You Cannot Feed Two Birds with One Score: the Accuracy-Naturalness Tradeoff in Translation

Gergely Flamich[1], David Vilar[2], Jan-Thorsten Peter[2], Markus Freitag[2]
g.flamich@imperial.ac.uk, [1]Imperial College London, [2]Google

TL;DR: It is not possible to quantify translation accuracy and naturalness using a single numeric metric.

Problem Setup and Motivation

Want to measure:

Accuracy: How much of the source's meaning does the translation carry?

Naturalness: How "good" does the translation sound in the target language?

How do we do it?

Linguistics: Multidimensional Quality Metrics (MQM) – Time-consuming, expensive.

Classical MT: BLEU, chrF etc. – Measure neither semantic accuracy nor naturalness.

Contemporary MT: MetricX, Comet, also QE variants – The way to go?

Questions

What *are* accuracy and naturalness? – **No formal definitions as yet.**

Folklore: there *should be some metric* that can measure both. – **Is this actually true?**

Proposed Definitions

$\mathcal{X}$ – source language $\quad P_{\mathbf{x},\mathbf{y}}$ – distribution over source-target pairs

$\mathcal{Y}$ – target language $\quad Q_{\mathbf{y}|\mathbf{x}}$ – translation system (human or machine)

Accuracy

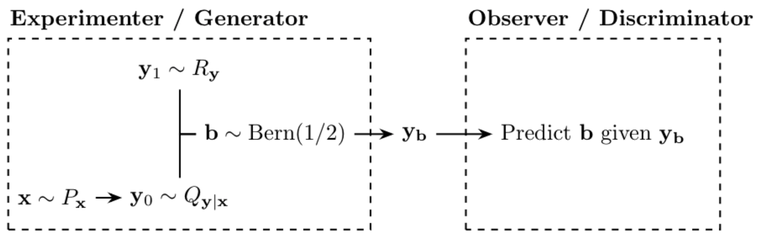
Ingredients: Distortion $\Delta : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ – bounded, minimal when translation correct.

$$A(Q_{\mathbf{y}|\mathbf{x}}) = -\Delta(Q_{\mathbf{y}|\mathbf{x}}) = -\mathbb{E}_{\mathbf{x},\mathbf{y}^r \sim P_{\mathbf{x},\mathbf{y}}}[\mathbb{E}_{\mathbf{y}^c \sim Q_{\mathbf{y}|\mathbf{x}}}[\Delta(\mathbf{x}, \mathbf{y}^r, \mathbf{y}^c)]]$$

Naturalness

Ingredients: Statistical metric D, marginal $Q_{\mathbf{y}}(\cdot) = \mathbb{E}_{\mathbf{x}\sim P_{\mathbf{x}}}[Q_{\mathbf{y}|\mathbf{x}}(\cdot)]$, reference $R_{\mathbf{y}}$.

$$N(Q_{\mathbf{y}|\mathbf{x}}) = -D(Q_{\mathbf{y}}, R_{\mathbf{y}})$$



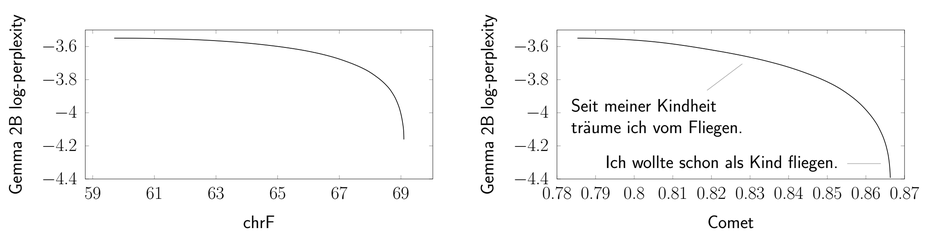
No Two Birds with One Score & the Accuracy-Naturalness Curve

Perfect naturalness ⇒ perfect accuracy? – No, set $Q_{\mathbf{y}|\mathbf{x}} \leftarrow R_{\mathbf{y}}$ as counterexample.

Perfect accuracy ⇒ perfect naturalness? – No, if $R_{\mathbf{y}}$ cannot depend on $P_{\mathbf{x},\mathbf{y}}$.

Accuracy-Naturalness Function

$$A(N) = \max_{Q_{\mathbf{y}|\mathbf{x}}} \{-\Delta(Q_{\mathbf{y}|\mathbf{x}})\} \quad \text{subject to} \quad -D(Q_{\mathbf{y}}, R_{\mathbf{y}}) \geq N$$



Theorem

1. $A(N)$ non-increasing – close to the curve, accuracy and naturalness **must anti-correlate**!
2. If D convex in first slot, then $A(N)$ concave – at the extremes, a small increase in one metric leads to large decrease in the other!

Evaluating the State-of-the-Art