The Basic

- **Translating with data**
  - how can computers learn from translated text?
  - what translated material is out there?
  - is it enough? how much is needed?

- **Statistical modeling**
  - framing translation as a generative statistical process

- **EM Training**
  - how do we automatically discover hidden data?

- **Decoding**
  - algorithm for translation
The Novel

- **Automatic evaluation methods**
  - can computers decide what are good translations?

- **Phrase-based models**
  - what are atomic units of translation?
  - the best method in statistical machine translation

- **Discriminative training**
  - what are the methods that directly optimize translation performance?

The Speculative

- **Syntax-based transfer models**
  - how can we build models that take advantage of syntax?
  - how can we ensure that the output is grammatical?

- **Factored translation models**
  - how can we integrate different levels of abstraction?
The Rosetta Stone

- Egyptian language was a mystery for centuries
- 1799 a stone with Egyptian text and its translation into Greek was found
  ⇒ Humans could learn how to translated Egyptian

Parallel Data

- Lots of translated text available: 100s of million words of translated text for some language pairs
  - a book has a few 100,000s words
  - an educated person may read 10,000 words a day
  → 3.5 million words a year
  → 300 million a lifetime
  → soon computers will be able to see more translated text than humans read in a lifetime
  ⇒ Machine can learn how to translated foreign languages
Statistical Machine Translation

- Components: Translation model, language model, decoder

Word-Based Models

• Translation process is decomposed into smaller steps, each is tied to words

• Original models for statistical machine translation [Brown et al., 1993]
Phrase-Based Models

- Foreign input is segmented in **phrases**
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered

Syntax-Based Models

- Foreign input is segmented in **phrases**
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered
Language Models

- **Language models** indicate, whether a sentence is **good English**
  - \( p(\text{Tomorrow I will fly to the conference}) = \text{high} \)
  - \( p(\text{Tomorrow fly me at a summit}) = \text{low} \)
  \( \rightarrow \) ensures fluent output by guiding word choice and word order

- Standard: **trigram language models**
  \[
p(\text{Tomorrow}|\text{START}) \times \nonumber \\
p(I|\text{START}, \text{Tomorrow}) \times \\
p(\text{will}|\text{Tomorrow}, I) \times \\
\ldots \\
p(\text{Canada}|\text{conference}, \text{in}) \times \\
p(\text{END}|\text{in}, \text{Canada})
\]

- Often estimated using additional **monolingual data** (billions of words)

Automatic Evaluation

- Why **automatic evaluation** metrics?
  - Manual evaluation is **too slow**
  - Evaluation on large test sets reveals minor improvements
  - **Automatic tuning** to improve machine translation performance

- History
  - **Word Error Rate**
  - **BLEU** since 2002

- **BLEU** in short: **Overlap with reference** translations
Automatic Evaluation

• Reference Translation
  – the gunman was shot to death by the police.

• System Translations
  – the gunman was police kill.
  – wounded police jaya of
  – the gunman was shot dead by the police.
  – the gunman arrested by police kill.
  – the gunmen were killed.
  – the gunman was shot to death by the police.
  – gunmen were killed by police ?SUB>0 ?SUB>0
  – al by the police.
  – the ringer is killed by the police.
  – police killed the gunman.

• Matches
  – green = 4 gram match (good!)
  – red = word not matched (bad!)

• BLEU correlates with human judgement
  – multiple reference translations may be used

[from George Doddington, NIST]
Correlation? [Callison-Burch et al., 2006]

- DARPA/NIST MT Eval 2005
  - Mostly statistical systems (all but one in graphs)
  - One submission manual post-edit of statistical system’s output
  → Good adequacy/fluency scores not reflected by BLEU

Comparison of
- good statistical system: high BLEU, high adequacy/fluency
- bad statistical sys. (trained on less data): low BLEU, low adequacy/fluency
- Systran: lowest BLEU score, but high adequacy/fluency
Automatic Evaluation: Outlook

- Research questions
  - why does BLEU fail Systran and manual post-edits?
  - how can this overcome with novel evaluation metrics?
- Future of automatic methods
  - automatic metrics too useful to be abandoned
  - evidence still supports that during system development, a better BLEU indicates a better system
  - final assessment has to be human judgement

Competitions

- Progress driven by MT Competitions
  - NIST/DARPA: Yearly campaigns for Arabic-English, Chinese-English, newstexts, since 2001
  - IWSLT: Yearly competitions for Asian languages and Arabic into English, speech travel domain, since 2003
  - WPT/WMT: Yearly competitions for European languages, European Parliament proceedings, since 2005
- Increasing number of statistical MT groups participate
- Competitions won by statistical systems
Competitions: Good or Bad?

- **Pro:**
  - *public forum* for demonstrating the state of the art
  - open data sets and evaluation metrics allow for *comparison of methods*
  - *credibility* for a new approach by doing well
  - *sharing* of ideas and implementation details

- **Con:**
  - winning competition is mostly due to better *engineering*
  - having *more data and faster machines* plays a role
  - *limit research* to few directions (re-engineering of other’s methods)

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*

- Basis of a new European Commission funded project
Quality of Translation Systems

- **Scores** for all 110 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>

[from Koehn, 2005: Europarl]

Clustering Languages

- **Clustering** languages based on how easy they translate into each other

⇒ Approximation of language families

[from Koehn, 2005, MT Summit]
Translation examples

• **Spanish-English**
  
  (1) the current situation, unsustainable above all for many self-employed drivers and in the area of agriculture, we must improve without doubt.
  
  (2) in itself, it is good to reach an agreement on procedures, but we have to ensure that this system is not likely to be used as a weapon policy.

• **Finnish-English**
  
  (1) the current situation, which is unacceptable, in particular, for many carriers and responsible for agriculture, is in any case, to be improved.
  
  (2) agreement on procedures in itself is a good thing, but there is a need to ensure that the system cannot be used as a political weapon lyömäaseena.

• **English reference**
  
  (1) the current situation, which is intolerable, particularly for many independent haulage firms and for agriculture, does in any case need to be improved.
  
  (2) an agreement on procedures in itself is a good thing, but we must make sure that the system cannot be used as a political weapon.

---

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>

[from Koehn, 2005: Europarl]

• **Morphologically rich languages** harder to generate (German, Finnish)
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 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>

[from Koehn, 2005: Europarl]
Available Data

- Available **parallel text**
  - **Europarl**: 30 million words in 11 languages http://www.statmt.org/europarl/
  - **Acquis Communautaire**: 8-50 million words in 20 EU languages
  - **Canadian Hansards**: 20 million words from Ulrich Germann, ISI
  - Chinese/Arabic to English: over 100 million words from LDC
  - lots more French/English, Spanish/French/English from LDC

- Available monolingual text (for language modeling)
  - 2.8 billion words of English from LDC
  - 100s of billions, trillions on the web

More Data, Better Translations

- **Log-scale improvements** on BLEU:
  Doubling the training data gives constant improvement (+1 %BLEU)
More LM Data, Better Translations

- Also log-scale improvements on BLEU: doubling the training data gives constant improvement (+0.5 %BLEU) (last addition is 218 billion words out-of-domain web data)

Decoding
- Statistical Modeling
- EM Algorithm
- Word Alignment
- Phrase-Based Translation
- Discriminative Training
- Syntax-Based Statistical MT
Decoding Process

- Build translation left to right
  - select foreign words to be translated

Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
Decoding Process

- Build translation left to right
  - select foreign words to be translated
  - find English phrase translation
  - add English phrase to end of partial translation
  - mark foreign words as translated

- One to many translation
Decoding Process

- Many to one translation

Decoding Process

- Many to one translation
Decoding Process

- **Reordering**

- **Translation finished**
Translation Options

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>hofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
</tr>
</tbody>
</table>

• Look up **possible phrase translations**
  – many different ways to **segment** words into phrases
  – many different ways to **translate** each phrase

Hypothesis Expansion

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>hofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
</tr>
<tr>
<td>did not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green</td>
<td>witch</td>
<td></td>
</tr>
</tbody>
</table>

• Start with **empty hypothesis**
  – e: no English words
  – f: no foreign words covered
  – p: probability 1
Hypothesis Expansion

<table>
<thead>
<tr>
<th>English</th>
<th>no</th>
<th>dio</th>
<th>una</th>
<th>bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>witch</td>
<td>green</td>
</tr>
<tr>
<td>English</td>
<td>did not</td>
<td>give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green witch</td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>did not give</td>
<td>a</td>
<td>slap</td>
<td>to</td>
<td>the</td>
<td>green witch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Pick translation option
- Create hypothesis
  - e: add English phrase Mary
  - f: first foreign word covered
  - p: probability 0.534

A Quick Word on Probabilities

- Not going into detail here, but...
- **Translation Model**
  - phrase translation probability \( p(\text{Mary}|\text{Maria}) \)
  - reordering costs
  - phrase/word count costs
  - ...
- **Language Model**
  - uses trigrams:
  - \( p(\text{Mary did not}) = p(\text{Mary}|\text{START}) \times p(\text{did}|\text{Mary,START}) \times p(\text{not}|\text{Mary did}) \)
Hypothesis Expansion

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio una bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not give a slap to the witch green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give a slap to the green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Add another hypothesis

Hypothesis Expansion

<table>
<thead>
<tr>
<th>Maria</th>
<th>no</th>
<th>dio una bofetada</th>
<th>a</th>
<th>la</th>
<th>bruja</th>
<th>verde</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>not give a slap to the witch green</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>did not give a slap to the green witch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

• Further hypothesis expansion
Hypothesis Expansion

• ... until all foreign words covered
  – find best hypothesis that covers all foreign words
  – backtrack to read off translation

Philipp Koehn SMT Tutorial 4 April 2006

Hypothesis Expansion

• Adding more hypothesis
  ⇒ Explosion of search space

Philipp Koehn SMT Tutorial 4 April 2006
Explosion of Search Space

- Number of hypotheses is exponential with respect to sentence length
  ⇒ Decoding is NP-complete [Knight, 1999]
  ⇒ Need to reduce search space
    - risk free: hypothesis recombination
    - risky: histogram/threshold pruning

Hypothesis Recombination

- Different paths to the same partial translation
Hypothesis Recombination

- Different paths to the same partial translation

⇒ **Combine paths**
  - **drop weaker** path
  - keep pointer from weaker path

- Recombined hypotheses do **not** have to **match completely**
- No matter what is added, weaker path can be dropped, if:
  - last two **English words** match (matters for language model)
  - foreign word coverage vectors match (effects future path)
Hypothesis Recombination

- Recombined hypotheses do not have to match completely
- No matter what is added, weaker path can be dropped, if:
  - last two English words match (matters for language model)
  - foreign word coverage vectors match (effects future path)
⇒ Combine paths

Pruning

- Hypothesis recombination is not sufficient
⇒ Heuristically discard weak hypotheses early
- Organize Hypothesis in stacks, e.g. by
  - same foreign words covered
  - same number of foreign words covered (Pharaoh does this)
  - same number of English words produced
- Compare hypotheses in stacks, discard bad ones
  - histogram pruning: keep top $n$ hypotheses in each stack (e.g., $n=100$)
  - threshold pruning: keep hypotheses that are at most $\alpha$ times the cost of best hypothesis in stack (e.g., $\alpha = 0.001$)
Hypothesis Stacks

- Organization of hypothesis into stacks
  - here: based on number of foreign words translated
  - during translation all hypotheses from one stack are expanded
  - expanded Hypotheses are placed into stacks

Comparing Hypotheses

- Comparing hypotheses with same number of foreign words covered

Maria no
e: Mary did not
f: **--------
p: 0.154

dio una bofetada
bruja verde
e: the
f: -----**--
p: 0.354

- Hypothesis that covers easy part of sentence is preferred
⇒ Need to consider future cost of uncovered parts
Future Cost Estimation

- **Estimate cost** to translate remaining part of input
- **Step 1**: estimate future cost for each **translation option**
  - look up translation model cost
  - estimate language model cost (no prior context)
  - ignore reordering model cost
  \[ LM \times TM = p(to) \times p(\text{the}|to) \times p(\text{to the}|a la) \]

Future Cost Estimation: Step 2

- **Step 2**: find **cheapest cost** among translation options
  - cost = 0.0372
  - cost = 0.0299
  - cost = 0.0354
Future Cost Estimation: Step 3

- **Step 3**: find **cheapest future cost path** for each span
  - can be done **efficiently** by dynamic programming
  - future cost for every span can be **pre-computed**

Future Cost Estimation: Application

- Use future cost estimates when **pruning** hypotheses
- For each **uncovered contiguous span**:
  - look up **future costs** for each maximal contiguous uncovered span
  - **add** to actually accumulated cost for translation option for pruning
Pharaoh

• A beam search decoder for phrase-based models
  – works with various phrase-based models
  – beam search algorithm
  – time complexity roughly linear with input length
  – good quality takes about 1 second per sentence
• Very good performance in DARPA/NIST Evaluation
• Freely available for researchers http://www.isi.edu/licensed-sw/pharaoh/
• Coming soon: open source version of Pharaoh

Running the decoder

• An example run of the decoder:
% echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini > out
Pharaoh v1.2.9, written by Philipp Koehn
a beam search decoder for phrase-based statistical machine translation models
(c) 2002-2003 University of Southern California
(c) 2004 Massachusetts Institute of Technology
(c) 2005 University of Edinburgh, Scotland
loading language model from europarl.srilm
loading phrase translation table from phrase-table, stored 21, pruned 0, kept 21
loaded data structures in 2 seconds
reading input sentences
translating 1 sentences
translated 1 sentences in 0 seconds
[3mm] % cat out
this is a small house
Phrase Translation Table

- Core model component is the phrase translation table:

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Translation</th>
<th>Probability</th>
</tr>
</thead>
</table>
| der ||| the ||| 0.3
| das ||| the ||| 0.4
| das ||| it ||| 0.1
| das ||| this ||| 0.1
| die ||| the ||| 0.3
| ist ||| is ||| 1.0
| ist ||| 's ||| 1.0
| das ist &&| it is ||| 0.2
| das ist &&| this is ||| 0.8
| es ist &&| it is ||| 0.8
| es ist &&| this is ||| 0.2
| ein &&| a ||| 1.0
| ein &&| an ||| 1.0
| klein ||| small ||| 0.8
| klein ||| little ||| 0.8
| kleines ||| small ||| 0.2
| kleines ||| little ||| 0.2
| haus ||| house ||| 1.0
| alt ||| old ||| 0.8
| altes ||| old ||| 0.2
| gibt ||| gives ||| 1.0
| es gibt &&| there is ||| 1.0

Trace

- Running the decoder with switch “-t”

```shell
% echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -t
[...]
this is |0.014086|0|1| a |0.188447|2|2| small |0.000706353|3|3|
house |1.46468e-07|4|4|
```

- **Trace** for each applied phrase translation:
  - output phrase (there is)
  - cost incurred by this phrase (0.014086)
  - coverage of foreign words (0-1)
Reordering Example

• Sometimes phrases have to be reordered:

% echo 'ein kleines haus ist das' | pharaoh -f pharaoh.ini -t -d 0.5

[...] this |0.000632805|4|4| is |0.13853|3|3| a |0.0255035|0|0|
small |0.000706353|1|1| house |1.46468e-07|2|2|

• First output phrase this is translation of the 4th word

Hypothesis Accounting

• The switch “-v” allows for detailed run time information:

% echo 'das ist ein kleins haus’ | pharaoh -f pharaoh.ini -v 2

[...] HYP: 114 added, 284 discarded below threshold, 0 pruned, 58 merged.
BEST: this is a small house -28.9234

• Statistics over how many hypothesis were generated
  – 114 hypotheses were added to hypothesis stacks
  – 284 hypotheses were discarded because they were too bad
  – 0 hypotheses were pruned, because a stack got too big
  – 58 hypotheses were merged due to recombination

• Probability of the best translation: exp(-28.9234)
Translation Options

- Even more run time information is revealed with "-v 3":

\[
\begin{align*}
\text{the} & \quad \text{pC}=-0.916291, \ c=-5.78855 \\
\text{it} & \quad \text{pC}=-2.30259, \ c=-8.0761 \\
\text{this} & \quad \text{pC}=-2.30259, \ c=-8.00205 \\
\text{is} & \quad \text{pC}=0, \ c=-4.92223 \\
\text{'s} & \quad \text{pC}=0, \ c=-6.11591 \\
\text{a} & \quad \text{pC}=0, \ c=-5.5151 \\
\text{an} & \quad \text{pC}=0, \ c=-6.41298 \\
\text{small} & \quad \text{pC}=-1.60944, \ c=-9.72116 \\
\text{little} & \quad \text{pC}=-1.60944, \ c=-10.0953 \\
\text{house} & \quad \text{pC}=0, \ c=-9.26607 \\
\text{it} & \quad \text{pC}=-1.60944, \ c=-10.207 \\
\text{this} & \quad \text{pC}=-0.223144, \ c=-10.2906
\end{align*}
\]

- Translation model cost (pC) and future cost estimates (c)

Future Cost Estimation

- Pre-computation of the future cost estimates:

\[
\begin{align*}
\text{future costs from 0 to 0 is } & -5.78855 \\
\text{future costs from 0 to 1 is } & -10.207 \\
\text{future costs from 0 to 2 is } & -15.7221 \\
\text{future costs from 0 to 3 is } & -25.4433 \\
\text{future costs from 0 to 4 is } & -34.7094 \\
\text{future costs from 1 to 1 is } & -4.92223 \\
\text{future costs from 1 to 2 is } & -10.4373 \\
\text{future costs from 1 to 3 is } & -20.1585 \\
\text{future costs from 1 to 4 is } & -29.4246 \\
\text{future costs from 2 to 2 is } & -5.5151 \\
\text{future costs from 2 to 3 is } & -15.2363 \\
\text{future costs from 2 to 4 is } & -24.5023 \\
\text{future costs from 3 to 3 is } & -9.72116 \\
\text{future costs from 3 to 4 is } & -18.9872 \\
\text{future costs from 4 to 4 is } & -9.26607
\end{align*}
\]
Hypothesis Expansion

• **Start** of beam search: First hypothesis \((\text{das} \rightarrow \text{the})\)

creating hypothesis 1 from 0 ( ... </s> <s> )  
basis score 0  
coverage 0-0: das  
translated as: the ⇒ translation cost -0.916291  
distance 0 ⇒ distortion cost 0  
language model cost for 'the' -2.03434  
word penalty -0  
score -2.95064 + futureCost -29.4246 = -32.3752  
new best estimate for this stack  
merged hypothesis on stack 1, now size 1

• Another hypothesis \((\text{das ist} \rightarrow \text{this is})\)

creating hypothesis 12 from 0 ( ... </s> <s> )  
basis score 0  
coverage 0-1: das ist  
translated as: this is ⇒ translation cost -0.223144  
distance 0 ⇒ distortion cost 0  
language model cost for 'this' -3.06276  
language model cost for 'is' -0.976669  
word penalty -0  
new best estimate for this stack  
merged hypothesis on stack 2, now size 2
Hypothesis Expansion

- **Hypothesis recombination**
  
  creating hypothesis 27 from 3 ( ... <s> this )
  base score -5.36535
  covering 1-1: ist
  translated as: is => translation cost 0
  distance 0 => distortion cost 0
  language model cost for ‘is’ -0.976669
  word penalty -0
  score -6.34202 + futureCost -24.5023 = -30.8443
  worse than existing path to 12, discarding

- **Bad hypothesis** that falls out of the beam

  creating hypothesis 52 from 6 ( ... <s> a )
  base score -6.65992
  covering 0-0: das
  translated as: this => translation cost -2.30259
  distance -3 => distortion cost -3
  language model cost for ‘this’ -8.69176
  word penalty -0
  score -20.6543 + futureCost -23.9095 = -44.5637
  estimate below threshold, discarding
Generating Best Translation

• Generating best translation
  – find best **final hypothesis** (442)
  – **trace back** path to initial hypothesis

```
best hypothesis 442
  [ 442 => 343 ]
  [ 343 => 106 ]
  [ 106 => 12 ]
  [ 12 => 0 ]
```

Beam Size

• **Trade-off** between **speed** and **quality** via beam size

```
% echo 'das ist ein kleines haus' | pharaoh -f pharaoh.ini -s 10 -v 2
...]
collected 12 translation options
HYP: 78 added, 122 discarded below threshold, 33 pruned, 20 merged.
BEST: *this is a small house -28.9234*
```

<table>
<thead>
<tr>
<th>Beam size</th>
<th>Threshold</th>
<th>Hyp. added</th>
<th>Hyp. discarded</th>
<th>Hyp. pruned</th>
<th>Hyp. merged</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>unlimited</td>
<td>634</td>
<td>0</td>
<td>0</td>
<td>1306</td>
</tr>
<tr>
<td>100</td>
<td>unlimited</td>
<td>557</td>
<td>32</td>
<td>199</td>
<td>572</td>
</tr>
<tr>
<td>100</td>
<td>0.00001</td>
<td>144</td>
<td>284</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>10</td>
<td>0.00001</td>
<td>78</td>
<td>122</td>
<td>33</td>
<td>20</td>
</tr>
<tr>
<td>1</td>
<td>0.00001</td>
<td>9</td>
<td>19</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Limits on Reordering

• Reordering may be limited
  – Monotone Translation: No reordering at all
  – Only phrase movements of at most $n$ words

• Reordering limits speed up search

• Current reordering models are weak, so limits improve translation quality

Word Lattice Generation

• Search graph can be easily converted into a word lattice
  – can be further mined for n-best lists
  → enables reranking approaches
  → enables discriminative training
Sample N-Best List

- **N-best list** from Pharaoh:

<table>
<thead>
<tr>
<th>Translation</th>
<th>Reordering LM</th>
<th>TM</th>
<th>WordPenalty</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>this is a small house</td>
<td>0</td>
<td>-27.0908</td>
<td>-1.83258</td>
<td>-28.9234</td>
</tr>
<tr>
<td>it is a small house</td>
<td>0</td>
<td>-27.108</td>
<td>-3.21888</td>
<td>-30.3268</td>
</tr>
<tr>
<td>this is a little house</td>
<td>0</td>
<td>-28.1791</td>
<td>-1.83258</td>
<td>-30.4623</td>
</tr>
<tr>
<td>it is a little house</td>
<td>0</td>
<td>-31.7294</td>
<td>-1.83258</td>
<td>-33.5622</td>
</tr>
<tr>
<td>it is an small house</td>
<td>0</td>
<td>-32.3094</td>
<td>-3.21888</td>
<td>-35.5283</td>
</tr>
<tr>
<td>this is an little house</td>
<td>0</td>
<td>-33.7639</td>
<td>-1.83258</td>
<td>-35.5965</td>
</tr>
<tr>
<td>this is an small house</td>
<td>-3</td>
<td>-31.4851</td>
<td>-1.83258</td>
<td>-36.3176</td>
</tr>
<tr>
<td>it is an little house</td>
<td>-3</td>
<td>-31.5022</td>
<td>-3.21888</td>
<td>-37.5628</td>
</tr>
<tr>
<td>this is an house small</td>
<td>-3</td>
<td>-31.586</td>
<td>-3.21888</td>
<td>-37.8049</td>
</tr>
<tr>
<td>it is a house small</td>
<td>-3</td>
<td>-32.9837</td>
<td>-1.83258</td>
<td>-38.1638</td>
</tr>
<tr>
<td>the house is a little</td>
<td>-7</td>
<td>-28.5107</td>
<td>-2.52573</td>
<td>-38.0364</td>
</tr>
<tr>
<td>the is a small house</td>
<td>0</td>
<td>-35.6899</td>
<td>-2.52573</td>
<td>-38.2156</td>
</tr>
<tr>
<td>it is a little house</td>
<td>-4</td>
<td>-30.3603</td>
<td>-3.91202</td>
<td>-38.2723</td>
</tr>
<tr>
<td>the house is a small</td>
<td>-7</td>
<td>-28.7983</td>
<td>-3.91202</td>
<td>-38.2944</td>
</tr>
<tr>
<td>it's a small house</td>
<td>0</td>
<td>-34.8557</td>
<td>-3.91202</td>
<td>-38.7677</td>
</tr>
<tr>
<td>this house is a little</td>
<td>-7</td>
<td>-28.0443</td>
<td>-3.91202</td>
<td>-38.9563</td>
</tr>
<tr>
<td>it's a little house</td>
<td>0</td>
<td>-36.1446</td>
<td>-3.91202</td>
<td>-39.0866</td>
</tr>
<tr>
<td>this house is a small</td>
<td>-7</td>
<td>-28.3019</td>
<td>-3.91202</td>
<td>-39.2139</td>
</tr>
</tbody>
</table>

---

XML Markup

Er erzielte <NUMBER english='17.55'>17,55</NUMBER> Punkte.

- **Add additional translation options**
  - number translation
  - noun phrase translation [Koehn, 2003]
  - name translation

- **Additional options**
  - provide multiple translations
  - provide probability distribution along with translations
  - allow bypassing of provided translations
- Decoding
- Statistical Modeling
- EM Algorithm
- Word Alignment
- Phrase-Based Translation
- Discriminative Training
- Syntax-Based Statistical MT

Statistical Modeling

Mary did not slap the green witch
Maria no daba una bofetada a la bruja verde
[from Knight and Knight, 2004, SMT Tutorial]

- Learn $P(f|e)$ from a parallel corpus
- Not sufficient data to estimate $P(f|e)$ directly
• **Decompose** the process into smaller steps

• Probabilities for **smaller steps** can be learned
Statistical Modeling (4)

- **Generate a story** how an English string \( e \) gets to be a foreign string \( f \)
  - choices in story are decided by reference to **parameters**
    - e.g., \( p(\text{bruja}|\text{witch}) \)
- **Formula** for \( P(f|e) \) in terms of parameters
  - usually long and hairy, but **mechanical to extract** from the story
- **Training** to obtain parameter estimates from possibly **incomplete data**
  - off-the-shelf **Expectation Maximization (EM)**

Parallel Corpora

... la maison ... la maison blue ... la fleur ...

... the house ... the blue house ... the flower ...

- **Incomplete data**
  - English and foreign words, but **no connections** between them
- Chicken and egg problem
  - if we had the **connections**, we could estimate the **parameters** of our generative story
  - if we had the **parameters**, we could estimate the **connections** in the data
• Decoding
• Statistical Modeling

\textbf{EM Algorithm}

• Word Alignment
• Phrase-Based Translation
• Discriminative Training
• Syntax-Based Statistical MT

\begin{itemize}
  \item Incomplete data
    \begin{itemize}
      \item if we had complete data, we could estimate model
      \item if we had model, we could fill in the gaps in the data
    \end{itemize}
  \item EM in a nutshell
    \begin{enumerate}
      \item \textbf{initialize model} parameters (e.g. uniform)
      \item \textbf{assign probabilities} to the missing data (the connections)
      \item \textbf{estimate model} parameters from completed data
      \item \textbf{iterate} steps 2 and 3
    \end{enumerate}
\end{itemize}
EM Algorithm (2)

... la maison ... la maison blue ... la fleur ...  

... the house ... the blue house ... the flower ...

- Initial step: all connections **equally likely**
- Model learns that, e.g., *la* is often connected with *the*

EM Algorithm (3)

... la maison ... la maison blue ... la fleur ...  

... the house ... the blue house ... the flower ...

- After one iteration
- Connections, e.g., between *la* and *the* are **more likely**
EM Algorithm (4)

... la maison ... la maison bleu ... la fleur ...
... the house ... the blue house ... the flower ...

- After another iteration
- It becomes apparent that connections, e.g., between *fleur* and *flower* are more likely *(pigeon hole principle)*

EM Algorithm (5)

... la maison ... la maison bleu ... la fleur ...
... the house ... the blue house ... the flower ...

- **Convergence**
- Inherent hidden structure *revealed* by EM
**EM Algorithm (6)**

... la maison ... la maison bleu ... la fleur ...

... the house ... the blue house ... the flower ...

\[
\begin{align*}
p(\text{la|the}) &= 0.453 \\
p(\text{le|the}) &= 0.334 \\
p(\text{maison|house}) &= 0.876 \\
p(\text{bleu|blue}) &= 0.563
\end{align*}
\]

- **Parameter estimation** from the connected corpus

---

**Flaws of Word-Based MT**

- **Multiple English words for one German word**

  **one-to-many** problem: Zeitmangel → lack of time

  German: Zeitmangel erschwert das Problem.
  Gloss: LACK OF TIME MAKES MORE DIFFICULT THE PROBLEM.
  Correct translation: Lack of time makes the problem more difficult.
  MT output: Time makes the problem.

- **Phrasal translation**

  **non-compositional phrase**: erübrigt sich → there is no point in

  German: Eine Diskussion erübrigt sich demnach.
  Gloss: A DISCUSSION IS MADE UNNECESSARY ITSELF THEREFORE.
  Correct translation: Therefore, there is no point in a discussion.
  MT output: A debate turned therefore.
Flaws of Word-Based MT (2)

- Syntactic transformations

**reordering, genitive NP**: der Sache → for this matter

German: Das ist der Sache nicht angemessen.
Gloss: THAT IS THE MATTER NOT APPROPRIATE.
Correct translation: That is not appropriate for this matter.
MT output: That is the thing is not appropriate.

**object/subject reordering**

German: Den Vorschlag lehnt die Kommission ab.
Gloss: THE PROPOSAL REJECTS THE COMMISSION OFF.
Correct translation: The commission rejects the proposal.
MT output: The proposal rejects the commission.

- Decoding
- Statistical Modeling
- EM Algorithm
- **Word Alignment**
- Phrase-Based Translation
- Discriminative Training
- Syntax-Based Statistical MT
### Word Alignment

- Notion of **word alignment** valuable
- Shared task at NAACL 2003 and ACL 2005 workshops

![Word Alignment Diagram]

### Word Alignment with IBM Models

- IBM Models create a **many-to-one** mapping
  - words are aligned using an **alignment function**
  - a function may return the same value for different input (one-to-many mapping)
  - a function can not return multiple values for one input (no **many-to-one** mapping)
- But we need **many-to-many** mappings
**Improved Word Alignments**

- *Intersection* of GIZA++ bidirectional alignments

- **Grow** additional alignment points [Och and Ney, CompLing2003]
Growing Heuristic

GROW-DIAG-FINAL(e2f,f2e):
neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
alignment = intersect(e2f,f2e);
GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
iterate until no new points added
for foreign word f = 0 ... fn
    for each neighboring point ( e-new, f-new ):
        if ( ( e-new not aligned and f-new not aligned ) and
            ( e-new, f-new ) in union( e2f, f2e )
        )
            add alignment point ( e-new, f-new )

FINAL(a):
for foreign word f-new = 0 ... fn
    for each neighboring point ( e-new, f-new ):
        if ( ( e-new not aligned or f-new not aligned ) and
            ( e-new, f-new ) in alignment a
        )
            add alignment point ( e-new, f-new )

• Decoding
• Statistical Modeling
• EM Algorithm
• Word Alignment

Phrase-Based Translation
• Discriminative Training
• Syntax-Based Statistical MT
Phrase-Based Translation

- Foreign input is segmented in phrases
  - any sequence of words, not necessarily linguistically motivated
- Each phrase is translated into English
- Phrases are reordered
- See [Koehn et al., NAACL2003] as introduction

Advantages of Phrase-Based Translation

- Many-to-many translation can handle non-compositional phrases
- Use of local context in translation
- The more data, the longer phrases can be learned
Phrase-Based Systems

• A number of research groups developed phrase-based systems
  – RWTH Aachen – Univ. of Southern California/ISI – CMU

• Systems differ in
  – training methods
  – model for phrase translation table
  – reordering models
  – additional feature functions

• Currently best method for SMT (MT?)
  – top systems in DARPA/NIST evaluation are phrase-based
  – best commercial system for Arabic-English is phrase-based

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     |                    |            |
How to Learn the Phrase Translation Table?

- Start with the word alignment:

  ![Alignment Table]

  - Collect all phrase pairs that are consistent with the word alignment

Consistent with Word Alignment

- Consistent with the word alignment :=
  phrase alignment has to contain all alignment points for all covered words

\[
(\overline{e}, \overline{f}) \in BP \iff \forall e_i \in \overline{e} : (e_i, f_j) \in A \rightarrow f_j \in \overline{f}
\]

AND \[
\forall f_j \in \overline{f} : (e_i, f_j) \in A \rightarrow e_i \in \overline{e}
\]
Word Alignment Induced Phrases

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Word Alignment Induced Phrases (2)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),

(bruja verde, green witch)
Word Alignment Induced Phrases (3)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green).

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch),  (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Philipp Koehn  SMT Tutorial  4 April 2006

Word Alignment Induced Phrases (4)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green).

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch),  (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),

(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)

Philipp Koehn  SMT Tutorial  4 April 2006
Word Alignment Induced Phrases (5)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green).

(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch).  (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde,
slap the green witch).  (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Probability Distribution of Phrase Pairs

• We need a probability distribution $\phi(f|e)$ over the collected phrase pairs

⇒ Possible choices

– relative frequency of collected phrases: $\phi(f|e) = \frac{\text{count}(f,e)}{\sum_{\mathcal{F}} \text{count}(f,e)}$

– or, conversely $\phi(e|f)$

– use lexical translation probabilities
Reordering

- **Monotone** translation
  - do not allow any reordering
  → worse translations
- **Limiting** reordering (to movement over max. number of words) helps
- **Distance-based** reordering cost
  - moving a foreign phrase over $n$ words: cost $\omega^n$
- **Lexicalized** reordering model

---

Lexicalized Reordering Models

- Three **orientation** types: monotone, swap, discontinuous
- Probability $p(\text{swap}|e, f)$ depends on foreign (and English) **phrase** involved

[from Koehn et al., 2005, IWSLT]
Training

- Orientation type is **learned during phrase extractions**
- **Alignment point** to the **top left** (monotone) or **top right** (swap)?
- For more, see [Tillmann, 2003] or [Koehn et al., 2005]

---

**Discriminative Training**

- Decoding
- Statistical Modeling
- EM Algorithm
- Word Alignment
- Phrase-Based Translation
- Syntax-Based Statistical MT
Log-Linear Models

- IBM Models provided mathematical justification for factoring components together
  \[ p_{LM} \times p_{TM} \times p_D \]
- These may be weighted
  \[ p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D} \]
- Many components \( p_i \) with weights \( \lambda_i \)
  \[ \prod_i p_i^{\lambda_i} = \exp(\sum_i \lambda_i \log(p_i)) \]
  \[ \log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i \log(p_i) \]

Knowledge Sources

- Many different knowledge sources useful
  - language model
  - reordering (distortion) model
  - phrase translation model
  - word translation model
  - word count
  - phrase count
  - drop word feature
  - phrase pair frequency
  - additional language models
  - additional features
Set Feature Weights

- Contribution of components $p_i$ determined by weight $\lambda_i$
- Methods
  - **manual setting** of weights: try a few, take best
  - **automate** this process
- Learn weights
  - set aside a **development corpus**
  - set the weights, so that **optimal translation performance** on this development corpus is achieved
  - requires **automatic scoring** method (e.g., BLEU)

Learn Feature Weights

- Model
  - **generate** n-best list
  - **score** translations
  - **find** feature weights that move up good translations
  - **change** feature weights
Discriminative vs. Generative Models

- **Generative models**
  - translation process is broken down to **steps**
  - each step is modeled by a **probability distribution**
  - each probability distribution is estimated from the data by **maximum likelihood**

- **Discriminative models**
  - model consist of a number of **features** (e.g. the language model score)
  - each feature has a **weight**, measuring its value for judging a translation as correct
  - feature weights are **optimized on development data**, so that the system output matches correct translations as close as possible

Discriminative Training (2)

- **Training set** (**development set**)
  - different from original training set
  - small (maybe 1000 sentences)
  - must be different from test set

- Current model **translates** this development set
  - **n-best list** of translations (n=100, 10000)
  - translations in n-best list can be **scored**

- Feature weights are **adjusted**

- N-Best list generation and feature weight adjustment repeated for a number of iterations
Learning Task

- Task: **find weights**, so that feature vector of the correct translations ranked first

<table>
<thead>
<tr>
<th>TRANSLATION</th>
<th>LM</th>
<th>TM</th>
<th>WP</th>
<th>SER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1  Mary not give slap witch green .</td>
<td>-17.2</td>
<td>-5.2</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>2  Mary not slap the witch green .</td>
<td>-16.3</td>
<td>-5.7</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>3  Mary not give slap of the green witch .</td>
<td>-18.1</td>
<td>-4.9</td>
<td>-9</td>
<td>1</td>
</tr>
<tr>
<td>4  Mary not give of green witch .</td>
<td>-16.5</td>
<td>-5.1</td>
<td>-8</td>
<td>1</td>
</tr>
<tr>
<td>5  Mary did not slap the witch green .</td>
<td>-20.3</td>
<td>-4.7</td>
<td>-8</td>
<td>1</td>
</tr>
<tr>
<td>6  Mary did not slap green witch .</td>
<td>-15.5</td>
<td>-3.2</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>7  Mary not slap of the witch green .</td>
<td>-19.2</td>
<td>-5.3</td>
<td>-8</td>
<td>1</td>
</tr>
<tr>
<td>8  Mary did not give slap of witch green .</td>
<td>-23.2</td>
<td>-5.8</td>
<td>-9</td>
<td>1</td>
</tr>
<tr>
<td>9  Mary did not give slap of the green witch .</td>
<td>-21.8</td>
<td>-4.4</td>
<td>-10</td>
<td>1</td>
</tr>
<tr>
<td>10 Mary did slap the witch green .</td>
<td>-15.5</td>
<td>-6.9</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>11 Mary did slap the green witch .</td>
<td>-17.4</td>
<td>-5.3</td>
<td>-8</td>
<td>0</td>
</tr>
<tr>
<td>12 Mary did slap witch green .</td>
<td>-16.5</td>
<td>-5.9</td>
<td>-6</td>
<td>1</td>
</tr>
<tr>
<td>13 Mary did slap the green witch .</td>
<td>-14.3</td>
<td>-7.1</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>14 Mary did not slap of green witch .</td>
<td>-24.2</td>
<td>-5.3</td>
<td>-9</td>
<td>1</td>
</tr>
<tr>
<td>15 Mary did not slap of the green witch .</td>
<td>-25.2</td>
<td>-5.5</td>
<td>-9</td>
<td>1</td>
</tr>
</tbody>
</table>

Methods to Adjust Feature Weights

- **Maximum entropy** [Och and Ney, ACL2002]
  - match **expectation** of feature values of model and data

- **Minimum error rate** training [Och, ACL2003]
  - try to rank best translations first in n-best list
  - can be adapted for various error metrics, even BLEU

- **Ordinal regression** [Shen et al., NAACL2004]
  - separate $k$: worst from the $k$: best translations
Discriminative Training: Outlook

- Many more features
- Discriminative training on entire training set
- Reranking vs. decoding
  - reranking: expensive, global features possible
  - decoding: integrating features in search reduces search errors
⇒ First decoding, then reranking

- Decoding
- Statistical Modeling
- EM Algorithm
- Word Alignment
- Phrase-Based Translation
- Discriminative Training

Syntax-Based Statistical MT
Syntax-based SMT

- Why Syntax?
- Yamada and Knight: translating into trees
- Wu: tree-based transfer
- Chiang: hierarchical transfer
- Collins, Kucerova, and Koehn: clause structure
- Koehn: factored translation models
- Other approaches

The Challenge of Syntax

- The classical machine translation pyramid

interlingua

foreign semantics

english semantics

foreign syntax

english syntax

foreign words

english words

foreign words
Advantages of Syntax-Based Translation

- **Reordering** for syntactic reasons
  - e.g., move German object to end of sentence
- Better explanation for **function words**
  - e.g., prepositions, determiners
- Conditioning to **syntactically related words**
  - translation of verb may depend on subject or object
- Use of **syntactic language models**
  - ensuring grammatical output

Syntactic Language Model

- **Good syntax tree** → good English
- Allows for **long distance constraints**

- Left translation preferred by syntactic LM
String to Tree Translation

- Use of English syntax trees [Yamada and Knight, 2001]
  - exploit rich resources on the English side
  - obtained with statistical parser [Collins, 1997]
  - flattened tree to allow more reorderings
  - works well with syntactic language model

Yamada and Knight [2001]

Kare ha ongaku wo kiku no ga daisuki desu

[from Yamada and Knight, 2001]
Reordering Table

| Original Order | Reordering       | p(reorder|original) |
|----------------|------------------|--------------|
| PRP VB1 VB2    | PRP VB1 VB2      | 0.074        |
| PRP VB1 VB2    | PRP VB2 VB1      | 0.723        |
| PRP VB1 VB2    | VB1 PRP VB2      | 0.061        |
| PRP VB1 VB2    | VB1 VB2 PRP      | 0.037        |
| PRP VB1 VB2    | VB2 PRP VB1      | 0.083        |
| PRP VB1 VB2    | VB2 VB1 PRP      | 0.021        |
| VB TO          | VB TO            | 0.107        |
| VB TO          | TO VB            | 0.893        |
| TO NN          | TO NN            | 0.251        |
| TO NN          | NN TO            | 0.749        |

Decoding as Parsing

- Chart Parsing

```
PRP
he

kare ha ongaku wo kiku no ga daisuki desu
```

- Pick Japanese words
- Translate into tree stumps
Decoding as Parsing

• Chart Parsing

```
PRP
he

NN
music

TO
to

kare ha ongaku wo kiku no ga daisuki desu
```

• Pick Japanese words

• Translate into tree stumps

Decoding as Parsing

```
PRP
he

NN
music

TO
to

PP

kare ha ongaku wo kiku no ga daisuki desu
```

• Adding some more entries...
Decoding as Parsing

\[
\begin{array}{c}
\text{PP} \\
\text{PRP} \quad \text{NN} \quad \text{TO} \quad \text{VB} \\
he \quad \text{music} \quad \text{to} \quad \text{listening}
\end{array}
\]

kare ha ongaku wo kiku no ga daisuki desu

- Combine entries
Decoding as Parsing

\[ k\text{a}r\text{e h}a \text{ o}n\text{g}a\kwidth{2pt}k\text{u w}o \text{k}iku \text{n}o \text{g}a \text{d}ais\text{u}ki \text{d}e\text{s}u \]

- **Finished** when all foreign words covered
Yamada and Knight: Training

- **Parsing** of the English side
  - using Collins statistical parser
- **EM training**
  - translation model is used to map training sentence pairs
  - EM training finds low-perplexity model
  → *unity of training and decoding* as in IBM models

Is the Model Realistic?

- Do English trees **match** foreign strings?
- Crossings between French-English [Fox, 2002]
  - 0.29-6.27 per sentence, depending on how it is measured
- Can be reduced by
  - **flattening tree**, as done by [Yamada and Knight, 2001]
  - detecting **phrasal** translation
  - **special treatment** for small number of constructions
- Most coherence between **dependency structures**
Inversion Transduction Grammars

• Generation of **both** English and foreign trees [Wu, 1997]

• Rules (binary and unary)
  - \( A \rightarrow A_1A_2\parallel A_1A_2 \)
  - \( A \rightarrow A_1A_2\parallel A_2A_1 \)
  - \( A \rightarrow e\parallel f \)
  - \( A \rightarrow e\parallel* \)
  - \( A \rightarrow *\parallel f \)

⇒ **Common binary tree** required
  - limits the complexity of reorderings

Syntax Trees

• English binary tree
Syntax Trees (2)

Maria no daba una bofetada a la bruja verde

- Spanish binary tree

Syntax Trees (3)

Mary did not slap Maria * no daba una bofetada a la verde bruja

- Combined tree with reordering of Spanish
Inversion Transduction Grammars

- Decoding by parsing (as before)
- Variations
  - may use real syntax on either side or both
  - may use multi-word units at leaf nodes

Chiang: Hierarchical Phrase Model

- Chiang [ACL, 2005] (best paper award!)
  - context free bi-grammar
  - one non-terminal symbol
  - right hand side of rule may include non-terminals and terminals
- Competitive with phrase-based models in 2005 DARPA/NIST evaluation
Types of Rules

- **Word** translation
  - \( X \rightarrow \textit{maison} \parallel \textit{house} \)

- **Phrasal** translation
  - \( X \rightarrow \textit{daba una bofetada} \parallel \textit{slap} \)

- **Mixed** non-terminal / terminal
  - \( X \rightarrow X \textit{bleue} \parallel \textit{blue} X \)
  - \( X \rightarrow \textit{ne X pas} \parallel \textit{not} X \)
  - \( X \rightarrow X1 X2 \parallel X2 \text{ of } X1 \)

- **Technical rules**
  - \( S \rightarrow S X \parallel S X \)
  - \( S \rightarrow X \parallel X \)

Learning Hierarchical Rules

Maria no daba una botefada a la bruja verde
Mary
slap
green
witch

\( X \rightarrow X \text{ verde} \parallel \text{green } X \)
**Learning Hierarchical Rules**

X → a la X || the X

**Details of Chiang’s Model**

- Too many rules
  → **filtering** of rules necessary
- **Efficient** parse decoding possible
  - hypothesis stack for each span of foreign words
  - only **one non-terminal** → hypotheses comparable
  - **length limit** for spans that do not start at beginning
Clause Level Restructuring [Collins et al.]

- 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

- Details see [Collins, Kucerova and Koehn, ACL 2005]

Syntax tree from German parser
  - statistical parser by Amit Dubay, 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
**Systematic Reordering German → English**

- 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>

**Improved Translations**

- we must also this criticism should be taken seriously.
  → we must also take this criticism seriously.

- i am with him that it is necessary, the institutional balance by means of a political revaluation of both the commission and the council to maintain.
  → i agree with him in this, that it is necessary to maintain the institutional balance by means of a political revaluation of both the commission and the council.

- thirdly, we believe that the principle of differentiation of negotiations note.
  → thirdly, we maintain the principle of differentiation of negotiations.

- perhaps it would be a constructive dialog between the government and opposition parties, social representative a positive impetus in the right direction.
  → perhaps a constructive dialog between government and opposition parties and social representative could give a positive impetus in the right direction.
Factored Translation Models

• **Factored representation** of words

  \[
  \begin{array}{c}
  \text{surface} \\
  \text{stem} \\
  \text{part-of-speech} \\
  \text{morphology} \\
  \text{word class} \\
  \ldots
  \end{array}
  \quad \Rightarrow 
  \begin{array}{c}
  \text{surface} \\
  \text{stem} \\
  \text{part-of-speech} \\
  \text{morphology} \\
  \text{word class} \\
  \ldots
  \end{array}
  \]

• **Goals**
  
  – **Generalization**, e.g. by translating stems, not surface forms
  – **Additional information** within model (using syntax for reordering, language modeling)

Decomposing Translation: Example

• **Translating** stem and syntactic information **separately**

  \[
  \begin{array}{c}
  \text{stem} \\
  \text{part-of-speech} \\
  \text{morphology}
  \end{array}
  \quad \Rightarrow 
  \begin{array}{c}
  \text{stem} \\
  \text{part-of-speech} \\
  \text{morphology}
  \end{array}
  \]

• **Generate surface** form on target side

  \[
  \begin{array}{c}
  \text{surface} \\
  \text{stem} \\
  \text{part-of-speech} \\
  \text{morphology}
  \end{array}
  \]

Factored Models: Open Questions

- What is the best decomposition into translation and generation steps?
- **What information** is useful?
  - translation: mostly lexical, or stems for richer statistics
  - reordering: syntactic information useful
  - language model: syntactic information for overall grammatical coherence

- Use of annotation tools
- Use of **automatically discovered** generalizations (word classes)
- **Back-off** models (use complex mappings, if available)

Other Syntax-Based Approaches

- ISI: extending work of Yamada/Knight
  - more complex rules
  - performance approaching phrase-based
- Prague: Translation via **dependency structures**
  - parallel Czech–English dependency treebank
  - tecto-grammatical translation model [EACL 2003]
- U.Alberta/Microsoft: **treelet translation**
  - translating from English into foreign languages
  - using dependency parser in English
  - project **dependency tree** into foreign language for training
  - map parts of the dependency tree ("treelets") into foreign languages
Other Syntax-Based Approaches (2)

- **Reranking** phrase-based SMT output with syntactic features
  - create n-best list with phrase-based system
  - POS tag and parse candidate translations
  - rerank with syntactic features
  - see [Koehn, 2003] and JHU Workshop [Och et al., 2003]

- JHU Summer workshop 2005
  - **Genpar**: tool for syntax-based SMT

Syntax: Does it help?

- **Not yet**
  - best systems still phrase-based, treat words as tokens

- **Well, maybe...**
  - work on reordering German
  - automatically trained tree transfer systems promising

- **Why not yet?**
  - if real syntax, we need **good parsers** — are they good enough?
  - syntactic annotations add a level of **complexity**
    → difficult to handle, slow to train and decode
  - few researchers good at statistical modeling and understand syntactic theories