmt.org/wmt14/translation-task.html

[3] We remove grammatical case and function information from the annotation obtained with BitPar.

system	dev		newstest2013		newstest2014	
	BLEU	TER	BLEU	TER	BLEU	TER
phrase-based	33.0	48.8	18.8	64.5	18.2	66.9
+ lexicalized reordering	34.2	48.1	19.2	64.5	18.3	67.1
string-to-string (syntax-directed extraction)	32.6	49.4	18.2	65.4	17.8	68.0
+ non-syntactic phrases	33.4	49.0	18.7 (+0.5)	65.0 (−0.4)	18.3 (+0.5)	67.6 (−0.4)
string-to-tree	33.6	48.7	19.5	63.9	18.6	66.9
+ non-syntactic phrases	34.3	48.0	19.8 (+0.3)	63.6 (−0.3)	19.1 (+0.5)	66.2 (−0.7)
tree-to-string	34.0	48.5	19.5	63.8	18.5	67.0
+ non-syntactic phrases	33.9	48.4	19.3 (−0.2)	64.0 (+0.2)	18.7 (+0.2)	66.6 (−0.4)
+ input tree constraints	33.7	48.4	19.3	63.9	18.3	67.0
+ non-syntactic phrases	34.2	48.2	19.7 (+0.4)	63.6 (−0.3)	18.7 (+0.3)	66.5 (−0.5)
+ input tree features	34.3	48.3	19.6	63.7	18.6	67.0
+ non-syntactic phrases	34.4	48.1	19.9 (+0.3)	63.4 (−0.3)	18.8 (+0.2)	66.5 (−0.5)

Table 1: English→German experimental results (truecase). BLEU scores are given in percentage.

tions, word and phrase penalty, an n-gram language model, a rule rareness penalty, and the monolingual PCFG probability of the tree fragment from which the rule was extracted (Williams et al., 2014). Phrase translation probabilities are smoothed via Good-Turing smoothing.

The language model (LM) is a large interpolated 5-gram LM with modified Kneser-Ney smoothing (Kneser and Ney, 1995; Chen and Goodman, 1998). The target side of the parallel corpus and the monolingual German News Crawl corpora are employed as training data. We use the SRILM toolkit (Stolcke, 2002) to train the LM and rely on KenLM (Heafield, 2011) for language model scoring during decoding.

Model weights are optimized to maximize BLEU (Papineni et al., 2002) with batch MIRA (Cherry and Foster, 2012) on 1000-best lists. We selected 2000 sentences from the newstest2008-2012 sets as a development set. The selected sentences obtained high sentence-level BLEU scores when being translated with a baseline phrase-based system, and do each contain less than 30 words for more rapid tuning. newstest2013 and newstest2014 are used as unseen test sets. Translation quality is measured in truecase with BLEU and TER (Snover et al., 2006).[4]

We apply a phrase length limit of five when extracting non-syntactic phrases for the fill-up of syntactic phrase tables.

[4]TER scores are computed with `tercom` version 0.7.25 and parameters `-N -s`.

5.2 Translation Results

Table 1 comprises the results of our empirical evaluation of the translation quality achieved by the different systems.

5.2.1 Phrase-based Baselines

We set up two phrase-based baselines for comparison. Their set of models is the same as for the syntax-based baselines, with the exception of the PCFG probability. One of the phrase-based systems moreover utilizes a lexicalized reordering model (Galley and Manning, 2008). No non-standard advanced features (like an operation sequence model or class-based LMs) are engrafted. The maximum phrase length is five, search is carried out with cube pruning at a k-best limit of 1000. A maximum number of 100 translation options per source side are taken into account.

5.2.2 String-to-String Contrastive System

A further contrastive experiment is done with a string-to-string system. The extraction method for this string-to-string system is GHKM syntax-directed with syntactic target-side annotation from BitPar, as in the string-to-tree setup. We actually extract the same rules but strip off the syntactic labels. The final grammar contains rules with a single generic non-terminal instead of syntactic ones. Note that a side effect of this is that the phrase inventory of the string-to-string system contains

a larger amount of hierarchical phrases[5] than the string-to-tree system, though the same rules are extracted. The reason is that we discard singleton hierarchical rules when we normalize the frequencies after extraction. Many rules that are singletons when the syntax decoration is taken into account have in fact been seen multiple times if syntactic labels are not distinguished, due to pooling of counts.

The string-to-string system is on newstest2013 1.0 points BLEU worse than the phrase-based system with lexicalized reordering and on newstest2014 0.5 points BLEU. We gain 0.5 points BLEU on both of the test sets if we augment the string-to-string system with non-syntactic phrases from the standard phrase-based extractor according to our procedure from Section 4.

5.2.3 String-to-Tree System

The translation quality of the string-to-tree system surpasses the translation quality of the better phrase-based baseline slightly (by 0.3 points BLEU on both test sets). The string-to-tree system is clearly superior to the string-to-string system, which verifies that syntactic non-terminals are indeed vital. We get a nice gain of 0.5 points BLEU and 0.7 points TER on newstest2014 if we augment the string-to-tree system with non-syntactic phrases. The phrase-based system is outperformed by 0.8 points BLEU.

5.2.4 Tree-to-String Systems

The tree-to-string baseline with text input performs at the level of the string-to-tree baseline, but augmenting it with non-syntactic phrases yields only a small improvement or even harms a little (on newstest2013).

Decoding with tree input and input tree constraints causes a minor loss in translation quality. We however observed a decoding speed-up. If we employ non-syntactic phrases to augment the tree-to-string setup with input tree constraints, we provide the new non-syntactic rules in the grammar with a particular property: their left-hand side non-terminal *X* can match any constituent span in the input sentence. The decoder would not be able to utilize non-syntactic phrases without this relaxation. Syntactic phrases amount to an increase of up to 0.4 points BLEU (newstest2013) and 0.5 points TER (newstest2014) in the tree-constrained setup.

Our best tree-to-string setup takes tree input, but involves soft matching features instead of hard input tree constraints. We incorporate two features, one that fires for matches and another one that fires for mismatches. The motivation for not relying on just one feature which would penalize mismatches is that the number of syntactic non-terminals in the derivation can differ between hypotheses. Not all constituent spans need to be matched (or mismatched) by non-terminals, some can be overlaid through larger rules.[6] Tree-to-string translation with input tree features benefits from being augmented with non-syntactic phrases by 0.2 to 0.3 points BLEU. The resulting system is minimally better than the best string-to-tree system on newstest2013, and slightly worse than it on newstest2014.

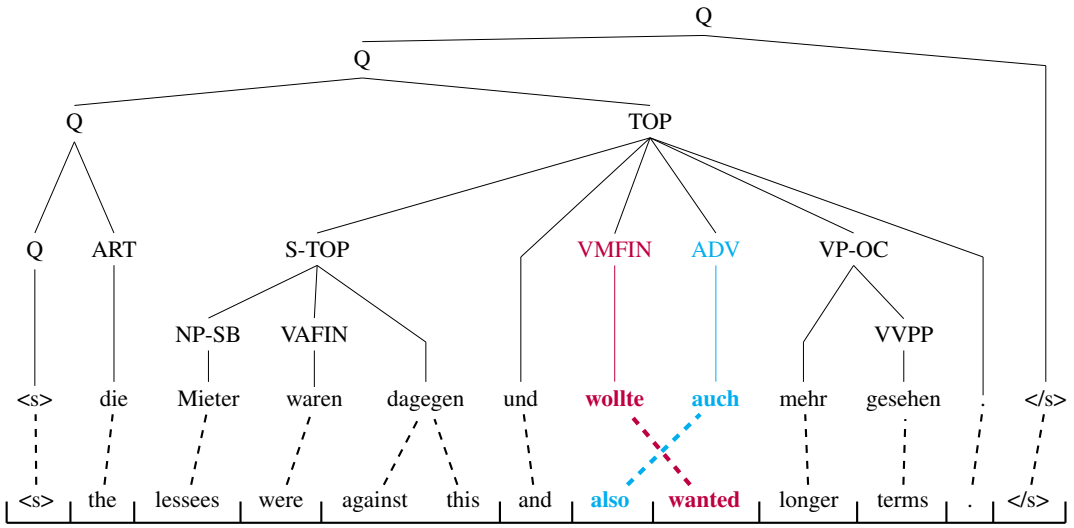
5.3 Translation Examples

We illustrate the differences between the syntax-based baseline systems and the setups augmented with non-syntactic phrases by means of two translation examples from newstest2014. Both examples are string-to-tree translations.

Figures 2 and 3 depict an example that corresponds well to the word-aligned training sentence pair with target-side syntactic annotation from Figure 1. Figure 2 shows the translation, segmentation, and parse tree derived by the string-to-tree baseline system as single-best output for the preprocessed input sentence: "*the lessees were against this and also wanted longer terms .*" The reference translation is: "*Die Pächter waren dagegen und wollten zudem längere Laufzeiten.*" Figure 3 shows the translation, segmentation, and parse tree derived by the string-to-tree system augmented with non-syntactic phrases. There are two word substitutions with respect to the reference in the latter translation, but they convey the same meaning. The baseline translation fails to convey the meaning, mostly because "*terms*" is translated to the verb "*gesehen*", which is a wrong syntactic analysis in the given context. Interestingly, the segmentation applied by the two systems is rather similar, apart from the interval "*also wanted*" which cannot be translated en bloc by the baseline. All rules in the baseline gram-

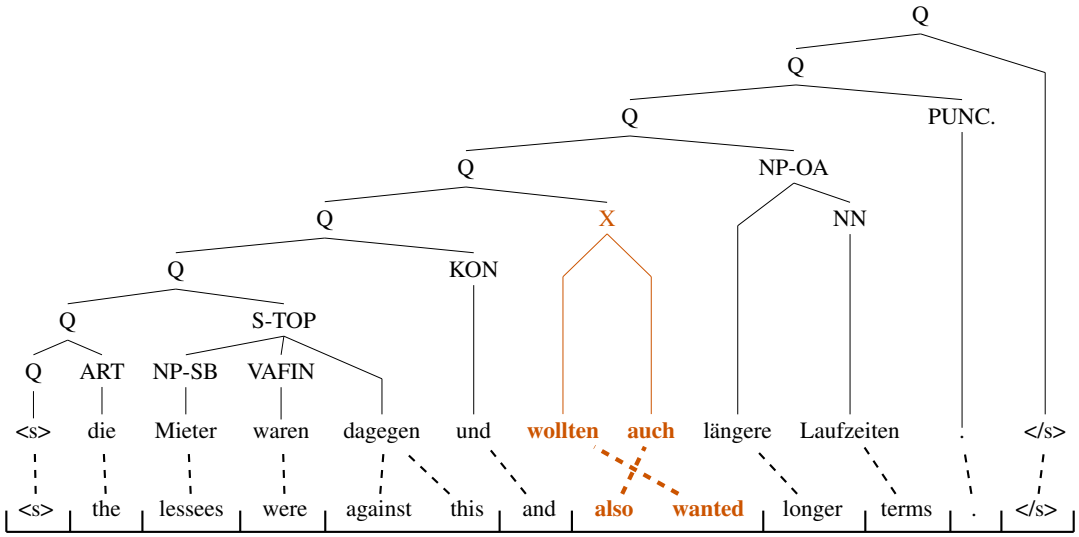
[5] We define *hierarchical phrases* as rules with non-terminals on their right-hand side, in contrast to *lexical phrases* which are continuous rules with right-hand sides that contain terminal symbols only.

[6] Also remember that we discarded the internal tree structure to obtain flat SCFG rules.



Reference: *Die Pächter waren dagegen und wollten zudem längere Laufzeiten.*

Figure 2: Translation and parse tree from the string-to-tree system.

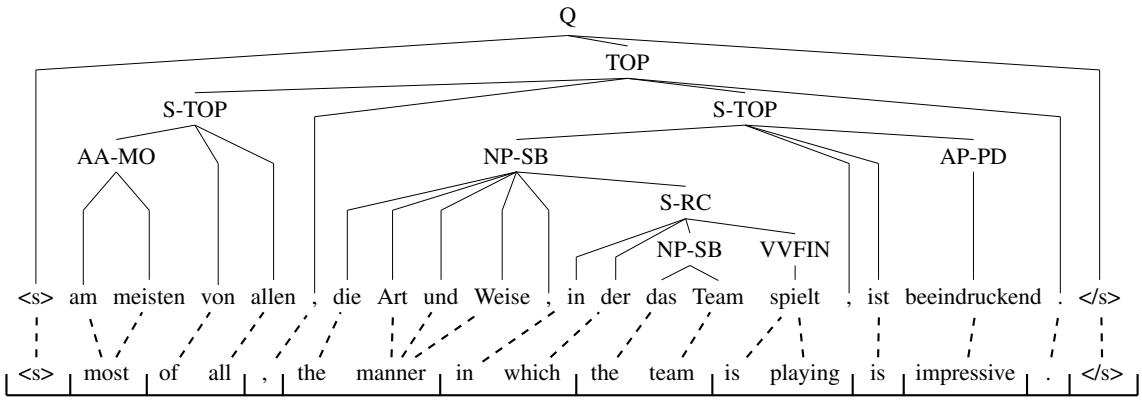


Reference: *Die Pächter waren dagegen und wollten zudem längere Laufzeiten.*

Figure 3: Translation and parse tree from the string-to-tree system augmented with non-syntactic phrases.

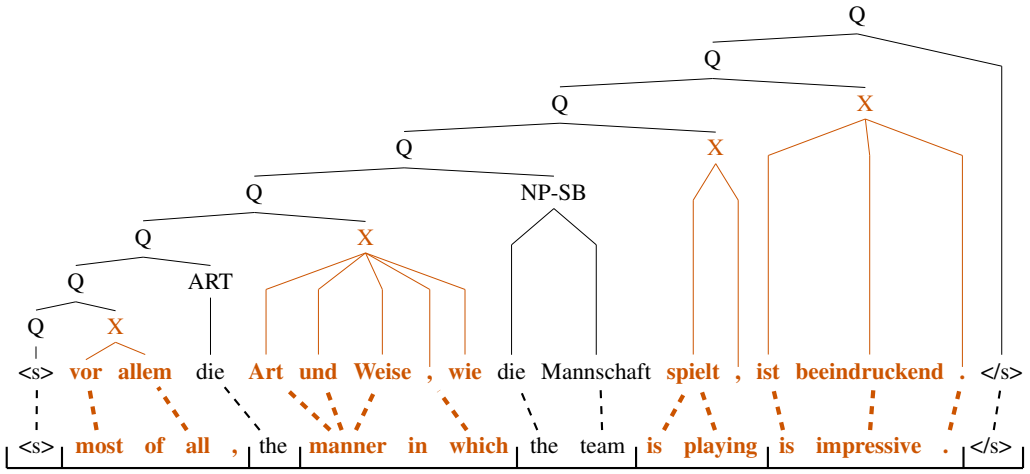
mar that contain "*also wanted*" as part of their source side imply a larger source-side lexical context that is not present in the given sentence. None of those rules matches the input. The baseline has to translate "*also*" and "*wanted*" separately and fails to translate the verb to a plural form German verb. The next rule in bottom-up order is already involved in the incorrect choice of a verb for "*terms*". The string-to-tree system augmented with non-syntactic phrases applies more glue rules, but this is beneficial in the present example, as it breaks apart the faulty syntactic derivation.

Figures 4 and 5 depict a second example. Compared to the baseline, filling up the phrase table with non-syntactic phrases had the effect of disassembling the originally nicely built syntactic tree structure over the translation nearly completely. Four non-syntactic phrases are applied, three of them span over target-side punctuation marks. The baseline translation is more literal and conveys the meaning, but the system augmented with non-syntactic phrases produces a more fluent output. Its translation seems more natural and happens to match the reference in this case.



Reference: *Vor allem die Art und Weise, wie die Mannschaft spielt, ist beeindruckend.*

Figure 4: Translation and parse tree from the string-to-tree system.



Reference: *Vor allem die Art und Weise, wie die Mannschaft spielt, ist beeindruckend.*

Figure 5: Translation and parse tree from the string-to-tree system augmented with non-syntactic phrases.

phrase table entries	**unfiltered**		**dev**		**newstest2013**		**newstest2014**	
	hier.	lexical	hier.	lexical	hier.	lexical	hier.	lexical
phrase-based	–	184.9 M	–	25.3 M	–	29.0 M	–	28.0 M
string-to-string	58.3 M	19.9 M	4.3 M	2.9 M	5.7 M	3.3 M	5.3 M	3.3 M
+ non-syntactic phrases	58.3 M	191.1 M	4.3. M	25.4 M	5.7 M	29.1 M	5.3 M	28.1 M
string-to-tree	39.7 M	21.2 M	4.9 M	3.4 M	5.7 M	3.8 M	5.5 M	3.7 M
+ non-syntactic phrases	39.7 M	192.4 M	4.9 M	25.8 M	5.7 M	29.6 M	5.5 M	28.6 M
tree-to-string	29.5 M	21.1 M	7.7 M	2.8 M	9.0 M	3.3 M	8.7 M	3.2 M
+ non-syntactic phrases	29.5 M	192.6 M	7.7 M	26.1 M	9.0 M	29.9 M	8.7 M	28.9 M

Table 2: Phrase inventory statistics for the different English→German translation systems. "*hier.*" denotes hierarchical phrases, i.e. rules with non-terminals on their right-hand side, "*lexical*" denotes continuous phrases.

5.4 Discussion

A drawback of our method is that it increases the size of the synchronous context-free grammar massively. Most phrase pairs from standard phrase-based extraction are actually not present in the GHKM rule set, even with composed rules. A large fraction of the extracted non-syntactic phrases is such added to the phrase inventory through phrase table fill-up. Table 2 shows the phrase inventory statistics for the different systems.

Another question relates to the glue rule applications. The application of a non-syntactic rule is always accompanied with a respective glue rule application in our implementation. The string-to-tree baseline utilizes glue rules on average 3.0 times in each single-best translation (measured on newstest2014), the string-to-tree system augmented with non-syntactic phrases utilizes glue rules on average 7.0 times. We considered an implementation that allows for embedding of non-syntactic rules into hierarchical rules (other than the glue rules) but did not see improvements with it as yet. Furthermore, efficiency concerns become more relevant in such an implementation.

6 Related Work

Issues with overly restrictive syntactic grammars for statistical machine translation, inadequate syntactic parses, and insufficient coverage have been tackled from several different directions in the literature.

A proposed approach to attain better syntactic phrase inventories is to restructure the syntactic parse trees in a preprocessing step (Wang et al., 2007; Wang et al., 2010; Burkett and Klein, 2012). This line of research aims at rearranging parse trees in a way that makes them a better fit for the requirements of the bilingual downstream application. Conversely, Fossum et al. (2008) retain the structure of the parse trees and modify the word alignments.

Marcu et al. (2006) relax syntactic phrase extraction constraints in their SPMT Model 2 to allow for phrases that do not match the span of one single constituent in the parse tree. SPMT Model 2 rules are created from spans that are consistent with the word alignment and covered by multiple constituents such that the union of the constituents matches the span. Pseudo non-syntactic non-terminals are introduced for the left-hand sides of SPMT Model 2 rules. Special additional rules allow for combination of those non-syntactic left-hand side non-terminals with genuine syntactic non-terminals on the right-hand sides of other rules during decoding.

Another line of research took the hierarchical phrase-based model (Chiang, 2005; Chiang, 2007) as a starting point and extended it with syntactic enhancements. In their SAMT system, Zollmann and Venugopal (2006) labeled the non-terminals of the hierarchical model with composite symbols derived from the syntactic tree annotation. Similar methods have been applied with CCG labels (Almaghout et al., 2012). Venugopal et al. (2009) and Stein et al. (2010) keep the grammar of the non-terminals of the hierarchical model unlabeled and apply the syntactic information in a separate model. Other authors added features which fire for phrases complying with certain syntactic properties while retaining all phrase pairs of the hierarchical model (Marton and Resnik, 2008; Vilar et al., 2008).

In a tree-to-tree translation setting, Chiang (2010) proposed techniques to soften the syntactic constraints. A fuzzy approach with complex non-terminal symbols as in SAMT is employed to overcome the limitations during phrase extraction. In decoding, substitutions of non-terminals are not restricted to matching ones. Any left-hand side non-terminal can substitute any right-hand side non-terminal. The decoder decides on the best derivation based on the tuned weights of a large number of binary features.

Joining phrase inventories that come from multiple origins is a common method in domain adaptation (Bertoldi and Federico, 2009; Niehues and Waibel, 2012) but has also been applied in the contexts of lightly-supervised training (Schwenk, 2008; Huck et al., 2011) and of forced alignment training (Wuebker et al., 2010). For our purposes, we apply a fill-up method in the manner of the one that has been shown to perform well for domain adaptation in earlier work (Bisazza et al., 2011).

Previous research that resembles our work most has been presented by Liu et al. (2006) and by Hanneman and Lavie (2009).

Liu et al. (2006) allow for application of non-syntactic phrase pairs in their tree-to-string alignment template (TAT) system. The translation probabilities for the non-syntactic phrases are obtained from a standard phrase-based extraction

pipeline. A non-syntactic phrase pair can however only be applied if its source side matches a subtree in the parsed input sentence. Syntactic and non-syntactic phrases are not distinguished, and overlap between the syntactic and non-syntactic part of the phrase inventory is not avoided. The decoder picks the entry with the higher phrase translation probability, which means that non-syntactic phrase table entries can supersede syntactic entries. The authors report improvements of 0.6 points BLEU on the 2005 NIST Chinese→English task with four reference translations.

Hanneman and Lavie (2009) examine non-syntactic phrases for tree-to-tree translation with the Stat-XFER framework as developed at Carnegie Mellon University (Lavie, 2008). They combine syntactic and non-syntactic phrase inventories and reestimate the probabilities for both types of phrase pairs by adding up the observed absolute frequencies. Two combination schemes are evaluated: combination with all extractable valid non-syntactic phrases ("direct combination") and combination with only those non-syntactic phrases whose source sides are not equal to the source side of any syntactic phrase ("syntax-prioritized combination"). On a French→English translation task, Hanneman and Lavie (2009) report improvements of around 2.6 points BLEU by adding non-syntactic phrases on top of their Stat-XFER syntactic baselines. Their best setup however does not reach the performance of a standard phrase-based system, which is still 1.6 points BLEU better.

Apart from the differences in the underlying syntax-based translation technology (string-to-tree/tree-to-string GHKM vs. TAT vs. Stat-XFER), our work also constitutes a novel contribution as compared to the previous approaches by Liu et al. (2006) and Hanneman and Lavie (2009) with respect to the following:

- The phrase inventory is augmented with non-syntactic phrases by means of a fill-up technique. Overlap is prevented, whereas not only new source sides, but also new target-side translation options can be added.
- The probabilities of syntactic phrase pairs are the same as in the syntax-based baseline, and the probabilities of the non-syntactic phrase pairs are the same as in a phrase-based system. Counts of syntactic and non-syntactic phrases are not summed up to obtain new estimates.
- Non-syntactic phrase pairs are distinguished from syntactic ones with an additional feature.

7 Conclusions

String-to-tree and tree-to-string translation systems can easily be augmented with non-syntactic phrases by means of phrase table fill-up, a special non-terminal symbol for left-hand sides of non-syntactic rules in the grammar, and an additional glue rule. A binary feature enables the system to distinguish non-syntactic phrases from syntactic ones and—on the basis of the respective feature weight—to favor syntactically motivated phrases during decoding.

Our results on an English→German translation task demonstrate the beneficial effect of augmenting GHKM translation systems with non-syntactic phrase pairs. Empirical gains in translation quality are up to 0.5 points BLEU and 0.7 points TER over the baseline on the recent test set of the shared translation task of the ACL 2014 Ninth Workshop on Statistical Machine Translation.

While GHKM-style syntactic translation has typically been utilized in string-to-tree settings in previous research, we have also adopted it to build tree-to-string systems in this work. Source syntax establishes interesting further directions for GHKM systems. We investigated two of them: input tree constraints and input tree features.

String-to-tree and tree-to-string GHKM systems perform roughly at the same level in terms of translation quality. Our best string-to-tree setup outperforms a phrase-based baseline by up to 0.8 points BLEU and 0.9 points TER (on newstest2014), our best tree-to-string setup outperforms the phrase-based baseline by up to 0.7 points BLEU and 1.1 points TER (on newstest2013).

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 287658 (EU-BRIDGE) and n° 288487 (MosesCore).

References

Hala Almaghout, Jie Jiang, and Andy Way. 2012. Extending CCG-based Syntactic Constraints in Hierarchical Phrase-Based SMT. In *Proc. of the Annual Conf. of the European Assoc. for Machine Translation (EAMT)*, pages 193–200, Trento, Italy, May.

Nicola Bertoldi and Marcello Federico. 2009. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 182–189, Athens, Greece, March.

Arianna Bisazza, Nick Ruiz, and Marcello Federico. 2011. Fill-up versus Interpolation Methods for Phrase-based SMT Adaptation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 136–143, San Francisco, CA, USA, December.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 1–44, Sofia, Bulgaria, August.

David Burkett and Dan Klein. 2012. Transforming Trees to Improve Syntactic Convergence. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, Jeju Island, South Korea, July.

Jean-Cédric Chappelier and Martin Rajman. 1998. A Generalized CYK Algorithm for Parsing Stochastic CFG. In *Proc. of the First Workshop on Tabulation in Parsing and Deduction*, pages 133–137, Paris, France, April.

Stanley F. Chen and Joshua Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Computer Science Group, Harvard University, Cambridge, MA, USA, August.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 427–436, Montréal, Canada, June.

David Chiang. 2005. A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 263–270, Ann Arbor, MI, USA, June.

David Chiang. 2007. Hierarchical Phrase-Based Translation. *Computational Linguistics*, 33(2):201–228, June.

David Chiang. 2010. Learning to Translate with Source and Target Syntax. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 1443–1452, Uppsala, Sweden, July.

Steve DeNeefe, Kevin Knight, Wei Wang, and Daniel Marcu. 2007. What Can Syntax-Based MT Learn from Phrase-Based MT? In *Proc. of the 2007 Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 755–763, Prague, Czech Republic, June.

Victoria Fossum, Kevin Knight, and Steven Abney. 2008. Using Syntax to Improve Word Alignment Precision for Syntax-Based Machine Translation. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 44–52, Columbus, OH, USA, June.

Michel Galley and Christopher D. Manning. 2008. A Simple and Effective Hierarchical Phrase Reordering Model. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 847–855, Honolulu, HI, USA, October.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 273–280, Boston, MA, USA, May.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable Inference and Training of Context-Rich Syntactic Translation Models. In *Proc. of the 21st International Conf. on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics*, pages 961–968, Sydney, Australia, July.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Columbus, OH, USA, June.

Greg Hanneman and Alon Lavie. 2009. Decoding with Syntactic and Non-syntactic Phrases in a Syntax-based Machine Translation System. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, SSST '09, pages 1–9, Boulder, CO, USA, June.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 187–197, Edinburgh, Scotland, UK, July.

Hieu Hoang and Philipp Koehn. 2010. Improved Translation with Source Syntax Labels. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 409–417, Uppsala, Sweden, July.

Hieu Hoang, Philipp Koehn, and Adam Lopez. 2009. A Unified Framework for Phrase-Based, Hierarchical, and Syntax-Based Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 152–159, Tokyo, Japan, December.

Matthias Huck, David Vilar, Daniel Stein, and Hermann Ney. 2011. Lightly-Supervised Training for Hierarchical Phrase-Based Machine Translation. In *Proc. of the EMNLP 2011 Workshop on Unsupervised Learning in NLP*, pages 91–96, Edinburgh, Scotland, UK, July.

Reinhard Kneser and Hermann Ney. 1995. Improved Backing-Off for M-gram Language Modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 181–184, Detroit, MI, USA, May.

Philipp Koehn, Franz Joseph Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 127–133, Edmonton, Canada, May/June.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of the MT Summit X*, Phuket, Thailand, September.

Alon Lavie. 2008. Stat-XFER: A General Search-Based Syntax-Driven Framework for Machine Translation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 362–375. Springer Berlin Heidelberg.

Yang Liu, Qun Liu, and Shouxun Lin. 2006. Tree-to-string Alignment Template for Statistical Machine Translation. In *Proc. of the 21st International Conf. on Computational Linguistics and the 44th Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 609–616, Sydney, Australia, July.

Daniel Marcu, Wei Wang, Abdessamad Echihabi, and Kevin Knight. 2006. SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proc. of the Conf. on Empirical Methods for Natural Language Processing (EMNLP)*, pages 44–52, Sydney, Australia.

Yuval Marton and Philip Resnik. 2008. Soft Syntactic Constraints for Hierarchical Phrased-Based Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 1003–1011, Columbus, OH, USA, June.

Maria Nadejde, Philip Williams, and Philipp Koehn. 2013. Edinburgh's Syntax-Based Machine Translation Systems. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 170–176, Sofia, Bulgaria, August.

Jan Niehues and Alex Waibel. 2012. Detailed Analysis of Different Strategies for Phrase Table Adaptation in SMT. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, San Diego, CA, USA, October/November.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 295–302, Philadelphia, PA, USA, July.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51, March.

Franz Josef Och, Christoph Tillmann, and Hermann Ney. 1999. Improved Alignment Models for Statistical Machine Translation. In *Proc. of the Joint SIGDAT Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, pages 20–28, University of Maryland, College Park, MD, USA, June.

Franz Josef Och. 2002. *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, October.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA, USA, July.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning Accurate, Compact, and Interpretable Tree Annotation. In *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Assoc. for Computational Linguistics*, pages 433–440, Sydney, Australia, July.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, Geneva, Switzerland, August.

Holger Schwenk. 2008. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 182–189, Waikiki, HI, USA, October.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, MA, USA, August.

Daniel Stein, Stephan Peitz, David Vilar, and Hermann Ney. 2010. A Cocktail of Deep Syntactic Features for Hierarchical Machine Translation. In *Proc. of the Conf. of the Assoc. for Machine Translation in the Americas (AMTA)*, Denver, CO, USA, October/November.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modeling Toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, volume 3, Denver, CO, USA, September.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference Grammars: Softening Syntactic Constraints to Improve Statistical Machine Translation. In *Proc. of the Human Language Technology Conf. / North American Chapter of the Assoc. for Computational Linguistics (HLT-NAACL)*, pages 236–244, Boulder, CO, USA, June.

David Vilar, Daniel Stein, and Hermann Ney. 2008. Analysing Soft Syntax Features and Heuristics for Hierarchical Phrase Based Machine Translation. In *Proc. of the Int. Workshop on Spoken Language Translation (IWSLT)*, pages 190–197, Waikiki, HI, USA, October.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007. Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 746–754, Prague, Czech Republic, June.

Wei Wang, Jonathan May, Kevin Knight, and Daniel Marcu. 2010. Re-structuring, Re-labeling, and Re-aligning for Syntax-based Machine Translation. *Computational Linguistics*, 36(2):247–277, June.

Philip Williams and Philipp Koehn. 2012. GHKM Rule Extraction and Scope-3 Parsing in Moses. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 388–394, Montréal, Canada, June.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, Baltimore, MD, USA, June.

Joern Wuebker, Arne Mauser, and Hermann Ney. 2010. Training Phrase Translation Models with Leaving-One-Out. In *Proc. of the Annual Meeting of the Assoc. for Computational Linguistics (ACL)*, pages 475–484, Uppsala, Sweden, July.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax Augmented Machine Translation via Chart Parsing. In *Proc. of the Workshop on Statistical Machine Translation (WMT)*, pages 138–141, New York City, NY, USA, June.