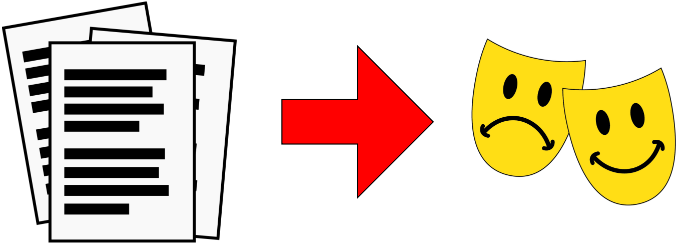
Evaluating Machine Translation Systems

Daniel Beck

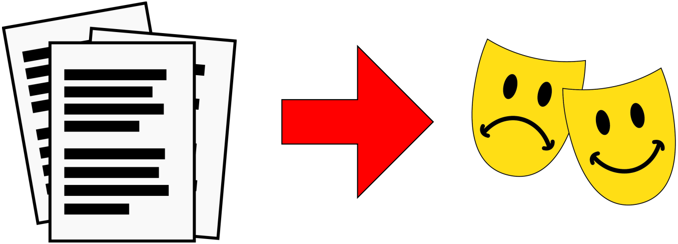
May 22, 2017

Sentiment Analysis



Evaluating Text Classifiers

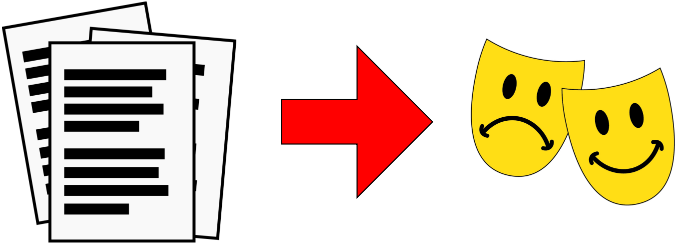
Sentiment Analysis



- Evaluation is easy: just count errors on a test set.

Evaluating Text Classifiers

Sentiment Analysis



- Evaluation is easy: just count errors on a test set.
- Accuracy, Precision, Recall, F-measure.

Evaluating Sequence Labellers

Part-of-speech Tagging

Input:	I	saw	her	duck
Output:	PRON	VERB	PRON	VERB
Truth:	PRON	VERB	DET	NOUN

Evaluating Sequence Labellers

Part-of-speech Tagging

Input:	I	saw	her	duck
Output:	PRON	VERB	PRON	VERB
Truth:	PRON	VERB	DET	NOUN

- Evaluation is similar to classification.

Evaluating Sequence Labellers

Part-of-speech Tagging

Input:	I	saw	her	duck
Output:	PRON	VERB	PRON	VERB
Truth:	PRON	VERB	DET	NOUN

- Evaluation is similar to classification.
- Count mistakes at the token level and use the same metrics.

Machine Translation Evaluation

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.

[Koehn, 2010, Figure 8.1]

Machine Translation Evaluation

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.

[Koehn, 2010, Figure 8.1]

Evaluation is hard!

A single sentence can admit multiple translations.

1. Manual Evaluation

2. Automatic Evaluation
   - Intrinsic metrics
   - Trained Metrics
   - Evaluating metrics

3. Task-based Evaluation

Manual Evaluation - Direct Assessment

- Define a scale of "correctness".

Manual Evaluation - Direct Assessment

- Define a scale of “correctness”.
  - 1 - Translation is completely wrong
  - ...
  - 5 - Translation is completely right

Manual Evaluation - Direct Assessment

- Define a scale of “correctness”.
  - 1 - Translation is completely wrong
  - ...
  - 5 - Translation is completely right
- Split into different aspects and define specific scales for each aspect

Manual Evaluation - Direct Assessment

- Define a scale of "correctness".
  - 1 - Translation is completely wrong
  - ...
  - 5 - Translation is completely right
- Split into different aspects and define specific scales for each aspect
  - Adequacy: does the translation convey the same meaning as the original sentence?
  - Fluency: is the output good English?

Manual Evaluation - Direct Assessment

- Define a scale of "correctness".
  - 1 - Translation is completely wrong
  - ...
  - 5 - Translation is completely right
- Split into different aspects and define specific scales for each aspect
  - Adequacy: does the translation convey the same meaning as the original sentence?
  - Fluency: is the output good English?

Prone to noise and scaling biases

Manual Evaluation - Ranking

Alternative: rank translations instead of assigning explicit scores.

Manual Evaluation - Ranking

Alternative: rank translations instead of assigning explicit scores.

- Given two or more translations, order them from the "worst" to the "best".

Manual Evaluation - Ranking

Alternative: rank translations instead of assigning explicit scores.

- Given two or more translations, order them from the “worst” to the “best”.
- Less information about the **absolute** quality of an MT system but is easier for the judges and more consistent. Useful to compare systems between themselves.

Manual Evaluation - Ranking

Alternative: rank translations instead of assigning explicit scores.

- Given two or more translations, order them from the “worst” to the “best”.
- Less information about the **absolute** quality of an MT system but is easier for the judges and more consistent. Useful to compare systems between themselves.
- Can also be done in more fine-grained ways, using aspects such as adequacy and fluency.

Manual Evaluation - Issues

Careful manual evaluation is arguably the best way to assess the quality of MT systems.

Manual Evaluation - Issues

Careful manual evaluation is arguably the best way to assess the quality of MT systems.

- Useful to compare systems with substantial differences in their architectures and/or training data.

Manual Evaluation - Issues

Careful manual evaluation is arguably the best way to assess the quality of MT systems.

- Useful to compare systems with substantial differences in their architectures and/or training data.

However, manual evaluation is **costly**, especially if one wants to be careful about consistency and minimise biases.

Manual Evaluation - Issues

Careful manual evaluation is arguably the best way to assess the quality of MT systems.

- Useful to compare systems with substantial differences in their architectures and/or training data.

However, manual evaluation is **costly**, especially if one wants to be careful about consistency and minimise biases.

- Judges need to be **paid** and take **time** to perform a good evaluation.
- We also need **lots** of judges to ensure consistency.

Manual Evaluation - Issues

Careful manual evaluation is arguably the best way to assess the quality of MT systems.

- Useful to compare systems with substantial differences in their architectures and/or training data.

However, manual evaluation is **costly**, especially if one wants to be careful about consistency and minimise biases.

- Judges need to be **paid** and take **time** to perform a good evaluation.
- We also need **lots** of judges to ensure consistency.

We usually want quick and cheap ways to assess quality when building MT systems, especially when **tuning** these systems.

Automatic Evaluation

Goal

Compare a MT output to a reference translation (or a list of references) **automatically**.

Automatic Evaluation

Goal

Compare a MT output to a reference translation (or a list of references) **automatically**.

- Design a metric that **mimics** human evaluation as close as possible...

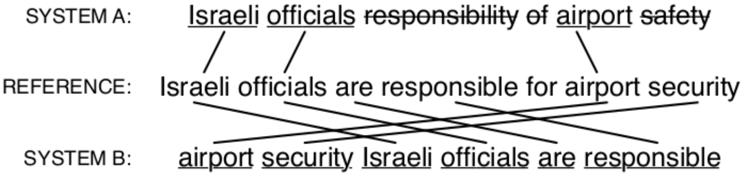
Goal

Compare a MT output to a reference translation (or a list of references) **automatically**.

- Design a metric that **mimics** human evaluation as close as possible...
- ...while also being cheap and fast.

Word Matches

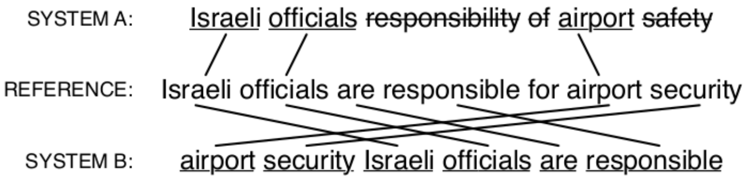
Count word matches and apply classification metrics.



[Koehn, 2010, Figure 8.4]

Word Matches

Count word matches and apply classification metrics.

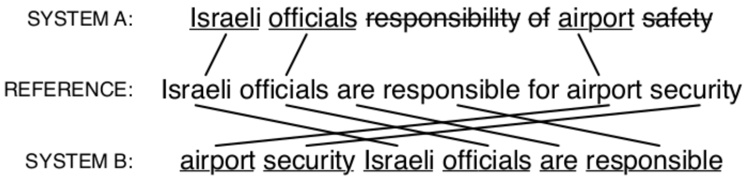


[Koehn, 2010, Figure 8.4]

	Precision	Recall	F-measure
System A	0.5	0.43	0.46
System B	1.0	0.86	0.92

Word Matches

Count word matches and apply classification metrics.



[Koehn, 2010, Figure 8.4]

	Precision	Recall	F-measure
System A	0.5	0.43	0.46
System B	1.0	0.86	0.92

Problem: metrics are agnostic to **word order**

Word Error Rate

We can take word order into account by using Word Error Rate (WER), which is based on the **edit distance** between strings.

$$\mathrm{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference length}}$$

Word Error Rate

We can take word order into account by using Word Error Rate (WER), which is based on the **edit distance** between strings.

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference length}}$$

- System A: 0.57
- System B: 0.71

Word Error Rate

We can take word order into account by using Word Error Rate (WER), which is based on the **edit distance** between strings.

$$\text{WER} = \frac{\textit{substitutions} + \textit{insertions} + \textit{deletions}}{\textit{reference length}}$$

- System A: 0.57
- System B: 0.71

Problem: sometimes the word order requirement is too harsh

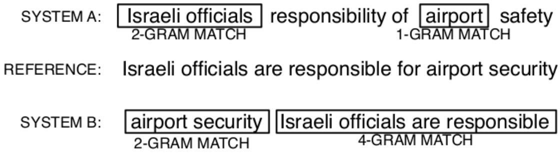
- Reference: "Israeli officials are responsible for airport security"
- Output: "This airport's security is responsibility of the Israeli security officials"
- WER: 1.29

BLEU - A Compromise

Bilingual Evaluation Understudy (BLEU) [Papineni et al., 2001] is by far the most used automatic metric in MT.

BLEU - A Compromise

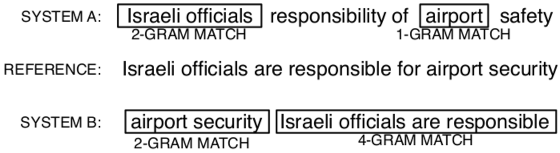
Bilingual Evaluation Understudy (BLEU) [Papineni et al., 2001] is by far the most used automatic metric in MT.



Based on precision scores but consider large n-gram matches as well (usually up to 4).

BLEU - A Compromise

Bilingual Evaluation Understudy (BLEU) [Papineni et al., 2001] is by far the most used automatic metric in MT.



Based on precision scores but consider large n-gram matches as well (usually up to 4).

	System A	System B
1-gram	3/6	6/6
2-gram	1/5	4/5
3-gram	0/4	2/4
4-gram	0/3	1/3

BLEU

- BLEU uses a geometric average of precisions:

$$\prod_{n=1}^{4} precision_n$$

- This reduces to zero if there are no n-gram matches (mostly a problem for 4-grams).

BLEU

- BLEU uses a geometric average of precisions:

$$\prod_{n=1}^{4} precision_n$$

- This reduces to zero if there are no n-gram matches (mostly a problem for 4-grams).

Key point

Evaluation procedures usually employ a whole set of sentences (test set).

BLEU

- BLEU uses a geometric average of precisions:

$$\prod_{n=1}^{4} precision_n$$

- This reduces to zero if there are no n-gram matches (mostly a problem for 4-grams).

Key point

Evaluation procedures usually employ a whole set of sentences (test set).

- BLEU takes advantage of that: all precisions are calculated with respect to the **whole test set** T.

$$precision_n = \frac{\sum_{s \in T} \sum_{\text{n-gram}} Match(\text{n-gram})}{\sum_{s \in T} \sum_{\text{n-gram}} Count(\text{n-gram})}$$

BLEU - Precision Issues

- Reference: "This airport's security is responsibility of the Israeli security officials"
- Output: "is responsiblity of the"

BLEU - Precision Issues

- Reference: “This airport’s security is responsibility of the Israeli security officials”
- Output: “is responsiblity of the”
- Perfect precision!

BLEU - Precision Issues

- Reference: "This airport's security is responsibility of the Israeli security officials"
- Output: "is responsiblity of the"
- Perfect precision!

BLEU introduces a brevity penalty term to penalise short translations, resulting in the formula:

$$\text{BLEU} = \min\left(1, \frac{\textit{output length}}{\textit{reference length}}\right) \prod_{n=1}^{4} \textit{precision}_n$$

Other Metrics

Translation Error Rate (TER) [Snover et al., 2006]

- Similar to WER but allow word **shifts**

Other Metrics

Translation Error Rate (TER) [Snover et al., 2006]

- Similar to WER but allow word **shifts**

METEOR [Banerjee and Lavie, 2005]

- Similar to BLEU but with focus on recall
- Also allow **soft** matches, using word stems, synonyms and paraphrases

Trained Metrics

Manual evaluation campaigns generate labels contain human judgements

Trained Metrics

Manual evaluation campaigns generate labels contain human judgements

This is valuable data! Can we use it to guide the design of new metrics?

Trained Metrics

Manual evaluation campaigns generate labels contain human judgements

This is valuable data! Can we use it to guide the design of new metrics?

Idea: use the data to **learn** a metric in a supervised learning setting.

- Inputs: MT output and reference translation
- Output: score (usually the higher, the better)

BEER

BEER [Stanojević and Sima'an, 2014] is a trained metric that uses a linear model. Given an MT output o and a reference r, it is defined as:

$$\text{BEER}(o, r) = \mathbf{w} \cdot \phi_{o,r}$$

BEER [Stanojević and Sima'an, 2014] is a trained metric that uses a linear model. Given an MT output o and a reference r, it is defined as:

$$\text{BEER}(o, r) = \mathbf{w} \cdot \phi_{o,r}$$

- $\mathbf{w}$ is a set of weights.
- $\phi_{o,r}$ is a set of features that compare o with r:
  - Word and character-level matches.
  - Word order features, obtained by counting permutations.

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$BEER(o_{good}, r) > BEER(o_{bad}, r)$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$BEER(o_{good}, r) > BEER(o_{bad}, r)$$
$$\mathbf{w} \cdot \phi_{good} > \mathbf{w} \cdot \phi_{bad}$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$BEER(o_{good}, r) > BEER(o_{bad}, r)$$

$$\mathbf{w} \cdot \phi_{good} > \mathbf{w} \cdot \phi_{bad}$$

$$\mathbf{w} \cdot \phi_{good} - \mathbf{w} \cdot \phi_{bad} > 0$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$
\begin{aligned}
BEER(o_{good}, r) &> BEER(o_{bad}, r) \\
\mathbf{w} \cdot \phi_{good} &> \mathbf{w} \cdot \phi_{bad} \\
\mathbf{w} \cdot \phi_{good} - \mathbf{w} \cdot \phi_{bad} &> 0 \\
\mathbf{w} \cdot (\phi_{good} - \phi_{bad}) &> 0
\end{aligned}
$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$
\begin{aligned}
BEER(o_{good}, r) &> BEER(o_{bad}, r) \\
\mathbf{w} \cdot \phi_{good} &> \mathbf{w} \cdot \phi_{bad} \\
\mathbf{w} \cdot \phi_{good} - \mathbf{w} \cdot \phi_{bad} &> 0 \\
\mathbf{w} \cdot (\phi_{good} - \phi_{bad}) &> 0 \\
\mathbf{w} \cdot (\phi_{bad} - \phi_{good}) &< 0
\end{aligned}
$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$\begin{aligned}
BEER(o_{good}, r) &> BEER(o_{bad}, r) \\
\mathbf{w} \cdot \phi_{good} &> \mathbf{w} \cdot \phi_{bad} \\
\mathbf{w} \cdot \phi_{good} - \mathbf{w} \cdot \phi_{bad} &> 0 \\
\mathbf{w} \cdot (\phi_{good} - \phi_{bad}) &> 0 \\
\mathbf{w} \cdot (\phi_{bad} - \phi_{good}) &< 0 \\
\mathbf{w} \cdot (\phi_1 - \phi_2) &\begin{cases} " > " & o_1 \text{ is better} \\ " < " & o_1 \text{ is worse} \end{cases}
\end{aligned}$$

BEER - Training

Labelled data is consisted of pairwise judgements:

- Given r, $o_1 > o_2$ or $o_1 < o_2$.

Goal of the metric is to mimic these judgments.

$$
\begin{aligned}
BEER(o_{good}, r) &> BEER(o_{bad}, r) \\
\mathbf{w} \cdot \phi_{good} &> \mathbf{w} \cdot \phi_{bad} \\
\mathbf{w} \cdot \phi_{good} - \mathbf{w} \cdot \phi_{bad} &> 0 \\
\mathbf{w} \cdot (\phi_{good} - \phi_{bad}) &> 0 \\
\mathbf{w} \cdot (\phi_{bad} - \phi_{good}) &< 0 \\
\mathbf{w} \cdot (\phi_1 - \phi_2) &\begin{cases} ">" & o_1 \text{ is better} \\ "<" & o_1 \text{ is worse} \end{cases}
\end{aligned}
$$

This is just binary classification! BEER uses logistic regression but other algorithms could be used as well.

Evaluating Metrics

How to evaluate an evaluation metric?

How to evaluate an evaluation metric?

- A perfect automatic metric would simulate manual evaluation
- We can assess a metric by its **correlation** with human judgements

Kendall's τ

For pairwise rankings, we can use Kendall's τ rank correlation.

Kendall's τ

For pairwise rankings, we can use Kendall's τ rank correlation.

- Given all pairs used in the manual evaluation, gather ranks according to the proposed **metric**.
- Concordant pairs (*Con*) are the ones that match the manual rank, while Discordant pairs (*Dis*) disagrees with the manual rank.

Kendall's τ

For pairwise rankings, we can use Kendall's τ rank correlation.

- Given all pairs used in the manual evaluation, gather ranks according to the proposed **metric**.
- Concordant pairs (Con) are the ones that match the manual rank, while Discordant pairs (Dis) disagrees with the manual rank.

$$\tau = \frac{|Con| - |Dis|}{|Con| + |Dis|}$$

Kendall's τ

For pairwise rankings, we can use Kendall's τ rank correlation.

- Given all pairs used in the manual evaluation, gather ranks according to the proposed **metric**.
- Concordant pairs (Con) are the ones that match the manual rank, while Discordant pairs (Dis) disagrees with the manual rank.

$$\tau = \frac{|Con| - |Dis|}{|Con| + |Dis|}$$

- A score of 1 gives perfect agreement, -1 gives perfect disagreement and 0 means no correlation.

Kendall's τ

Direction	en-fr	en-de	en-hi	en-cs	en-ru
BEER	**.295**	**.258**	.250	**.344**	**.440**
METEOR	.278	.233	.264	.318	.427
AMBER	.261	.224	**.286**	.302	.397
BLEU-NRC	.257	.193	.234	.297	.391
APAC	.255	.201	.203	.292	.388

Table 3: Kendall τ correlations on the WMT14 human judgements when translating out of English.

[Stanojević and Sima'an, 2014]

Task-based Evaluation

Suppose we need to choose between System A or System B and manual evaluation is not an option.

Task-based Evaluation

Suppose we need to choose between System A or System B and manual evaluation is not an option.

Option 1: we can choose the system by assessing it through an automatic metric (BLEU, for instance).

Task-based Evaluation

Suppose we need to choose between System A or System B and manual evaluation is not an option.

Option 1: we can choose the system by assessing it through an automatic metric (BLEU, for instance).

- Since the metric is not perfect we risk choosing the wrong system.

Suppose we need to choose between System A or System B and manual evaluation is not an option.

Option 1: we can choose the system by assessing it through an automatic metric (BLEU, for instance).

- Since the metric is not perfect we risk choosing the wrong system.

Option 2: employ **both** systems and choose the best one at production ("test") time.

Task-based Evaluation

Suppose we need to choose between System A or System B and manual evaluation is not an option.

Option 1: we can choose the system by assessing it through an automatic metric (BLEU, for instance).

- Since the metric is not perfect we risk choosing the wrong system.

Option 2: employ **both** systems and choose the best one at production ("test") time.

- No reference translations!
- No human judgements!

Task-based Evaluation

In practice, MT is not the **end goal** but rather a tool to solve a problem.

Task-based Evaluation

In practice, MT is not the **end goal** but rather a tool to solve a problem.

- Gisting (did I get the information I needed?)

Task-based Evaluation

In practice, MT is not the **end goal** but rather a tool to solve a problem.

- Gisting (did I get the information I needed?)
- Product localisation (did I sell more items?)

Task-based Evaluation

In practice, MT is not the **end goal** but rather a tool to solve a problem.

- Gisting (did I get the information I needed?)
- Product localisation (did I sell more items?)
- Post-editing in human translation (did I translate faster by post-editing the MT output?)

Task-based Evaluation

In practice, MT is not the **end goal** but rather a tool to solve a problem.

- Gisting (did I get the information I needed?)
- Product localisation (did I sell more items?)
- Post-editing in human translation (did I translate faster by post-editing the MT output?)

All these tasks can be measured in some way

Quality Estimation

Evaluating translations with respect to a task is usually referred as Quality Estimation [Blatz et al., 2004, Specia, 2011].

Quality Estimation

Evaluating translations with respect to a task is usually referred as Quality Estimation [Blatz et al., 2004, Specia, 2011].

- Main goal is to predict quality at **test time**, when the system is used in production.

Quality Estimation

Evaluating translations with respect to a task is usually referred as Quality Estimation [Blatz et al., 2004, Specia, 2011].

- Main goal is to predict quality at **test time**, when the system is used in production.
- Idea is similar to trained metrics: we collect data and train supervised ML models.

Quality Estimation

Evaluating translations with respect to a task is usually referred as Quality Estimation [Blatz et al., 2004, Specia, 2011].

- Main goal is to predict quality at **test time**, when the system is used in production.
- Idea is similar to trained metrics: we collect data and train supervised ML models.
  - Except that we have no access to reference translations.
  - And the labels are directly related to end tasks.

Post-editing

Scenario: an MT system is used to preprocess texts for translation. Main idea is that translators will spend less time post-editing MT outputs than translate the source text from scratch.

Post-editing

Scenario: an MT system is used to preprocess texts for translation. Main idea is that translators will spend less time post-editing MT outputs than translate the source text from scratch.
How to evaluate post-editing effort?

- Post-edition operations
- Post-editing time
- Keystrokes
- ...

Post-editing

Scenario: an MT system is used to preprocess texts for translation. Main idea is that translators will spend less time post-editing MT outputs than translate the source text from scratch.
How to evaluate post-editing effort?

- Post-edition operations
- Post-editing time
- Keystrokes
- ...

Metric choice depends on the application. For instance, post-editing time can be useful not only for quality prediction but also to give productivity estimates.

Post-editing

Given a metric and collected data, we can build supervised ML models and use the predictions as a measure of quality.

Post-editing

Given a metric and collected data, we can build supervised ML models and use the predictions as a measure of quality.

- Similar to what we do with trained metrics, but now features come from the MT output and the source sentence (no reference).

Post-editing

Given a metric and collected data, we can build supervised ML models and use the predictions as a measure of quality.

- Similar to what we do with trained metrics, but now features come from the MT output and the source sentence (no reference).
  - Source sentence length;
  - MT output length;
  - Language model probabilities;
  - Word alignment scores;
  - ...

Quality Estimation - Challenges

Many issues arises in the data.

Quality Estimation - Challenges

Many issues arises in the data.

- Different users/tasks have different biases.
- Measurement noise (post-editing time, for instance).
- Hard to transfer data between domains.

Quality Estimation - Challenges

Many issues arises in the data.

- Different users/tasks have different biases.
- Measurement noise (post-editing time, for instance).
- Hard to transfer data between domains.

Some recent work try to tackle these problem.s

- Learning from multiple users [Cohn and Specia, 2013].
- Uncertainty estimates for predictions [Beck et al., 2016].

Summary

MT evaluation is **hard**!

Summary

MT evaluation is **hard**!

- Manual evaluation is ideal but very expensive.

Summary

MT evaluation is hard!

- Manual evaluation is ideal but very expensive.
- Automatic intrinsic metrics are cheap but imperfect.
  - Still, BLEU is largely employed and accepted as the standard evaluation metric.

Summary

MT evaluation is **hard**!

- Manual evaluation is ideal but very expensive.
- Automatic intrinsic metrics are cheap but imperfect.
  - Still, BLEU is largely employed and accepted as the standard evaluation metric.
- Trained metrics can better correlate with human judgements but require data and feature/system engineering.

Summary

MT evaluation is hard!

- Manual evaluation is ideal but very expensive.
- Automatic intrinsic metrics are cheap but imperfect.
  - Still, BLEU is largely employed and accepted as the standard evaluation metric.
- Trained metrics can better correlate with human judgements but require data and feature/system engineering.
- Task-based evaluation (Quality Estimation) can provide a way to estimate quality **after** system is in production.

Summary

MT evaluation is hard!

- Manual evaluation is ideal but very expensive.
- Automatic intrinsic metrics are cheap but imperfect.
  - Still, BLEU is largely employed and accepted as the standard evaluation metric.
- Trained metrics can better correlate with human judgements but require data and feature/system engineering.
- Task-based evaluation (Quality Estimation) can provide a way to estimate quality **after** system is in production.

Further reading: [Koehn, 2010, Chap. 8]

Banerjee, S. and Lavie, A. (2005).
METEOR: An automatic metric for MT evaluation with improved correlation with human judgments.
Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, (June):65.

Beck, D., Specia, L., and Cohn, T. (2016).
Exploring Prediction Uncertainty in Machine Translation Quality Estimation.
In *Proceedings of CoNLL*.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004).
Confidence estimation for machine translation.
In *Proceedings of the 20th Conference on Computational Linguistics*, pages 315–321.

Cohn, T. and Specia, L. (2013).
Modelling Annotator Bias with Multi-task Gaussian Processes: An Application to Machine Translation Quality Estimation.

In *Proceedings of ACL*, pages 32–42.

Koehn, P. (2010).
Statistical Machine Translation.
Cambridge University Press, 1st edition.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001).
Bleu: a method for automatic evaluation of machine translation.
In *Proceedings of ACL*, pages 311–318.

Rosti, A.-V. I., Zhang, B., Matsoukas, S., and Schwartz, R. (2010).
BBN System Description for WMT10 System Combination Task.
In *Proceedings of WMT*, pages 321–326.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Makhoul, J. (2006).
A study of translation edit rate with targeted human annotation.

In *Proceedings of AMTA*.

Specia, L. (2011).
Exploiting Objective Annotations for Measuring Translation Post-editing Effort.
In *Proceedings of EAMT*, pages 73–80.

Stanojević, M. and Sima'an, K. (2014).
BEER : BEtter Evaluation as Ranking.
In *Proceedings of WMT*, pages 414–419.