Effective Incorporation of Source Syntax into Hierarchical Phrase-based Translation

Tong Xiao†, Adrià de Gispert‡, Jingbo Zhu†, Bill Byrne‡

†Northeastern University (CN) ‡University of Cambridge
Hierarchical Phrase-based vs. Tree-to-String

Hierarchical phrase-based models are popular but they typically require constraints
• span limit
• # of non-terminals
• no consecutive non-terminals (except glue rules)

Ref: After North Korea demanded concessions from U.S. again before the start of a new round of six-nation talks, ...
Hiero: In the new round of six-nation talks on North Korea again demanded that U.S. in the former promise concessions ...

this span (15) is larger than the span limit (10). X → ⟨ 在 X₁ 后, after X₁ ⟩ cannot be used.

Input: 在₁ 北韩₂ 再度₃ 要求₄ 美国₅ 于₆ 新₇ 回合₈ 六₉ 国₁₀ 会谈₁₁ 前₁₂ 承诺₁₃ 让步₁₄ 后₁₅ ,₁₆ ⋯
Hierarchical phrase-based vs. Tree-to-String

Hierarchical phrase-based models are popular but they typically require constraints
- span limit
- # of non-terminals
- no consecutive non-terminals (except glue rules)...

Tree-to-string models do not have such constraints but they can have problems with robustness
- quality of parse trees
- sparsity and limited coverage

Ref: After North Korea demanded concessions from U.S. again before the start of a new round of six-nation talks, ...

Hiero: In the new round of six-nation talks on North Korea again demanded that U.S. in the former promise concessions ...

Input: 在1 北韩2 再度3 要求4 美国5 于6 新7 回合8 六9 国10 会谈11 前12 承诺13 让步14 后15 16 ...

no span limit for constituents
Hierarchical Phrase-based vs. Tree-to-String

Ref: After North Korea demanded concessions from U.S. again before the start of a new round of six-nation talks, ... 

Hiero: In the new round of six-nation talks on North Korea again demanded that U.S. in the former promise concessions ... 

Hiero+t2s: After North Korea again demanded that U.S. promised concessions before the new round of six-nation talks ... 

Input: 在 北韩 再度 要求 美国 于 新回合 六九 国会 前 承诺 让步 后 ...
Overview

- **Our approach**: Introducing source language syntax into hierarchical phrase-based translation
  - **More rules**
    - Hiero-style rules + GHKM rules
  - **More syntax-sensitive derivations**
    - Decoding the input string and the parse tree
- **Good results**
  - Improvements of +0.8~1.2 BLEU on the NIST data
Rule Transformation and Categorization

(Hiero-style) SCFG rules

- string → string
- no syntactic labels

\[ X \rightarrow \langle \text{在 } X_1 \text{ 后, after } X_1 \rangle \]
Rule Transformation and Categorization

(Hiero-style) SCFG rules
- string → string
- no syntactic labels

X → ⟨在 X₁ 后, after X₁⟩

(GHKM-style) Tree-to-string rules
- tree-fragment → string
- syntactic labels

```
   VP
  /   \
PP  x₂:VP → x₂ with x₁
    /   \
  x₁:NP
```

```
P x₁:NP
```

Rule Transformation and Categorization

(Hiero-style) SCFG rules
- string → string
- no syntactic labels

\[ X \rightarrow \langle 在 X_1 后, after X_1 \rangle \]

(GHKM-style) Tree-to-string rules
- tree-fragment → string
- syntactic labels

\[ X \rightarrow \langle 对 X_1 X_2, X_2 with X_1 \rangle \] (consecutive non-terminals)
Rule Transformation and Categorization

(Hiero-style) SCFG rules
- string → string
- no syntactic labels

\[ X \rightarrow \langle \text{在 X}_1 \text{ 后}, \text{after X}_1 \rangle \]

Type 1 rules:
all rules from Hiero extraction

\[ X \rightarrow \langle \text{对 X}_1 \text{ X}_2, \text{X}_2 \text{ with X}_1 \rangle \]

Type 2 rules:
otherwise (only from GHKM)

(GHKM-style) Tree-to-string rules
- tree-fragment → string
- syntactic labels

\[ VP \rightarrow x_2:VP \rightarrow x_2 \text{ with } x_1 \]
\[ PP \rightarrow x_1:NP \]
\[ \text{对 } \rightarrow \text{对 X}_1 \text{ with X}_1 \]

(consecutive non-terminals)
**Hiero and GHKM Rule Extraction**

![Diagram of SCFG rules (Hiero)]

**SCFG rules (Hiero):**

- \( X \rightarrow \langle \text{他, he} \rangle \)
- \( X \rightarrow \langle \text{对, with} \rangle \)
- \( X \rightarrow \langle \text{回答, the answer} \rangle \)
- \( X \rightarrow \langle \text{表示 满意, was satisfied} \rangle \)
- \( X \rightarrow \langle X_1 \text{ 表示 满意, was satisfied} X_1 \rangle \)

...
Hiero and GHKM Rule Extraction

SCFG rules (Hiero):
\[ X \rightarrow \langle \text{他}, \text{he} \rangle \]
\[ X \rightarrow \langle \text{对}, \text{with} \rangle \]
\[ X \rightarrow \langle \text{回答}, \text{the answer} \rangle \]
\[ X \rightarrow \langle \text{表示 满意}, \text{was satisfied} \rangle \]
\[ X \rightarrow \langle X_1 \text{ 表示 满意,} \text{was satisfied} X_1 \rangle \]

Tree-to-string rules (GHKM):
\[ \text{NP(PN(他))} \rightarrow \text{he} \]
\[ \text{P(对)} \rightarrow \text{with} \]
\[ \text{NP(NN(回答))} \rightarrow \text{the answer} \]
\[ \text{VP(VV(表示) NN(满意))} \rightarrow \text{was satisfied} \]
\[ \text{VP}(x_1: \text{PP} \text{ VP(VV(表示) NN(满意)))} \rightarrow \text{was satisfied} x_1 \]
\[ \text{VP}(\text{PP}(\text{P(对) } x_1: \text{NP}) x_2: \text{VP}) \rightarrow \]
\[ x_2 \text{ with } x_1 \]
Hiero and GHKM Rule Extraction

SCFG rules (Hiero):

\[
\begin{align*}
X & \rightarrow \langle \text{他, he} \rangle \\
X & \rightarrow \langle \text{对, with} \rangle \\
X & \rightarrow \langle \text{回答, the answer} \rangle \\
X & \rightarrow \langle \text{表示 满意, was satisfied} \rangle \\
X & \rightarrow \langle X_1 \text{ 表示 满意, was satisfied } X_1 \rangle \\
X & \rightarrow \langle \text{对 } X_1 X_2, \text{ with } X_1 \rangle \\
\ldots
\end{align*}
\]

Tree-to-string rules (GHKM):

\[
\begin{align*}
\text{NP}(\text{PN(他)}) & \rightarrow \text{he} \\
\text{P(对)} & \rightarrow \text{with} \\
\text{NP}(\text{NN(回答)}) & \rightarrow \text{the answer} \\
\text{VP(VV(表示) NN(满意))} & \rightarrow \text{was satisfied} \\
\text{VP}(x_1:\text{PP VP(VV(表示) NN(满意))}) & \rightarrow \text{was satisfied } x_1 \\
\text{VP}(\text{PP(P(对) } x_1:\text{NP}) x_2:\text{VP}) & \rightarrow \text{with } x_1 \\
\ldots
\end{align*}
\]
**Hiero and GHKM Rule Extraction**

**A "Larger" SCFG:**

- Type 1: $X \rightarrow \langle \text{他}, \text{he} \rangle$
- Type 1: $X \rightarrow \langle \text{对}, \text{with} \rangle$
- Type 1: $X \rightarrow \langle \text{回答}, \text{the answer} \rangle$
- Type 1: $X \rightarrow \langle \text{表示 满意, was satisfied} \rangle$
- Type 1: $X \rightarrow \langle X_1 \text{ 表示 满意, was satisfied } X_1 \rangle$
- Type 2: $X \rightarrow \langle \text{对 } X_1 X_2, \text{ with } X_1 \rangle$

**Tree-to-string rules (GHKM):**

- NP(PN(他)) → he
- P(对) → with
- NP(NN(回答)) → the answer
- VP(VV(表示) NN(满意)) → was satisfied
- VP($x_1$:PP VP(VV(表示) NN(满意))) → was satisfied $x_1$
- VP(PP(P(对) $x_1$:NP) $x_2$:VP) → $x_2$ with $x_1$
- ...
Decoding Paradigm (HiFST)

HierSCFG \rightarrow \text{larger SCFG} \rightarrow t2s rules

input string\&tree \rightarrow \text{rule match} \rightarrow \text{decoding} \rightarrow output string
Decoding Paradigm (HiFST)

- **Hiero SCFG**
- **larger SCFG**
- **t2s rules**
- **input string & tree**
- **rule match**
- **decoding**
- **output string**

- CYK cells
- RTN
- WFSA

- rule match
- construct
- expand
- 1best path
Decoding Paradigm (HiFST)

HierSCFG \rightarrow \text{larger SCFG} \rightarrow \text{t2s rules} \rightarrow \text{decoding} \rightarrow \text{output string}

input string&tree \rightarrow \text{rule match} \rightarrow \text{decoding}

Modifying HiFST (this work)

Re-using the components of HiFST
Decoding with Type 1 Rules - String Matching

- **Type 1 rules are used as in standard Hiero style systems**
  - **constraint:** \(\text{span} \leq \text{span limit} (=10)\)

![Chart Cells]

X → ⟨X₁ 满意, satisfied with X₁⟩

阿都拉₁ 对₂ 自己₃ 四₄ 个₅ 多₆ 月₇ 以来₈ 的₉ 施政₁₀ 表现₁₁ 感到₁₂ 满意₁₃

4 words < span limit
Decoding with Type 1 Rules - String Matching

- Type 1 rules are used as in standard Hiero style systems
  - constraint: $span \leq \text{span limit} (=10)$
- This work: no span limit for tree constituents
Decoding with Type 1 Rules - String Matching

Type 1 rules are used as in standard Hiero style systems
  - constraint: $\text{span} \leq \text{span limit} (=10)$

This work: no span limit for tree constituents

Rules otherwise cannot be applied
Decoding with Type 2 Rules - Tree Matching

- Type 2 rule are applied via tree matching.
  - Tree-to-string rule $\rightarrow$ input tree fragments

Tree-to-string rule

$\text{VP(PP(P(对) } x_1 \text{:NP) } x_2 \text{:VP)}$

$\rightarrow x_2 \text{ with } x_1$

SCFG rule

$X \rightarrow \langle \text{对 X}_1 X_2, \text{X}_2 \text{ with X}_1 \rangle$

Chart Cells

Tree Matching
Features

- Baseline features (See Juan et al., 2013)
  - a 4-gram language model
  - a strong 5-gram language model
  - bidirectional translation probabilities
  - bidirectional lexical weights
  - a word count
  - a phrase count
  - a glue rule count
  - frequency-1/frequency-2/larger-than-frequency-2 rule counts

- Additional features
  - Rule type indicators
    - tree-to-string rules
    - lexicalized tree-to-string rules
    - rules with consecutive non-terminals
    - non-lexicalized rules.
  - Features in syntactic MT
    - phrase-based conditional translation probabilities
    - syntax-based conditional probabilities
Experiments

- NIST MT12 Chinese-English task
- **Baseline system**: HiFST (de Gispert et al. 2010)
  - [http://ucam-smt.github.io/](http://ucam-smt.github.io/)
  - one of the top academic systems in NIST MT12
  - Standard SCFG extraction, with span limit = 10
  - first past decoding with 4-gram LM + second pass rescoring with 5-gram LM

- Data
  - **Bilingual**: 9.2 million sentence pairs
  - **Tune**: 1,755 sentences (newswire)/2,160 sentences (web)
  - **Test**: all evaluation data of MT08-MT12
    - 1,779 sentences (newswire)/1,768 sentences (web)
  - **Monolingual**: all English data of MT12 + Google counts

- Tree-to-string Rules
  - GHKM extraction from 600K sentences
  - Discarding low probability rules
## BLEU Results

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune</td>
<td>Test</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(1755)</td>
<td>(1779)</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84</td>
<td>35.63</td>
</tr>
<tr>
<td>exp02</td>
<td>+= no span limit</td>
<td>36.05</td>
<td>35.79</td>
</tr>
</tbody>
</table>

* means significant different from exp01. += means incrementally adding methods/features to the previous system

- **Baseline and span limits** (exp01 and exp02)
  Relaxing span limits for tree constituents helps (a little)
## BLEU Results

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune (1755)</td>
<td>Test (1779)</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84</td>
<td>35.63</td>
</tr>
<tr>
<td>exp02</td>
<td>+= no span limit</td>
<td>36.05</td>
<td>35.79</td>
</tr>
<tr>
<td>exp03</td>
<td>+= t-to-s rules</td>
<td>36.63</td>
<td>36.25*</td>
</tr>
<tr>
<td>exp04</td>
<td>+= t-to-s features</td>
<td>36.82</td>
<td>36.38*</td>
</tr>
</tbody>
</table>

* means significant different from exp01. += means incrementally adding methods/features to the previous system

- **Baseline and span limits** (exp01 and exp02)
  Relaxing span limits for tree constituents helps (a little)

- **t2s + Hiero** (exp03 and exp04)
  Our method significantly improves the system
  - exp03: +0.6 and +0.5 BLEU improvements
  - exp04: +0.7 and +0.6 BLEU improvements
<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune 1755</td>
<td>Test 1779</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84 35.63</td>
<td>29.98 25.33</td>
</tr>
<tr>
<td>exp04</td>
<td>+= our approach</td>
<td>36.82 36.38*</td>
<td>30.91 25.98*</td>
</tr>
<tr>
<td>exp05</td>
<td>t-to-s baseline</td>
<td>34.63 34.25*</td>
<td>28.30 23.56*</td>
</tr>
</tbody>
</table>

* means significant different from exp01. += means incrementally adding methods/features to the previous system

- **Impact of search space (t-to-s rules only) (exp05)**
  Using a smaller space of hypotheses is harmful
## BLEU Results (2)

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune (1755)</td>
<td>Test (1779)</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84</td>
<td>29.98</td>
</tr>
<tr>
<td>exp04</td>
<td>+= our approach</td>
<td>36.82</td>
<td>30.91</td>
</tr>
<tr>
<td>exp05</td>
<td>t-to-s baseline</td>
<td>34.63</td>
<td>28.30</td>
</tr>
<tr>
<td>exp06</td>
<td>exp04 on spans &gt; 10</td>
<td>36.17</td>
<td>30.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35.63</td>
<td>25.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36.38*</td>
<td>30.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.25*</td>
<td>28.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>35.92</td>
<td>23.56*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>29.98</td>
<td>25.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.91</td>
<td>25.98*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>28.30</td>
<td>23.56*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.18</td>
<td>25.45</td>
</tr>
</tbody>
</table>

* means significant different from exp01. += means incrementally adding methods/features to the previous system

- **Impact of search space** (t-to-s rules only) (exp05)
  Using a smaller space of hypotheses is harmful
- **t2s + Hiero, span > 10 only** (exp06)
  t2s rules on smaller spans also help
### BLEU Results (2)

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune  (1755)</td>
<td>Test (1779)</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84</td>
<td>35.63</td>
</tr>
<tr>
<td>exp04</td>
<td>+= our approach</td>
<td>36.82</td>
<td>36.38*</td>
</tr>
<tr>
<td>exp05</td>
<td>t-to-s baseline</td>
<td>34.63</td>
<td>34.25*</td>
</tr>
<tr>
<td>exp06</td>
<td>exp04 on spans $&gt; 10$</td>
<td>36.17</td>
<td>35.92</td>
</tr>
<tr>
<td>exp07</td>
<td>exp04 with null trans.</td>
<td>36.10</td>
<td>35.42</td>
</tr>
</tbody>
</table>

* means significant different from exp01. += means incrementally adding methods/features to the previous system.

- **Impact of search space (t-to-s rules only)** (exp05)
  Using a smaller space of hypotheses is harmful.
- **t2s + Hiero, span $> 10$ only** (exp06)
  t2s rules on smaller spans also help.
- **Impact of failed parses** (exp07)
  Empty translations for which the parser fails.
  Our approach is more robust than standard t2s approaches.
## BLEU Results (3)

<table>
<thead>
<tr>
<th>Entry</th>
<th>System</th>
<th>Newswire</th>
<th>Web</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Tune</td>
<td>Test</td>
</tr>
<tr>
<td>exp01</td>
<td>Baseline</td>
<td>35.84</td>
<td>35.63</td>
</tr>
<tr>
<td>exp04</td>
<td>+= our approach</td>
<td>36.82</td>
<td>36.38*</td>
</tr>
<tr>
<td>exp08</td>
<td>exp04 + left binariz.</td>
<td>37.11</td>
<td>36.91*</td>
</tr>
<tr>
<td>exp09</td>
<td>exp04 + right binariz.</td>
<td>36.58</td>
<td>36.20*</td>
</tr>
<tr>
<td>exp10</td>
<td>exp04 + forest binariz.</td>
<td>37.03</td>
<td>36.98*</td>
</tr>
</tbody>
</table>

* means significant different from exp01.

### Results on Binarization (exp08-10)

- Left-binarization improves the system significantly
- Right-binarization is not very promising
- Additional trees introduced in binarization forest are not sufficient to make a big impact on BLEU scores
Rule Usage

<table>
<thead>
<tr>
<th>Type 1</th>
<th>Type 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>15%</td>
</tr>
<tr>
<td>85%</td>
<td>13%</td>
</tr>
<tr>
<td>87%</td>
<td>13%</td>
</tr>
</tbody>
</table>

Newswire
- Baseline (exp01): 100%
- Our approach (exp04): 85%

Web
- Baseline (exp01): 100%
- Our approach (exp04): 87%
Example Translation

A rule with three non-terminals (some of which are adjacent) perfectly performs a syntactic movement of the required tree constituents.

Reference: The Chinese star performance troupe presented a wonderful Peking opera as well as singing and dancing performance to Hong Kong audience.

Baseline: Star troupe of China, highlights of Peking opera and dance show to the audience of Hong Kong.

Hiero+t2s: Chinese star troupe presented a wonderful Peking opera singing and dancing to Hong Kong audience.

A tree-to-string rule is applied:

$$\left( \text{VP BA(将) } x_1: \text{NP} \ x_2: \text{VP PP(P(给) } x_3: \text{NP} \right) \rightarrow x_2 \ x_1 \ \text{to} \ x_3$$

Input:

```
中国_ 明星_ 艺术团_
```

```
将_ 台_ 精彩_ 的_ 京剧_ 歌舞
```

```
将_ 一_ 台_ 精彩_ 的_ 京剧_ 歌舞
```

```
BA
```

```
NP
```

```
VP
```

```
PP
```

```
P
```

```
NP
```

```
IP
```

```
香_ 港_ 观众
```

```
给
```

```
呈_ 现
```

```
香港_ 观众
```

```
中国_ 明星_ 艺术团
```

```
将_ 一_ 台_ 精彩_ 的_ 京剧_ 歌舞
```

```
中国_ 明星_ 艺术团
```

```
将_ 一_ 台_ 精彩_ 的_ 京剧_ 歌舞
```

```
中国_ 明星_ 艺术团
```

```
将_ 一_ 台_ 精彩_ 的_ 京剧_ 歌舞
```

```
中国_ 明星_ 艺术团
```

```
将_ 一_ 台_ 精彩_ 的_ 京剧_ 歌舞
```
Note

- **Differences** from system combination, e.g., joint decoding (Liu et al., 2009)
  - Our approach is much **simpler**: no need to change the decoding or its parameter optimisation
  - We introduce tree-to-string rules where they can contribute (using only 10% of the training data)

- **Differences** from soft/hard syntactic constraints (Marton and Resnik, 2008; Li et al. 2013)
  - These features do not work well in our system, which motivates this work

- **Differences** from using syntactic labels as non-terminals (Zollmann and Venugopal, 2006; Zhao and Al-Onaizan, 2008; Chiang, 2010; Hoang and Koehn, 2010)
  - They resort to Hiero-style extraction while we use GHKM
  - We consider the hierarchical syntactic tree structure which is ignore in SCFG-based systems
Summary

• We improve HiFST with a tree-to-string model
  ▶ Extracting more rules:
    Hiero extraction + GHKM
  ▶ Introducing more hypotheses into decoding:
    No span limit + decoding with source parse tree

• Our approach yields 0.8~1.2 BLEU improvements

• We plan to make it open source
Thanks!

HierO+t2s