The Statistical Machine Translation System of the University of Edinburgh

Philipp Koehn
pkoehn@inf.ed.ac.uk

School of Informatics
University of Edinburgh

Outline

- **Overview: SMT at Edinburgh**
- Baseline System
- Improvements
- Evaluation
- Related Recent Work in SMT
People Working On SMT at Edinburgh

- Philipp Koehn (lecturer)
- Miles Osborne (lecturer)
- Amittai Axelrod (graduate student)
- Alexandra Birch Mayne (graduate student)
- Chris Callison-Burch (graduate student, Linear-B)
- David Talbot (graduate student)
- Michael White (researcher)

MT Eval 2005 Effort

- 3-month effort building on previous work at MIT
  - improved system performance
  - introduced other researchers to the system
- Focus on Arabic-English:
  - deal with more data
  - various feature improvements

⇒ It is never finished...
  - did not train on new data
  - some changes not completed on time
Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
- Evaluation
- Related Recent Work in SMT

Phrase-Based Translation

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

- Phrase model similar to other groups’ model
  - word align corpus, using GIZA++ and Och’s refined method
  - collect phrase pairs consistent with word alignment
  - log-linear model to combine model components
  - parameter tuning by minimum error rate training
  - decoder Pharaoh (http://www.isi.edu/licensed-sw/pharaoh/)
System Components

- reordering model linear reordering cost, max. 4 word movement
- language model trigram LM trained using SRILM toolkit
- phrase translation model f→e
- phrase translation model e→f
- word translation model f→e
- word translation model e→f
- word penalty
- phrase penalty

Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
  - more training data (+2% BLEU)
  - bigger language model (+2% BLEU)
  - minor model improvements (+2% BLEU)
- Evaluation
- Related Recent Work in SMT
More Training Data

- All of the data (instead of half)
  - maximum sentence length 40 words
  - break up corpus in 2-3 parts
  - run snt2cooc separately, merge
  - combined GIZA++ run (3-5 days CPU time)

- Chunking

- Splitting

---

**Philipp Koehn, University of Edinburgh**

The Statistical Machine Translation System of the University of Edinburgh

**Chunking**

- Break up along comma, semicolon, colon, etc.

- Sentence-align smaller units

- 63.9 → 100.3 million words used
Splitting

- Break up longer sentences
  - minimum number of crossed word alignments
  - cut sentences in the middle third
  - cut as central as possible

- 100.3 → 130.3 million words used

Splitting II

- Aligned sentences using lexical t-table with $p > 0.03$ threshold,
  eliminate multiple aligned words
Splitting III

- Good and bad (2 crossings) split points

Splitting IV

- Quality of split points in the middle third
Splitting V

- Find most central best split point

---

Bigger Language Model

- Dealing with memory limitations in training
- Dealing with memory limitations in decoding
- Multiple language models
Memory Limitations in Training

- A lot of monolingual English text is available
  - English half of parallel text: 130 million words
  - English gigaword corpus: 1.78 billion words
  - the web: 1 trillion words?

- SRILM training keeps all n-grams in memory (2-4 GB limit)

- Practically limited to:
  - 800 million words (training + part of Gigaword)
  - ignored trigram singletons
  - digits ('0'-9') replaced by '5'

---

Memory Limitations in Decoding

- Pruning possible?
  - only need to consider words that can be produced
  - translation model can be cut down to a few (1-2) percent

<table>
<thead>
<tr>
<th></th>
<th>Unigrams</th>
<th>Bigrams</th>
<th>Trigrams</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire LM (trained on 130m)</td>
<td>291,767</td>
<td>4,991,346</td>
<td>7,881,122</td>
</tr>
<tr>
<td>1000 sent.</td>
<td>13,792</td>
<td>2,850,983</td>
<td>6,540,940</td>
</tr>
<tr>
<td>1000 sent, top 20 transl.</td>
<td>9,860</td>
<td>2,251,111</td>
<td>5,590,783</td>
</tr>
<tr>
<td>10 sent, top 20 transl.</td>
<td>871</td>
<td>127,552</td>
<td>488,694</td>
</tr>
</tbody>
</table>

⇒ High overhead in filtering LM
Multiple Language Models

- Pharaoh allows multiple language models:
  - Large LM
    - trained on 800 million words (training + part of Gigaword)
    - ignored trigram singletons
    - digits ('0'-'9') replaced by '5'
  - Specialized LM
    - trained on 1.1 million words (news training corpus)
    - including all singletons
    - no special treatment of numbers
  - Weights of LM determined by discriminative training

Minor Model Improvements

- dropping unknown words during decoding
- delete word feature
- limited changes to the recapitalizer
- limited post-editing of the output
- limited changes to the tokenization of Arabic
Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
- Evaluation
- Related Recent Work in SMT

Evaluation for Arabic-English

- Improvements for Arabic-English:

<table>
<thead>
<tr>
<th>Eval set</th>
<th>'04 system</th>
<th>'05 system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval 2002 (partial)</td>
<td>34.4% BLEU</td>
<td>40.4% BLEU</td>
</tr>
<tr>
<td>Eval 2004</td>
<td>34.1% BLEU</td>
<td>34.3% BLEU</td>
</tr>
<tr>
<td>Eval 2005</td>
<td>35.6% BLEU</td>
<td>40.5% BLEU</td>
</tr>
</tbody>
</table>
Why so Little Improvement on Eval 2004?

- Model optimized on first 300 sentences of Eval 2002
  ⇒ very short output (length ratio 0.905)

- Word penalty feature allows tuning of output length:

  ![Graph showing BLEU scores vs. length ratio output/reference.](image)

  - Manual adjustment: 34.3% ⇒ 37.7% BLEU

Evaluation for Chinese-English

- Improvements for Chinese-English

  - System changes:
    - bigger language model (800 million words)
    - debugged number translator

<table>
<thead>
<tr>
<th>Eval set</th>
<th>'04 system</th>
<th>'05 system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eval 2002 (partial)</td>
<td>26.1% BLEU</td>
<td>27.2% BLEU</td>
</tr>
<tr>
<td>Eval 2004</td>
<td>27.1% BLEU</td>
<td>28.1% BLEU</td>
</tr>
<tr>
<td>Eval 2005</td>
<td>24.4% BLEU</td>
<td>25.1% BLEU</td>
</tr>
</tbody>
</table>

Philipp Koehn, University of Edinburgh

The Statistical Machine Translation System of the University of Edinburgh
Outline

- Overview: SMT at Edinburgh
- Baseline System
- Improvements
- Evaluation
- Related Recent Work in SMT
  - clause restructuring [Collins, Koehn, Kucerova, 2005]
  - Euromatrix [Koehn, 2005]
  - shared task at ACL workshop [Koehn and Monz, 2005]

Clause Level Restructuring

- Why clause structure?
  - languages differ vastly in their clause structure
    (English: SVO, Arabic: VSO, German: fairly free order;
    a lot details differ: position of adverbs, sub clauses, etc.)
  - large-scale restructuring is a problem for phrase models

- Restructuring
  - reordering of constituents (main focus)
  - add/drop/change of function words

- Ongoing work
  - collaboration with Michael Collins and Ivona Kucerova
  - currently German-English
  - see ACL paper for details
Clause Structure

- Syntax tree from German parser
  - statistical parser by Amit Dubey, trained on TIGER treebank

Reordering When Translating

- Reordering when translating into English
  - tree is flattened
  - clause level constituents line up
Clause Level Reordering

Clause level reordering is a well defined task
   - label German constituents with their English order
   - done this for 300 sentences, two annotators, high agreement

Many types of reorderings are systematic
   - move verb group together
   - subject - verb - object
   - move negation in front of verb

⇒ Write rules by hand
   - apply rules to test and training data
   - train standard phrase-based SMT system

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline system</td>
<td>25.2%</td>
</tr>
<tr>
<td>with manual rules</td>
<td>26.8%</td>
</tr>
</tbody>
</table>
Euromatrix

- Proceedings of the European Parliament
  - translated into 11 official languages
  - entry of new members in May 2004: more to come...
- Europarl corpus
  - collected 20-30 million words per language
  - 110 language pairs
- 110 Translation systems
  - 3 weeks on 16-node cluster computer
  - 110 translation systems

<table>
<thead>
<tr>
<th></th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>es</th>
<th>fr</th>
<th>fi</th>
<th>it</th>
<th>nl</th>
<th>pt</th>
<th>sv</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>-</td>
<td>18.4</td>
<td>21.1</td>
<td>28.5</td>
<td>26.4</td>
<td>28.7</td>
<td>14.2</td>
<td>22.2</td>
<td>21.4</td>
<td>24.3</td>
<td>28.3</td>
</tr>
<tr>
<td>de</td>
<td>22.3</td>
<td>-</td>
<td>20.7</td>
<td>25.3</td>
<td>25.4</td>
<td>27.7</td>
<td>11.8</td>
<td>21.3</td>
<td>23.4</td>
<td>23.2</td>
<td>20.5</td>
</tr>
<tr>
<td>el</td>
<td>22.7</td>
<td>17.4</td>
<td>-</td>
<td>27.2</td>
<td>31.2</td>
<td>32.1</td>
<td>11.4</td>
<td>26.8</td>
<td>20.0</td>
<td>27.6</td>
<td>21.2</td>
</tr>
<tr>
<td>en</td>
<td>25.2</td>
<td>17.6</td>
<td>23.2</td>
<td>-</td>
<td>30.1</td>
<td>31.1</td>
<td>13.0</td>
<td>25.3</td>
<td>21.0</td>
<td>27.1</td>
<td>24.8</td>
</tr>
<tr>
<td>es</td>
<td>24.1</td>
<td>18.2</td>
<td>28.3</td>
<td>30.5</td>
<td>-</td>
<td>40.2</td>
<td>12.5</td>
<td>32.3</td>
<td>21.4</td>
<td>35.9</td>
<td>23.9</td>
</tr>
<tr>
<td>fr</td>
<td>23.7</td>
<td>18.5</td>
<td>26.1</td>
<td>30.0</td>
<td>38.4</td>
<td>-</td>
<td>12.6</td>
<td>32.4</td>
<td>21.1</td>
<td>35.3</td>
<td>22.6</td>
</tr>
<tr>
<td>fi</td>
<td>20.0</td>
<td>14.5</td>
<td>18.2</td>
<td>21.8</td>
<td>21.1</td>
<td>22.4</td>
<td>-</td>
<td>18.3</td>
<td>17.0</td>
<td>19.1</td>
<td>18.8</td>
</tr>
<tr>
<td>it</td>
<td>21.4</td>
<td>16.9</td>
<td>24.8</td>
<td>27.8</td>
<td>34.0</td>
<td>36.0</td>
<td>11.0</td>
<td>-</td>
<td>20.0</td>
<td>31.2</td>
<td>20.2</td>
</tr>
<tr>
<td>nl</td>
<td>20.5</td>
<td>18.3</td>
<td>17.4</td>
<td>23.0</td>
<td>22.9</td>
<td>24.6</td>
<td>10.3</td>
<td>20.0</td>
<td>-</td>
<td>20.7</td>
<td>19.0</td>
</tr>
<tr>
<td>pt</td>
<td>23.2</td>
<td>18.2</td>
<td>26.4</td>
<td>30.1</td>
<td>37.9</td>
<td>39.0</td>
<td>11.9</td>
<td>32.0</td>
<td>20.2</td>
<td>-</td>
<td>21.9</td>
</tr>
<tr>
<td>sv</td>
<td>30.3</td>
<td>18.9</td>
<td>22.8</td>
<td>30.2</td>
<td>28.6</td>
<td>29.7</td>
<td>15.3</td>
<td>23.9</td>
<td>21.9</td>
<td>25.9</td>
<td>-</td>
</tr>
</tbody>
</table>
Translate into vs. out of a Language

- Some languages are easier to translate into that out of

<table>
<thead>
<tr>
<th>Language</th>
<th>From</th>
<th>Into</th>
<th>Diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>23.4</td>
<td>23.3</td>
<td>0.0</td>
</tr>
<tr>
<td>de</td>
<td>22.2</td>
<td>17.7</td>
<td>-4.5</td>
</tr>
<tr>
<td>el</td>
<td>23.8</td>
<td>22.9</td>
<td>-0.9</td>
</tr>
<tr>
<td>en</td>
<td>23.8</td>
<td>27.4</td>
<td>+3.6</td>
</tr>
<tr>
<td>es</td>
<td>26.7</td>
<td>29.6</td>
<td>+2.9</td>
</tr>
<tr>
<td>fr</td>
<td>26.1</td>
<td>31.1</td>
<td>+5.1</td>
</tr>
<tr>
<td>fi</td>
<td>19.1</td>
<td>12.4</td>
<td>-6.7</td>
</tr>
<tr>
<td>it</td>
<td>24.3</td>
<td>25.4</td>
<td>+1.1</td>
</tr>
<tr>
<td>nl</td>
<td>19.7</td>
<td>20.7</td>
<td>+1.1</td>
</tr>
<tr>
<td>pt</td>
<td>26.1</td>
<td>27.0</td>
<td>+0.9</td>
</tr>
<tr>
<td>sv</td>
<td>24.8</td>
<td>22.1</td>
<td>-2.6</td>
</tr>
</tbody>
</table>

Backtranslations

- Checking translation quality by back-transliteration

- “The spirit is willing, but the flesh is weak“

- English → Russian → English

- “The vodka is good but the meat is rotten“
Backtranslations II

- Does not correlate well with unidirectional performance

<table>
<thead>
<tr>
<th>Language</th>
<th>From</th>
<th>Into</th>
<th>Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>da</td>
<td>28.5</td>
<td>25.2</td>
<td>56.6</td>
</tr>
<tr>
<td>de</td>
<td>25.3</td>
<td>17.6</td>
<td>48.8</td>
</tr>
<tr>
<td>el</td>
<td>27.2</td>
<td>23.2</td>
<td>56.5</td>
</tr>
<tr>
<td>es</td>
<td>30.5</td>
<td>30.1</td>
<td>52.6</td>
</tr>
<tr>
<td>fi</td>
<td>21.8</td>
<td>13.0</td>
<td>44.4</td>
</tr>
<tr>
<td>it</td>
<td>27.8</td>
<td>25.3</td>
<td>49.9</td>
</tr>
<tr>
<td>nl</td>
<td>23.0</td>
<td>21.0</td>
<td>46.0</td>
</tr>
<tr>
<td>pt</td>
<td>30.1</td>
<td>27.1</td>
<td>53.6</td>
</tr>
<tr>
<td>sv</td>
<td>30.2</td>
<td>24.8</td>
<td>54.4</td>
</tr>
</tbody>
</table>

Shared Task at ACL 2005 Workshop

- Given
  - parallel text, word alignment
  - language model
  - decoder Pharaoh

- Task:
  - build SMT system (at least: probabilistic phrase table)
  - French-English, Spanish-English, Finnish-English, German-English

- Participation
  - 11 teams from 8 institutions
  - several new research groups
Thank You!

- Questions?