Machine Translation

- Automatically translate one natural language into another.

Mary didn’t slap the green witch.

Maria no dio una bofetada a la bruja verde.

(Mary do not gave a slap to the witch green.)

Thousands of Languages Are Spoken

<table>
<thead>
<tr>
<th>Language</th>
<th>Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mandarin</td>
<td>885,000,000</td>
</tr>
<tr>
<td>Spanish</td>
<td>332,000,000</td>
</tr>
<tr>
<td>English</td>
<td>322,000,000</td>
</tr>
<tr>
<td>Bengali</td>
<td>189,000,000</td>
</tr>
<tr>
<td>Hindi</td>
<td>182,000,000</td>
</tr>
<tr>
<td>Portuguese</td>
<td>170,000,000</td>
</tr>
<tr>
<td>Russian</td>
<td>170,000,000</td>
</tr>
<tr>
<td>Japanese</td>
<td>125,000,000</td>
</tr>
<tr>
<td>German</td>
<td>98,000,000</td>
</tr>
<tr>
<td>Wuu (China)</td>
<td>77,175,000</td>
</tr>
<tr>
<td>Japanese</td>
<td>75,500,800</td>
</tr>
<tr>
<td>Korean</td>
<td>75,000,000</td>
</tr>
<tr>
<td>French</td>
<td>72,000,000</td>
</tr>
<tr>
<td>Vietnamese</td>
<td>67,662,000</td>
</tr>
<tr>
<td>Telugu</td>
<td>66,350,000</td>
</tr>
<tr>
<td>Yee (China)</td>
<td>66,000,000</td>
</tr>
<tr>
<td>Marathi</td>
<td>64,783,000</td>
</tr>
<tr>
<td>Tamil</td>
<td>63,075,000</td>
</tr>
<tr>
<td>Turkish</td>
<td>59,000,000</td>
</tr>
<tr>
<td>Urdu</td>
<td>58,000,000</td>
</tr>
<tr>
<td>Min Nan (C)</td>
<td>49,000,000</td>
</tr>
<tr>
<td>Jinyu (C)</td>
<td>45,000,000</td>
</tr>
<tr>
<td>Gujrati</td>
<td>44,000,000</td>
</tr>
<tr>
<td>Polish</td>
<td>44,000,000</td>
</tr>
<tr>
<td>Arabic</td>
<td>42,500,000</td>
</tr>
<tr>
<td>Ukranian</td>
<td>41,000,000</td>
</tr>
<tr>
<td>Italian</td>
<td>37,000,000</td>
</tr>
<tr>
<td>Xiang (C)</td>
<td>36,015,000</td>
</tr>
<tr>
<td>Malayalam</td>
<td>34,022,000</td>
</tr>
<tr>
<td>Hakka (C)</td>
<td>34,000,000</td>
</tr>
<tr>
<td>Kannada</td>
<td>33,663,000</td>
</tr>
<tr>
<td>Oriya</td>
<td>31,000,000</td>
</tr>
<tr>
<td>Panjabi</td>
<td>30,000,000</td>
</tr>
<tr>
<td>Sunda</td>
<td>27,000,000</td>
</tr>
</tbody>
</table>

Translation Quality: what’s the current status?

- Achieving literary quality translation is very difficult.
- Existing MT systems can generate rough translations that frequently at least convey the gist of a document.
- High quality translations possible when specialized to narrow domains, e.g. weather forecasts.
- Some MT systems used in computer-aided translation in which a bilingual human post-edits the output to produce more readable accurate translations.

Outline

- Issues in machine translation (MT)
  - Direct transfer and syntactic transfer
  - Statistical MT and noisy channel model
  - MT evaluation
Ambiguity Resolution is Required for Translation

- Syntactic and semantic ambiguities must be properly resolved for correct translation:
  - “John plays the guitar.” → “John toca la guitarra.”
  - “John plays soccer.” → “John juega al fútbol.”
- An apocryphal story is that an early MT system gave the following results when translating from English to Russian and then back to English:
  - “The spirit is willing but the flesh is weak.” ≈ “The liquor is good but the meat is spoiled.”
  - “Out of sight, out of mind.” ≈ “Invisible idiot.”

Issues: Lexical Gaps

- Some words in one language do not have a corresponding term in the other.
  - Rivière (river that flows into ocean) and fleuve (river that does not flow into ocean) in French
  - Schadenfraude (feeling good about another’s pain) in German.
  - Oyakoko (filial piety) in Japanese

Issues: Differing Word Orders

- English word order is subject – verb – object (SVO)
- Japanese word order is subject – object – verb (SOV)

<table>
<thead>
<tr>
<th>Language</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>IBM bought Lotus</td>
</tr>
<tr>
<td>Japanese</td>
<td>IBM Lotus bought</td>
</tr>
<tr>
<td>English</td>
<td>Sources said that IBM bought Lotus yesterday</td>
</tr>
<tr>
<td>Japanese</td>
<td>Sources yesterday IBM Lotus bought that said</td>
</tr>
</tbody>
</table>

Issues: Syntactic Structure is not Preserved Across Translations

The bottle floated into the cave

La botella entro a la cueva flotando
(the bottle entered the cave floating)

Outline

- Issues in machine translation (MT)
  - Direct transfer and syntactic transfer
  - Statistical MT and noisy channel model
- MT evaluation
Vauquois Triangle

Direct Transfer/Translation

- Translation is word-by-word
- Very little analysis of the source text (e.g., no syntactic or semantic analysis)
- Relies on a large bilingual dictionary. For each word in the source language, the dictionary specifies a set of rules for translating that word.

13

14

15

16

CLASSIC SOUPS

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Price</th>
<th>Cal.</th>
</tr>
</thead>
<tbody>
<tr>
<td>57</td>
<td>House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)</td>
<td>1.50</td>
<td>2.75</td>
</tr>
<tr>
<td>58</td>
<td>Chicken Noodle Soup</td>
<td>1.85</td>
<td>3.25</td>
</tr>
<tr>
<td>59</td>
<td>Chinese Noodle Soup</td>
<td>1.50</td>
<td>2.75</td>
</tr>
<tr>
<td>60</td>
<td>Tomato Clear Egg Drop Soup</td>
<td>1.65</td>
<td>2.95</td>
</tr>
<tr>
<td>61</td>
<td>Lebanese Vegetable Soup</td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>62</td>
<td>Rice &amp; Soup Soup</td>
<td>1.10</td>
<td>2.10</td>
</tr>
<tr>
<td>63</td>
<td>Egg Drop Soup</td>
<td>1.20</td>
<td>2.30</td>
</tr>
<tr>
<td>64</td>
<td>Tomato Clear Egg Drop Soup</td>
<td>1.65</td>
<td>2.95</td>
</tr>
<tr>
<td>65</td>
<td>Chicken Corn Cereal Soup</td>
<td>NA</td>
<td>3.50</td>
</tr>
<tr>
<td>66</td>
<td>Crabe C nelle Cereal Soup</td>
<td>NA</td>
<td>3.50</td>
</tr>
<tr>
<td>67</td>
<td>Sausage Soup</td>
<td>NA</td>
<td>3.50</td>
</tr>
</tbody>
</table>

Direct Transfer/Translation

- Morphological Analysis
  - Mary didn't slap the green witch. → Mary DO:PAST not slap the green witch.
  - Relies on a large bilingual dictionary. For each word in the source language, the dictionary specifies a set of rules for translating that word.

17

18

An Example of a set of Direct Translation Rules

Rules for translating much or many into Russian:

if preceding word is how return slei/ko
if preceding word is at return sto/ko she
if word is much
if preceding word is very return oil
if following word is a noun return mango
if (word is much)
if preceding word is a preposition and following word is noun return mango
else return mango

Lack of any analysis of the source language causes several problems

- Difficult or impossible to capture long-range reorderings
- Words are translated without disambiguation of their syntactic role e.g., that can be a complementizer or determiner, and will often be translated differently for these two cases

English: Sources said that IBM bought Lotus yesterday
Japanese: Sources yesterday IBM Lotus bought that said

They said that...
They like that ice-cream
Possible Solution

- Analysis: Analyze the source language sentence; for example, build a syntactic analysis of the source language sentence.
- Transfer: Convert the source-language parse tree to a target-language parse tree.
- Generation: Convert the target-language parse tree to an output sentence.

Syntactic Transfer

- Simple lexical reordering does not adequately handle more dramatic reordering such as that required to translate from an SVO to an SOV language.
- Need syntactic transfer rules that map parse tree for one language into one for another.
  - English to Spanish:
    - NP → ADJ Nom Þ NP → Nom ADJ
  - English to Japanese:
    - VP → V NP Þ VP → NP V
    - PP → P NP Þ PP → NP P

Outline

- Issues in machine translation (MT)
- Direct transfer and syntactic transfer
  - Statistical MT and noisy channel model
  - MT evaluation

Statistical MT

- Manually encoding comprehensive bilingual lexicons and transfer rules is difficult.
- SMT acquires knowledge needed for translation from a parallel corpus or bitext that contains the same set of documents in two languages.
- The Canadian Hansards (parliamentary proceedings in French and English) is a well-known parallel corpus.
- First align the sentences in the corpus based on simple methods that use coarse cues like sentence length to give bilingual sentence pairs.
- Then align the words in parallel sentences
**Word Alignment**

Mary didn’t slap the green witch.

*Maria no dio una bofetada a la bruja verde.*

(Mary do not gave a slap to the witch green.)

---

**Basic idea:** co-occurrence between words and phrases (like a bipartite matching)

• The IBM models (will not be discussed in class, but reference here:

---

**Next: Picking a Good Translation**

• A good translation should be **faithful** and correctly convey the information and tone of the original source sentence.

• A good translation should also be **fluential**, grammatically well structured and readable in the target language.

• Final objective:

\[
T_{wor} = \arg\max_{T \in \text{Target}} \text{fluency}(T) \times \text{faithfulness}(T)
\]

---

**Noisy Channel Model**

• Based on analogy to information-theoretic model used to decode messages transmitted via a communication channel that adds errors.

• Assume that source sentence was generated by a "noisy" transformation of some target language sentence and then use Bayesian analysis to recover the most likely target sentence that generated it.

---

**Bayesian Analysis of Noisy Channel**

\[
\hat{E} = \arg\max_{E \in \text{English}} P(E | F) = \arg\max_{E \in \text{English}} P(F | E) P(E)
\]

A decoder determines the most probable translation \(\hat{E}\) given \(F\).
Translation from Spanish to English, candidate translations based on $p(\text{Spanish} \mid \text{English})$ alone:

Que hambre tengo yo

- What hunger have $p(s|e) = 0.000014$
- Hungry I am so $p(s|e) = 0.000001$
- I am so hungry $p(s|e) = 0.0000015$
- Have i that hunger $p(s|e) = 0.000020$

... (This is where the translation table comes in!)

With $p(\text{Spanish} \mid \text{English}) \times p(\text{English})$:

Que hambre tengo yo

→

What hunger have $p(s|e)p(e) = 0.000014 \times 0.000001$

Hungry I am so $p(s|e)p(e) = 0.000001 \times 0.0000014$

I am so hungry $p(s|e)p(e) = 0.0000015 \times 0.00001$

Have i that hunger $p(s|e)p(e) = 0.000020 \times 0.0000098$

Outline

- Issues in machine translation (MT)
- Direct transfer and syntactic transfer
- Statistical MT and noisy channel model

Evaluating MT

- Human subjective evaluation is the best but is time-consuming and expensive.
- Automated evaluation comparing the output to multiple human reference translations is cheaper and correlates with human judgements.

Human Evaluation of MT

- Ask humans to estimate MT output on several dimensions.
  - Fluency: Is the result grammatical, understandable, and readable in the target language.
  - Fidelity: Does the result correctly convey the information in the original source language.

Computer-Aided Translation Evaluation

- Edit cost: Measure the number of changes that a human translator must make to correct the MT output.
  - Number of words changed
  - Amount of time taken to edit
  - Number of keystrokes needed to edit
Automatic Evaluation of MT

• Collect one or more human reference translations of the source.
• Compare MT output to these reference translations.
• Score result based on similarity to the reference translations.
• BLEU

BLEU

• Determine number of n-grams of various sizes that the MT output shares with the reference translations.
• Compute a modified precision measure of the n-grams in MT result.

BLEU Example

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.
Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 1 Unigram Precision: 5/6

Cand 1 Bigram Precision: 1/5

BLEU Example

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.
Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Cand 2 Unigram Precision: 7/10

Cand 2 Bigram Precision: 4/9
Modified N-Gram Precision

- Average n-gram precision over all n-grams up to size N (typically 4, 2 in this example) using geometric mean.

\[ p_n = \frac{\sum \text{count}(n\text{-gram})}{\sum \text{count}(n\text{-gram})} \]

\[ p = \sqrt{\prod p_n} \]

- Cand 1: \( p = \sqrt{\frac{4}{7}} = 0.408 \)
- Cand 2: \( p = \sqrt{\frac{4}{10}} = 0.558 \)

Brevity Penalty

- Not easy to compute recall to complement precision since there are multiple alternative gold-standard references and don’t need to match all of them.
- Instead, use a penalty for translations that are shorter than the reference translations.
- Define effective reference length, \( r \), for each sentence as the length of the reference sentence with the largest number of n-gram matches. Let \( c \) be the candidate sentence length.

\[ BP = \begin{cases} 1 & \text{if } c > r \\ \exp^{-1.5 \left( \frac{c}{r} - 1 \right)} & \text{if } c \leq r \end{cases} \]

BLEU Score

- Final BLEU Score: \( \text{BLEU} = BP \cdot p \)
- Cand 1: Mary no slap the witch green.
  - Best Ref: Mary did not slap the green witch.
  - \( c = 6, \ r = 7, \ BP = e^{1.5} = 0.846 \)
  - \( \text{BLEU} = 0.846 \cdot 0.408 = 0.345 \)
- Cand 2: Mary did not give a smack to a green witch.
  - Best Ref: Mary did not smack the green witch.
  - \( c = 10, \ r = 7, \ BP = 1 \)
  - \( \text{BLEU} = 1 \cdot 0.558 = 0.558 \)

BLEU Score Issues

- BLEU has been shown to correlate with human evaluation when comparing outputs from different SMT systems.
- However, it does not correlate with human judgments when comparing SMT systems with manually developed MT (Systran) or MT with human translations.
- Other MT evaluation metrics have been proposed that claim to overcome some of the limitations of BLEU (e.g. METEOR, NIST, etc.).