Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model

Kay Rottmann (UKA), Stephan Vogel (CMU)

September 7, 2007
1 Motivation
   - Word Order Problem
   - Current Approaches
   - Goals

2 The Model
   - Using POS Information
   - Learning the Rules
   - Application of the Rules
   - Reordering of Training Corpus

3 Experiments
   - Setup
   - Results

4 Conclusion

5 Translation Examples
Problem of Word Order

- Different languages differ in word order
Problem of Word Order

- Different languages differ in word order
- Differences within small context

<table>
<thead>
<tr>
<th>Example: ADJ NN → NN ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>An important agreement</td>
</tr>
<tr>
<td>Un acuerato importante</td>
</tr>
</tbody>
</table>
Problem of Word Order

- Different languages differ in word order
- Differences within small context
  
<table>
<thead>
<tr>
<th>Example: ADJ NN → NN ADJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>An important agreement</td>
</tr>
<tr>
<td>Un acuerto importante</td>
</tr>
</tbody>
</table>

- Long range reorderings

<table>
<thead>
<tr>
<th>Example: auxiliary verb and infinite verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ich werde morgen nachmittag ... ankommen</td>
</tr>
<tr>
<td>I will arrive tomorrow afternoon ...</td>
</tr>
</tbody>
</table>
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB\textsuperscript{+}05] ...
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB+05] . . .
- Reordering of source sentence [ChCF06], [PoNe06], [CrMa06]
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB+05] ...
- Reordering of source sentence [ChCF06], [PoNe06], [CrMa06]
  - Reordering before translation process
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB\textsuperscript{+}05] . . .
- Reordering of source sentence [ChCF06], [PoNe06], [CrMa06]
  - Reordering before translation process
  - monotone decoding
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB+05] . . .
- Reordering of source sentence [ChCF06], [PoNe06], [CrMa06]
  - Reordering before translation process
  - monotone decoding
  - more than one word order coded in lattice structure
Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB+05] ...
- Reordering of source sentence [ChCF06], [PoNe06], [CrMa06]
  - Reordering before translation process
  - monotone decoding
  - more than one word order coded in lattice structure

⇒ our work based on this approach
Goals

- Restriction of search to make it fast
Goals

- Restriction of search to make it fast
- Correct reorderings in different contexts
Goals

- Restriction of search to make it fast
- Correct reorderings in different contexts
- Better translations of long range reorderings
How the System works

- Reorderings based on rules extracted prior to translation from corpus
How the System works

- Reorderings based on rules extracted prior to translation from corpus
- Use of POS-Tags for generalization
  - POS-Tagger are available for many languages
How the System works

- Reorderings based on rules extracted prior to translation from corpus
- Use of POS-Tags for generalization
  - POS-Tagger are available for many languages
- Assign probabilities to rules
  - as a guide for the decoding process
How the System works

- Reorderings based on rules extracted prior to translation from corpus
- Use of POS-Tags for generalization
  - POS-Tagger are available for many languages
- Assign probabilities to rules
  - as a guide for the decoding process
- Create a lattice with possible reorderings
How the System works

- Reorderings based on rules extracted prior to translation from corpus
- Use of POS-Tags for generalization
  - POS-Tagger are available for many languages
- Assign probabilities to rules
  - as a guide for the decoding process
- Create a lattice with possible reorderings
- Decoder finds best monotone translation path through the lattice
What is a Rule

- A rule consists of three parts:
  - Left hand side: Sequence of POS on the source side
What is a Rule

- A rule consists of three parts:
  - Left hand side: Sequence of POS on the source side
  - Right hand side: Permutation on that word order
What is a Rule

A rule consists of three parts:
- Left hand side: Sequence of POS on the source side
- Right hand side: Permutation on that word order
- Score for the rule: Relative frequency
What is a Rule

- A rule consists of three parts:
  - Left hand side: Sequence of POS on the source side
  - Right hand side: Permutation on that word order
  - Score for the rule: Relative frequency

Example: ADJ NN → 1 0 : 0.72
Context Dependency of Rules

- Left hand side is the POS-Sequence that needs to be reordered
Context Dependency of Rules

- Left hand side is the POS-Sequence that needs to be reordered
- Problem: different reorderings for the same POS sequence

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td>He will come.</td>
<td>Er wird kommen.</td>
<td></td>
</tr>
<tr>
<td>He says that he will come.</td>
<td>Er sagt, dass er kommen wird.</td>
<td></td>
</tr>
</tbody>
</table>
Context Dependency of Rules

- Left hand side is the POS-Sequence that needs to be reordered.
- Problem: different reorderings for the same POS sequence.

\[
\begin{align*}
&\text{He will come.} \\
&\text{Er wird kommen.} \\
&\text{He says that \textit{he will come}.} \\
&\text{Er sagt, dass \textit{er kommen wird}.}
\end{align*}
\]

- Idea: Use more complex left hand side that indicates the context ⇒
  - Usage of POS-Tags to the left and / or right of sequence
  - Usage of words to the left and / or right of sequence
  - Usage of words as the sequence
**Example Rules with Context Information**

<table>
<thead>
<tr>
<th>source sequence</th>
<th>rule</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.60</td>
</tr>
<tr>
<td>VVFIN :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.71</td>
</tr>
<tr>
<td>moechte :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table: Example rules for German to English translation with no context, with one tag of context to the left and one word of context to the left.
Example Rules with Context Information

<table>
<thead>
<tr>
<th>source sequence</th>
<th>rule</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.60</td>
</tr>
<tr>
<td>VVFIN :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.71</td>
</tr>
<tr>
<td>moechte :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table: Example rules for German to English translation with no context, with one tag of context to the left and one word of context to the left

- "Ich moechte diese Gelegenheit nutzen, …"
Example Rules with Context Information

<table>
<thead>
<tr>
<th>source sequence</th>
<th>rule</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.60</td>
</tr>
<tr>
<td>VVFIN :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.71</td>
</tr>
<tr>
<td>moechte :: PDAT NN VVINF</td>
<td>3 1 2</td>
<td>0.92</td>
</tr>
</tbody>
</table>

Table: Example rules for German to English translation with no context, with one tag of context to the left and one word of context to the left

- ”Ich moechte diese Gelegenheit nutzen , . . .”
- becomes ”Ich moechte nutzen diese Gelegenheit , . . .”
Learning the Rules

- Use aligned corpus with a tagged source side
- whenever there is a crossing of alignments in a sentence
- store rules for different context types and count them
Learning the Rules

- Use aligned corpus with a tagged source side
- whenever there is a crossing of alignments in a sentence
- store rules for different context types and count them
- But only if the rule occurs without being part of a larger reordering that will be learned
  - This reduces the number of rules - allows longer reorderings without getting problems in decoding time
  - Significant rules will still be extracted
Learning the Rules

- Use aligned corpus with a tagged source side
- Whenever there is a crossing of alignments in a sentence
- Store rules for different context types and count them
- But only if the rule occurs without being part of a larger reordering that will be learned
  - This reduces the number of rules - allows longer reorderings without getting problems in decoding time
  - Significant rules will still be extracted
- Compute relative frequency for every rule
Learning the Rules

- Use aligned corpus with a tagged source side
- whenever there is a crossing of alignments in a sentence
- store rules for different context types and count them
- But only if the rule occurs without being part of a larger reordering that will be learned
  - This reduces the number of rules - allows longer reorderings without getting problems in decoding time
  - Significant rules will still be extracted
- Compute relative frequency for every rule
- Throw away rules seen less than a given threshold
Building the Lattice (Basics)

- Start with monotone path of the sentence, weight of every edge $= 1.0$
Building the Lattice (Basics)

- Start with monotone path of the sentence, weight of every edge $= 1.0$
- Test for subsequences of the sentence, if a rule for that exists
  - Start with longest subsequences
  - adjust score of first edge according to monotone path
  - before testing rules that are shorter adjust score for monotone path
Building the Lattice (Basics)

- Start with monotone path of the sentence, weight of every edge = 1.0
- Test for subsequences of the sentence, if a rule for that exists
  - Start with longest subsequences
  - Adjust score of first edge according to monotone path
  - Before testing rules that are shorter adjust score for monotone path
- BUT: This works only for one rule type!
Building the Lattice (Advanced)

- For more rule types: Combination is needed
Building the Lattice (Advanced)

- For more rule types: Combination is needed
- Use of all individual scores is bad
  - Same reorderings get different scores because of context
  - Scores will contradict each other
  - Optimization will lead to a preferred single type
Building the Lattice (Advanced)

- For more rule types: Combination is needed
- Use of all individual scores is bad
  - Same reorderings get different scores because of context
  - Scores will contradict each other
  - Optimization will lead to a preferred single type
  - ⇒ For same reorderings use max score of all rule types
- For monotone Path:
  - use minimum score over all individual scores for the monotone path
Phrases from reordered corpus were shown to perform better [PoNe06]

Idea: phrases match the situation in the lattice better than before
Phrases from reordered corpus were shown to perform better [PoNe06]

Idea: phrases match the situation in the lattice better than before

Question: How should the training be corpus reordered?

Usage of alignment information to monotonize alignment
  new alignment should be nearly monotone
Phrases from reordered corpus were shown to perform better [PoNe06]

Idea: phrases match the situation in the lattice better than before

Question: How should the training be corpus reordered?

Usage of alignment information to monotonize alignment
  - new alignment should be nearly monotone

Usage of the rules to reorder corpus
  - better fits the decoding situation
Setup

- English → Spanish (TC-Star 07)
  - Training Corpus: Europarl Corpus 33M Words
  - Development Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References
**Setup**

- **English → Spanish (TC-Star 07)**
  - Training Corpus: Europarl Corpus 33M Words
  - Development Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References

- **German ↔ English (WMT 06)**
  - Training Corpus: Europarl Corpus 34M Words
  - Development Set: 2K Sentences / (306 / 62) OOV
  - Test Set: 2K Sentences / (551 / 250) OOV
  - 1 Reference
Setup

- English → Spanish (TC-Star 07)
  - Training Corpus: Europarl Corpus 33M Words
  - Development Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References

- German ↔ English (WMT 06)
  - Training Corpus: Europarl Corpus 34M Words
  - Development Set: 2K Sentences / (306 / 62) OOV
  - Test Set: 2K Sentences / (551 / 250) OOV
  - 1 Reference

- Brill Tagger for English (36 Tags)
- Stuttgart Tree-Tagger for German (57 Tags)
Combination of all Ruletypes

- Addition of different context types to the rules

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (RO3)</td>
<td>48.51</td>
<td>17.69</td>
<td>23.70</td>
</tr>
<tr>
<td>no Context</td>
<td>49.52</td>
<td>17.78</td>
<td>24.79</td>
</tr>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
</tbody>
</table>
Combination of all Ruletypes

- Addition of different context types to the rules

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (RO3)</td>
<td>48.51</td>
<td>17.69</td>
<td>23.70</td>
</tr>
<tr>
<td>no Context</td>
<td>49.52</td>
<td>17.78</td>
<td>24.79</td>
</tr>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
</tbody>
</table>

- Why is further improvement sometimes so low?
Combination of all Ruletypes

- Addition of different context types to the rules

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (RO3)</td>
<td>48.51</td>
<td>17.69</td>
<td>23.70</td>
</tr>
<tr>
<td>no Context</td>
<td>49.52</td>
<td>17.78</td>
<td>24.79</td>
</tr>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
</tbody>
</table>

- Why is further improvement sometimes so low?
  - Spanish and English Translations already very good
Combination of all Ruletypes

- Addition of different context types to the rules

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline(RO3)</td>
<td>48.51</td>
<td>17.69</td>
<td>23.70</td>
</tr>
<tr>
<td>no Context</td>
<td>49.52</td>
<td>17.78</td>
<td>24.79</td>
</tr>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
</tbody>
</table>

- Why is further improvement sometimes so low?
  - Spanish and English Translations already very good
  - AND: Phrases did not match lexical reorderings anymore

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>no Lexical Reorderings</td>
<td>49.83</td>
<td>18.21</td>
<td>24.88</td>
</tr>
</tbody>
</table>
Reordering of Source Corpus

- Reordering via GIZA++ alignment information
Reordering of Source Corpus

- Reordering via GIZA++ alignment information

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
<tr>
<td>no Lex Reorderings</td>
<td>49.83</td>
<td>18.21</td>
<td>24.88</td>
</tr>
<tr>
<td>all Rules GIZA++</td>
<td>49.78</td>
<td>18.23</td>
<td>24.09</td>
</tr>
</tbody>
</table>

Reordering via GIZA++ did not help for us!
Phrases do not match decoding situation

Rule Reordering

Kay Rottmann (UKA), Stephan Vogel (CMU)

Word Reordering in Statistical Machine Translation with a POS
Reordering of Source Corpus

- Reordering via GIZA++ alignment information

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
<tr>
<td>no Lex Reorderings</td>
<td>49.83</td>
<td>18.21</td>
<td>24.88</td>
</tr>
<tr>
<td>all Rules GIZA++</td>
<td>49.78</td>
<td>18.23</td>
<td>24.09</td>
</tr>
</tbody>
</table>

- Reordering via GIZA++ did not help for us!
  - Phrases do not match decoding situation
Reordering of Source Corpus

- Reordering via GIZA++ alignment information

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combination</td>
<td>49.58</td>
<td>18.27</td>
<td>24.85</td>
</tr>
<tr>
<td>no Lex Reorderings</td>
<td>49.83</td>
<td>18.21</td>
<td>24.88</td>
</tr>
<tr>
<td>all Rules GIZA++</td>
<td>49.78</td>
<td>18.23</td>
<td>24.09</td>
</tr>
</tbody>
</table>

- Reordering via GIZA++ did not help for us!
  - Phrases do not match decoding situation

- Reordering: Most probable word order according to Reordering Rules

<table>
<thead>
<tr>
<th>System</th>
<th>en → es</th>
<th>en → de</th>
<th>de → en</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule Reordering</td>
<td>49.75</td>
<td>18.42</td>
<td>25.06</td>
</tr>
</tbody>
</table>

Kay Rottmann (UKA), Stephan Vogel (CMU)
Conclusion

- Addition of context leads to improved translation quality
Addition of context leads to improved translation quality

BUT: some context types help for some languages, some hurt performance for other languages
Addition of context leads to improved translation quality

BUT: some context types help for some languages, some hurt performance for other languages

Reordering source side of training corpus before phrase extraction can help
Conclusion

- Addition of context leads to improved translation quality
- BUT: some context types help for some languages, some hurt performance for other languages
- Reordering source side of training corpus before phrase extraction can help
- BUT: reordered corpus has to be similar to decoding situation
Addition of context leads to improved translation quality

BUT: some context types help for some languages, some hurt performance for other languages

Reordering source side of training corpus before phrase extraction can help

BUT: reordered corpus has to be similar to decoding situation

$\approx 1.3$ improvement on English to Spanish
Addition of context leads to improved translation quality

BUT: some context types help for some languages, some hurt performance for other languages

Reordering source side of training corpus before phrase extraction can help

BUT: reordered corpus has to be similar to decoding situation

\( \approx 1.3 \) improvement on English to Spanish

\( \approx 0.7 \) improvement on English to German
Conclusion

- Addition of context leads to improved translation quality
- BUT: some context types help for some languages, some hurt performance for other languages
- Reordering source side of training corpus before phrase extraction can help
- BUT: reordered corpus has to be similar to decoding situation
- $\approx 1.3$ improvement on English to Spanish
- $\approx 0.7$ improvement on English to German
- $\approx 1.4$ improvement on German to English
Translation Examples

- German Source: bessere Erkenntnisse und moderne Technik bieten die Chance, die Umwelt in Europas Städten zu verbessern.
German Source: bessere Erkenntnisse und moderne Technik bieten die Chance, die Umwelt in Europas Städten zu verbessern.

Baseline: better knowledge and modern technology offer the chance of the environment in Europe’s cities to improve.
German Source: bessere Erkenntnisse und moderne Technik bieten die Chance, die Umwelt in Europas Städten zu verbessern.

- Baseline: better knowledge and modern technology offer the chance of the environment in Europe’s cities to improve.
- Combination: better knowledge and modern technology offers the opportunity to improve the urban environment in Europe.
The Lattice
Future Work

- Test on other language pairs (Arabic, Japanese, Farsi...)

Future Work

- Test on other language pairs (Arabic, Japanese, Farsi...)

Kay Rottmann (UKA), Stephan Vogel (CMU)
Future Work

- Test on other language pairs (Arabic, Japanese, Farsi...)
- Additional internal reordering
Future Work

- Test on other language pairs (Arabic, Japanese, Farsi...)
- Additional internal reordering
- Long range reorderings (more general)
Future Work

- Test on other language pairs (Arabic, Japanese, Farsi...)
- Additional internal reordering
- Long range reorderings (more general)
- Dealing with languages without reliable POS-Tagger (using word clustering techniques)
Thank you for your attention
A maximum entropy approach to natural language processing.

B. Chen, M. Cettolo und M. Federico.
Reordering rules for phrase-based statistical machine translation.

Josep M. Crego und Jose B. Marino.
Reordering Experiments for N-Gram-Based SMT.
In *Spoken Language Technology Workshop, Palm Beach, Aruba, 2006*, S. 242–245.

Edinburgh system description for the 2005 IWSLT speech translation evaluation.


M. Popovic und H. Ney.

POS-based word reorderings for statistical machine translation.

In *Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC)*, Genoa, Italy, 2006. S. 1278.

D. Wu.

A polynomial-time algorithm for statistical machine translation.