02 – Statistical machine translation
IA161 Advanced Techniques of Natural Language Processing

V. Baisa

NLP Centre, FI MU, Brno

September 25, 2019
1 Introduction

2 Machine Translation

3 Phrase-based translation model

4 Machine translation evaluation
Moses is an implementation of the statistical (or data-driven) approach to machine translation (MT). This is the dominant approach in the field at the moment, and is employed by the online translation systems deployed by the likes of Google and Microsoft.

1. Mojžíš je implementace statistické (nebo řízené daty) přístupu k strojového překladu (MT). To je převládajícím přístupem v oblasti v současné době, a je zaměstnán pro on-line překladatelských systémů nasazených likes Google a Microsoft.

2. Moses je implementace statistického (nebo daty řízeného) přístupu k strojovému překladu (MT). V současné době jde o převažující přístup v rámci strojového překladu, který je použit online překladovými systémy nasazenými Googlem a Microsoftem.

3. Mojžíš je provádění statistické (nebo aktivovaný) přístup na strojový překlad (mt). To je dominantní přístup v oblasti v tuto chvíli, a zaměstnává on - line překlad systémů uskutečněné takové, Google a Microsoft.

4. Mojžíš je implementace statistického (nebo datově řízeného) přístupu k strojovému překladu (MT). To je v současné době dominantní přístup v oboru a je využíván online překladatelskými systémy, které používají společnosti Google a Microsoft.
Statistical Machine Translation

- rule-based systems motivated by linguistics
- SMT inspired by information theory and statistics
- Google Translate (before 2016), Bing Translator, Moses
- **gisting**: the most frequent usage of MT on Internet
- in fact, MT output is always post-edited
- neural networks: boom in the last few years (state-of-the-art)
Machine translation: what is translated

- web pages
- technical manuals, how-tos
- scientific documents, papers, articles
- leaflets, flyers, catalogues
- texts from limited domains in general
- Wikipedia articles (CS–SK)
intense collecting of data
development of systems driven by evaluation metrics
EU: 24 official languages (EuroMatrix)
software companies focus on English as source language (i18n)
large language pairs (En↔Sp, En↔Fr): fairly high-quality translation
Google Translate as a gold standard
morphologically rich languages: worse results
En-* and *-En pairs prevail
Moses: freely available statistical machine translation [Koehn, 2007]
Data: parallel corpora

- OPUS: parallel texts of various source, one of the biggest resources [Tiedemann and Nygaard, 2004]
- Acquis Communautaire: EU laws [Steinberger et al., 2006]
- EUR-Lex: access to European Union law
- DGT translation memory [Steinberger et al., 2013], MyMemory
- freely available corpora are usually of size of 10–100 million words
- multilingual webpages (Wikipedia)
- comparable corpora: texts from the same domain
Sentence alignment

- 1:1, 1:0, 0:1, 1:2, 2:1, ... alignments
- Gale-Church (sentence lengths)
- Hunalign (with a dictionary, G-Ch is a fallback)
- BLEUalign (MT-based sentence alignment)
- cognates
Schema of SMT

Spanish/English Bilingual Text

Statistical Analysis

Spanish → Translation Model → Broken English

Decoding Algorithm
\[ \text{argmax } P(e) \times p(s|e) \]

English Text

Statistical Analysis

Broken English → Language Model → English
**SMT – noisy channel**

Developed by Shannon (1948) [Shannon, 1956] for self-correcting codes, for corrections of coded signals transferred through noisy channels based on information about a source message and types of errors occurring in the channels.

Another application: OCR, Optical Character Recognition. It is messy, but we can estimate what was in the source text from a language model and frequent errors: 1-1-I, rn-m etc.

\[
    e^* = \arg \max_e p(e|f) \\
    = \arg \max_e \frac{p(e)p(f|e)}{p(f)} \\
    = \arg \max_e p(e)p(f|e).
\]

We will speak about language models later.
Lexical translation

Standard translation dictionary does not contain translation probabilities for word meanings.

\[\text{key} \rightarrow \text{klíč, tónina, klávesa}\]

How often are the individual equivalents used?

\[\text{key} \rightarrow \text{klíč (0.7), tónina (0.18), klávesa (0.08), ...}\]

We need a lexical probability distribution \(p_f\) with the property:

\[
\sum_{e} p_f(e) = 1 \\
\forall e : 0 \leq p_f(e) \leq 1
\]

\[p_{\text{klíč}}(\text{key}) \neq p_{\text{mrkev}}(\text{carrot})\]
Word alignment

GIZA++ is the most widely used tool. [Och and Ney, 2003]

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

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

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

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

\[ p(\text{la|the}) = 0.453 \]
\[ p(\text{le|the}) = 0.334 \]
\[ p(\text{maison|house}) = 0.876 \]
\[ p(\text{bleu|blue}) = 0.563 \]
<table>
<thead>
<tr>
<th></th>
<th>michael</th>
<th>geht</th>
<th>davon</th>
<th>aus</th>
<th>,</th>
<th>dass</th>
<th>er</th>
<th>in</th>
<th>haus</th>
<th>bleibt</th>
</tr>
</thead>
<tbody>
<tr>
<td>house</td>
<td></td>
<td>assume</td>
<td>that</td>
<td>he</td>
<td>will</td>
<td>stay</td>
<td>in</td>
<td>the</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

V. Baisa

IA161 Advanced NLP

02 – Statistical machine translation
Lexical translation problems

- John kicked the bucket.
- Does John live here?

<table>
<thead>
<tr>
<th>John</th>
<th>Biss</th>
<th>Ins</th>
<th>Grass</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kicked</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bucket</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>John</th>
<th>Wohnt</th>
<th>Hier</th>
<th>Nicht</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Does</td>
<td>?</td>
<td>?</td>
<td></td>
</tr>
<tr>
<td>Not</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Live</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Here</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phrase-based translation model

State-of-the-art of SMT. Not only words, but whole phrases are translated at a time. [Koehn et al., 2003] [Chiang, 2005]

Phrases are not linguistically motivated. German *am* is usually not translated by one word *with*. Statistically significant context *spass am* helps with a proper translation. Common phrases would be segmented in a different way: *(fun (with (the game)))*. 
Advantages of PBMT

- we often translate $n : m$ words, a word is unsuitable element
- the translation of groups of words helps with translation ambiguity
- and also fluency
- systems can learn longer phrases, ad infinitum, if data is available
- the model is simpler: fertility, NULL tokens are not needed
Phrase extraction

michael
assumes
that
he
will
stay
in
the
house

gibt
aus
dass
er
in
haus
bleibt
Automatic evaluation of translation

- advantages: speed, price; disadvantages: do we measure quality of translation?
- gold standard: manually prepared reference translations
- candidate $c$ is compared with $n$ reference translations $r_i$
- various approaches: n-gram agreement between $c$ and $r_i$, edit distance, . . .
- BLEU: the most widely used [Papineni et al., 2002]
- METEOR: correlates best with human evaluation [Banerjee and Lavie, 2005]
BLEU

- the most popular (a standard), the most widely used, the oldest (2001)
- IBM, Papineni [Papineni et al., 2002]
- n-gram agreement between references and candidates
- precision for 1–4-grams
- brevity penalty

\[
\text{BLEU} = \min \left(1, \frac{\text{output-length}}{\text{reference-length}} \right) \left( \prod_{i=1}^{4} \text{precision}_i \right)^{\frac{1}{4}}
\]
**BLEU – an example**

**SYSTEM A:**

<table>
<thead>
<tr>
<th>Israeli officials</th>
<th>responsibility of</th>
<th>airport</th>
<th>safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-GRAM MATCH</td>
<td>1-GRAM MATCH</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**REFERENCE:**

Israeli officials are responsible for airport security

**SYSTEM B:**

<table>
<thead>
<tr>
<th>airport security</th>
<th>Israeli officials are responsible</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-GRAM MATCH</td>
<td>4-GRAM MATCH</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>metrics</th>
<th>system A</th>
<th>system B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision (1gram)</td>
<td>3/6</td>
<td>6/6</td>
</tr>
<tr>
<td>precision (2gram)</td>
<td>1/5</td>
<td>4/5</td>
</tr>
<tr>
<td>precision (3gram)</td>
<td>0/4</td>
<td>2/4</td>
</tr>
<tr>
<td>precision (4gram)</td>
<td>0/3</td>
<td>1/3</td>
</tr>
<tr>
<td>brevity penalty</td>
<td>6/7</td>
<td>6/7</td>
</tr>
<tr>
<td>BLEU</td>
<td>0 %</td>
<td>52 %</td>
</tr>
</tbody>
</table>
Translation quality according to language pairs

<table>
<thead>
<tr>
<th>input language</th>
<th>output language</th>
<th>Czech</th>
<th>German</th>
<th>English</th>
<th>Finnish</th>
<th>French</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Czech</td>
<td>26.2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>German</td>
<td>29.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td>18.8</td>
<td>24.9</td>
<td></td>
<td></td>
<td>15.5</td>
<td>33.6</td>
<td>24.3</td>
</tr>
<tr>
<td>Finnish</td>
<td>19.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>French</td>
<td>33.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Russian</td>
<td>27.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://matrix.statmt.org/ [Koehn, 2007]


References II


References III


References IV
