Empirical Methods in Natural Language Processing
Lecture 14
Machine translation (I): Introduction

Philipp Koehn

21 February 2008
Machine translation

- Task: make sense of foreign text like

- One of the oldest problems in Artificial Intelligence

- AI-hard: reasoning and world knowledge required
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**

- **foreign/English parallel text**
  - statistical analysis
  - Translation Model

- **English text**
  - statistical analysis
  - Language Model

- **Decoding Algorithm**
The machine translation pyramid

interlingua

foreign
semantics

foreign
syntax

foreign
words

english
semantics

english
syntax

english
words
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

Kare ha ongaku wo kiku no ga daisuki desu

[from Yamada and Knight, 2001]
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!)
Automatic evaluation

- 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

[from Callison-Burch et al., 2006, EACL]
Correlation? [Callison-Burch et al., 2006]

- 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

[from Callison-Burch et al., 2006, EACL]
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, newsstexts, 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
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
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 based on how easy they translate into each other

⇒ Approximation of language families
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-translation**

- *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/](http://www.statmt.org/europarl/)
  - **Acquis Communitaire**: 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](http://www.ldc.upenn.edu/)
  - lots more French/English, Spanish/French/English from [LDC](http://www.ldc.upenn.edu/)

- Available monolingual text (for language modeling)
  - 2.8 billion words of English from [LDC](http://www.ldc.upenn.edu/)
  - 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)
In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.
Partially excellent translations

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.
Mangled grammar

In the First Two Months Guangdong’s Export of High-Tech Products 3.76 Billion US Dollars
Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong’s export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province’s total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong’s export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province’s total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion us dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem
Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.