Introduction to Machine Translation

CMSC 723 / LING 723 / INST 725

Marine Carpuat

Slides & figure credits: Philipp Koehn
mt-class.org
Today’s topics
Machine Translation

• Historical Background
  • Machine Translation is an old idea

• Machine Translation Today
  • Use cases and method

• Machine Translation Evaluation
When I look at an article in Russian, I say to myself: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.

Warren Weaver
1950s-1960s

• 1954 Georgetown-IBM experiment
  • 250 words, 6 grammar rules

• 1966 ALPAC report
  • Skeptical in research progress
  • Led to decreased US government funding for MT
Rule based systems

• Approach
  • Build dictionaries
  • Write transformation rules
  • Refine, refine, refine

• Meteo system for weather forecasts (1976)

• Systran (1968), …

"have" :=

if
  subject(animate)
  and object(owned-by-subject)
then
  translate to "kade... aahe"
if
  subject(animate)
  and object(kinship-with-subject)
then
  translate to "laa... aahe"
if
  subject(inanimate)
then
  translate to "madhye... aahe"
A Statistical Approach to Machine Translation

IBM
Thomas J. Watson Research Center
Yorktown Heights, NY

In this paper, we present a statistical approach to machine translation. We describe the application of our approach to translation from French to English and give preliminary results.

The COLING Paper Review

The validity of statistical (information theoretic) approach to MT has indeed been recognized, as the authors mention, by Weaver as early as 1949. And was universally recognized as mistaken by 1950. (cf. Hutchins, MT: Past, Present, Future, Ellis Horwood, 1986, pp. 30ff. and references therein) The crude force of computers is not science. The paper is simply beyond the scope of COLING.

More about the IBM story: 20 years of bitext workshop
Statistical Machine Translation

• 1990s: increased research

• Mid 2000s: phrase-based MT
  • (Moses, Google Translate)

• Around 2010: commercial viability

• Since mid 2010s: neural network models
MT History: Hype vs. Reality

- Georgetown experiment
- Expert systems / 5th generation AI
- Statistical MT
- Neural MT
Reporters learned from the Ministry of Environmental Protection, "Water 10" requirements before the end of this year before the municipality, the provincial capital city, plans to build a separate city to solve the basic black and black water. Up to now, the country's 224 prefecture-level and above cities were identified to confirm the black and white water 2082, of which 34.9% to complete the renovation, 28.4% is remediation, 22.8% is carrying out the project early.
How Good is Machine Translation?
French > English

A l’orée de ce débat télévisé inédit dans l’histoire de la Ve République, on attendait une forme de «Tous sur Macron» mais c’est la candidate du Front national qui s’est retrouvée au cœur des premières attaques de ses quatre adversaires d’un soir, favorisées par le premier thème abordé, les questions de société et donc de sécurité, d’immigration et de laïcité.

At the beginning of this televised debate, which was unheard of in the history of the Fifth Republic, a "Tous sur Macron" was expected, but it was the candidate of the National Front who found itself at the heart of the first attacks of its four Opponents of one evening, favored by the first theme tackled, the issues of society and thus security, immigration and secularism.
The Vauquois Triangle
Learning from Data

• What is the best translation?
  
  \[
  \begin{aligned}
  \text{Sicherheit} & \rightarrow \text{security} \; 14,516 \\
  \text{Sicherheit} & \rightarrow \text{safety} \; 10,015 \\
  \text{Sicherheit} & \rightarrow \text{certainty} \; 334
  \end{aligned}
  \]

• Counts in parallel corpus (aka bitext)
  • Here European Parliament corpus
Learning from Data

• What is most fluent?

  a problem for translation
  a problem of translation
  a problem in translation

• A language modeling problem!
Word Alignment

<table>
<thead>
<tr>
<th>Michael</th>
<th>geht</th>
<th>davon</th>
<th>aus</th>
<th>dass</th>
<th>er</th>
<th>im</th>
<th>haus</th>
<th>bleibt</th>
</tr>
</thead>
<tbody>
<tr>
<td>assumes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>that</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>he</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>will</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stay</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>house</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Phrase-based Models

• Input segmented in phrases
• Each phrase is translated in output language
• Phrases are reordered
Neural MT
What is MT good (enough) for?

• **Assimilation:** reader initiates translation, wants to know content
  • User is tolerant of inferior quality
  • Focus of majority of research

• **Communication:** participants in conversation don’t speak same language
  • Users can ask questions when something is unclear
  • Chat room translations, hand-held devices
  • Often combined with speech recognition

• **Dissemination:** publisher wants to make content available in other languages
  • High quality required
  • Almost exclusively done by human translators
## Applications

<table>
<thead>
<tr>
<th>HTER</th>
<th>assessment</th>
<th>application examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>publishable</td>
<td>Seamless bridging of language divide</td>
</tr>
<tr>
<td>10%</td>
<td>editable</td>
<td>Automatic publication of official announcements</td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td>Increased productivity of human translators</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Access to official publications</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Multi-lingual communication (chat, social networks)</td>
</tr>
<tr>
<td>30%</td>
<td>gistable</td>
<td>Information gathering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Trend spotting</td>
</tr>
<tr>
<td>40%</td>
<td>triagable</td>
<td>Identifying relevant documents</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
State of the Art (rough estimates)

<table>
<thead>
<tr>
<th>HTER</th>
<th>assessment</th>
<th>language pairs and domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>publishable</td>
<td>French-English restricted domain</td>
</tr>
<tr>
<td>10%</td>
<td>editable</td>
<td>French-English technical document localization</td>
</tr>
<tr>
<td>20%</td>
<td></td>
<td>French-English news stories</td>
</tr>
<tr>
<td>30%</td>
<td>gistable</td>
<td>English-German news stories</td>
</tr>
<tr>
<td>40%</td>
<td>triagable</td>
<td>English-Czech open domain</td>
</tr>
<tr>
<td>50%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Today’s topics
Machine Translation

• Historical Background
  • Machine Translation is an old idea

• Machine Translation Today
  • Use cases and method

• Machine Translation Evaluation
How good is a translation?
Problem: no single right answer

This airport’s security is the responsibility of the Israeli security officials.
Evaluation

• How good is a given machine translation system?

• Many different translations acceptable

• Evaluation metrics
  • Subjective judgments by human evaluators
  • Automatic evaluation metrics
  • Task-based evaluation
Adequacy and Fluency

• Human judgment
  • Given: machine translation output
  • Given: input and/or reference translation
  • Task: assess quality of MT output

• Metrics
  • Adequacy: does the output convey the meaning of the input sentence? Is part of the message lost, added, or distorted?
  • Fluency: is the output fluent? Involves both grammatical correctness and idiomatic word choices.
## Fluency and Adequacy: Scales

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>flawless English</td>
</tr>
<tr>
<td>4</td>
<td>good English</td>
</tr>
<tr>
<td>3</td>
<td>non-native English</td>
</tr>
<tr>
<td>2</td>
<td>disfluent English</td>
</tr>
<tr>
<td>1</td>
<td>incomprehensible</td>
</tr>
</tbody>
</table>

- 5: all meaning
- 4: most meaning
- 3: much meaning
- 2: little meaning
- 1: none
Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l'ue.

Reference: rather, the two countries form a laboratory needed for the internal working of the eu.

<table>
<thead>
<tr>
<th>Translation</th>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>both countries are a necessary laboratory the internal operation of the eu.</td>
<td><img src="image" alt="Score" /></td>
<td><img src="image" alt="Score" /></td>
</tr>
<tr>
<td>both countries are a necessary laboratory at internal functioning of the eu.</td>
<td><img src="image" alt="Score" /></td>
<td><img src="image" alt="Score" /></td>
</tr>
<tr>
<td>the two countries are rather a laboratory necessary for the internal workings of the eu.</td>
<td><img src="image" alt="Score" /></td>
<td><img src="image" alt="Score" /></td>
</tr>
<tr>
<td>the two countries are rather a laboratory for the internal workings of the eu.</td>
<td><img src="image" alt="Score" /></td>
<td><img src="image" alt="Score" /></td>
</tr>
<tr>
<td>the two countries are rather a necessary laboratory internal workings of the eu.</td>
<td><img src="image" alt="Score" /></td>
<td><img src="image" alt="Score" /></td>
</tr>
</tbody>
</table>

Annotator: Philipp Koehn  Task: WMT06 French-English

Instructions:

5= All Meaning
4= Most Meaning
3= Much Meaning
2= Little Meaning
1= None

5= Flawless English
4= Good English
3= Non-native English
2= Disfluent English
1= Incomprehensible
Let’s try:
rate fluency & adequacy on 1-5 scale

- Source:
  N’y aurait-il pas comme une vague hypocrisie de votre part ?

- Reference:
  Is there not an element of hypocrisy on your part?

- System1:
  Would it not as a wave of hypocrisy on your part?

- System2:
  Is there would be no hypocrisy like a wave of your hand?

- System3:
  Is there not as a wave of hypocrisy from you?
Challenges in MT evaluation

• No single correct answer

• Human evaluators disagree
Automatic Evaluation Metrics

• Goal: computer program that computes quality of translations

• Advantages: low cost, optimizable, consistent

• Basic strategy
  • Given: MT output
  • Given: human reference translation
  • Task: compute similarity between them
Precision and Recall of Words

SYSTEM A: Israeli officials responsibility of airport safety

REFERENCE: Israeli officials are responsible for airport security

Precision
\[
\frac{\text{correct}}{\text{output-length}} = \frac{3}{6} = 50\%
\]

Recall
\[
\frac{\text{correct}}{\text{reference-length}} = \frac{3}{7} = 43\%
\]

F-measure
\[
\frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%
\]
Precision and Recall of Words

SYSTEM A: Israeli officials responsibility of airport safety
REFERENCE: Israeli officials are responsible for airport security
SYSTEM B: airport security Israeli officials are responsible

<table>
<thead>
<tr>
<th>Metric</th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>precision</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>recall</td>
<td>43%</td>
<td>100%</td>
</tr>
<tr>
<td>f-measure</td>
<td>46%</td>
<td>100%</td>
</tr>
</tbody>
</table>

flaw: no penalty for reordering
Word Error Rate

Minimum number of editing steps to transform output to reference

**match**: words match, no cost

**substitution**: replace one word with another

**insertion**: add word

**deletion**: drop word

Levenshtein distance

\[
\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference-length}}
\]
WER example

```
<table>
<thead>
<tr>
<th></th>
<th>System A</th>
<th>System B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metric</td>
<td>word error rate (WER)</td>
<td>57%</td>
</tr>
</tbody>
</table>
```
BLEU

Bilingual Evaluation Understudy

N-gram overlap between machine translation output and reference translation

Compute precision for n-grams of size 1 to 4

Add brevity penalty (for too short translations)

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

Typically computed over the entire corpus, not single sentences
Multiple Reference Translations

To account for variability, use multiple reference translations

- n-grams may match in any of the references
- closest reference length used

Example

SYSTEM:

<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>2-GRAM MATCH</td>
<td>1-GRAM</td>
<td></td>
</tr>
</tbody>
</table>

Israeli officials are responsible for airport security
Israel is in charge of the security at this airport

REFERENCES:

The security work for this airport is the responsibility of the Israeli government
Israeli side was in charge of the security of this airport
**BLEU examples**

**SYSTEM A:**
- Israeli officials responsibility of airport safety
  - 2-GRAM MATCH

**REFERENCE:**
- Israeli officials are responsible for airport security

**SYSTEM B:**
- airport security Israeli officials are responsible
  - 2-GRAM MATCH

<table>
<thead>
<tr>
<th>Metric</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>
Semantics-aware metrics: e.g., METEOR

Partial credit for matching stems

SYSTEM: Jim went home
REFERENCE: Joe goes home

Partial credit for matching synonyms

SYSTEM: Jim walks home
REFERENCE: Joe goes home

Use of paraphrases
Drawbacks of Automatic Metrics

• All words are treated as equally relevant

• Operate on local level

• Scores are meaningless (absolute value not informative)

• Human translators score low on BLEU
Yet automatic metrics such as BLEU correlate with human judgement.
Caveats: bias toward statistical systems
Automatic metrics

• Essential tool for system development

• Use with caution: not suited to rank systems of different types

• Still an open area of research
  • Connects with semantic analysis
Task-Based Evaluation
Post-Editing Machine Translation

Measuring time spent on producing translations

– baseline: translation from scratch
– post-editing machine translation

But: time consuming, depend on skills of translator and post-editor

Metrics inspired by this task

– TER: based on number of editing steps
  Levenshtein operations (insertion, deletion, substitution) plus movement
– HTER: manually construct reference translation for output, apply TER
  (very time consuming, used in DARPA GALE program 2005-2011)
Task-Based Evaluation
Content Understanding Tests

Given machine translation output, can monolingual target side speaker answer questions about it?

1. basic facts: who? where? when? names, numbers, and dates
2. actors and events: relationships, temporal and causal order
3. nuance and author intent: emphasis and subtext

Very hard to devise questions

Sentence editing task (WMT 2009–2010)

- person A edits the translation to make it fluent (with no access to source or reference)
- person B checks if edit is correct
  → did person A understand the translation correctly?
Today’s topics
Machine Translation

• Historical Background
  • Machine Translation is an old idea

• Machine Translation Today
  • Use cases and method

• Machine Translation Evaluation