Multi-metric Optimization using Ensemble Tuning

Baskaran Sankaran\textsuperscript{1}
Anoop Sarkar\textsuperscript{1}
Kevin Duh\textsuperscript{2}

\textsuperscript{1}Simon Fraser University
\textsuperscript{2}NAIST
MT Evaluation Metrics
MT Evaluation Metrics

- AMBER
- METER
- TER
- BLEU
- RIBES
- Block
- Err
- Cats
- WER
- posF
- SimpBLEU
- SVM
- Rank
- TerrorCat
- TESLA
- WordBlockEC
- P
- SAGAN
- STS
- Badger
- NIST
- mNCD

Multi-Metric Optimization for SMT
MT Evaluation Metrics

To Optimize:

BLEU
MT Evaluation Metrics

AMBER
TER
Meteor
WER
BLEU
RIBES
Block
Err
Cats
posF
Simp
BLEU
SVM
Rank
Terror
Cat
TESLA
Word
Block
EC
SAGAN-STS
TERp
NCD
Sem
Pos
NIST
mNCD

Multi-Metric Optimization for SMT
MT Evaluation Metrics

• Each captures certain *unique* aspect of translation
  – Does the translation quality improve if we consider multiple metrics?
MT Evaluation Metrics

- Meta-Validation: Human evaluation
  - Effort required to post-edit for publication: HTER
  - Could we do better by optimizing multiple metrics?

Multi-Metric Optimization for SMT
Multi-Metric Optimization

Multi-Metric Optimization for SMT
Multi-Metric Optimization for SMT
$w^* = \arg\max_w g\left(M_1(H), \ldots, M_k(H)\right)$
Multi-Metric Optimization

\[ w^* = \arg \max_w g\left(M_1(H), \ldots, M_k(H)\right) \]

- \( M_i \) scores hypotheses \( H \) for \( i^{th} \) metric
- \( g \) decides how the metrics are combined
Multi-Metric Optimization

$$w^* = \arg \max_w g \left( M_1(H), \ldots, M_k(H) \right)$$

- $M_i$ scores hypotheses $H$ for $i^{th}$ metric
- $g$ decides how the metrics are combined

Lateen, Union
PMO-Ensemble, Ensemble Tuning
Single-Metric Optimization

PRO (Hopkins and May, 2011)

Multi-Metric Optimization for SMT
Single-Metric Optimization

PRO (Hopkins and May, 2011)

Positive examples

Model Score

Multi-Metric Optimization for SMT
Single-Metric Optimization

- Consider top-$k$ positive and negative candidates
- Pairwise rankings: $h^+ - h^-$ and $M_{\text{Bleu}}(h^+) - M_{\text{Bleu}}(h^-)$
Single-Metric Optimization

- Consider top-\(k\) positive and negative candidates
- Pairwise rankings: \(h^+ - h^-\) and \(M_{\text{Bleu}}(h^+) - M_{\text{Bleu}}(h^-)\)

PRO (Hopkins and May, 2011)
**Single-Metric Optimization**

- Consider top-$k$ positive and negative candidates
- Pairwise rankings: $h^+ - h^-$ and $M_{Bleu}(h^+) - M_{Bleu}(h^-)$

**PRO (Hopkins and May, 2011)**

Multi-Metric Optimization for SMT
Single-Metric Optimization

- Consider top-\(k\) positive and negative candidates
- Pairwise rankings: \(h^+ - h^-\) and \(M_{\text{Bleu}}(h^+) - M_{\text{Bleu}}(h^-)\)

PRO (Hopkins and May, 2011)
Experimental Setup

• Arabic-English
  – Training: ISI corpus, > 1M sentence pairs

• Hiero system: Kriya (Sankaran et al. 2012)
  – https://github.com/sfu-natlang/Kriya

• Tuning with PRO (Hopkins and May 2011)

• 4-Metrics: BLEU, METEOR, RIBES, TER

Multi-Metric Optimization for SMT
Baseline MMO
Baseline MMO

• Linear combination
  – $g$ defines the weighted sum/avg
  – $(\text{TER} - \text{BLEU})/2$ and variants

$$g_{\text{avg}} = \sum_{1}^{k} \lambda_k M_k(H)$$

– Manually tune $\lambda_k$

(Cer et al. ’10)
(Servan et al. ’11)
Baseline MMO

• Linear combination
  – $g$ defines the weighted sum/ avg
  – $(\text{TER} - \text{BLEU})/2$ and variants

$$g_{\text{avg}} = \sum_{1}^{k} \lambda_k M_k(H)$$

– Manually tune $\lambda_k$

(Cer et al. ’10)
(Servan et al. ’11)

Multi-Metric Optimization for SMT
MMO Approaches

• Lateen
• Union
• PMO Ensemble
• Ensemble Tuning
Lateen Optimization
Lateen Optimization

• Uses two objective functions  (Spitkovsky et al. 2011)
  – A secondary objective as a course-correction
  – Supports a primary objective function
  – Effective in moving away from local optima and possibly towards better point
  – Successfully used for dependency parsing

• Lateen MMO
  – Optimize towards two (or more) metrics
Lateen MMO

Multi-Metric Optimization for SMT
Lateen MMO

Multi-Metric Optimization for SMT
Lateen MMO

Multi-Metric Optimization for SMT
• $g$ fires only one metric at each iteration $j$

$$I_i = \begin{cases} 
1 & \text{if } i \mod j = 0, \\
0 & \text{otherwise}
\end{cases}$$

$$g(H) = [I_1, \ldots, I_k].[M_1(H), \ldots, M_k(H)]$$

$$w^* = \arg \max_w g(H)$$
• Optimize all metrics jointly
• Optimize all metrics jointly
• Optimize all metrics jointly
• Optimize all metrics jointly

\[ g(H) = M_1(H) \cup \ldots \cup M_k(H) \]

\[ w^* = \arg \max_w g(H) \]
Results

- **RIBES**
  - Rank-based correlation metric (Isozaki et al. 2010)
  - Tracks *word-order* differences between ref and output
  - *Precision* term to additionally enforce adequacy
  - Similar to LRScore (Birch and Osborne, 2010)

- Complements **BLEU** in multi-metric setting
  - **BLEU-RIBES** was used in Duh et al. (2012)
Multi-Metric Optimization for SMT
Bleu-Ribes

Multi-Metric Optimization for SMT
Points on the upper-right quadrant.
MMO Approaches

• Lateen
• Union
• PMO Ensemble
• Ensemble Tuning
Pareto-Multi Objective

(Duh et al. 2012)

Metric-1 vs. Metric-2
Pareto-Multi Objective

(Duh et al. 2012)

Metric-1 vs. Metric-2

Multi-Metric Optimization for SMT
Pareto-Multi Objective

(Duh et al. 2012)
Pareto-Multi Objective

(Duh et al. 2012)

\[ h_i \text{ dominates } h_j : h_i \triangleright h_j, \text{ if} \]
\[ \forall M_l, M_l(h_i) \geq M_l(h_j) \text{ and} \]
\[ \exists M_k, M_k(h_i) > M_k(h_j) \]
Pareto-Multi Objective

• Pareto domination

\[ h_i \text{ dominates } h_j : h_i \triangleright h_j, \text{ if } \]
\[ \forall M_i, M_i(h_i) \geq M_i(h_j) \text{ and } \]
\[ \exists M_k, M_k(h_i) > M_k(h_j) \]

• Pareto-Optimal

– \( h^* \) is Pareto-optimal, \( \text{iff} \)

\[ h' \nsubseteq h^*, \forall h' \in H \text{ and } h' \neq h^* \]

(Duh et al. 2012)
Pareto-Multi Objective

• Pareto domination

\[
h_i \text{ dominates } h_j \iff h_i \triangleright h_j, \text{ if } \forall M_l, M_l(h_i) \geq M_l(h_j) \text{ and } \exists M_k, M_k(h_i) > M_k(h_j)
\]

• Pareto-Optimal

— \(h^*\) is Pareto-optimal, \(iff\)

\[
h' \not\triangleright h^*, \forall h' \in H \text{ and } h' \neq h^*
\]

(Duh et al. 2012)

Multi-Metric Optimization for SMT
PMO-PRO

(Duh et al. 2012)

Multi-Metric Optimization for SMT
PMO-PRO

• Pareto-frontier
  – Set of Pareto-optimal hypotheses: \( \{h_f\} \)
  – All hypotheses on the frontier are equally good

• PRO: Optimize to increase points on the Pareto-frontier

(Duh et al. 2012)
PMO-PRO

(Duh et al. 2012)

- Pareto-frontier
  - Set of Pareto-optimal hypotheses: \( \{h\} \)
  - All hypotheses on the Pareto-frontier are equally good

\[
g_{PMO} = \int_{1,...,k} [\lambda_k M_k(H)]
\]

- PRO: Optimize to increase points on the Pareto-frontier
PMO-PRO

- Pareto-frontier
  - Set of Pareto-optimal hypotheses: \( \{ h_f \} \)
  - All hypotheses on the frontier are equally good

\[
g_{\text{PMO}} = \int_{1,...,k} [\lambda_k M_k(H)]
\]

- PRO: Optimize to increase points on the Pareto-frontier

(Duh et al. 2012)
PMO-PRO

• Run PMO with different meta-weights \( \{\lambda\} \)
  – Each setting yields a Pareto solution

• User is forced to choose one solution
  – Hard trade-off choice
  – Not possible to exploit all the solutions

• Our approach
  – PMO Ensemble
PMO-Ensemble

Ensemble of weights

Pareto Sol #1
- TM
- LM

Pareto Sol #2
- TM
- LM

Ensemble decoding (Razmara et al. 2012)

Multi-Metric Optimization for SMT
PMO-Ensemble

Ensemble of weights

Pareto Sol #1
- TM
- LM

il $\rightarrow$ he (-4)
il $\rightarrow$ for (-5)
il $\rightarrow$ it (-6)

Ensemble decoding (Razmara et al. 2012)

Pareto Sol #2
- TM
- LM

il $\rightarrow$ it (-5)
il $\rightarrow$ he (-6)
il $\rightarrow$ for (-10)

Multi-Metric Optimization for SMT
PMO-Ensemble

\[ p(\bar{e}|\bar{f}) \propto \sum_m \beta_m \exp(w_m \cdot \phi_m) \]

Ensemble of weights

Ensemble decoding (Razmara et al. 2012)

Multi-Metric Optimization for SMT
PMO-Ensemble

\[ p(\bar{e}|\bar{f}) \propto \sum_m \beta_m \exp(w_m \cdot \phi_m) \]

Multi-Metric Optimization for SMT
Results: PMO-Ensemble
Results: PMO-Ensemble

Multi-Metric Optimization for SMT
Results: PMO-Ensemble

Again pushes the knee of the frontier curve
MMO Approaches

• Lateen
• Union
• PMO Ensemble
• Ensemble Tuning
Why?
Why?

• Linear combination/ PMO-PRO:
  – Meta-weights are manually tuned (hard trade-off)

• Lateen/ Union:
  – No meta-weights

• Union:
  – Inherent *conflict of interest* in candidate generation
Why?

- **Linear combination/ PMO-PRO:**
  - Meta-weights are manually tuned (hard trade-off)

- **Lateen/ Union:**
  - No meta-weights

- **Union:**
  - Inherent *conflict of interest* in candidate generation
Why?

• **Linear combination/ PMO-PRO:**
  – Meta-weights are manually tuned (hard trade-off)

• **Lateen/ Union:**
  – No meta-weights

• **Union:**
  – Inherent *conflict of interest* in candidate generation
Why?

- **Linear combination/ PMO-PRO:**
  - Meta-weights are manually tuned (hard trade-off)
- **Lateen/ Union:**
  - No meta-weights
- **Union:**
  - Inherent *conflict of interest* in candidate generation
Ensemble Tuning

Multi-Metric Optimization for SMT
Ensemble Tuning

Algorithm

• Ensemble decode tuning set to get $H_{ens}$

• Two-step tuning:
  – Optimize metrics $M_k(H_{ens})$ separately
  – Optimize the meta-weights $\lambda$ (Pareto-PMO)

Multi-Metric Optimization for SMT
Ensemble Tuning

Algorithm

• Ensemble decode tuning set to get $H_{ens}$

• Two-step tuning:
  – Optimize metrics $M_k(H_{ens})$ separately
  – Optimize the meta-weights $\lambda$ (Pareto-PMO)

Multi-Metric Optimization for SMT
Ensemble Tuning

Algorithm

• Ensemble decode tuning set to get $H_{ens}$

• Two-step tuning:
  – Optimize metrics $M_k(H_{ens})$ separately
  – Optimize the meta-weights $\lambda$ (Pareto-PMO)

Metric-1 vs. Metric-2

Multi-Metric Optimization for SMT
Ensemble Tuning

Algorithm

• Ensemble decode tuning set to get $H_{ens}$

• Two-step tuning:
  – Optimize metrics $M_k(H_{ens})$ separately
  – Optimize the meta-weights $\lambda$ (Pareto-PMO)

Multi-Metric Optimization for SMT
Results

Multi-Metric Optimization for SMT
Results

Pushes the knee of the frontier curve

Multi-Metric Optimization for SMT
Results

Multi-Metric Optimization for SMT
Results

Improves scores on individual dimensions

Multi-Metric Optimization for SMT
Results

Improves scores on individual dimensions

- More results in the paper
- Including the results for your favourite metric
Human Evaluation
Human Evaluation

• Post-editing with PET tool (Aziz et al. 2012)
• 100 randomly chosen sentences
  – At least 15 words long
  – BLEU only, METEOR only and Ens. tuning (B-M-R)
• Compute “1 – HTER”
  – 6+% improvement for Ensemble tuned B-M-R

Multi-Metric Optimization for SMT
Metric dichotomy

Multi-Metric Optimization for SMT
Metric dichotomy

Multi-Metric Optimization for SMT
Metric dichotomy

What metrics are optimized together matters

Multi-Metric Optimization for SMT
Summary

• Four different approaches for MMO
• Ensemble tuning performs well
  – Higher scores along BLEU and RIBES dimensions
  – Substantially easy to post-edit: better HTER
• Metric dichotomy
  – Groups of metrics are amenable to be optimized together
The military's southern command said in a statement *that the jailed* “guards found unconscious, not breathing in his cell”.

**Bleu:**

The military's southern command said in a statement *that the jailed* “guards found unconscious, not breathing in his cell”.

**Meteor:**

The military southern command said in a statement *that the jailed* “guards found the unconscious, not breathing in his cell”.

**BMR EnsTune:**

The military southern command said in a statement *that the* “guards found the *detainee* unconscious, not breathing in his cell”.

Questions?
Multi-Metric Optimization for SMT
Improves scores on BLEU dimension
BLeU-nTER

Improves scores on BLEU dimension

Multi-Metric Optimization for SMT
Bleu vs. # of metrics

Multi-Metric Optimization for SMT
Stable BLEU across sets of varying # of metrics
Results: 1-ref setting

• Arabic-English
  – ISI tuning/test set

• Similar trend as seen in 4-ref setting
  – Improvement along BLEU and RIBES dimensions
  – Moderate loss in METEOR and TER
Multimetric Optimization for SMT