Bilingual Markov Reordering Labels for Hierarchical SMT

Gideon Maillette de Buy Wenniger and **Khalil Sima'an**
Institute for Logic, Language and Computation
University of Amsterdam
Science Park 107, 1098 XG Amsterdam, The Netherlands
gemdbw AT gmail.com, k.simaan AT uva.nl

Abstract

Earlier work on labeling Hiero grammars with monolingual syntax reports improved performance, suggesting that such labeling may impact phrase reordering as well as lexical selection. In this paper we explore the idea of inducing *bilingual labels* for Hiero grammars without using any additional resources other than original Hiero itself does. Our bilingual labels aim at capturing salient patterns of phrase reordering in the training parallel corpus. These bilingual labels originate from hierarchical factorizations of the word alignments in Hiero's own training data. In this paper we take a Markovian view on synchronous top-down derivations over these factorizations which allows us to extract 0^{th}- and 1^{st}-order bilingual reordering labels. Using exactly the same training data as Hiero we show that the Markovian interpretation of word alignment factorization offers major benefits over the unlabeled version. We report extensive experiments with strict and soft bilingual labeled Hiero showing improved performance up to 1 BLEU points for Chinese-English and about 0.1 BLEU points for German-English.

Phrase reordering in Hiero (Chiang, 2007) is modelled with synchronous rules consisting of phrase pairs with at most two nonterminal gaps, thereby embedding ITG permutations (Wu, 1997) in lexical context. It is by now recognized that Hiero's reordering can be strengthened either by labeling (e.g., (Zollmann and Venugopal, 2006)) or by supplementing the grammar with extra-grammatical reordering models, e.g., (Xiao et al., 2011; Huck et al., 2013; Nguyen and Vogel, 2013). In this paper we concentrate on labeling approaches.

Conceptually, labeling Hiero rules aims at introducing preference in the SCFG derivations for frequently occurring lexicalized ordering constellations over rare ones which also affects lexical selection. In this paper, we present an approach for distilling phrase reordering labels directly from alignments (hence *bilingual labels*).

To extract bilingual labels from word alignments we must first interpret the alignments as a hierarchy of phrases. Luckily, every word alignment factorizes into Normalized Decomposition Trees (NDTs) (Zhang et al., 2008), showing explicitly how the word alignment recursively decomposes into phrase pairs. Zhang et al. (2008) employ NDTs for extracting Hiero grammars. In this work, we extend NDTs with explicit phrase permutation operators also extracted from the original word alignment (Sima'an and Maillette de Buy Wenniger, 2013); Every node in the NDT is equipped with a *node operator* that specifies how the order of the target phrases (children of this node) is produced from the corresponding source phrases. Subsequently, we cluster the node operators in these enriched NDTs according to their complexity, e.g., monotone (straight), inverted, non-binary but one-to-one, and the more complex case of discontinuous (Maillette de Buy Wenniger and Sima'an, 2013).

Inspired by work on parsing (Klein and Manning, 2003), we explore a vertical Markovian labeling approach: intuitively, 0^{th}-order labels signify the reordering of the sub-phrases inside the phrase pair (Zhang et al., 2008), 1^{st}-order labels signify reordering aspects of the direct context (an embedding, parent phrase pair) of the phrase pair, and so on. Like the phrase orientation models this labeling approach does not employ external resources (e.g., taggers, parsers) beyond the training data used by Hiero.

We empirically explore this bucketing for 0^{th}-

Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, pages 11–21,
October 25, 2014, Doha, Qatar.

and 1^{st}-order labels both as hard and soft labels. In experiments on German-English and Chinese-English we show that this extension of Hiero often significantly outperforms the unlabeled model while using no external data or monolingual labeling mechanisms. This suggests the viability of automatically inducing bilingual labels following the Markov labeling approach on operator-labelled NDTs as proposed in this paper.

1 Hierarchical models and related work

Hiero SCFGs (Chiang, 2005; Chiang, 2007) allow only up to two (pairs of) nonterminals on the right-hand-side (RHS) of synchronous rules. The types of permissible Hiero rules are:

$$X \rightarrow \langle \alpha,\ \gamma \rangle \quad (1)$$

$$X \rightarrow \langle \alpha\ X_{\boxed{1}} \beta,\ \delta\ X_{\boxed{1}} \zeta \rangle \quad (2)$$

$$X \rightarrow \langle \alpha\ X_{\boxed{1}} \beta\ X_{\boxed{2}} \gamma\ ,\ \delta\ X_{\boxed{1}}\ \zeta\ X_{\boxed{2}} \eta\ \rangle \quad (3)$$

$$X \rightarrow \langle \alpha\ X_{\boxed{1}} \beta\ X_{\boxed{2}} \gamma\ ,\ \delta\ X_{\boxed{2}}\ \zeta\ X_{\boxed{1}} \eta\ \rangle \quad (4)$$

Here $\alpha, \beta, \gamma, \delta, \zeta, \eta$ are terminal sequences, possibly empty. Equation 1 corresponds to a normal phrase pair, 2 to a rule with one gap and 3 and 4 to the monotone- and inverting rules respectively.

Given an Hiero SCFG G, a source sentence $\mathbf{s}$ is translated into a target sentence $\mathbf{t}$ by synchronous derivations $\mathbf{d}$, each is a finite sequence of well-formed substitutions of synchronous productions from G, see (Chiang, 2006). Existing phrase-based models score a derivation *der* with linear interpolation of a finite set of feature functions ($\Phi(\mathbf{d})$) of the derivation $\mathbf{d}$, mostly working with local feature functions ϕ_i of individual productions, the target side yield string t of $\mathbf{d}$ (target language model features) and other features (see experimental section): $\arg\max_{\mathbf{d}\in G} P(\mathbf{t}, \mathbf{d} \mid \mathbf{s}) \approx \arg\max_{\mathbf{d}\in G} \sum_{i=1}^{|\Phi(\mathbf{d})|} \lambda_i \times \phi_i$. The parameters $\{\lambda_i\}$ are optimized on a held-out parallel corpus by direct error-minimization (Och, 2003).

A range of (distantly) related work exploits syntax for Hiero models, e.g. (Liu et al., 2006; Huang et al., 2006; Mi et al., 2008; Mi and Huang, 2008; Zollmann and Venugopal, 2006; Wu and Hkust, 1998). In terms of labeling Hiero rules, SAMT (Zollmann and Venugopal, 2006; Mylonakis and Sima'an, 2011) exploits a "softer notion" of syntax by fitting the CCG-like syntactic labels to non-constituent phrases. The work of (Xiao et al., 2011) adds a lexicalized orientation model to Hiero, akin to (Tillmann, 2004) and achieves significant gains. The work of (Huck et al., 2013; Nguyen and Vogel, 2013) overcomes technical limitations of (Xiao et al., 2011), making necessary changes to the decoder, which involves delayed (re-)scoring at hypernodes up in the derivation of nodes lower in the chart whose orientations are affected by them. This goes to show that phrase-orientation models are not mere labelings of Hiero.

Soft syntactic constraints has been around for some time now (Zhou et al., 2008; Venugopal et al., 2009; Chiang, 2010). In (Zhou et al., 2008) Hiero is reinforced with a linguistically motivated prior. This prior is based on the level of syntactic homogeneity between pairs of non-terminals and the associated syntactic forests rooted at these nonterminals, whereby tree-kernels are applied to efficiently measure the amount of overlap between all pairs of sub-trees induced by the pairs of syntactic forests. Crucially, the syntactic prior encourages derivations that are more syntactically coherent but does not block derivations when they are not. In (Venugopal et al., 2009) the authors associate distributions over compatible syntactic labelings with grammar rules, and combine these preference distributions during decoding, thus achieving a summation rather than competition between compatible label configurations. The latter approach requires significant changes to the decoder and comes at a considerable computational cost. An alternative approach (Chiang, 2010) uses labels similar to (Zollmann and Venugopal, 2006) together with boolean features for rule-label and substituted-label combinations; using discriminative training (MIRA) it is learned what combinations are associated with better translations.

The labeling approach presented next differs from existing approaches. It is inspired by soft labeling but employs novel, non-linguistic bilingual labels. And it shares the bilingual intuition with phrase orientation models but it is based on a Markov approach for SCFG labeling, thereby remaining within the confines of Hiero SCFG, avoiding the need to make changes inside the decoder.[1]

[1] Soft constraint decoding can easily be implemented without adapting the decoder, through a smart application of "label bridging" unary rules. In practice however, adapting the decoder turns out to be computationally more efficient, therefore we used this solution in our experiments.

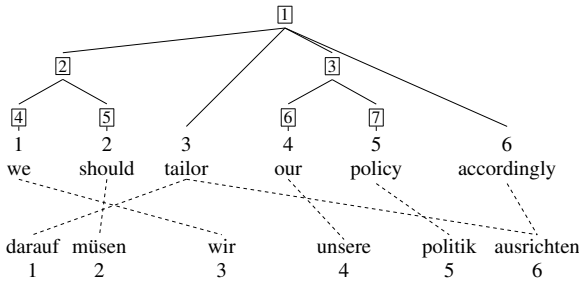

Figure 1: Example alignment from Europarl

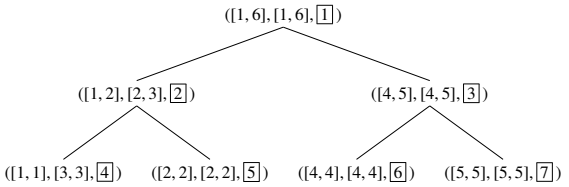

Figure 2: Normalized Decomposition Tree (Zhang et al., 2008) extended with pointers to original alignment structure from Figure 1

2 Bilingual reordering labels for Hiero

Figure 1 shows an alignment from Europarl German-English (Koehn, 2005) along with a tree showing corresponding maximally decomposed phrase pairs. Phrase pairs can be grouped into a maximally decomposed tree (called Normalized Decomposition Tree – NDT) (Zhang et al., 2008). Figure 2 shows the NDT for Figure 1, extended with pointers to the original alignment structure in Figure 2. The numbered boxes indicate how the phrases in the two representations correspond. In an NDT every phrase pair is recursively split up at every level into a minimum number (two or greater) of contiguous parts. In this example the root node splits into three phrase pairs, but these phrase pairs together do not cover the entire parent phrase pair because of the discontinuity: "tailor ... accordingly/ darauf ... ausrichten".

Following (Zhang et al., 2008), we use the NDT factorizations of word alignments in the training data for extracting phrases. Every NDT shows the hierarchical structuring into phrases embedded in larger phrases, which together with the context of the original alignment exposes the reordering complexity of every phrase (Sima'an and Maillette de Buy Wenniger, 2013). We will exploit these elaborate distinctions based on the complexity of reordering for Hiero rule labels as explained next.

Phrase-centric (0^{th}-order) labels are based on the view of looking inside a phrase pair to see how it decomposes into sub-phrase pairs. The operator signifying how the sub-phrase pairs are reordered (target relative to source) is bucketted into a number of "permutation complexity" categories. Straightforwardly, we can start out by using the two well known cases of Inversion Transduction Grammars (ITG) {*Monotone*, *Inverted*} and label everything[2] that falls outside these two category with a default label "*X*" (leaving some Hiero nodes unlabeled). This leads to the following *coarse* phrase-centric labeling scheme, which we name $0^{th}{}_{ITG+}$: (1) *Monotonic(Mono)*: binarizable, fully monotone plus non-decomposable phrases (2) *Inverted(Inv)*: binarizable, fully inverted (3) *X*: decomposable phrases that are not binarizable.

A clear limitation of the above ITG-like labeling approach is that all phrase pairs that decompose into complex non-binarizable reordering patterns are not further distinguished. Furthermore, non-decomposable phrases are lumped together with decomposable monotone phrases, although they are in fact quite different. To overcome these problems we extend ITG in a way that further distinguishes the non-binarizable phrases and also distinguishes non-decomposable phrases from the rest. This gives a labeling scheme we will call simply 0^{th}-order labeling, abbreviated 0^{th}, consisting of a more fine-grained set of five cases, ordered by increasing complexity (see examples in Figure 4): (1) *Atomic*: non-decomposable phrases, (2) *Monotonic(Mono)*: binarizable, fully monotone, (3) *Inverted(Inv)*: binarizable, fully inverted (4) *Permutation(Perm)*: factorizes into a permutation of four or more sub-phrases (5) *Complex(Comp)*: does not factorize into a permutation and contains at least one embedded phrase.

In Figure 3, we show a phrase-complexity labeled derivation for the example of Figure 1. Observe how the phrase-centric labels reflect the relative reordering at the node. For example, the

[2]Non-decomposable phrases will still be grouped together with Monotone, since they are more similar to this category than to the catchall "*X*" category.

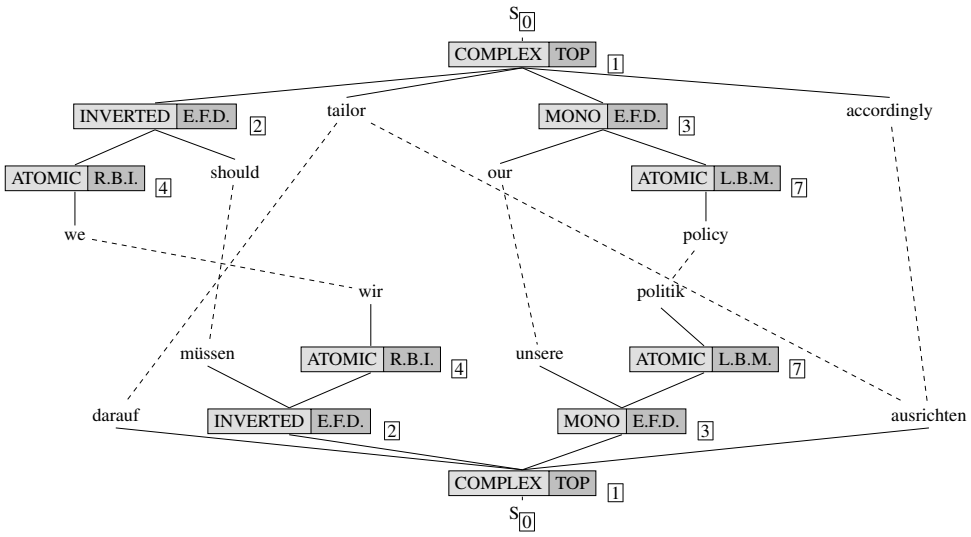

Figure 3: Synchronous trees (implicit derivations end results) based on differently labelled Hiero grammars. The figure shows alternative labeling for every node: *Phrase-Centric (0^{th}-order)* (light gray) and *Parent-Relative (1^{st}-order)* (dark gray).

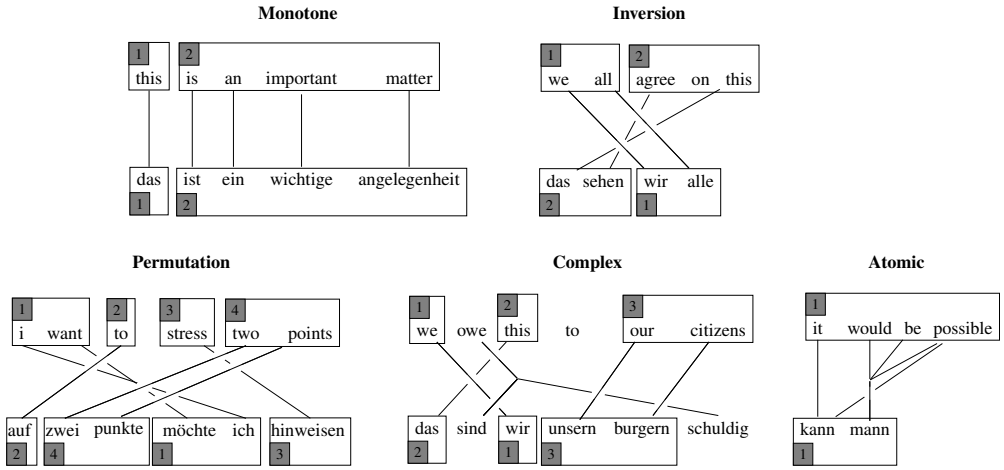

Figure 4: Different types of Phrase-Centric Alignment Labels

Inverted label of node-pair 2 corresponds to the inversion in the alignment of ⟨we should, müsen wir⟩; in contrast, node-pair 1 is complex and discontinuous and the label is *Complex*.

Parent-relative (1^{st}-order) labels capture the reordering that a phrase undergoes relative to an embedding parent phrase.

1. For a binarizable mother phrase with orientation $X_o \in \{Mono, Inv\}$, the phrase itself can either group to the left only *Left-Binding-X_o*, right only *Right-Binding-X_o*, or with both sides (*Fully-X_o*).

2. *Fully-Discontinuous*: Any phrase within a non-binarizable permutation or complex alignment containing discontinuity.

3. *Top*: phrases that span the entire aligned sentence pair.

In cases were multiple labels are applicable, the simplest applicable label is chosen according to the following preference order:
{*Fully-Monotone*, *Left/Right-Binding-Monotone*, *Fully-Inverted*, *Left/Right-Binding-Inverted*, *Fully-Discontinuous*, *TOP*}.

In Figure 3 the parent-relative labels in the derivation reflect the reordering taking place at the phrases with respect to their parent node. Node 4 has a parent node that inverts the order and the sibling node it binds is on the right, therefore it

is labeled "right-binding inverted" (R.B.I.); E.F.D. and L.B.M. are similar abbreviations for "embedded fully discontinuous" and "left-binding monotone" respectively. As yet another example node 7 in Figure 3 is labeled "left-binding monotone" (L.B.M.) since it is monotone, but the alignment allows it only to bind to the left at the parent node, as opposed to only to the right or to both sides which cases would have yielded "right-binding monotone" R.B.M. and "(embedded) fully monotone" (E.F.M.) parent-relative reordering labels respectively.

Note that for parent-relative labels the binding direction of monotone and inverted may not be informative. We therefore also form a set of *coarse* parent-relative labels ("$1^{st}{}_{Coarse}$") by collapsing the label pairs *Left/Right-Binding-Mono* and *Left/Right-Binding-Inverted* into single labels *One-Side-Binding-Mono* and *One-Side-Binding-Inv*[3].

3 Features for soft bilingual labeling

Labels used in hierarchical Statistical Machine Translation (SMT) are typically adapted from external resources such as taggers and parsers. Like in our case, these labels are typically not fitted to the training data – with very few exceptions e.g., (Mylonakis and Sima'an, 2011; Mylonakis, 2012; Hanneman and Lavie, 2013). Unfortunately this means that the labels will either overfit or underfit, and when they are used as strict constraints on SCFG derivations they are likely to underperform. Experience with mismatch between syntactic labels and the data is abundant (Venugopal et al., 2009; Marton et al., 2012; Chiang, 2010), and using soft constraint decoding with suitable label substitution features has been shown to be an effective workaround solution. The intuition behind soft constraint decoding is that even though heuristic labels are not perfectly tailored to the data, they do provide useful information provided the model is "allowed to learn" to use them only in as far as they can improve the final evaluation metric (usually BLEU).

[3]We could also further coarsen the 1^{st}labels by removing entirely all sub-distinctions of binding-type for the binarizable cases, but that would make the labeling essentially equal to the earlier mentioned $0^{th}{}_{ITG+}$ except for looking at the reordering occurring at the parent rather than inside the phrase itself. We did not explore this variant in this work, as the high similarity to the already explored $0^{th}{}_{ITG+}$ variant made it not seem to add much extra information.

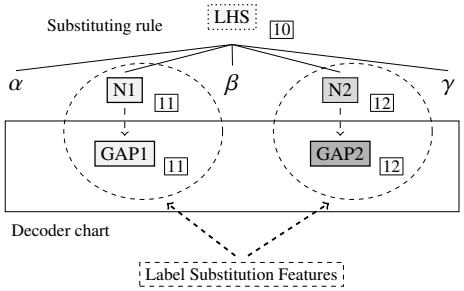

Figure 5: Label substitution features, schematic view. Labels/Gaps with same filling in the figures correspond to the situation of a nonterminal/gap whose labels correspond (for N1/GAP1). Fillings of different shades (as for N2/GAP2 on the right in the two figures) indicates the situation were the label of the nonterminal and the gap is different.

Next we introduce the set of label substitution features used in our experiments.

Label substitution features consist of a unique feature for every pair of labels $\langle L_\alpha, L_\beta \rangle$ in the grammar, signifying a rule with left-hand-side label L_β substituting on a gap labeled L_α. These features are combined with two more coarse features, "*Match*" and "*Nomatch*", indicating if the substitution involves labels that match or not.

Figure 5 illustrates the concept of label substitution features schematically. In this figure the substituting rule is substituted onto two gaps in the chart, which induces two label substitution features indicated by the two ellipses. The situation is analogous for rules with just one gap. To make things concrete, lets assume that both the first nonterminal of the rule $N1$ as well as the first gap it is substituted onto $GAP1$ have label $MONO$. Furthermore lets assume the second nonterminal $N2$ has label $COMPLEX$ while the label of the gap $GAP2$ it substitutes onto is INV. This situation results in the following two specific label substitution features:

- subst($MONO$,$MONO$)
- subst(INV,$COMPLEX$)

Canonical labeled rules. Typically when labeling Hiero rules there can be many different labeled variants of every original Hiero rule. With soft constraint decoding this leads to prohibitive computational cost. This also has the effect of making tuning the features more difficult. In practice, soft constraint decoding usually exploits

Systen Name	Matching Type	Label Order	Label Granularity
Hiero-$0^{th}{}_{ITG+}$	Strict	0^{th} order	Coarse
Hiero-0^{th}	Strict	0^{th} order	Fine
Hiero-$1^{st}{}_{Coarse}$	Strict	1^{th} order	Coarse
Hiero-1^{st}	Strict	1^{th} order	Fine
Hiero-$0^{th}{}_{ITG+}$-Sft	Soft	0^{th} order	Coarse
Hiero-0^{th}-Sft	Soft	0^{th} order	Fine
Hiero-$1^{st}{}_{Coarse}$-Sft	Soft	1^{th} order	Coarse
Hiero-1^{st}-Sft	Soft	1^{th} order	Fine

Table 1: Experiment names legend

System Name	DEV				TEST			
	BLEU ↑	METEOR ↑	TER ↓	KRS ↑	BLEU ↑	METEOR ↑	TER ↓	KRS ↑
	German-English							
Hiero	**27.90**	32.69	58.22	66.37	**28.39**	32.94	58.01	67.44
SAMT	27.76	32.67	58.05	**66.84**$^{\blacktriangle}$	28.32	32.88	**57.70**$^{\blacktriangle\blacktriangle}$	**67.63**
Hiero-$0^{th}{}_{ITG+}$	27.85	32.70	58.04$^{\blacktriangle\blacktriangle}$	66.27	28.36	32.90$^{\blacktriangledown}$	57.83$^{\blacktriangle\blacktriangle}$	67.30
Hiero-0^{th}	27.82	**32.75**	**57.92**$^{\blacktriangle\blacktriangle}$	66.66	**28.39**	**33.03**$^{\blacktriangle\blacktriangle}$	57.75$^{\blacktriangle\blacktriangle}$	67.55
Hiero-$1^{st}{}_{Coarse}$	27.86	32.66	58.23	66.37	28.22$^{\blacktriangledown}$	32.90	57.93	67.47
Hiero-1^{st}	27.74$^{\blacktriangledown}$	32.60$^{\blacktriangledown\blacktriangledown}$	58.11	66.44	28.27	32.80$^{\blacktriangledown\blacktriangledown}$	57.95	67.39
	Chinese-English							
Hiero	31.70	30.72	**61.21**	58.28	31.63	30.56	**59.28**	58.03
Hiero-$0^{th}{}_{ITG+}$	31.54	**30.97**$^{\blacktriangle\blacktriangle}$	62.79$^{\blacktriangledown\blacktriangledown}$	59.54$^{\blacktriangle\blacktriangle}$	**31.94**$^{\blacktriangle\blacktriangle}$	**30.84**$^{\blacktriangle\blacktriangle}$	60.76$^{\blacktriangledown\blacktriangledown}$	59.45$^{\blacktriangle\blacktriangle}$
Hiero-0^{th}	31.66	30.95$^{\blacktriangle\blacktriangle}$	62.20$^{\blacktriangledown\blacktriangledown}$	60.00$^{\blacktriangle\blacktriangle}$	31.90$^{\blacktriangle\blacktriangle}$	30.79$^{\blacktriangle\blacktriangle}$	60.11$^{\blacktriangledown\blacktriangledown}$	59.68$^{\blacktriangle\blacktriangle}$
Hiero-$1^{st}{}_{Coarse}$	31.64	30.75	61.37	59.48$^{\blacktriangle\blacktriangle}$	31.57	30.57	59.58$^{\blacktriangledown\blacktriangledown}$	59.13$^{\blacktriangle\blacktriangle}$
Hiero-1^{st}	**31.74**	30.79	61.94$^{\blacktriangledown\blacktriangledown}$	**60.22**$^{\blacktriangle\blacktriangle}$	31.77	30.62	60.13$^{\blacktriangledown\blacktriangledown}$	**59.89**$^{\blacktriangle\blacktriangle}$

Table 2: Mean results bilingual labels with strict matching.[4]

a single labeled version per Hiero rule, which we call the "canonical labeled rule". Following (Chiang, 2010), this canonical form is the most frequent labeled variant.

4 Experiments

We evaluate our method on two language pairs: using German/Chinese as source and English as target. In all experiments we decode with a 4-gram language model smoothed with modified Knesser-Ney discounting (Chen and Goodman, 1998). The data used for training the language models differs per language pair, details are given in the next paragraphs. All data is lowercased as a last pre-processing step. In all experiments we use our own grammar extractor for the generation of all grammars, including the baseline Hiero grammars. This enables us to use the same features (as far as applicable given the grammar formalism) and assure true comparability of the grammars under comparison.

German-English

The data for our German-English experiments is derived from parliament proceedings sourced from the Europarl corpus (Koehn, 2005), with WMT-07 development and test data. We used a maximum sentence length of 40 for filtering the training data. We employ 1M sentence pairs for training, 1K for development and 2K for testing (single reference per source sentence). Both source and target of all datasets are tokenized using the Moses(Hoang et al., 2007) tokenization script. For these experiments both the baseline and our method use a language model trained on the target side of the full original training set (approximately 1M sentences).

Chinese-English

The data for our Chinese-English experiments is derived from a combination of *MultiUn*(Eisele and Chen, 2010; Tiedemann, 2012)[5] data and *Hong Kong Parallel Text* data from the Linguistic Data Consortium[6]. The *Hong Kong Parallel Text* data is in *traditional Chinese* and is thus first converted to *simplified Chinese* to be compatible

[4]Statistical significance is dependent on variance of resampled scores, and hence sometimes different for same mean scores across different systems.

[5]Freely available and downloaded from http://opus.lingfil.uu.se/

[6]The LDC catalog number of this dataset is LDC2004T08

System Name	DEV				TEST			
	BLEU ↑	METEOR ↑	TER ↓	KRS ↑	BLEU ↑	METEOR ↑	TER ↓	KRS ↑
	German-English							
Hiero	27.90	32.69	58.22	66.37	28.39	32.94	58.01	67.44
SAMT	27.76	32.67	58.05	**66.84**$^{\blacktriangle}$	28.32	32.88	**57.70**$^{\blacktriangle\blacktriangle}$	**67.63**
Hiero-$0^{th}{}_{ITG+}$-Sft	28.00$^{\blacktriangle}$	32.76$^{\blacktriangle\blacktriangle}$	**57.90**$^{\blacktriangle\blacktriangle}$	66.17	**28.48**	32.98	57.79$^{\blacktriangle\blacktriangle}$	67.32
Hiero-0^{th}-Sft	28.01$^{\blacktriangle}$	32.71	57.95$^{\blacktriangle\blacktriangle}$	66.24	28.45	32.98	57.73$^{\blacktriangle\blacktriangle}$	67.51
Hiero-$1^{st}{}_{Coarse}$-Sft	27.94	32.69	57.91$^{\blacktriangle\blacktriangle}$	66.26	28.45$^{\blacktriangle}$	32.94	57.75$^{\blacktriangle\blacktriangle}$	67.36
Hiero-1^{st}-Sft	**28.13**$^{\blacktriangle\blacktriangle}$	**32.80**$^{\blacktriangle\blacktriangle}$	57.92$^{\blacktriangle\blacktriangle}$	66.32	28.45	**33.00**$^{\blacktriangle}$	57.79$^{\blacktriangle\blacktriangle}$	67.45
	Chinese-English							
Hiero	31.70	30.72	61.21	58.28	31.63	30.56	59.28	58.03
Hiero-$0^{th}{}_{ITG+}$-Sft	31.88$^{\blacktriangle}$	30.46$^{\blacktriangledown\blacktriangledown}$	**60.64**$^{\blacktriangle\blacktriangle}$	57.82$^{\blacktriangledown}$	31.93$^{\blacktriangle\blacktriangle}$	30.37$^{\blacktriangledown\blacktriangledown}$	**58.86**$^{\blacktriangle\blacktriangle}$	57.60$^{\blacktriangledown}$
Hiero-0^{th}-Sft	32.04$^{\blacktriangle\blacktriangle}$	30.90$^{\blacktriangle\blacktriangle}$	61.47$^{\blacktriangledown\blacktriangledown}$	59.36$^{\blacktriangle\blacktriangle}$	32.20$^{\blacktriangle\blacktriangle}$	30.74$^{\blacktriangle\blacktriangle}$	59.45$^{\blacktriangledown}$	58.92$^{\blacktriangle\blacktriangle}$
Hiero-$1^{st}{}_{Coarse}$-Sft	32.39$^{\blacktriangle\blacktriangle}$	31.02$^{\blacktriangle\blacktriangle}$	61.56$^{\blacktriangledown\blacktriangledown}$	59.51$^{\blacktriangle\blacktriangle}$	32.55$^{\blacktriangle\blacktriangle}$	30.86$^{\blacktriangle\blacktriangle}$	59.57$^{\blacktriangledown\blacktriangledown}$	59.03$^{\blacktriangle\blacktriangle}$
Hiero-1^{st}-Sft	**32.63**$^{\blacktriangle\blacktriangle}$	**31.22**$^{\blacktriangle\blacktriangle}$	62.00$^{\blacktriangledown\blacktriangledown}$	**60.43**$^{\blacktriangle\blacktriangle}$	**32.61**$^{\blacktriangle\blacktriangle}$	**30.98**$^{\blacktriangle\blacktriangle}$	60.19$^{\blacktriangledown\blacktriangledown}$	**59.84**$^{\blacktriangle\blacktriangle}$

Table 3: Mean results bilingual labels with soft matching.[4]

with the rest of the data [7]. We used a maximum sentence length of 40 for filtering the training data. The combined dataset has 7.34M sentence pairs. The *MulitUN* dataset contains translated documents from the United Nations, similar in genre to the parliament domain. The *Hong Kong Parallel Text* in contrast contains a richer mix of domains, namely Hansards, Laws and News. For the dev and test set we use the *Multiple-Translation Chinese* datasets from LDC, part 1-4[8], which contain sentences from the News domain. We combined part 2 and 3 to form the dev set (1813 sentence pairs) and part 1 and 4 to form the test set (1912 sentence pairs). For both development and testing we use 4 references. The Chinese source side of all datasets is segmented using the Stanford Segmenter(Chang et al., 2008)[9]. The English target side of all datasets is tokenized using the Moses tokenization script.

For these experiments both the baseline and our method use a language model trained on 5.4M sentences of *domain specific*[10] news data taken from the "Xinhua" subcorpus of the English Gigaword corpus of LDC. [11]

[7]Using a simple conversion script downloaded from http://www.mandarintools.com/zhcode.html

[8]LDC catalog numbers: LDC2002T01, DC2003T17, LDC2004T07 and LDC2004T07

[9]Downloaded from http://nlp.stanford.edu/software/segmenter.shtml

[10]For Chinese-English translation the different domain of the train data (mainly parliament) and dev/test data (news) requires usage of a domain specific language model to get optimal results. For German-English, all data is from the the parliament domain, so a language model trained on the (translation model) training data is already domain-specific.

[11]The LDC catalog number of this dataset is LDC2003T05

4.1 Experimental Structure

In our experiments we explore the influence of three dimensions of bilingual reordering labels on translation accuracy. These dimensions are:

- *label granularity* : granularity of the labeling {Coarse,Fine}
- *label order* : the type/order of the labeling {0^{th}, 1^{st}}
- *matching type* : the type of label matching performed during decoding {Strict,Soft}

Combining these dimensions gives 8 different reordering labeled systems per language pair. On top of that we use two baseline systems, namely Hiero and Syntax Augmented Machine Translation (SAMT) to measure these systems against. An overview of the naming of our reordering labeled systems is given in Table 1.

Training and decoding details Our experiments use Joshua (Ganitkevitch et al., 2012) with Viterbi best derivation. Baseline experiments use normal decoding whereas soft labeling experiments use soft constraint decoding. For training we use standard Hiero grammar extraction constraints (Chiang, 2007) (phrase pairs with source spans up to 10 words; abstract rules are forbidden). During decoding maximum span 10 on the source side is maintained. Following common practice, we use relative frequency estimates for phrase probabilities, lexical probabilities and generative rule probability.

We train our systems using (batch-kbest) Mira as borrowed by Joshua from the Moses codebase, allowing up to 30 tuning iterations. Following

standard practice, we tune on BLEU, and after tuning we use the configuration with the highest scores on the dev set with actual (corpus level) BLEU evaluation. We report lowercase BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2011) and TER (Snover et al., 2006) scores for the tuned test set and also for the tuned dev set, the latter mainly to observe any possible overfitting. We use Multeval version 0.5.1.[12] for computing these metrics. We also use MultEval's implementation of statistical significance testing between systems, which is based on multiple optimizer runs and approximate randomization. Multeval (Clark et al., 2011) randomly swaps outputs between systems and estimates the probability that the observed score difference arose by chance. Differences that are statistically significant and correspond to improvement/worsening with respect to the baseline are marked with $^{\blacktriangle}/^{\blacktriangledown}$at the $p \leq .05$ level and $^{\blacktriangle\blacktriangle}/^{\blacktriangledown\blacktriangledown}$at the $p \leq .01$ level. We also report the Kendall Reordering Score (KRS), which is the reordering-only variant of the LR-score (Birch and Osborne, 2010) (without the optional interpolation with BLEU) and which is a sentence-level score. For the computation of statistical significance of this metric we use our own implementation of the *sign test* [13] (Dixon and Mood, 1946), as also described in (Koehn, 2010).

In our experiments we repeated each experiment three times to counter unreliable conclusions due to optimizer variance. Scores are averages over three runs of tuning plus testing. Scores marked with $^{\blacktriangle}$ are significantly better than the baseline, those marked with $^{\blacktriangledown}$ are significantly worse; according to the resampling test of Multeval (Clark et al., 2011).

Preliminary experiment with strict matching Initial experiments concerned 0^{th}-order reordering labels in a *strict matching* approach (no soft constraints). The results are shown in Table 2 for both language pairs. The results for the Hiero and SAMT[14] baselines (Hiero and SAMT) are shown in the first rows. Below it results for the 0^{th}-order (phrase-centric) bilingual labeled systems with either the *Coarse* (Hiero-$0^{th}{}_{ITG+}$) or *Fine* label variant (Hiero-0^{th}) are shown, followed by the results for *Coarse* and *Fine* variant of the 1^{th}-order (parent-relative) bilingual labeled systems (Hiero-$1^{st}{}_{Coarse}$ and Hiero-1^{st}). All these systems use the default decoding with strict label matching.

For German-English the effect of strict bilingual labels is mostly positive: although we have no improvement for BLEU we do achieve significant improvements for METEOR and TER on the test set. For Chinese-English, overall Hiero-$0^{th}{}_{ITG+}$ shows the biggest improvements, namely significant improvements of +0.31 BLEU, +0.28 METEOR and +1.42 KRS. TER is the only metric that worsens, and considerably so with +1.48 point. Hiero-1^{st}achieves the highest improvement of KRS, namely 1.86 point higher than the Hiero baseline. Overall, this preliminary experiment shows that strict labeling sometimes gives improvements over Hiero, but sometimes it leads to worsening in terms of some of the metrics.

Results with soft bilingual constraints Our initial experiments with strict bilingual labels in combination with strict matching by the decoder gave some hope such constraints could be useful. At the same time the results showed no stable improvements across language pairs, and thus does not allow us to draw definite conclusions about the merit of bilingual labels.

Results for experiments with soft bilingual labeling are shown in Table 3. Here *Hiero* corresponds to the Hiero baseline. Below it are shown the systems that use soft constraint decoding (SCD). *Hiero*-$0^{th}{}_{ITG+}$-*Sft* and *Hiero*-0^{th}-*Sft* using phrase-centric labels (0^{th}-order) in *Coarse* or *Fine* form. Similarly, *Hiero*-$1^{st}{}_{Coarse}$-*Sft* and *Hiero*-1^{st}-*Sft* correspond to the analog systems with 1^{st}-order, parent-relative labels. For German-English there are only minor improvements for BLEU and METEOR, with somewhat bigger improvements for TER. For Chinese-English however the improvements are considerable, +0.98 BLEU improvement over the Hiero baseline for Hiero-1^{st}-Sft as well as +0.42 METEOR and +1.81 KRS. TER is worsening with +0.85 for this system. For Chinese-English the *Fine* version of the labels gives overall superior results for both 0^{th}-order and 1^{st}-order labels.

Discussion Our best soft bilingual labeling system for German-English shows small but significant improvements of METEOR and TER while im-

[12] https://github.com/jhclark/multeval

[13] To make optimal usage of the 3 runs we computed equally weighted improvement/worsening counts for all possible 3×3 baseline output / system output pairs and use those weighted counts in the sign test.

[14] SAMT could only be ran for German-English and not for Chinese-English, due to memory constraints.

proving BLEU and KRS as well, but not significantly. The results with soft-constraint matching are better than those for strict-matching in general, while there is no clear winner between the *Coarse* and *Fine* variant of labels.

For Chinese-English we see considerable improvements and overall the best results for the combination of soft-constraint matching, with the *Fine* 1^{st}-order variant of the labeled systems (Hiero-1^{st}-Sft). For Chinese-English the improvement of the word-order is also particularly clear as indicated by the +1.81 KRS improvement for this best system. Furthermore the negative effects in terms of worsening of TER are also reduced in the soft-matching setting, dropping from +1.48 TER to +0.85 TER. The results for Hiero-0^{th}-Sft are also competitive, since though it gives somewhat lower improvements of BLEU and METEOR, it gives an improvement of +1.89 KRS, while TER only worsens by +0.17 for this system.

We conclude that *bilingual Markov labels* can make a big difference in improvement of hierarchical SMT. We observe that going beyond the basic reordering labels of ITG, refining the cases not captured by ITG and even more effective: taking a 1^{st}-order rather than o^{th}-order perspective on reordering are major factors for the success of including reordering information to hierarchical SMT through labeling. Crucial to the success of this undertaking is also the usage of a soft-constraint approach to label matching, as opposed to strict-matching. Finally, comparison of the German-English results with results for Syntax-Augmented Machine Translation (SAMT) reveals that SAMT loses performance compared to the Hiero baseline for BLEU, the metric upon which tuning is done, as well as METEOR, while only TER and KRS show improvement. Since the best bilingual labeled system for German-English (Hiero-1^{st}-Sft) improves METEOR and TER significantly, while also improving BLEU and KRS, though not significant, we believe our labeling is highly competitive with syntax-based labeling approaches, without the need for any additional resources in the form of parsers or taggers, as syntax-based systems require. Likely complementarity of reordering information, and (target) syntax, which improves fluency, makes combining both a promising possibility we would like to explore in future work.

5 Conclusion

We presented a novel method to enrich Hierarchical Statistical Machine Translation with bilingual labels that help to improve the translation quality. Considerable and significant improvements of the BLEU, METEOR and KRS are achieved simultaneously for Chinese-English translation while tuning on BLEU, where the Kendall Reordering Score is specifically designed to measure improvement of reordering in isolation. For German-English more modest, statistically significant improvements of METEOR and TER (simultaneously) or BLEU (separately) are achieved. Our work differs from related approaches that use syntactic or part-of-speech information in the formation of reordering constraints in that it needs no such additional information. It also differs from related work on reordering constraints based on lexicalization in that it uses no such lexicalization but instead strives to achieve more globally coherent translations, afforded by global, holistic constraints that take the local reordering history of the derivation directly into account. Our experiments also once again reinforce the established wisdom that soft, rather than strict constraints, are a necessity when aiming to include new information to an already strong system without the risk of effectively worsening performance through constraints that have not been directly tailored to the data through a proper learning approach. While lexicalized constraints on reordering have proven to have great potential, un-lexicalized soft bilingual constraints, which are more general and transcend the rule level have their own place in providing another agenda of improving translation which focusses more on the global coherence direction by directly putting soft alignment-informed constraints on the combination of rules. Finally, while more research is necessary in this direction, there are strong reasons to believe that in the right setup these different approaches can be made to further reinforce each other.

Acknowledgements

This work is supported by The Netherlands Organization for Scientific Research (NWO) under grant nr. 612.066.929. The authors would like to thank Matt Post and Juri Ganitkevitch, for their support with respect to the integration of *Fuzzy Matching Decoding* into the Joshua codebase.

References

Alexandra Birch and Miles Osborne. 2010. Lrscore for evaluating lexical and reordering quality in mt. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 327–332.

Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. 2008. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224–232.

Stanley F. Chen and Joshua T. Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Computer Science Group, Harvard University.

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

David Chiang. 2006. An introduction to synchronous grammars.

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

David Chiang. 2010. Learning to translate with source and target syntax. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1443–1452.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: HLTTechnologies: Short Papers - Volume 2*, pages 176–181.

Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 85–91.

W. J. Dixon and A. M. Mood. 1946. The statistical sign test. *Journal of the American Statistical Association*, pages 557–566.

Andreas Eisele and Yu Chen. 2010. Multiun: A multilingual corpus from united nation documents. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2868–2872.

Juri Ganitkevitch, Yuan Cao, Jonathan Weese, Matt Post, and Chris Callison-Burch. 2012. Joshua 4.0: Packing, pro, and paraphrases. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 283–291, Montréal, Canada, June. Association for Computational Linguistics.

Greg Hanneman and Alon Lavie. 2013. Improving syntax-augmented machine translation by coarsening the label set. In *HLT-NAACL*, pages 288–297.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondrej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 177–180.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. A syntax-directed translator with extended domain of locality. In *Proceedings of the Workshop on Computationally Hard Problems and Joint Inference in Speech and Language Processing*, pages 1–8.

Matthias Huck, Joern Wuebker, Felix Rietig, and Hermann Ney. 2013. A phrase orientation model for hierarchical machine translation. In *ACL 2013 Eighth Workshop on Statistical Machine Translation*, pages 452–463.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 423–430.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proc. of MT Summit*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA.

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.

Gideon Maillette de Buy Wenniger and Khalil Sima'an. 2013. Hierarchical alignment decomposition labels for hiero grammar rules. In *Proceedings of the Seventh Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 19–28.

Yuval Marton, David Chiang, and Philip Resnik. 2012. Soft syntactic constraints for arabic—english hierarchical phrase-based translation. *Machine Translation*, 26(1-2):137–157.

Haitao Mi and Liang Huang. 2008. Forest-based translation rule extraction. In *Proceedings of EMNLP*.

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

Markos Mylonakis and Khalil Sima'an. 2011. Learning hierarchical translation structure with linguistic annotations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 642–652.

Markos Mylonakis. 2012. *Learning the Latent Structure of Translation*. Ph.D. thesis, University of Amsterdam.

ThuyLinh Nguyen and Stephan Vogel. 2013. Integrating phrase-based reordering features into a chart-based decoder for machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1587–1596.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, pages 160–167.

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.

Khalil Sima'an and Gideon Maillette de Buy Wenniger. 2013. Hierarchical alignment trees: A recursive factorization of reordering in word alignments with empirical results. Internal Report.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *In Proceedings of Association for Machine Translation in the Americas*, pages 223–231.

Jrg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2868–2872.

Christoph Tillmann. 2004. A unigram orientation model for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 101–104.

Ashish Venugopal, Andreas Zollmann, Noah A. Smith, and Stephan Vogel. 2009. Preference grammars: softening syntactic constraints to improve statistical machine translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–244.

Dekai Wu and Hongsing Wong Hkust. 1998. Machine translation with a stochastic grammatical channel. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, pages 1408–1415.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23:377–404.

Xinyan Xiao, Jinsong Su, Yang Liu, Qun Liu, and Shouxun Lin. 2011. An orientation model for hierarchical phrase-based translation. In *Proceedings of the 2011 International Conference on Asian Language Processing*, pages 165–168.

Hao Zhang, Daniel Gildea, and David Chiang. 2008. Extracting synchronous grammar rules from word-level alignments in linear time. In *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, pages 1081–1088.

Bowen Zhou, Bing Xiang, Xiaodan Zhu, and Yuqing Gao. 2008. Prior derivation models for formally syntax-based translation using linguistically syntactic parsing and tree kernels. In *Proceedings of the ACL-08: HLT Second Workshop on Syntax and Structure in Statistical Translation (SSST-2)*, pages 19–27.

Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In *NAACL 2006 - Workshop on statistical machine translation*, June.