Machine Translation

Dan Klein
UC Berkeley

Translation Task

• Text is both the input and the output.
• Input and output have roughly the same information content.
• Output is more predictable than a language modeling task.
• Lots of naturally occurring examples (but not much metadata).

Translation Examples

English-German News Test 2013 (a standard dev set)

Republican leaders justified their policy by the need to combat electoral fraud.

Die Führungskräfte der Republikaner rechtfertigen ihre Politik mit der Notwendigkeit, den Wahlbetrug zu bekämpfen.

The Executives of the republican justify your politics with the need to fight the election fraud.
Variety in Human-Generated Translations

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomers got to know this incident 4 days later. This small planet is 50m in diameter. The astronomers are hard to find it for it comes from the direction of sun.

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

Variety in Machine Translations

A small planet, whose is as big as could destroy a middle sized city, passed by the earth with a distance of 463 thousand kilometers. This was not found in advance. The astronomers got to know this incident 4 days later. This small planet is 50m in diameter. The astronomers are hard to find it for it comes from the direction of sun.

A volume enough to destroy a medium city small planet is big, flit earth within 463,000 kilometres of close however were not in advance discovered, astronomer just knew this matter after four days. This small planet diameter is about 50 metre, from the direction at sun, therefore astronomer very hard to discovers it.

Evaluation

BLEU Score

\[
\text{BLEU} = \frac{\mathbf{P}_{\text{BLEU}}}{\mathbf{B}}\left(\prod_{n=1}^{4} \mathbf{F}_{\text{BLEU}}\right)
\]

\[
\mathbf{P}_{\text{BLEU}} = \frac{1}{N} \sum_{i=1}^{N} \max_{j} \left\{ C_{0}(t_i, \max C_{\text{BLEU}}^j(t_i)) \right\}
\]

\[
\mathbf{B} = \exp \left( \min_{j} \left( \frac{B - L}{m} \right) \right)
\]

\[
\text{BLEU} \text{ is a mean of clipped precisions, scaled down by the brevity penalty.}
\]

BLEU score: geometric mean of 1-, 2-, 3-, and 4-gram precision vs. a reference, multiplied by brevity penalty (harmlessly penalizes translations shorter than the reference).
Evaluation with BLEU

BLEU = 26.52, 75.4/74.8/21.4/57.7 (BP=3.088, ratio=1.143, hyp_len=16, ref_len=54)

In this sense, the measures will partially undermine the American democratic system.

Corpus BLEU Correlations with Average Human Judgments

These are ecological correlations over multiple segments; segment-level BLEU scores are noisy.

Commercial machine translation providers seem to all perform human evaluations of some sort.

Commercial machine translation providers seem to all perform human evaluations of some sort.

Results of the WMT19 Metrics Shared Task: Segment-Level and Strong MT Systems Pose Big Challenges

Human Evaluations

Direct assessment: adequacy & fluency
- Monolingual: Ask humans to compare machine translation to a human-generated reference.
  (Easier to source annotators)
- Bilingual: Ask humans to compare machine translation to the source sentence that was translated. (Compares to human quality)
- Annotators can assess segments (sentences) or whole documents.
- Segments can be assessed with or without document context.

Ranking assessment:
- Raters are presented with 2 or more translations.
- A human-generated reference may be provided, along with the source.
- “In a pairwise ranking experiment, human raters assessing adequacy and fluency show a stronger preference for human over machine translation when evaluating documents as compared to isolated sentences.” (Laubli et al., 2018)

Editing assessment: How many edits required to reach human quality

Translationese and Evaluation

Translated text can: (Baker et al., 1993; Graham et al., 2019)
- be more explicit than the original source
- be less ambiguous
- be simplified (lexical, syntactically and stylistically)
- display a preference for conventional grammaticality
- avoid repetition
- exaggerate target language features
- display features of the source language

“If we consider only original source text (i.e. not translated from another language, or translationese), then we find evidence showing that human parity has not been achieved.” (Toral et al., 2018)
WMT 2019 Evaluation

2019 segment-in-context direct assessment (Barrault et al, 2019):
- English to German: many systems are tied with human performance.
- English to Chinese: all systems are outperformed by the human translator.
- English to Czech: all systems are outperformed by the human translator.
- English to Finnish: all systems are outperformed by the human translator.
- English-German Facebook FAIR achieves superb human translation performance; several systems are tied with human performance.

14

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."
Warren Weaver (1949)
Data-Driven Machine Translation

Target language corpus gives examples of well-formed sentences

I will do it later
See you later
He will do it

Parallel corpus gives translation examples

I will do it gladly
Yo lo haré de muy buen grado
You will see later
Después lo verás

Machine translation system:

Source language

Yo lo haré después

Model of translation

I will do it later

Target language

Evolution of the Noisy Channel Model

\[ P(e|f) \propto P(f|e) \cdot P(e) \]

\[ \max_e P(e|f) = \max_e P(f|e) \cdot P(e) \]

\[ P(e|f) \propto P(f|e)^{obs} \cdot P(e)^{p_e} \]

\[ P(e|f) \propto \exp \left( \sum_{e} \log P(e|f) \right) \]

Word Alignment

\[ \frac{1}{A} \]

Stitching Together Fragments

Parallel corpus gives translation examples

I will do it gladly
Yo lo haré de muy buen grado
You will see later
Después lo verás

Machine translation system:

Source language

Yo lo haré después

Model of translation

I will do it later

Target language
2/22/21

Extracting Translation Rules

Counting Aligned Phrases

Interlude: Lexical Translation Models

HMM Alignment Model
Alignment Link Posteriors
\[ \psi(e, f, \epsilon) = \frac{\hat{\delta}(e, f) \cdot \psi(f, \epsilon)}{\sum_j \hat{\delta}(e, f) \cdot \psi(f, \epsilon) - \hat{\delta}(e, f) \cdot \psi(f, \epsilon)} \]

Non-zero for any alignment vector (for sentence pair \( e, f \)) that has word \( e \) aligned to word \( f \)

\[ \phi(e, f) = \sum_j \hat{\delta}(e, f) \cdot \psi(f, \epsilon) \cdot P(a(j) = i | e, f) \]

\[ = \sum_j \hat{\delta}(e, f) \cdot \psi(f, \epsilon) \cdot P(a(j) = i | e, f) \]

Non-zero for any alignment vector (for sentence pair \( e, f \)) that has position \( j \) aligned to position \( i \)

Model 1 Posteriors
\[ P(a(j) = i | e, f) = \frac{\sum \psi_i(j)}{\sum \psi_i(j)} \]

HMM Alignment Model
\[ P(a(e, f)) = \prod_j P(a(j) | a(j-1)) \]

\[ \text{the house is very small} \]

HMM Alignment Model Posteriors
\[ P(a(e, f)) = \frac{\sum \psi_i(j) \cdot P(a(j-1))}{\sum \psi_i(j) \cdot P(a(j-1))} \]

Non-zero for alignments where \( j \) is aligned to \( i \)

Words up to \( i \) (including the aligned word)
\[ \alpha_i(0) = \frac{P(e) \cdot \psi_i(0) \cdot P(f)}{P(e) \cdot \psi_i(0) \cdot P(f)} \]

Forward-Backward algorithm
\[ \alpha_i(j) = \frac{P(e) \cdot \psi_i(j) \cdot P(f | \epsilon) \cdot \beta_{j-1}(i)}{P(e) \cdot \psi_i(j) \cdot P(f | \epsilon) \cdot \beta_{j-1}(i)} \]

Words after \( i \)
\[ \beta_i(j) = \frac{P(e) \cdot \psi_i(j) \cdot P(f | \epsilon) \cdot \alpha_{j+1}(i)}{P(e) \cdot \psi_i(j) \cdot P(f | \epsilon) \cdot \alpha_{j+1}(i)} \]

\[ a_i(j) = P(e_1, e_2, \ldots, e_j, a(j) = i | f) \]

\[ \beta_i(j) = P(e_{j+1}, e_{j+2}, \ldots, e_n | a(j) = i, f) \]
Interlude: Phrase-Based Models

What’s Next?

Neural models: attention and the transformer architecture

Tricks of the trade: back-translation, knowledge distillation, subword models, and coverage vectors