Phrase-level System Combination for Machine Translation Based on Target-to-Target Decoding

Columbia University
Wei-Yun Ma
Kathleen McKeown
Motivation of Combination

• Many successful MT approaches with very different techniques.
• Want to take advantage of the individual strengths and avoid the individual weakness of them
Motivation of Phrase-level Combination

• In Translation, Phrase-based MT is useful: several words should be translated as a whole.
• Similarly, in combination, several words should be substituted as a whole
• we develop a new phrase-level lattice decoding approach instead of phrase-based re-decoding approach (Rosti et al. 2007, Chen et al. 2008)
  • To utilize syntactic structure (or word order) of the best MT output
Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Related Work: Confusion Network

(Matusov et al., 2006; He et al. 2008; Rosti et al. 2007; Leusch and Ney, 2010)

Sys1: I feel like fruit
Sys2: I prefer apples
Sys3: I am fond of apples

Select backbone

Sys1: I feel like fruit
Sys2: I prefer apples
Sys3: I am fond of apples

Get alignment (TERp or IHHH)

Sys1: I feel like fruit
Sys2: I prefer apples
Sys3: I am fond of apples

Build confusion network

I feel like fruit

prefer

am fond of apples

apples
I feel like fruit

Confusion Network considers too many hypotheses.

→ Sometimes several words should be substituted as a whole with several other words

Decoding using word consensus and LM

Considering:
I feel like of apples
I feel like of fruit
I feel like apples
I feel like fruit
I prefer apples
I prefer fruit
I feel prefer apples
I am fond apples
I feel prefer apples.
I like apples
...

I prefer apples
Outline

• Motivation
• Related Work
• **Basic Idea**
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Basic Idea

• Phrase-level Lattice decoding
  – Robust Paraphrase Extraction Strategy
    • Independent with alignment approaches
  – Combination using Log-linear-model
    • Various phrase scoring functions
    • Use soft syntactic constraints

• Target-to-Target Decoding
  – “Translation” from the best MT output (backbone) to the combination result
  – Any MT decoder can serve this mission:
    • Ex, using Moses, Paraphrase Lattice can be modeled as Phrase Table (Target-to-Target pairs), and input is backbone
  – Capability of reordering
Outline

• Motivation
• Relevant Work
• Basic Idea
• **Methodology**
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Select the backbone

• Features
  – Sentence-level consensus by using TER
  – A general LM
  – Length smoothing

\[
\log p(E_i) =
\sum_{s=1}^{N_s} \left( \lambda_s \log(1 - \text{TER}(E_i, E_s)) \right) + \lambda^l \log(LM(E_i)) + \lambda^w \cdot \text{Length}(E_i)
\]
Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Monolingual Word Alignment: 
TERp tool

\[ E_b : w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ w_6 \ w_7 \ w_8 \ w_9 \ w_{10} \ w_{11} \]
\[ E_h : \bar{w}_1 \ \bar{w}_2 \ \bar{w}_3 \ \bar{w}_4 \ \bar{w}_5 \ \bar{w}_6 \ \bar{w}_7 \ \bar{w}_8 \ \bar{w}_9 \ \bar{w}_{10} \]

TERp tool reorders \( E_h \) and generate word alignment :

\[ E_b : w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ \varepsilon \ \varepsilon \ [w_6 \ w_7 \ w_8] \ w_9 \ [w_{10} \ w_{11}] \]
\[ E_h : \bar{w}_2 \ \bar{w}_1 \ \bar{w}_3 \ \varepsilon \ \bar{w}_4 \ \bar{w}_5 \ \bar{w}_6 \ [\bar{w}_8 \ \bar{w}_7] \ \bar{w}_{10} \ [\bar{w}_9] \]

After we reorder \( E_h \) back to its original order :

\[ E_b : w_1 \ w_2 \ w_3 \ w_4 \ w_5 \ w_6 \ w_7 \ w_8 \ w_9 \ w_{10} \ w_{11} \]
\[ E_h : \bar{w}_1 \ \bar{w}_2 \ \bar{w}_3 \ \bar{w}_4 \ \bar{w}_5 \ \bar{w}_6 \ \bar{w}_7 \ \bar{w}_8 \ \bar{w}_9 \ \bar{w}_{10} \]
Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment

\[ E_b : w_1 \quad w_2 \quad w_3 \quad w_4 \quad w_5 \quad w_6 \quad w_7 \quad w_8 \quad w_9 \quad w_{10} \quad w_{11} \]

\[ E_h : \overline{w}_1 \quad \overline{w}_2 \quad \overline{w}_3 \quad \overline{w}_4 \quad \overline{w}_5 \quad \overline{w}_6 \quad \overline{w}_7 \quad \overline{w}_8 \quad \overline{w}_9 \quad \overline{w}_{10} \]
Paraphrase Extraction

• Extract all phrases that are word-continuous and consistent with the monolingual word alignment
Paraphrase Extraction

• Extract all phrases that are word-continuous and consistent with the monolingual word alignment
Paraphrase Extraction

• Extract all phrases that are word-continuous and consistent with the monolingual word alignment
Paraphrase Extraction

- Extract all phrases that are word-continuous and consistent with the monolingual word alignment

\[ E_b : \overline{w}_1 \overline{w}_2 \overline{w}_3 \overline{w}_4 \overline{w}_5 \overline{w}_6 \overline{w}_7 \overline{w}_8 \overline{w}_9 \overline{w}_{10} \overline{w}_{11} \]

\[ E_h : \overline{w}_1 \overline{w}_2 \overline{w}_3 \overline{w}_4 \overline{w}_5 \overline{w}_6 \overline{w}_7 \overline{w}_8 \overline{w}_9 \overline{w}_{10} \]
Paraphrase Extraction

• Extract all phrases that are word-continuous and consistent with the monolingual word alignment
Sys1: I feel like fruit
Sys2: I prefer apples

Sys1: I feel like fruit
Sys2: I am fond of apples

I feel like fruit
prefer apples

am fond of apples
prefer apples

Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Combination Model

• Phrase Scoring Functions
  – Paraphrase confidence scores (cs)
  – Lexical weighting (lex)
  – Syntactic indicators of whether paraphrases are syntactic constituents (syn)
  – Word and phrase penalty
  – Reordering model (r)
  – General language model
  – System-specific LMs for employing N-gram consensus information. (sl)
Combination Model

• Phrase Scoring Functions
  – Paraphrase confidence scores (cs)
  – Lexical weighting (lex)
  – Syntactic indicators of whether paraphrases are syntactic constituents (syn)
  – Word and phrase penalty
  – Reordering model (r)
  – General language model
  – System-specific LMs for employing N-gram consensus information. (sl)
What do you think about this phrase?
What do you think about this phrase?

Sounds good. I also have that in my output

What do you think about this phrase?

Paraphrase confidence scores
What do you think about this phrase?

Sounds ok. Although I don’t have that in my output, I found a lot of words in common.

What do you think about this phrase?

Lexical weighting
What do you think about this phrase?

Sounds great. I have that in my output, and this is a constituent in my output.

What do you think about this phrase?
Paraphrase confidence scores (cs)

\[
cs_s(\bar{e} | e) = \begin{cases} 
\frac{\text{MTYP# of }(e, \bar{e})}{\text{MTYP# of }(e, \bar{e}) + \text{IDS# of }(e, \bar{e})} & \text{if } (e, \bar{e}) \text{ can be extracted in system } s \\
0 & \text{otherwise}
\end{cases}
\]

Lexical weighting (lex)

\[
lex_s(\bar{e} | e) = \frac{\text{Common } \text{Word# of } \bar{e} \text{ and } A_s(e)}{|\bar{e}| + |A_s(e)|}
\]

Syntactic indicator (syn)

\[
syn_s(\bar{e}_i | e_i) = \begin{cases} 
1 & \text{if } (e, \bar{e}) \text{ can be extracted in system } s \text{ and } e \text{ and } \bar{e} \text{ are both syntactic constituents} \\
0 & \text{otherwise}
\end{cases}
\]
Consensus Model:

*Paraphrase confidence scores
*Lexical weighting

*Paraphrase confidence scores
*Lexical weighting
System-specific LM

Consensus Model:

\[
\begin{align*}
\text{phrase}_i & \quad \text{phrase}_{i+1} \\
\text{Paraphrase confidence scores} & \quad \text{Paraphrase confidence scores} \\
\text{Lexical weighting} & \quad \text{Lexical weighting}
\end{align*}
\]

We also want to model consensus between phrases.
System-specific LM

Consensus Model:

\[ \text{phrase}_i \quad \text{phrase}_{i+1} \]

*Paraphrase confidence scores
*Lexical weighting

We also want to model consensus between phrases

LM from sys1  LM from sys2  LM from sys3

For each single system, we build its system-specific LM based on the whole tuning/test corpus of all translation
\[
\log p(\overline{E} \mid E) = \sum_{i=1}^{I} \left( \sum_{s=1}^{N_s} \left( \lambda_{s}^{pc} \ast cs_s(\overline{e}_i \mid e_i) + \lambda_{s}^{lex} \ast lex_s(\overline{e}_i \mid e_i) + \lambda_{s}^{syn} \ast syn_s(\overline{e}_i \mid e_i) \right) \right) + \sum_{s=1}^{N_s} \left( \lambda_{s}^{sl} \ast \log(LM_s(\overline{E})) \right) + \sum_{i=1}^{I} \left( \lambda_{d}^{d} \ast d(start_i, end_{i-1}) \right) + \lambda_{l}^{l} \ast \log(LM(\overline{E})) + \lambda_{w}^{w} \ast length(\overline{E}) + \lambda_{p}^{p} \ast I
\]

Paraphrase confidence scores
Lexical weighting
Syntactic indicators of whether paraphrases are syntactic constituents
System-specific LM
Reordering model

Target-to-Target Decoding Outline

• Motivation
• Related Work
• Basic Idea
• Methodology
  – Select the backbone
  – Monolingual Word Alignment
  – Paraphrase Extraction
  – Combination Model
• Experiments
Two Environments

• Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)

• German-English combination shared task held by the WMT in 2011
Two Environments

- Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)
- German-English combination shared task held by the WMT in 2011
Chinese-English of NIST 2008

- Four human reference translations and corresponding machine translations for the NIST Open MT08 test sets
- Manually select Top5 systems out of 23 systems
- Tuning: 524 sentences
- Testing: 788 sentences
## Result of backbone selection

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>System 03</strong></td>
<td>30.16</td>
<td>63.04</td>
<td>51.94</td>
</tr>
<tr>
<td>System 15</td>
<td>30.06</td>
<td>62.82</td>
<td>51.80</td>
</tr>
<tr>
<td>System 20</td>
<td>28.15</td>
<td>65.39</td>
<td>49.72</td>
</tr>
<tr>
<td>System 22</td>
<td>29.94</td>
<td>63.19</td>
<td>51.51</td>
</tr>
<tr>
<td>System 31</td>
<td>29.52</td>
<td>61.70</td>
<td>51.89</td>
</tr>
<tr>
<td><strong>backbone</strong></td>
<td><strong>30.89</strong></td>
<td><strong>61.28</strong></td>
<td><strong>52.65</strong></td>
</tr>
</tbody>
</table>
# Impact of Feature Combination

<table>
<thead>
<tr>
<th>Feature Setting</th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 03</td>
<td>30.16</td>
<td>63.04</td>
<td>51.94</td>
</tr>
<tr>
<td>backbone</td>
<td>30.89</td>
<td>61.28</td>
<td>52.65</td>
</tr>
<tr>
<td>P+cs</td>
<td>31.74</td>
<td>60.11</td>
<td>53.59</td>
</tr>
<tr>
<td>P+cs+sl</td>
<td>32.63</td>
<td>60.49</td>
<td>53.53</td>
</tr>
<tr>
<td>P+cs+lex</td>
<td>31.81</td>
<td>60.32</td>
<td>53.53</td>
</tr>
<tr>
<td>P+cs+syn</td>
<td>31.74</td>
<td>60.22</td>
<td>53.55</td>
</tr>
<tr>
<td>P+cs+sl+lex+syn</td>
<td>32.85</td>
<td>60.32</td>
<td>53.76</td>
</tr>
</tbody>
</table>

P: phrase
W: word

Best feature setting: P+cs+sl+lex+syn

Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)
### Impact of Using Phrase and Word under no re-ordering

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 03</td>
<td>30.16</td>
<td>63.04</td>
<td>51.94</td>
</tr>
<tr>
<td>backbone</td>
<td>30.89</td>
<td>61.28</td>
<td>52.65</td>
</tr>
<tr>
<td><strong>W+cs</strong></td>
<td><strong>30.98</strong></td>
<td><strong>60.98</strong></td>
<td><strong>52.90</strong></td>
</tr>
<tr>
<td><strong>W+cs+sl</strong></td>
<td><strong>31.29</strong></td>
<td><strong>61.36</strong></td>
<td><strong>52.70</strong></td>
</tr>
<tr>
<td><strong>P+cs</strong></td>
<td><strong>31.74</strong></td>
<td><strong>60.11</strong></td>
<td><strong>53.59</strong></td>
</tr>
<tr>
<td><strong>P+cs+sl</strong></td>
<td><strong>32.63</strong></td>
<td><strong>60.49</strong></td>
<td><strong>53.53</strong></td>
</tr>
<tr>
<td>P+cs+lex</td>
<td>31.81</td>
<td>60.32</td>
<td>53.53</td>
</tr>
<tr>
<td>P+cs+syn</td>
<td>31.74</td>
<td>60.22</td>
<td>53.55</td>
</tr>
<tr>
<td>P+cs+sl+lex+syn</td>
<td>32.85</td>
<td>60.32</td>
<td>53.76</td>
</tr>
</tbody>
</table>

Under the same settings, phrase is always better than word

Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)
## Impact of Word Reordering

<table>
<thead>
<tr>
<th>Feature Setting</th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 03</td>
<td>30.16</td>
<td>63.04</td>
<td>51.94</td>
</tr>
<tr>
<td>backbone</td>
<td>30.89</td>
<td>61.28</td>
<td>52.65</td>
</tr>
<tr>
<td>P+r+cs</td>
<td>31.80</td>
<td>60.21</td>
<td>53.71</td>
</tr>
<tr>
<td>P+r+cs+sl</td>
<td>32.80</td>
<td>60.13</td>
<td>53.86</td>
</tr>
<tr>
<td>P+r+cs+lex</td>
<td>31.76</td>
<td>60.12</td>
<td>53.54</td>
</tr>
<tr>
<td>P+r+cs+syn</td>
<td>31.72</td>
<td>60.37</td>
<td>53.38</td>
</tr>
<tr>
<td>P+r+cs+sl+lex+syn</td>
<td>32.75</td>
<td>60.48</td>
<td>53.63</td>
</tr>
</tbody>
</table>

Best feature setting: P+cs+sl+lex+syn

Re-ordering (r)
Paraphrase confidence scores (cs)
Lexical weighting (lex)
Syntactic indicators (syn)
System-specific LMs (sl)
## Impact of Using Phrase and Word under re-ordering

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>System 03</td>
<td>30.16</td>
<td>63.04</td>
<td>51.94</td>
</tr>
<tr>
<td>backbone</td>
<td>30.89</td>
<td>61.28</td>
<td>52.65</td>
</tr>
<tr>
<td><strong>W+r+cs</strong></td>
<td>31.13</td>
<td>60.99</td>
<td>53.01</td>
</tr>
<tr>
<td><strong>W+r+cs+sl</strong></td>
<td>31.33</td>
<td>61.72</td>
<td>52.55</td>
</tr>
<tr>
<td><strong>P+r+cs</strong></td>
<td>31.80</td>
<td>60.21</td>
<td>53.71</td>
</tr>
<tr>
<td><strong>P+r+cs+sl</strong></td>
<td>32.80</td>
<td>60.13</td>
<td>53.86</td>
</tr>
<tr>
<td><strong>P+r+cs+lex</strong></td>
<td>31.76</td>
<td>60.12</td>
<td>53.54</td>
</tr>
<tr>
<td><strong>P+r+cs+syn</strong></td>
<td>31.72</td>
<td>60.37</td>
<td>53.38</td>
</tr>
<tr>
<td><strong>P+r+cs+sl+lex+syn</strong></td>
<td>32.75</td>
<td>60.48</td>
<td>53.63</td>
</tr>
</tbody>
</table>

Under the same settings, phrase is always better than word

Re-ordering (r)  
Paraphrase confidence scores (cs)  
Lexical weighting (lex)  
Syntactic indicators (syn)  
System-specific LMs (sl)
Two Environments

• Chinese-English of NIST 2008 (Selected Reference and System Translations-LDC2010T01)

• German-English combination shared task held by the WMT in 2011
German-English combination shared task (WMT 2011)

• One human reference translations and corresponding machine translations for the WMT 2011 test sets
• 10 combination system results are provided
• Manually select Top 6 systems out of 26 systems
• Tuning: 524 sentences
• Testing: 788 sentences
Result of backbone selection (WMT 2011)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>cmu-dyer</td>
<td>22.72</td>
<td>60.89</td>
<td>55.09</td>
</tr>
<tr>
<td>dfki-xu</td>
<td>22.44</td>
<td>62.31</td>
<td>53.89</td>
</tr>
<tr>
<td>kit</td>
<td>22.75</td>
<td>60.82</td>
<td>54.81</td>
</tr>
<tr>
<td>online-A</td>
<td>23.16</td>
<td>58.96</td>
<td>56.34</td>
</tr>
<tr>
<td><strong>online-B</strong></td>
<td><strong>24.27</strong></td>
<td><strong>57.89</strong></td>
<td><strong>56.93</strong></td>
</tr>
<tr>
<td>rwth-fre-c</td>
<td>21.86</td>
<td>62.82</td>
<td>53.46</td>
</tr>
<tr>
<td><strong>backbone</strong></td>
<td><strong>25.38</strong></td>
<td><strong>57.05</strong></td>
<td><strong>57.72</strong></td>
</tr>
</tbody>
</table>
## Result of Combination (WMT 2011)

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>TERp</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Online B</strong></td>
<td>24.27</td>
<td>57.89</td>
<td>56.93</td>
</tr>
<tr>
<td><strong>backbone</strong></td>
<td>25.38</td>
<td>57.05</td>
<td>57.72</td>
</tr>
<tr>
<td>koc-combo</td>
<td>23.41</td>
<td>61.83</td>
<td>54.08</td>
</tr>
<tr>
<td>quaero-combo</td>
<td>23.37</td>
<td>60.86</td>
<td>55.03</td>
</tr>
<tr>
<td>rwth-leusch-combo</td>
<td>25.62</td>
<td>57.44</td>
<td>57.20</td>
</tr>
<tr>
<td>jhu-combo</td>
<td>25.08</td>
<td>57.81</td>
<td>56.87</td>
</tr>
<tr>
<td>jhu-combo-contrastive</td>
<td>24.46</td>
<td>57.20</td>
<td>57.26</td>
</tr>
<tr>
<td><strong>bbn-combo</strong></td>
<td>26.73</td>
<td>56.13</td>
<td>58.30</td>
</tr>
<tr>
<td>cmu-heafield-combo</td>
<td>25.31</td>
<td>57.27</td>
<td>57.71</td>
</tr>
<tr>
<td>cmu-heafield-combo-contrastive</td>
<td>25.24</td>
<td>57.37</td>
<td>57.68</td>
</tr>
<tr>
<td>upv-prhlt-combo</td>
<td>24.65</td>
<td>59.25</td>
<td>56.24</td>
</tr>
<tr>
<td>uzh-combo</td>
<td>24.55</td>
<td>58.47</td>
<td>56.76</td>
</tr>
<tr>
<td><strong>P+r+cs+sl</strong></td>
<td>25.81</td>
<td>56.89</td>
<td>57.88</td>
</tr>
<tr>
<td><strong>P+cs+sl+lex+syn</strong></td>
<td>25.96</td>
<td>57.18</td>
<td>57.64</td>
</tr>
</tbody>
</table>

We try our best two settings, TTD is Top 2 out of 11 combination systems.
Discussion

• Under same feature setting, feature is better than word
• Effect of phrase confidence score and System-specific LM is significant
• Effect of Lexical weighting, syntactic indicator and reordering is not very significant
Conclusion

• A new phrase-level combination technique
  – A novel perspective: “translation” from one target (backbone) to another target (combination output)
  – Several phrase confidence estimations are presented, such as phrase confidence score, lexical weighting and syntactic indicator
  – Introduce the capability of word re-ordering
  – System-specific LM is proposed
Future Work

• Exploit information from the source
  – What do you think about this phrase?
    Ask the source.
Mrs. Source

What do you think about this phrase?

It seems ok. The phrase in either sys1 or sys3 preserves the semantics (relation) of the source.

What do you think about this phrase?
Future Work

• “Translation” from backbone to the combination result motivates that we can try other more comprehensive “translation” model than Moses
  – Ex: Hierarchical phrase-based model
    Syntax-oriented phrase-based model