Model with Minimal Translation Units, But Decode with Phrases

Nadir Durrani
University of Edinburgh

Joint Work With:

Alex Fraser and Helmut Schmid
Ludwigs Maximilian University of Munich
Introduction

- N-gram-based translation models learn Markov chains over Minimal Translation Units

- Why N-gram-based translation models are better than Phrase-based model?

- Why Phrase-based search is better than N-gram-based decoding framework?

- This Work: Combine N-gram-based Translation Model with Phrase-based Decoding

- Follow-up Work: Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? (ACL-13)
Phrase-based SMT

Lexical Step

Diesen Woche ist die grüne Hexe zu Haus

this week is the green witch at home

Reordering Step

the green witch is at home this week
Benefits of the Phrase-based Model

1. Local Reordering

Dann fliege ich nach Kanada
Then I will fly to Canada

1. Local Reordering

2. Discontinuities in Phrases

er hat ein Buch gelesen
he read a book

2. Discontinuities in Phrases

3. Idioms

in den sauren apfel beißen
to bite the bullet

3. Idioms

4. Insertions and Deletions

lesen Sie mit
read with me
Problems in Phrase-based Model

- Strong phrasal independences assumption

Example:

<table>
<thead>
<tr>
<th>Sie würden gegen sie stimmen</th>
</tr>
</thead>
<tbody>
<tr>
<td>They would vote against you</td>
</tr>
</tbody>
</table>
Problems in Phrase-based Model

- Strong phrasal independences assumption

- Can not capture the dependency between parts of verb predicate
  - würden ... stimmen – vote against
Problems in Phrase-based Model

- Spurious Phrasal Segmentation

\[
\begin{align*}
\text{Sie würden} & \quad \text{gegen Ihre Kampagne stimmen} \\
\text{They would} & \quad \text{vote against your campaign}
\end{align*}
\]

\[
\begin{align*}
\text{Sie} & \quad \text{würden} & \quad \text{gegen Ihre Kampagne} & \quad \text{stimmen} \\
\text{They} & \quad \text{would} & \quad \text{vote} & \quad \text{against my foreign policy}
\end{align*}
\]
Problems in Phrase-based Model

• Long Distance Reordering

Sie würden gegen Ihre Kampagne stimmen
They would vote against your campaign
Problems in Phrase-based Model

• Unable to perform long distance reordering

Sie würden gegen Ihre Kampagne stimmen

They would vote against your campaign

die menschen würden gegen meine außenpolitik stimmen

People would vote against my foreign policy

9
N-gram-based SMT
(Marino et. al 2006)

Sie würden gegen Sie stimmen
They would vote against you
N-gram-based SMT

Sie würden gegen Sie stimmen

They would vote against you

Minimal Translation Units (MTUs)

$t_1$: Sie $\rightarrow$ They
$t_2$: würden $\rightarrow$ would
$t_3$: stimmen $\rightarrow$ vote
$t_4$: gegen $\rightarrow$ against
$t_5$: Sie $\rightarrow$ you
N-gram-based SMT

Sie würden gegen Sie stimmen

They would vote against you

$\text{Sie} \quad \text{würden} \quad \text{gegen Sie stimmen}$

$\text{They} \quad \text{would} \quad \text{vote against you}$
N-gram-based SMT

Sie würden gegen Sie stimmen
They would vote against you

\[ p(t_1) \times p(t_2 | t_1) \]
N-gram-based SMT

Sie würden gegen Sie stimmen

They would vote against you

$\Pr(t_1) \times \Pr(t_2 \mid t_1) \times \Pr(t_3 \mid t_1 t_2)$
N-gram-based SMT

Sie würden gegen Sie stimmen

They would vote against you

\[ P(t_1) \times P(t_2 | t_1) \times P(t_3 | t_1 t_2) \times P(t_4 | t_1 t_2 t_3) \]
Model

- Joint model over sequences of minimal units

\[ p_{tsm}(F, E, A) = p(t_1^J) = \prod_{j=1}^{J} p(t_j | t_{j-n+1}, \ldots, t_{j-1}) \]

Context Window: 4-gram Model
Operation Sequence Model (OSM)
(Durrani et. al 2011)

• Generate the MTUs as a sequence of operations
• Lexical operation or reordering operation
• Learn a Markov model over the sequence of operations
Example

Er würde gegen Sie stimmen

He would vote against you
Example

\[
\begin{align*}
\text{Er würde gegen Sie stimmen} & \\
\text{He would vote against you}
\end{align*}
\]

- Operations
  - \( o_1 \): Generate (Er – He)
Example

Er würde gegen Sie stimmen

He would vote against you

- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
Example

Er würde gegen Sie stimmen
He would vote against you

• Operations
  – $o_1$ Generate (Er, He)
  – $o_2$ Generate (würde, would)
  – $o_3$ Insert Gap
Example

Er würde gegen Sie stimmen

He would vote against you

- Operations
  - $o_1$ Generate (Er, He)
  - $o_2$ Generate (würde, would)
  - $o_3$ Insert Gap
  - $o_4$ Generate (stimmen, vote)
Example

Er würde gegen Sie stimmen
He would vote against you

• Operations
  – $o_1$ Generate (Er, He)
  – $o_2$ Generate (würde, would)
  – $o_3$ Insert Gap
  – $o_4$ Generate (stimmen, vote)
  – $o_5$ Jump Back (1)
Example

Er würde **gegen** Sie stimmen

He would vote **against** you

• Operations
  – \( o_1 \) Generate (Er, He)
  – \( o_2 \) Generate (würde, would)
  – \( o_3 \) Insert Gap
  – \( o_4 \) Generate (stimmen, vote)
  – \( o_5 \) Jump Back (1)
  – \( o_6 \) Generate (gegen, against)
Example

Er würde gegen Sie stimmen

He would vote against you

- $o_1$ Generate (Er, He)
- $o_2$ Generate (würde, would)
- $o_3$ Insert Gap
- $o_4$ Generate (stimmen, vote)
- $o_5$ Jump Back (1)
- $o_6$ Generate (gegen, against)
- $o_7$ Generate (Sie, you)
Model

- Joint-probability model over operation sequences

\[
p_{osm}(F, E, A) = p(o_1^J) = \prod_{j=1}^{J} p(o_j | o_{j-n+1}, \ldots, o_{j-1})
\]

Context Window: 9-gram Model
Benefits of N-gram-based SMT

• Considers both contextual information across phrasal boundaries
  – Source and target context
  – Reordering context

• Does not have spurious ambiguity in the model

• Memorizes reordering patterns
  – Consistently handles local and non-local reorderings
### MTU-based versus Phrase-based Search

<table>
<thead>
<tr>
<th>Stack 1</th>
<th>Stack 2</th>
<th>Stack 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td></td>
<td>Wie heisst du</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is your name</td>
</tr>
</tbody>
</table>

- Decoding with phrases
  - Access to infrequent translations such as “Wie – What is”
  - Covers multiple source words in a single step
  - Better future cost estimates
MTU-based versus Phrase-based Search

Stack 1: S

Stack 2: Wie, as, such as, like, what is

Stack 3: Wie heisst du, What is your name

Ranked 124th
Future Cost Estimation

Source: \( F_1 \), \( F_2 \), \( F_3 \), \( F_4 \), \( F_5 \), \( F_6 \), …, \( F_k \)

- Better Future Cost Estimates using Phrases
  - Contextual information is available
Future Cost Estimation for Translation Model

Wie heisst du

What is your name

Generate (Wie – what is) → Insert Gap → Generate (du – your) → Jump Back (1) → Generate (heisst – name)
Future Cost Estimation for Translation Model

Wie heisst du

What is your name

Generate (Wie – what is) $\rightarrow$ Insert Gap $\rightarrow$ Generate (du – your) $\rightarrow$ Jump Back (1) $\rightarrow$ Generate (heisst – name)

**MTU-based FC estimate:** $p($Generate (Wie – what is)$) \times p($Generate (du – your)$) \times p($Generate (heisst – name)$)$
Future Cost Estimation for Translation Model

Wie heisst du

What is your name

Generate (Wie – what is) → Insert Gap → Generate (du – your) → Jump Back (1) → Generate (heisst – name)

**MTU-based FC estimate:** \[ p(\text{Generate (Wie – what is)}) \times p(\text{Generate (du – your)}) \times p(\text{Generate (heisst – name)}) \]

**Phrase-based FC estimate:** \[ p(\text{Generate (Wie – what is)}) \times p(\text{Insert Gap | Context}) \times p(\text{Generate (du – your) | Context}) \times ... \times p(\text{Generate (heisst – name) | Context}) \]
Improving Search through Phrases

• Use phrase pairs rather than MTUs for decoding

• Load the MTU-based decoding with
  – i) Future cost estimated from phrases
  – ii) Infrequent translation options from phrasal units
Search Accuracy

- Hold the model constant and change the search
  - Tune the parameters on dev
  - For test change the search to use
    - OSM MTU decoder (baseline)
    - OSM MTU decoder (using information extracted from phrases)
    - OSM Phrase-based decoder

- Compute search accuracy
  - Create a list of best scoring hypotheses $h_0$ by selecting 1 hypothesis per test sentence
  - Compute the percentage of hypotheses contributed from each decoding run into $h_0$
Search Accuracy

Best Scoring Hyps $h_0$

1: --------------- -12.4
2: --------------- -74.2
3: --------------- -24.8
          ---------------
n: --------------- -45.1

1: --------------- -10.4
2: --------------- -74.2
3: --------------- -23.8
          ---------------
n: --------------- -44.6

1: --------------- -10.4
2: --------------- -75.1
3: --------------- -25.8
          ---------------
n: --------------- -45.1

1: --------------- -12.4
2: --------------- -74.2
3: --------------- -23.8
          ---------------
n: --------------- -45.5
Results on WMT-09b (Search Accuracy)

MTU = Baseline decoder (uses minimal translation units)
MTU.FC = Future cost estimated from phrases
MTU.FC.T = Translation list is appended with missing translations
PB = Phrase-based Decoder
Results on WMT-09b (BLEU)

MTU = Baseline decoder (uses minimal translation units)
MTU.FC = Future cost estimated from phrases
MTU.FC.T = Translation list is appended with missing translations
PB = Phrase-based Decoder
Experimental Setup

- Language Pairs: DE-EN, FR-EN, ES-EN

- Baseline System
  - OSM Cept-based Decoder (Stack Size: 500)
  - Moses (Koehn et. al 2007)
  - Phrasal (Galley and Manning 2010)
  - NCode (Crego et. al 2011)

- Data
  - 8th Version of the Europarl Corpus
  - Bilingual Data: Europarl + News Commentary ~2M Sent
  - Monolingual Data: 22M Sentences News08
  - Standard WMT08 set for tuning WMT[09-12] for test
German-to-English

- Significant improvements over all baselines

<table>
<thead>
<tr>
<th>Test</th>
<th>Moses</th>
<th>Phrasal</th>
<th>Ncode</th>
<th>$O_{\text{mtu}}$</th>
<th>$O_{\text{phrase}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT09</td>
<td>20.47*</td>
<td>20.78*</td>
<td>20.52</td>
<td>21.17*</td>
<td>21.47</td>
</tr>
<tr>
<td>WMT10</td>
<td>21.37*</td>
<td>21.91*</td>
<td>21.53</td>
<td>22.29*</td>
<td>22.73</td>
</tr>
<tr>
<td>WMT11</td>
<td>20.40*</td>
<td>20.96*</td>
<td>20.21</td>
<td>21.05*</td>
<td>21.43</td>
</tr>
<tr>
<td><strong>Avg</strong></td>
<td>20.77</td>
<td>21.18</td>
<td>20.76</td>
<td>21.47</td>
<td>21.90</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>+1.13</td>
<td>+0.72</td>
<td>+1.14</td>
<td>+0.43</td>
<td></td>
</tr>
</tbody>
</table>
French-to-English

- Significant improvement in 11/16 cases

<table>
<thead>
<tr>
<th>Test</th>
<th>Moses</th>
<th>Phrasal</th>
<th>Ncode</th>
<th>$\text{OS}_{\text{mtu}}$</th>
<th>$\text{OS}_{\text{phrase}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT09</td>
<td>25.78*</td>
<td>25.87*</td>
<td>26.15*</td>
<td>26.22*</td>
<td>26.51</td>
</tr>
<tr>
<td>WMT11</td>
<td>27.37*</td>
<td>27.62*</td>
<td>27.46*</td>
<td>27.75</td>
<td>27.91</td>
</tr>
<tr>
<td>WMT12</td>
<td>27.15*</td>
<td>27.76</td>
<td>27.55*</td>
<td>27.66*</td>
<td>27.98</td>
</tr>
<tr>
<td>Avg</td>
<td>26.74</td>
<td>27.03</td>
<td>27.01</td>
<td>27.05</td>
<td>27.32</td>
</tr>
<tr>
<td>$\Delta$</td>
<td>+0.58</td>
<td>+0.29</td>
<td>+0.31</td>
<td>+0.27</td>
<td></td>
</tr>
</tbody>
</table>
Spanish-to-English

- Significant improvement in 12/16 cases

<table>
<thead>
<tr>
<th>Test</th>
<th>Moses</th>
<th>Phrasal</th>
<th>Ncode</th>
<th>OS\textsubscript{mtu}</th>
<th>OS\textsubscript{phrase}</th>
</tr>
</thead>
<tbody>
<tr>
<td>WMT09</td>
<td>25.90*</td>
<td>26.13</td>
<td>25.91*</td>
<td>25.90*</td>
<td>26.18</td>
</tr>
<tr>
<td>WMT10</td>
<td>28.91*</td>
<td>28.89*</td>
<td>29.02*</td>
<td>28.82*</td>
<td>29.37</td>
</tr>
<tr>
<td>WMT11</td>
<td>28.84*</td>
<td>28.98*</td>
<td>28.93*</td>
<td>28.95*</td>
<td>29.66</td>
</tr>
<tr>
<td>WMT12</td>
<td>31.28*</td>
<td>31.57</td>
<td>31.42</td>
<td>30.86*</td>
<td>31.52</td>
</tr>
<tr>
<td>Avg</td>
<td>28.73</td>
<td>28.89</td>
<td>28.82</td>
<td>28.63</td>
<td>29.18</td>
</tr>
<tr>
<td>(\Delta)</td>
<td>+0.45</td>
<td>+0.29</td>
<td>+0.36</td>
<td>+0.55</td>
<td></td>
</tr>
</tbody>
</table>
Summary

• Model based on minimal units
  – Do not make phrasal independence assumption
  – Avoid spurious segmentation ambiguities

• Phrase-based search provides better
  – Future cost estimation
  – Translation coverage
  – Robust performance at lower beams

• Model based on MTUs + Phrase-based Decoding
  – Improve both search accuracy and BLEU
  – Information in phrases can be used indirectly to improve MTU-based decoding

• Statistically Significant improvements over state-of-the-art Phrase-based and N-gram-based baseline systems
Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT? (ACL 2013)
Can Markov Models Over Minimal Translation Units Help Phrase-Based SMT?

<table>
<thead>
<tr>
<th>LP</th>
<th>WMT-12</th>
<th></th>
<th></th>
<th>WMT-13</th>
<th></th>
<th></th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Moses</td>
<td>+OSM</td>
<td>∆</td>
<td>Moses</td>
<td>+OSM</td>
<td>∆</td>
<td>Moses</td>
</tr>
<tr>
<td>es-en</td>
<td>34.02</td>
<td>34.51</td>
<td>0.49*</td>
<td>30.04</td>
<td>30.94</td>
<td>0.9*</td>
<td>32.03</td>
</tr>
<tr>
<td>en-es</td>
<td>33.87</td>
<td>34.44</td>
<td>0.57*</td>
<td>29.66</td>
<td>30.10</td>
<td>0.44*</td>
<td>31.77</td>
</tr>
<tr>
<td>fr-en</td>
<td>30.72</td>
<td>30.96</td>
<td>0.24</td>
<td>31.06</td>
<td>31.46</td>
<td>0.40</td>
<td>30.89</td>
</tr>
<tr>
<td>en-fr</td>
<td>28.76</td>
<td>29.36</td>
<td>0.6*</td>
<td>30.03</td>
<td>30.39</td>
<td>0.36*</td>
<td>29.40</td>
</tr>
<tr>
<td>cs-en</td>
<td>22.70</td>
<td>23.03</td>
<td>0.33*</td>
<td>25.70</td>
<td>25.79</td>
<td>0.09</td>
<td>24.2</td>
</tr>
<tr>
<td>en-cs</td>
<td>15.81</td>
<td>16.16</td>
<td>0.35*</td>
<td>18.35</td>
<td>18.62</td>
<td>0.27</td>
<td>17.08</td>
</tr>
<tr>
<td>ru-en</td>
<td>31.87</td>
<td>32.33</td>
<td>0.46*</td>
<td>24.0</td>
<td>24.33</td>
<td>0.33</td>
<td>27.94</td>
</tr>
<tr>
<td>en-ru</td>
<td>23.75</td>
<td>24.11</td>
<td>0.36*</td>
<td>18.44</td>
<td>18.84</td>
<td>0.4*</td>
<td>21.1</td>
</tr>
<tr>
<td>Avg.</td>
<td>27.69</td>
<td>28.11</td>
<td>0.43</td>
<td>25.91</td>
<td>26.31</td>
<td>0.4</td>
<td>26.80</td>
</tr>
</tbody>
</table>
Learned Pattern

- **Operations**
  - Generate (würde, would)
  - Insert Gap
  - Generate (stimmen, vote)

- **Can generalize to**
  - die menschen würden gegen meine außenpolitik stimmen