Advances in Statistical Machine Translation: Phrases, Noun Phrases and Beyond

Philipp Koehn
koehn@isi.edu

Information Sciences Institute
Department of Computer Science
University of Southern California
Outline

● Statistical Machine Translation

● Phrase-Based Methods

● Syntactic Structure
  – Noun Phrase Translation
  – Clause Structure
Machine Translation

- Translating text in a foreign language into English
- One of the oldest problems in Artificial Intelligence
- AI-hard: reasoning and world knowledge required
- State of the art:

The United States and India May Will Be Held in the Past 40 Years the First Joint Military Exercises

(Afp report from new Delhi) India and U. S. will be held in the past 39 years the first joint military exercises in the world’s two biggest democracies the cooperative relationship between making milestone.
The Defense Ministry said in a class Indian paratrooper Brigade mid-May and the US Pacific Command of the special units in the well-known far and near the Thai women Maha tomb near joint military exercises.
The two countries will provide air support.

(Chinese-English, statistical machine translation system)
Statistical Machine Translation

- Learn translation from parallel text
  
  Den Vorschlag zur Steuererhöhung lehnte die Kommission ab.

  The proposal for the tax hike was rejected by the commission.

- Currently corpora for many language pairs available
  
  - up to 20-200 million words
  - e.g., Europarl [http://www.isi.edu/~koehn/europarl/],
    11 European languages, 20 million words each
Phrase-Based Methods

- Currently best performing methods map phrases

```
Den Vorschlag zur Steuererhoehung lehnte die Kommission ab .
```

```
The proposal for the tax hike was rejected by the commission .
```

- “Phrases”
  - any sequences of words
  - reordering of phrases possible

```
Den Vorschlag lehnte die Kommission ab .
```

```
The commission rejected the proposal .
```
Phrase Translation Table

Phrase Translations for “den Vorschlag”:

| English          | \(\phi(e|f)\) | English          | \(\phi(e|f)\) |
|------------------|---------------|------------------|---------------|
| the proposal     | 0.6227        | the suggestions  | 0.0114        |
| ’s proposal      | 0.1068        | the proposed     | 0.0114        |
| a proposal       | 0.0341        | the motion       | 0.0091        |
| the idea         | 0.0250        | the idea of      | 0.0091        |
| this proposal    | 0.0227        | the proposal ,   | 0.0068        |
| proposal         | 0.0205        | its proposal     | 0.0068        |
| of the proposal  | 0.0159        | it               | 0.0068        |
| the proposals    | 0.0159        | ...              | ...           |
Noisy Channel Model

- Bayes rule: $p(e|f) \sim p(f|e) \cdot p(e)$
Decoding

- Translation ("decoding") is NP-complete [Knight, 1999]

⇒ Various heuristic search methods
  - dynamic programming beam search [Och et al., 2001]
  - greedy search [Marcu and Wong, 2002]
  - finite state transducers [Kumar and Byrne, 2003]

- My decoder “Pharaoh” freely available (soon)
Learning Phrase Translation Tables

- Comparison of various methods [Koehn et. al, 2003]
  - directly aligning phrases in parallel corpora
  - aided by word alignments
  - syntactic constraints

![Graph showing BLEU scores for different training corpus sizes](chart.png)
Syntax and Statistical MT

- Why syntax?
  - many transformations can be best explained in syntactic terms
  - syntactic annotation on the foreign input adds additional knowledge
  - syntactic annotation on the English output aids grammatical output

- Machine Translation Pyramid
Previous Syntax-Based Transfer Models

- Various attempts as using syntax:
  - string to tree translation [Yamada and Knight, 2001]
  - using syntactic chunks [Schafer and Yarowsky, 2003]
  - loosely tree-based [Gildea, 2003]
  - syntactic features [Koehn and Knight, 2003] [Och et al., 2003]

- None showed significant improvement over phrase models
Syntactically Structured Statistical MT

Focus on two main syntactic categories:

- **Noun phrases**
  - contain most of vocabulary
  - can be translated in separation
  - subject of my PhD thesis [Koehn, 2003]

- **Clauses**
  - syntactic restructuring not well captured by phrase models
  - ongoing work
Noun Phrase Translation

- **Definition NP/PP**
  - the maximal noun phrases and prepositional phrases attached to the clause level
  - do not contain relative clauses

- **Examples**
  - *(The proposal for the tax hike)* was rejected *(by the commission)*.

- Cover roughly half of all words, all nouns, most of vocabulary
Translatability

- Study on German-English Europarl corpus
- Are NP/PPs translated as NP/PPs?
  - 75% are translated, 98% can be
  - exceptions
    - merge with verb: *make an observation*
    - PP translated as adverb: *in der Hauptsache = mainly*
- Translation in Isolation?
  - human translation w/o sentence context
  - 89% NP/PPs correctly translated
  - 9% wrong leading preposition
  - 2% wrong content word
Framework

- NP/PPs translated by modular subsystem
Translation as Reranking

- Base model proposes candidate
- Reranking with additional features
  - maximum entropy
  - similar to [Och and Ney, 2002]
Translation as Reranking: Why Possible?

- 60% of NP/PPs translated correctly
- 90% of NP/PPs have correct translation in 100-best list
- Advantage of reranking: global features
## Error Analysis: Not in n-Best List

<table>
<thead>
<tr>
<th>Error</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unknown Word</td>
<td>34%</td>
</tr>
<tr>
<td>Tagging or parsing error</td>
<td>28%</td>
</tr>
<tr>
<td>Unknown translation</td>
<td>14%</td>
</tr>
<tr>
<td>Complex syntactic restructuring</td>
<td>7%</td>
</tr>
<tr>
<td>Too long</td>
<td>6%</td>
</tr>
<tr>
<td>Untranslatable</td>
<td>2%</td>
</tr>
<tr>
<td>Other</td>
<td>9%</td>
</tr>
</tbody>
</table>

- 10% of NP/PPs: no acceptable translation in list
- Main problem: unknown words, translations
Special Modeling for NP/PP Translation

- Compound splitting
- Web n-Grams
- Syntactic features
## Compound Splitting

- **Compounding occurs in German, Finnish, Greek, ...**
  - increased vocabulary size
  - leads to sparse data problems and unknown words

- **Frequency based method for compound splitting**
  - break up, if parts are more frequent than whole
  - geometric mean: \( S_{\text{best}} = \arg\max_S (\prod_{p_i \in S} \text{count}(p_i))^{\frac{1}{n}} \)

- More detail in [Koehn and Knight, EACL 2003]
Web n-Grams

- Web as language model
  - 3 billion documents indexed by Google
  - 97% of trigrams seen on web
  - 30% of 7-grams seen on web

- Occurrence on web as feature
  - does phrase occur on web?
  - do all its n-grams occur on web? (n=2...7)
  - at least once? at least ten times?
  - using Google to collect counts
Syntactic Features

- Keep foreign syntactic parse tree
- Annotate English candidate translation with syntax
Syntactic Features (2)

- Given two syntactic parse trees

⇒ Any computable property between pair can be feature

- We implemented three features
  - preservation of number of a noun (singular stays singular)
  - preservation of preposition (no dropping of preposition, except if there is movement)
  - number agreement in baseNPs (not: \textit{this nice green flowers})

- Many more conceivable
Evaluation

- German-English translation task
- Europarl corpus
  - extracted NP/PP pairs using Giza++ for word alignment, Collins’ parser for English and Lopar parser for German
- Training
  - base model trained on 743,370 aligned NP/PPs
  - feature values trained on 683 NP/PPs
- Test set
  - 1362 NP/PPs from 534 sentences
### Accuracy (Human Judgment)

<table>
<thead>
<tr>
<th>System</th>
<th>NP/PP Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word-Based Model</td>
<td>724</td>
</tr>
<tr>
<td>Phrase-Based Model</td>
<td>800</td>
</tr>
<tr>
<td>Compound Splitting</td>
<td>838</td>
</tr>
<tr>
<td>Re-Estimated Parameters</td>
<td>858</td>
</tr>
<tr>
<td>Web Count Features</td>
<td>881</td>
</tr>
<tr>
<td>Syntactic Features</td>
<td>892</td>
</tr>
</tbody>
</table>

- Overall +12.3% improvement
- 95% Statistical significance interval 2.5%
### Error Analysis: Reranking Failure

<table>
<thead>
<tr>
<th>Error</th>
<th>Freq.</th>
<th>Error</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Involving content words</td>
<td>44%</td>
<td>Involving only function words</td>
<td>56%</td>
</tr>
<tr>
<td>Wrong word choice</td>
<td>16%</td>
<td>Wrong phrase start</td>
<td>38%</td>
</tr>
<tr>
<td>Content word mistranslated</td>
<td>4%</td>
<td>Internal preposition choice</td>
<td>4%</td>
</tr>
<tr>
<td>Wrong phrase choice</td>
<td>3%</td>
<td>Pronoun / anaphora</td>
<td>4%</td>
</tr>
<tr>
<td>Content dropped</td>
<td>13%</td>
<td>Pronoun added or dropped</td>
<td>3%</td>
</tr>
<tr>
<td>Content added</td>
<td>2%</td>
<td>Determiner added, dropped, wrong</td>
<td>2%</td>
</tr>
<tr>
<td>Number of noun wrong</td>
<td>2%</td>
<td>Function word phrase choice</td>
<td>2%</td>
</tr>
<tr>
<td>Other</td>
<td>2%</td>
<td>Function word mistranslated or dropped</td>
<td>2%</td>
</tr>
<tr>
<td>Reordering wrong</td>
<td>1%</td>
<td>Preposition dropped</td>
<td>1%</td>
</tr>
<tr>
<td>Other</td>
<td>1%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- 25% of NP/PPs: acceptable translation in list, not picked
- Main problems: wrong phrase start, word choice, dropping of content
Integration

- Translations passed to full sentence translation system
  - using XML markup
  - allow passing of reranked list (with probabilities)
Evaluation of Integration

- Performance on full-sentence translation (BLEU score)

<table>
<thead>
<tr>
<th>System</th>
<th>Word-Based MT</th>
<th>Phrase-Based MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline system</td>
<td>0.176</td>
<td>0.220</td>
</tr>
<tr>
<td>with NP/PP subsystem</td>
<td>0.199</td>
<td>0.224</td>
</tr>
</tbody>
</table>

- Why little improvement for phrase-based MT?
  - cuts around NP/PP disable overlapping phrase translations
  - parsing errors force hard decisions
Conclusions on NP/PP Translation

- It is possible to separate out NP/PP translation
- Improved NP/PP translation performance
- Improved overall sentence translation performance
  - still needs better integration
  - still needs better conditioning on sentence context
Clause Level

● Phrase-based system mistranslated the example:
  – German: *Den Vorschlag zur Steuererhöhung lehnt die Kommission ab.*
  – English: *The proposal for the tax hike is rejected by the commission.*
  – Phrase-MT: *The tax increase proposal opposes the commission.*

● Why?
  – semantic reasons: proposals usually don’t reject, they get rejected
  – syntactic reason: *Den Vorschlag* is accusative case, therefore object

● Syntactic information is ignored by phrase-based system
  – only indicated by the determiner *den*
Clause Level Transformations

- Required: clause level transformation
  
  \( G: \text{object} - \text{verb} - \text{subject} \)
  
  \( \Rightarrow \text{subject} - \text{verb} - \text{object} \)
  
  \( \Rightarrow \text{object} - \text{passive} \text{ verb} - \text{subject} \)

- There are more such transformations
  
  - some systematic due to different syntax in foreign and English
  
  - some driven by verbs whose translation has different subcategorization

- How do we add this to statistical machine translation?
  
  - future work...
Conclusion

- Reviewed phrase-based MT
- Syntactic structure
  - improved translation quality by special handling of noun phrases
  - clause level transformations as research challenge
Thank You!

- Questions?