1. A short history of MT
2. MT paradigms
   – Rule-based MT (RBMT)
   – Statistical MT (SMT)
   – Neural MT (NMT)
3. Evaluation of MT
A short history of MT
Early days of MT

- The US was not following developments in the USSR because most people couldn’t read Russian
  - Teaching everyone Russian takes too long and it’s too expensive
- Let’s get computers to automatically translate Russian to English
  - There was optimism and it was thought that the MT problem would be solved in a few years (!)
  - But sooner than after it was realised that it was not such an easy task...
Early days of MT

• Given the sentence "Spring is a season", which translation is correct?
  - «La primavera es una estación»
  - «El muelle es una estación»
Early days of MT

- Given the sentence "Spring is a season", which translation is correct?
  - «La primavera es una estación»
  - «El muelle es una estación»

spring = a season AND a coil

Word Sense Disambiguation is essential for translation
History overview

• Early 1950s to 1966
  – Decade of optimism. Goal: fully automatic high-quality machine translation (FAHQMT)
• 1966
  – ALPAC report. FAHQMT not possible → MT not a research priority any more
• 1970s to early 1990s
  – New more realistic goal: MT to assist translators, instead of to replace them
  – Rule-based MT
• 1990s to 2014
  – Statistical MT (SMT)
• From 2014
  – Neural MT (NMT)
2 MT paradigms
2.1 MT paradigms

Rule-based MT
MT paradigms

• Transfer Rule-based
  - Monolingual and bilingual dictionaries
  - Rules
  - Morphological analyser and generator
  - 3 person-months to build a basic MT system between related languages
Rule-based MT

- Bilingual dictionary
  - Word correspondences + morphological information (e.g. *part-of-speech, gender, number*)

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojo &lt;adj, masc, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ roja &lt;adj, fem, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojos &lt;adj, masc, plu&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojas &lt;adj, fem, plu&gt;</td>
</tr>
<tr>
<td>house&lt;noun, ø, sing&gt;</td>
<td>→ casa&lt;noun, fem, sing&gt;</td>
</tr>
<tr>
<td>houses&lt;noun, ø, plu&gt;</td>
<td>→ casas&lt;noun, fem, plu&gt;</td>
</tr>
</tbody>
</table>

*The relevant type of morphological information depends on the language pair*
Rule-based MT

• Rules
  - Transformations between languages

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojo &lt;adj, masc, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>roja &lt;adj, fem, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojos &lt;adj, masc, plu&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojas &lt;adj, fem, plu&gt;</td>
</tr>
<tr>
<td>house&lt;noun, ø, sing&gt;</td>
<td>casa&lt;noun, fem, sing&gt;</td>
</tr>
<tr>
<td>houses&lt;noun, ø, plu&gt;</td>
<td>casas&lt;noun, fem, plu&gt;</td>
</tr>
</tbody>
</table>

<Adj,ø,ø> <noun,ø,num>  →  <Adj, gen, num> <noun, gen, num>  →  noun adj
Rule-based MT

red house → ?
red house → ?

1: analyse input and translate with bilingual dictionary

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojo &lt;adj, masc, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ roja &lt;adj, fem, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojos &lt;adj, masc, plu&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>→ rojas &lt;adj, fem, plu&gt;</td>
</tr>
<tr>
<td>house&lt;noun, ø, sing&gt;</td>
<td>→ casa&lt;noun, fem, sing&gt;</td>
</tr>
<tr>
<td>houses&lt;noun, ø, plu&gt;</td>
<td>→ casas&lt;noun, fem, plu&gt;</td>
</tr>
</tbody>
</table>
### Rule-based MT

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojo &lt;adj, masc, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>roja &lt;adj, fem, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojos &lt;adj, masc, plu&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojas &lt;adj, fem, plu&gt;</td>
</tr>
<tr>
<td>house&lt;noun, ø, sing&gt;</td>
<td>casa&lt;noun, fem, sing&gt;</td>
</tr>
<tr>
<td>houses&lt;noun, ø, plu&gt;</td>
<td>casas&lt;noun, fem, plu&gt;</td>
</tr>
</tbody>
</table>
### Rule-based MT

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojo &lt;adj, masc, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>roja  &lt;adj, fem, sing&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojos &lt;adj, masc, plu&gt;</td>
</tr>
<tr>
<td>red&lt;adj, ø, ø&gt;</td>
<td>rojas &lt;adj, fem, plu&gt;</td>
</tr>
<tr>
<td>house&lt;noun, ø, sing&gt;</td>
<td>casa &lt;noun, fem, sing&gt;</td>
</tr>
<tr>
<td>houses&lt;noun, ø, plu&gt;</td>
<td>casas &lt;noun, fem, plu&gt;</td>
</tr>
</tbody>
</table>
Rule-based MT

red<adj, Ø, Ø>
- rojo<adj, masc, sing>
- roja<adj, fem, sing>
- rojos<adj, masc, plu>
- rojas<adj, fem, plu>

house<noun, Ø, sing>
- casa<noun, fem, sing>

2: apply rules
Rule-based MT

red<adj, ø, ø>  house<noun, ø, sing>  
-rojo<adj, masc, sing>  casa<noun, fem, sing>  
-roja<adj, fem, sing>  
-rojos<adj, masc, plu>  
-rojas<adj, fem, plu>  

2: apply rules

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
</table>
| <Adj,ø,ø> <noun,ø,num> → <Adj, gen, num> <noun, gen, num>  
Adj noun  
→ noun adj |
Rule-based MT

red<adj, ø, ø> house<noun, ø, sing>
-rojo<adj, masc, sing> casa<noun, fem, sing>
-roja<adj, fem, sing>
-rojos<adj, masc, plu>
-rojas<adj, fem, plu>

2: apply rules

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Adj,ø,ø&gt; &lt;noun,ø,num&gt; Adj noun</td>
<td>→ &lt;Adj, gen, num&gt; &lt;noun, gen, num&gt;</td>
</tr>
<tr>
<td>→ noun adj</td>
<td>→ noun adj</td>
</tr>
</tbody>
</table>
**Rule-based MT**

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;$Adj,$\varnothing$,$\varnothing$ $&gt;$ $&lt;$noun,$\varnothing$,$\text{num}$ $&gt;$</td>
<td>$&lt;$Adj, gen, num $&gt;$ $&lt;$noun, gen, num $&gt;$</td>
</tr>
<tr>
<td>Adj noun</td>
<td>noun adj</td>
</tr>
</tbody>
</table>

2: apply rules

red$<$adj, $\varnothing$, $\varnothing$$>$  
roja$<$adj, fem, sing$$>$  
house$<$noun, $\varnothing$, sing$$>$  
casa$<$noun, fem, sing$$>$
Rule-based MT

red<adj, ø, ø>  house<noun, ø, sing>
roja<adj, fem, sing>  casa<noun, fem, sing>

2: apply rules

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;Adj,ø,ø&gt; &lt;noun,ø,num&gt; Adj noun</td>
<td>→ &lt;Adj, gen, num&gt; &lt;noun, gen, num&gt; noun adj</td>
</tr>
</tbody>
</table>
Rule-based MT

3: output translation

Red house → casa roja
Rule-based MT

- Architecture of Apertium, a RBMT system ([Tyers and Nordfalk 2009](#))

**Figure 1:** The eight modules of the shallow-transfer machine translation system
2.2 MT paradigms
Statistical MT
MT paradigms

- **Rule-based**
  - Dictionaries and rules
  - 3 person months to build a basic MT system between related languages

- **Statistical**
  - Pairs of equivalent sentences (millions)

Europarl (official EU languages)
~50 million words per language
MT paradigms

- Rule-based
  - Dictionaries and rules
  - 3 person months to build a basic MT system between related languages

- Statistical
  - Pairs of equivalent sentences (millions)

Europarl (official EU languages)
~50 million words per language
MT paradigms

• Rule-based
  – Dictionaries and rules
  – 3 person months to build a basic MT system between related languages

• Statistical
  – Pairs of equivalent sentences (millions)

Europarl (official EU languages)
~50 million words per language
red house →

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>the house</td>
<td>la casa</td>
</tr>
<tr>
<td>the houses</td>
<td>las casas</td>
</tr>
<tr>
<td>red chair</td>
<td>silla roja</td>
</tr>
<tr>
<td>red doors</td>
<td>puertas rojas</td>
</tr>
</tbody>
</table>
red house

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>the house</td>
<td>la casa</td>
</tr>
<tr>
<td>the houses</td>
<td>las casas</td>
</tr>
<tr>
<td>red chair</td>
<td>silla roja</td>
</tr>
<tr>
<td>red doors</td>
<td>puertas rojas</td>
</tr>
</tbody>
</table>

Translation Model
red house → roja
→ rojas
red house → roja
→ rojas

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>the house</td>
<td>la casa</td>
</tr>
<tr>
<td>the houses</td>
<td>las casas</td>
</tr>
<tr>
<td>red chair</td>
<td>silla roja</td>
</tr>
<tr>
<td>red doors</td>
<td>puertas rojas</td>
</tr>
</tbody>
</table>
red house → roja casa
→ rojas casa

English                  Spanish
the house → la casa
the houses → las casas
red chair → silla roja
red doors → puertas rojas
### Statistical MT

<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>the house</td>
<td>la casa</td>
</tr>
<tr>
<td>the houses</td>
<td>las casas</td>
</tr>
<tr>
<td>red chair</td>
<td>silla roja</td>
</tr>
<tr>
<td>red doors</td>
<td>puertas rojas</td>
</tr>
</tbody>
</table>

The reordering model takes into account the word order in English and Spanish.
<table>
<thead>
<tr>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>the house</td>
<td>la casa</td>
</tr>
<tr>
<td>the houses</td>
<td>las casas</td>
</tr>
<tr>
<td>red chair</td>
<td>silla roja</td>
</tr>
<tr>
<td>red doors</td>
<td>puertas rojas</td>
</tr>
</tbody>
</table>
Statistical MT

Parallel data → Translation and Reordering Models

Input

Output

Language Model (Fluency) → Monolingual Data (target language)
## MT Paradigms: Pros & Cons

<table>
<thead>
<tr>
<th>Paradigm</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-based</td>
<td>Fast and efficient, very low hardware requirements</td>
<td>At least 3 months of an expert (linguist) required</td>
</tr>
<tr>
<td>Statistical</td>
<td>Doesn't require knowledge from an expert*</td>
<td>It requires:</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Vast amount of parallel sentences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Hardware</td>
</tr>
</tbody>
</table>
Translation Options

• Language is highly combinatorial
  – Chances of seeing a new short sentence: 1 word (1-gram) or 2 words (2-grams) in a big corpus such as Europarl are high
  – But chances of seeing a new 4-gram are already low!
  – In statistical MT a new sentence is typically translated by putting together short snippets from known sentences
Translation Options

- If the fragments known are short this can lead to a "word salad effect"
Translation Options

- If the fragments known are too short this can lead to a “word salad” translation
  - “he is yes not after house” vs “he is not home”
- A good language model can prevent word salads
  - It would score such sequences as not very plausible
2.3 Neural MT
Neural MT

• Not that recent approach!
  – Theoretically introduced in the 90s (e.g. Forcada and Ñeco 1997)
  – But first practical application from 2010s (e.g. Bahdanau et al. 2014)

• As SMT, NMT is a data-driven approach: it still requires parallel data, but treats words as concepts
  – Words in SMT: surface forms
    • dog -> dog
    • cat -> cat
  – Words in NMT: vectors of real numbers
    • dog -> [0, 0.9, 0.1, ... 0]
    • cat -> [0, 0.8, 0.2, ... 0]
Neural MT. Continuous space

In SMT (and RBMT) there are no relations between words as the model works with the strings.

In NMT, words are represented as numerical vectors. Therefore you can do mathematical operations with them, e.g. compute the distance between two words.

(Blitzer et al 2005)
Neural MT. Continuous space

(Blitzer et al 2005)

(Forcada 2019)
Neural MT
recurrent sequence to sequence

\( f = \) (La, croissance, économique, s'est, ralentie, ces, dernières, années, .)

\[ e = \) (Economic, growth, has, slowed, down, in, recent, years, .) \]

(Cho, 2015)
Neural MT. Encoder

\[ e = (\text{Economic, growth, has, slowed, down, in, recent, years, .}) \]

(Cho, 2015)
Neural MT. Decoder

\( f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .}) \)

\( e = (\text{Economic, growth, has, slowed, down, in, recent, years, .}) \)

(Cho, 2015)
Neural MT. Advanced methods

- Attention in recurrent architecture (Bahdanau et al. 2014)
  - Pay more attention to some parts of the source sentence when producing a word in the target
- Transformer architecture (Vaswani et al. 2017)
  - Words connected by shorter paths than in the recurrent architecture
- Unsupervised models (Artetxe et al. 2017)
  - Learn to translate without parallel data
### NMT: Pros & Cons

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better at generalisation (continuous space)</td>
<td>Computationally very expensive</td>
</tr>
<tr>
<td>Can model long-distance relations (beyond n-gram language models used in SMT)</td>
<td>Adequacy issues.</td>
</tr>
<tr>
<td>Fluent translations</td>
<td>- Misstranslation of proper nouns</td>
</tr>
<tr>
<td></td>
<td>- Allucinations</td>
</tr>
<tr>
<td></td>
<td>Repeated word phenomenon and omissions</td>
</tr>
<tr>
<td></td>
<td>Less predictable translations than with SMT or RBMT</td>
</tr>
</tbody>
</table>
MT Evaluation
Not an easy problem

这个 机场 的 安全 工作 由 以色列 方面 负责．

- Israeli officials are responsible for airport security.
- Israel is in charge of the security at this airport.
- The security work for this airport is the responsibility of the Israel government.
- Israeli side was in charge of the security of this airport.
- Israel is responsible for the airport’s security.
- Israel is responsible for safety work at this airport.
- Israel presides over the security of the airport.
- Israel took charge of the airport security.
- The safety of this airport is taken charge of by Israel.
- This airport’s security is the responsibility of the Israeli security officials.

(Source: 2001 NIST evaluation)
Desiderata for MT evaluation

• Meaningful
  − Score should give intuitive interpretation of translation quality
• Consistent
  − Repeated use of metric should lead to same results
• Correct
  − Metric must rank better systems higher
• Low cost
  − Reduce time and money spent to carry out evaluation
Types of MT evaluation

1) Subjective judgments by human evaluators
2) Automatic evaluation metrics
3) Task-based evaluation, e.g.
   - How much post-editing effort?
   - Does the information come across?
3.1 Human evaluation
Given

- MT output
- Source and/or reference translation
  - Reference translation: a translation produced by a human translator

Task: assess the quality of the MT output
Human evaluation

English-to-Spanish example

• MT output: «El muelle es mi estación favorita.»
• Source: "Spring is my favourite season."
• Reference translation: «La primavera es mi estación favorita.»
Method 1/3: adequacy and fluency

Adequacy
- Does the output convey the same meaning as the source sentence?
- Is part of the message lost, added, or distorted?

Fluency
- Is the output a good fluent sentence in the target language?
- This involves both grammatical correctness and idiomatic word choices
Method 1/3: adequacy and fluency

Using Likert scales, e.g.

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 all meaning</td>
<td>5 flawless target language</td>
</tr>
<tr>
<td>4 most meaning</td>
<td>4 good target language</td>
</tr>
<tr>
<td>3 much meaning</td>
<td>3 non-native target language</td>
</tr>
<tr>
<td>2 little meaning</td>
<td>2 disfluent target language</td>
</tr>
<tr>
<td>1 none</td>
<td>1 incomprehensible</td>
</tr>
</tbody>
</table>
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 rather a necessary laboratory the internal operation of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>both countries are a necessary laboratory at internal functioning of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a laboratory necessary for the internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a laboratory for the internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</td>
</tr>
<tr>
<td>the two countries are rather a necessary laboratory internal workings of the eu.</td>
<td>1 2 3 4 5</td>
<td>1 2 3 4 5</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
Method 2/3: relative ranking

Problem with adequacy and fluency: low inter-annotator agreement

Relative ranking: given two translations A and B, decide which one is better, i.e. A>B or A<B or A=B

• The agreement is higher than for adequacy and fluency
Method 2/3: relative ranking

Example with 5 systems (Appraise toolkit)
Method 3/3: direct assessment

Problem with relative ranking: one can know which translation is better but not by how much (relative instead of absolute)

Direct assessemnt: given a translation provide a score, e.g. between 0 and 100

- If enough evaluators assess a translation then the average score becomes stable
Method 3/3: direct assessment

- Example of direct assessment from WMT 2019

For the pair of sentences below: Read the text and state how much you agree that:

The black text adequately expresses the meaning of the gray text in German (deutsch).

North Korea says 'no way' will disarm unilaterally without trust
— Source text

Nordkorea sagt, Sprünge ohne Vertrauen entwaffnen ohne Vertrauen.
— Candidate translation

0% 100%
Reset Submit
3.2 Automatic evaluation
Automatic evaluation metrics

Basic strategy

• Inputs
  – MT output
  – Human reference translation

• Output: a score which represents the similarity between the MT output and the human reference
The most popular automatic evaluation metric

- Calculates how many sequences of up to 4 words in the MT output and in the reference translation match
- If the MT output is shorter than the reference, the score is multiplied by a 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

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police.</td>
</tr>
</tbody>
</table>

1-grams: The, gunman, was, shot, dead, by, police, .

- Overlap: $o_1 = 7/8$
- Brevity Penalty:
- Final Score:
BLEU: an example

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police .</td>
</tr>
</tbody>
</table>

2-grams: *The gunman, gunman was, was shot, shot dead, dead by, by police, police.*

- Overlap: $o1=7/8$, $o2=4/7$
- Brevity Penalty:
- Final Score:
BLEU: an example

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police.</td>
</tr>
</tbody>
</table>

3-grams: The gunman was, gunman was shot, was shot dead, shot dead by, dead by police, by police.

- Overlap: $o_1=7/8$, $o_2=4/7$, $o_3=2/6$
- Brevity Penalty:
- Final Score:
BLEU: an example

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police.</td>
</tr>
</tbody>
</table>

4-grams: *The gunman was shot*, *gunman was shot dead*, *was shot dead by*, *shot dead by police*, *dead by police*.

- Overlap: $o_1=7/8$, $o_2=4/7$, $o_3=2/6$, $o_4=1/5$
- Brevity Penalty:
- Final Score:
BLEU: an example

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police.</td>
</tr>
</tbody>
</table>

4-grams: The gunman was shot, gunman was shot dead, was shot dead by, shot dead by police, dead by police.

- Overlap: $o_1=7/8$, $o_2=4/7$, $o_3=2/6$, $o_4=1/5$
- Brevity Penalty: output length=8, reference length=10 → 8/10
- Final Score:
# BLEU: an example

<table>
<thead>
<tr>
<th>MT output</th>
<th>The gunman was shot dead by police.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ref</td>
<td>The gunman was shot to death by the police.</td>
</tr>
</tbody>
</table>

- **Overlap:**  $o1=7/8$, $o2=4/7$, $o3=2/6$, $o4=1/5$
- **Brevity Penalty:** output length=8, reference length=10 → $8/10$
- **Final Score:**  $8/10 \times (7/8 \times 4/7 \times 2/6 \times 1/5)^{1/4} = 0.34$
Issues with BLEU

1) All sequences given the same importance, while some may be more informative than others
2) A word would not match if the suffix is wrong, problematic for morphologically-rich languages
3) Synonyms are not given any credit

Other metrics address these issues, e.g.
• NIST 1)
• CHRF and METEOR 2)
• METEOR 3)
3.3 Task-based evaluation
Motivation

• Machine translation is a means to an end
• Does machine translation output help accomplish a task?
  – Producing high-quality translations by post-editing MT output (MT for publishing)
  – Information gathering from foreign language sources (MT for gisting)
Post-editing effort

Measuring time spent on producing translations

• baseline: translation from scratch
• post-editing the output of an MT system

Note: time depends not only on MT quality, also on post-editor's skills → one need to take translators’ variability into account
Productivity per Language – Translation vs Post-Editing

For all languages tested – in fact for all 37 test participants –, post-editing productivity was significantly higher than translation productivity.
Given MT output, can monolingual speakers (target language) answer questions about it?

- Basic facts: who? where? when? names, numbers, and dates
- Actors and events: relationships, temporal and causal order
- Nuance and author intent: emphasis and subtext
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?
License

- This work may be distributed under the terms of the Creative Commons Attribution-ShareAlike 4.0 International licence
- E-mail me to obtain the PDF and/or LibreOffice sources