OUTLINE

• Some basic concepts in machine translation design
• Evaluating translation quality using BLEU score
• The generative models underlying Candide, the influential statistical machine translation system
BASICS
THE NOISY CHANNEL MODEL OF TRANSLATION (REDUX)

Warren Weaver, 1949 Rockefeller Foundation memorandum

Translation:

“When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

\[
\arg\max_e P(e) \quad P(r \mid e)
\]
Machine translation received massive US government funding in the ‘50s and early ‘60s, but made next to no progress on the core problems.
The ALPAC report (1964) recommended that government-funded MT research focus on:

1. practical methods for evaluation of translations

... 

3. evaluation of quality and cost of various sources of translations

... 

9. production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation
This (ultimately) lead researchers to adopt a clearer problem statement, the modeling of *translator behavior*. 
Effective domain-general machine translation systems consist of...

data-driven, language-agnostic models of *translator behavior*...

paired with language-specific models of *linguistic analysis* and *generation*. 
VAUQUOIS TRIANGLE
THE QUADRATIC GROWTH PROBLEM

As the number of languages a system supports \((n)\) increases, the number of translation models needed grows quadratically \((n^2 - n)\)*

Thus, when developing multilingual translation systems, we place language-specific methods in the monolingual analysis and generation models so the translation model is as language-independent as possible.

*Note that translation models need not be invertible.*
In the early 1990s, a team at IBM Research built **Candide**, the first modern *statistical* machine translation system. We will be reviewing the intuitions behind Candide in great detail.
EVALUATION
(HANDOUT)
Candidate:
Many will lose their right to a pension in their own name because of their husband ’s income .

Reference: Many will lose their right to draw a pension with their own name because of the income of their husband .
Candidate: Many will lose their right to a pension in their own name because of their husband’s income.

Reference: Many will lose their right to draw a pension with their own name because of the income of their husband.
Candidate:
Many will lose their right to a pension in their own name because of their husband ‘s income .

Reference: Many will lose their right to draw a pension with their own name because of the income of their husband .

\[ p_1: \frac{17}{19} = .895 \]
\[ p_2: \frac{12}{18} = .667 \]
\[ p_3: \frac{8}{17} = .471 \]
\[ p_4: \frac{5}{16} = .313 \]

\[ GM_n = .544 \]
\[ BP = .900 \]
\[ BLEU = .490 \]
BLEU is one of the first evaluation metrics which is well-correlated with human judgements of translation quality.
THE CANDIDE STATISTICAL MACHINE TRANSLATION MODELS

[Sources: Brown et al. 1990, 1993, Knight 1999]
TRANSLATION STORY ELEMENTS

• The translation model $P(t \mid s)$ helps to select likely translations:

$$P(\text{house} \mid \text{maison}) > P(\text{dog} \mid \text{maison})$$

• The language model $P(t)$ helps with source-to-target polysemy:

$$P(\text{in the end zone}) > P(\text{on the end zone})$$

• It also helps to sort out word order:

$$P(\text{the dog runs}) > P(\text{runs dog the})$$

• Decoding helps us find “likely” “stories”.
MODEL I: BASIC STORY

1. Given a source $S$ of length $|S|$, select a target length $|T|$ according to $P_L(|T| \mid |S|)$

2. Populate $T$ with tokens $t$ according to $P(t \mid s)$

3. Reorder the tokens in $T$ to maximize $P_L(t_0 \ldots t_{|T|})$
MODEL I: TRANSLATION MODEL
ESTIMATION VIA THE E.M. ALGORITHM

1. Compute $P(t | s)$, the MLE conditional probability distribution of $s$ and $t$ co-occurring

2. For $n$ iterations:
   1. Initialize $a(s, t) = 0, Z(t) = 0$ for all $s \in S, t \in T$.
   2. For all pairs of sentences $S, T$:
      
      For all $s \in S, t \in T$,
      
      $$a(s, t) = a(s, t) + P(t | s)$$
      $$Z(t) = Z(t) + P(t | s) .$$

   3. For all $s, t$, let
      
      $$P(t | s) = a(s, t) / Z(t)$$
      
      then normalize $P(t | s)$. 
Source:

LA MAISON BLEUE
LA MAISON
MAISON

Target:

THE BLUE HOUSE
THE HOUSE
HOUSE
ITERATION 0 (MLE ONLY)

\[ P(\text{HOUSE} \mid \text{MAISON}) = .500 \]
\[ P(\text{BLUE} \mid \text{MAISON}) = .167 \]
\[ P(\text{THE} \mid \text{MAISON}) = .333 \]
ITERATION 1

\[ P(\text{HOUSE} \mid \text{MAISON}) = .440 \]
\[ P(\text{BLUE} \mid \text{MAISON}) = .233 \]
\[ P(\text{THE} \mid \text{MAISON}) = .327 \]
ITERATION 2

\[ P(\text{HOUSE} \mid \text{MAISON}) = .478 \]

\[ P(\text{BLUE} \mid \text{MAISON}) = .196 \]

\[ P(\text{THE} \mid \text{MAISON}) = .325 \]
ITERATION 5

\[ P(\text{HOUSE} \mid \text{MAISON}) = 0.572 \]
\[ P(\text{BLUE} \mid \text{MAISON}) = 0.117 \]
\[ P(\text{THE} \mid \text{MAISON}) = 0.311 \]
ITERATION 10

\[ P(\text{HOUSE} \mid \text{MAISON}) = 0.643 \]

\[ P(\text{BLUE} \mid \text{MAISON}) = 0.077 \]

\[ P(\text{THE} \mid \text{MAISON}) = 0.280 \]
MODEL II: BASIC STORY

1. Given a source $S$ of length $|S|$, select a target length $|T|$ according to $P_L(|T| \mid |S|)$

2. For each source token $s_i$ and the null token, “align” it with some $t_j$ according to $P_d(i, j)$

3. Translate all aligned source/target $s_i, t_j$ pairs according to $P(t_j \mid s_i)$. 
MODEL III

Distortion parameters $P_d$ are now sensitive to lengths:

$$P_d(i | j, |S|, |T|)$$ is the probability that source token $j$ corresponds with (i.e., is aligned to and is translated by) target token $i$ when the source is $|S|$ tokens long and the target is $|T|$ tokens long.

Each source word has a fertility parameter $P_f$:

$$P_f(3 | s)$$ is the probability that $s$ aligns to exactly 3 target words.
MODEL III

Mary did not slap the green witch

Mary not slap slap slap the green witch \((P_i)\)

Mary not slap slap slap \(\emptyset\) the green witch \((P_\emptyset)\)

Maria no dió una bofetada a la verde bruja \((P)\)

Maria no dio una bofetada a la bruja verde \((P_d)\)

[h/t: Kevin Knight]
OPEN-SOURCE MACHINE TRANSLATION SOFTWARE

• MT engines:
  
  • EGYPT (CSLP/JHU 1999 team): IBM models I-V
  
  • GIZA++ (Och & Ney 2003): optimized IBM models
  
  • MOSES (Koehn 2009): also supports more advanced models

• MT evaluation tools:
  
  • NIST/BLEU: like BLEU but n-grams weighted by informativity
  
  • METEOR (Banerjee & Lavie 2005): n-gram $F_1$-score
