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

Kay Rottmann (UKA), Stephan Vogel (CMU)

September 7, 2007

Problem of Word Order

- Different languages differ in word order

Problem of Word Order

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

Example: ADJ NN $\rightarrow$ NN ADJ
An important agreement
Un acuerto importante

Problem of Word Order

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

Example: ADJ NN $\rightarrow$ NN ADJ
An important agreement
Un acuerto importante

- Long range reorderings

Example: auxiliary verb and infinite verb
Ich *werde* morgen nachmittag ... *ankommen*
I *will arrive* tomorrow afternoon ...

Current Approaches

- IBM constraints [BePP96], ITG [Wu96], lexicalised block oriented model [KAMCB$^+$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$^+$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
- $\Rightarrow$ 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 probabilies 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 probabilies 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 probabilies 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 $\rightarrow$ 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

He will come.
Er wird kommen.
He says that *he will come.*
Er sagt, dass *er kommen wird.*

Context Dependency of Rules

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

He will come.
Er wird kommen.
He says that *he will come.*
Er sagt, dass *er kommen wird.*

- Idea: Use more complex left hand side that indicates the context $\Rightarrow$
  - 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

source sequence	rule	freq.
PDAT NN VVINF	3 1 2	0.60
VVFIN :: PDAT NN VVINF	3 1 2	0.71
moechte :: PDAT NN VVINF	3 1 2	0.92

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

source sequence	rule	freq.
PDAT NN VVINF	3 1 2	0.60
VVFIN :: PDAT NN VVINF	3 1 2	0.71
moechte :: PDAT NN VVINF	3 1 2	0.92

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

source sequence	rule	freq.
PDAT NN VVINF	3 1 2	0.60
VVFIN :: PDAT NN VVINF	3 1 2	0.71
moechte :: PDAT NN VVINF	3 1 2	0.92

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
- $\Rightarrow$ For same reorderings use max score of all rule types
- For monotone Path:
  - use minimum score over all individual scores for the monotone path

Reordering of the Training Corpus

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

Reordering of the Training Corpus

- 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

Reordering of the Training Corpus

- 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 $\rightarrow$ Spanish (TC-Star 07)
  - Training Corpus: Europarl Corpus $33M$ Words
  - Developement Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References

Setup

- English $\rightarrow$ Spanish (TC-Star 07)
  - Training Corpus: Europarl Corpus 33M Words
  - Developement Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References
- German $\leftrightarrow$ English (WMT 06)
  - Training Corpus: Europarl Corpus 34M Words
  - Developement Set: 2K Sentences / (306 / 62) OOV
  - Test Set: 2K Sentences / (551 / 250) OOV
  - 1 Reference

Setup

- English $\rightarrow$ Spanish (TC-Star 07)
  - Training Corpus: Europarl Corpus 33M Words
  - Developement Set: 1.2K Sentences / 79 OOV
  - Test Set: 1.1K Sentences / 105 OOV
  - 2 References
- German $\leftrightarrow$ English (WMT 06)
  - Training Corpus: Europarl Corpus 34M Words
  - Developement 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

System	en → es	en → de	de → en
Baseline(RO3)	48.51	17.69	23.70
no Context	49.52	17.78	24.79
Combination	49.58	18.27	24.85

Combination of all Ruletypes

- Addition of different context types to the rules

System	en → es	en → de	de → en
Baseline(RO3)	48.51	17.69	23.70
no Context	49.52	17.78	24.79
Combination	49.58	18.27	24.85

- Why is further improvement sometimes so low?

Combination of all Ruletypes

- Addition of different context types to the rules

System	en → es	en → de	de → en
Baseline(RO3)	48.51	17.69	23.70
no Context	49.52	17.78	24.79
Combination	49.58	18.27	24.85

- 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

System	en → es	en → de	de → en
Baseline(RO3)	48.51	17.69	23.70
no Context	49.52	17.78	24.79
Combination	49.58	18.27	24.85

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

System	en → es	en → de	de → en
no Lexical Reorderings	49.83	18.21	24.88

Reordering of Source Corpus

- Reordering via GIZA++ alignment information

Reordering of Source Corpus

- Reordering via GIZA++ alignment information

System	en → es	en → de	de → en
Combination	49.58	18.27	24.85
no Lex Reorderings	49.83	18.21	24.88
all Rules GIZA++	49.78	18.23	24.09

Reordering of Source Corpus

- Reordering via GIZA++ alignment information

System	en → es	en → de	de → en
Combination	49.58	18.27	24.85
no Lex Reorderings	49.83	18.21	24.88
all Rules GIZA++	49.78	18.23	24.09

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

Reordering of Source Corpus

- Reordering via GIZA++ alignment information

System	en → es	en → de	de → en
Combination	49.58	18.27	24.85
no Lex Reorderings	49.83	18.21	24.88
all Rules GIZA++	49.78	18.23	24.09

- Reordering via GIZA++ did not help for us!
  - Phrases do not match decoding situation
- Reordering: Most probable word order according to Reordering Rules

System	en → es	en → de	de → en
Rule Reordering	49.75	18.42	25.06

Conclusion

- Addition of context leads to improved translation quality

Conclusion

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

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

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

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

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

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 Staedten zu verbessern .

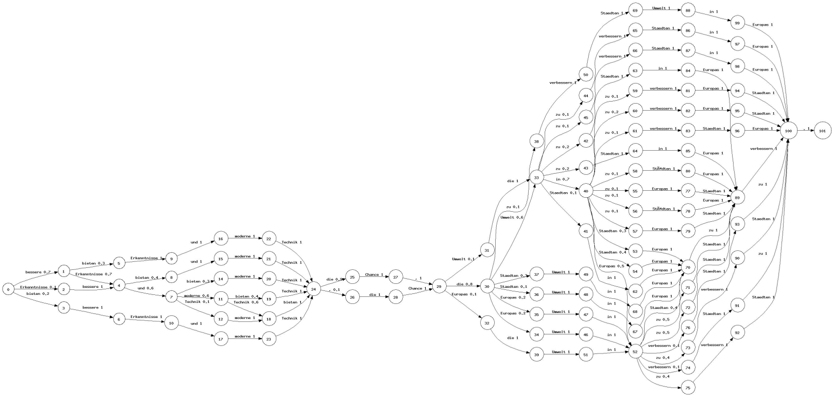
Translation Examples

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

Translation Examples

- German Source: bessere Erkenntnisse und moderne Technik bieten die Chance , die Umwelt in Europas Staedten 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...)

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)

The End

Thank you for your attention


A. L. Berger, S. A. Della Pietra und V. J. Della Pietra.
A maximum entropy approach to natural language processing.
Computational Linguistics, 22(1), 1996, S. 39.


B. Chen, M. Cettolo und M. Federico.
Reordering rules for phrase-based statistical machine translation.
In *Int. Workshop on Spoken Language Translation Evaluation Campaign on Spoken Language Translation*, 2006, S. 1–15.


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.


P. Koehn, A. Axelrod, A. B. Mayne, C. Callison-Burch, M. Osborne und D. Talbot.

Edinburgh system description for the 2005 IWSLT speech translation evaluation.
In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, PA, 2005.

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.
Proc. 34th Annual Meeting of the Assoc. for Computational Linguistics, 1996, S. 152.