Statistical Machine Translation

Evaluation

Jakub Waszczuk
Heinrich-Heine-Universität Düsseldorf

Winter Semester 2018/19

Overview

Question: How do we rate the quality of an (automatically) translated sentence/text?

Applications:
– paper publication
– evaluation campaigns where teams submit their SMT systems
– control development of an SMT system
– system optimisation (tuning)

Evaluation options:
– Manual evaluation
– Automatic evaluation: w.r.t *reference translation(s)*
– Downstream evaluation: for instance, information extraction from a foreign-language text

Complications

– there is no *single right* answer
  → possible multiple reference translations
– practically impossible to capture *all* acceptable translations
  (e.g., word order often very variable)
– evaluation by humans is very expensive and time consuming

Manual translations of a Chinese sentence by different translators:

- `Israeli officials are responsible for airport security.`
- `Israel is in charge of the security of 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.`

...

Manaul Evaluation

Several human experts rate the same set of translations

Ideal: **bilingual translator**
– but hard to get

In practice: mostly monolingual evaluators, who compare to reference translation(s)

Manual evaluation is very **subjective**.
– for example: some translations initially make no sense, but become clear when one reads the reference or the input sentence in advance
– cause (partial): sentences without context (or background knowledge) are generally difficult to understand

Individual criteria

Fluency und Adequacy, each on a scale of 1-5

Adequacy:

5 - entire content reproduced

4 - most of the content available

3 - decent portion of the content present

2 - little content available

1 - almost nothing

Fluency:

5 - flawless

4 - good

3 - non-native level

2 - heavily distorted / disfluent

1 - incomprehensible

Example

German:	`aber ich will nicht nach hause gehen !`
Reference:	`but i don't want to go home !`
Hypothesis:	`i want not go home but !`

Adequacy: ca. 4

Fluency: ca. 2

Reasons for automatic evaluation

Evaluation extremely important for tuning

Tuning = adjustment of parameters to optimize translation quality

Necessary: evaluation scores for thousands of (reference, translation hypothesis) pairs each time parameters change

⇒ Scores must be calculated automatically.

In addition: human evaluators have to be paid, while automatic evaluation requires virtually no cost.

Desired: metric that correlates well with human-produced scores
Basis: comparison between hypothesis and reference

In the following: translation hypothesis **h**, reference **r**

n-gram-based metrics

Basis: number of correctly identified n-grams for different n

For hypothesis **h**, reference **r**:

n-gram Precision:

$$\frac{\#n\texttt{-grams present in } \mathbf{h} \texttt{ and } \mathbf{r}}{\#n\texttt{-grams present in } \mathbf{h}}$$

n-gram Recall:

$$\frac{\#n\texttt{-grams present in } \mathbf{h} \texttt{ and } \mathbf{r}}{\#n\texttt{-grams present in } \mathbf{r}}$$

F-measure: combination of precision and recall (rarely used in SMT, though)

Example

Reference	`Israeli officials are responsible for airport security`
Hypothesis A	`Israeli officials` ~~`responsibility of`~~ `airport` ~~`safety`~~
Hypothesis B	`airport security Israeli officials are responsible`

For hypothesis A:

1-gram precision: $\frac{3}{6}$ 2-gram precision: $\frac{1}{5}$ 3-gram precision: $\frac{0}{4}$

1-gram recall: $\frac{3}{7}$ 2-gram recall: $\frac{1}{6}$ 3-gram recall: $\frac{0}{5}$

For hypothesis B:

1-gram precision: $\frac{6}{6}$ 2-gram precision: $\frac{4}{5}$ 3-gram precision: $\frac{2}{4}$

1-gram recall: $\frac{6}{7}$ 2-gram recall: $\frac{4}{6}$ 3-gram recall: $\frac{2}{5}$

BLEU: A bilingual evaluation understudy

n-gram precisions for different n + length penalty

$$\texttt{BLEU-}n : \min\left(1, \frac{H}{R}\right) \exp\left(\sum_{k=1}^{n} \lambda_k \log\left(k\texttt{-precision}\right)\right)$$

Common: $\lambda_k = 1$, BLEU-4

H = length of the hypothesis, R = length of the reference

Note: higher value = better translation

BLEU: Example

Refence	`Israeli officials are responsible for airport security`
Hypo A	`Israeli officials` ~~`responsibility of`~~ `airport` ~~`safety`~~
Hypo B	`airport security Israeli officials are responsible`

	n	1	2	3	4
Hypo A	n-gram prec.	$\frac{3}{6}$	$\frac{1}{5}$	0	0
	BLEU-n	$\frac{6}{7} \cdot \frac{3}{6} \approx$ **0,42**	$\frac{6}{7} \cdot \frac{3}{6} \cdot \frac{1}{5} \approx$ **0,09**	**0**	**0**
Hypo B	n-gram prec.	$\frac{6}{6}$	$\frac{4}{5}$	$\frac{2}{4}$	$\frac{1}{3}$
	BLEU-n	**0,86**	**0,69**	**0,34**	**0,11**

Problem: score = 0 as soon as one n-gram precision is 0.

→ Adaptation for several reference translations

→ Evaluation/normalization on corpus-level

Criticism of BLEU

– Words either completely wrong or completely correct
– But: `responsibility` and `responsible` are similar
⇒ sentence content partially available

METEOR:

Handle similarity/synonyms via stemming and WordNet

Problems of METEOR:

– Many parameters involved (how to set them up?)
– WordNet is work-in-progress, not available for some languages
– Difficult to create a generic program for all languages

Edit-distance-based metrics

Principle: The reference translation is gradually transformed by elementary operations into the given hypothesis

Word Error Rate (WER): elementary operations:
– substitute (replace) one word with another
– insert a word
– delete a word

WER for:
– a sentence: minimal number of operations required to transform the reference into the hypothesis, normalized by reference length
– a set of sentences: mean of WERs for the individual sentences

Note: smaller value = better translation

Determining WER (sentence level)

Determining a monotonic alignment (called *Levenshtein-Alignment*) between the hypothesis and the reference:

– Maching identical words: score unchanged

– Matching non-identical words: increase score by 1.

– Reference word without alignment (= delete): increase score by 1.

– Hypothesis word without alignment (= insert): increase score by 1.

→ WER: Levenshtein-Alignment with minimal score

Minimal Levenshtein-Alignment

Dynamic Programming:
Table $Q(i, j)$ with $0 \leq i \leq R, 0 \leq j \leq H$.

Base case: $Q(0, 0) = 0$

Otherwise ($i \geq 1$ or $j \geq 1$):

$$
\begin{aligned}
Q(i,j) = \min\{ & \\
& Q(i-1, j-1) \quad \text{if } r_i = h_j, \quad \texttt{\% match} \\
& Q(i-1, j-1) + 1, \quad \texttt{\% substitute} \\
& Q(i-1, j) + 1, \quad \texttt{\% delete} \\
& Q(i, j-1) + 1 \quad \texttt{\% insert} \\
& \}
\end{aligned}
$$

(where scores for $i = -1$ or $j = -1$ are defined as ∞)

Minimal Levenshtein-Alignment: Example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

Calculating WER

$\text{WER}_{r,h} = \frac{Q(R,H)}{R}$

If desired, the corresponding alignment can be determined by traceback through the dynamic programming table.

Improving WER

Problem of WER:

– Rearrangements not explicitely modeled, thus heavily penalized

Reference 1	`Israeli officials are responsible for airport security`
Reference 2	`This airport's security is the responsibility of the Israeli security officicals`

Handling rearrangements: → Translation Edit Rate (TER)

Basis: **Block-Moves** in additional to normal edit operations

Discussion (1)

- BLEU-4 currently accepted standard (also popular: TER)
- BLEU scores correlate with manual scores (Arabic-English, NIST 2002)

However:

- For BLEU, all words are equally relevant: negation, content words vs. articles, punctuation?
- Nobody knows, what BLEU of, let's say, $0,34$ actually means.
- BLEU operates on very local level $\rightarrow$ this may unfairly bias the metric in favor of phrase-based systems (and against more syntactically-oriented ones)

Discussion (2)

Experiments:

- Rule-based vs. statistical systems:
  Statistical got higher BLEU scores, but manual scores similar (sometimes even higher – see e.g. WMT16 shared task results)
- (monolingual) manually improved translations got only slighly better BLEU scores, but much better manual scores

Similar arguments can be found for the other automatic evaluation metrics.

Other evaluation criteria

In additiona to quality:

- Speed
- System/model size ($\rightarrow$ server vs. smartphone)
- Ease of integration in an application environment
- Customization (other domains, customer requests, etc.)