Machine Translation

15: Tidbits and Open Challenges

Rico Sennrich

University of Edinburgh

Training Objectives

- traditionally, NMT models are trained to minimize cross-entropy (equivalent to minimizing perplexity, and maximizing the likelihood of the training data)
- we (to often) measure model performance via BLEU
- can we directly optimize towards BLEU, or some other reward?

minimum risk training [Shen et al., 2016]

- minimize the *risk* (expected loss) of the model
- key ingredients:
  - a loss function Δ (e.g. negative sentence-level BLEU)
  - a set of translations S obtained via
    - sampling [Shen et al., 2016]
    - beam search [Edunov et al., 2017]
  - (using the full set of translations $\mathcal{Y}$ is intractable)

Minimum risk

maximum likelihood

$$\hat{\boldsymbol{\theta}}_{\mathrm{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}) \right\}$$

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)} | \mathbf{x}^{(s)}; \boldsymbol{\theta}) \\
&= \sum_{s=1}^{S} \sum_{n=1}^{N^{(s)}} \log P(\mathbf{y}_n^{(s)} | \mathbf{x}^{(s)}, \mathbf{y}_{<n}^{(s)}; \boldsymbol{\theta})
\end{aligned}$$

Minimum risk

minimum risk

$$\hat{\boldsymbol{\theta}}_{\mathrm{MRT}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Big\{ \mathcal{R}(\boldsymbol{\theta}) \Big\}.$$

$$\begin{aligned}\mathcal{R}(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta}} \Big[\Delta(\mathbf{y}, \mathbf{y}^{(s)}) \Big] \\ &= \sum_{s=1}^{S} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(s)})} P(\mathbf{y}|\mathbf{x}^{(s)}; \boldsymbol{\theta}) \Delta(\mathbf{y}, \mathbf{y}^{(s)})\end{aligned}$$

Minimum risk

minimum risk

$$\hat{\boldsymbol{\theta}}_{\mathrm{MRT}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Big\{ \mathcal{R}(\boldsymbol{\theta}) \Big\}.$$

$$\begin{aligned}\tilde{\mathcal{R}}(\boldsymbol{\theta}) &= \sum_{s=1}^{S} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta},\alpha} \Big[\Delta(\mathbf{y}, \mathbf{y}^{(s)}) \Big] \\ &= \sum_{s=1}^{S} \sum_{\mathbf{y} \in \mathcal{S}(\mathbf{x}^{(s)})} Q(\mathbf{y}|\mathbf{x}^{(s)}; \boldsymbol{\theta}, \alpha) \Delta(\mathbf{y}, \mathbf{y}^{(s)})\end{aligned}$$

Domain Adaptation

Different text collections can be different in:

- topic
- genre
- style
- level of formality
- ...

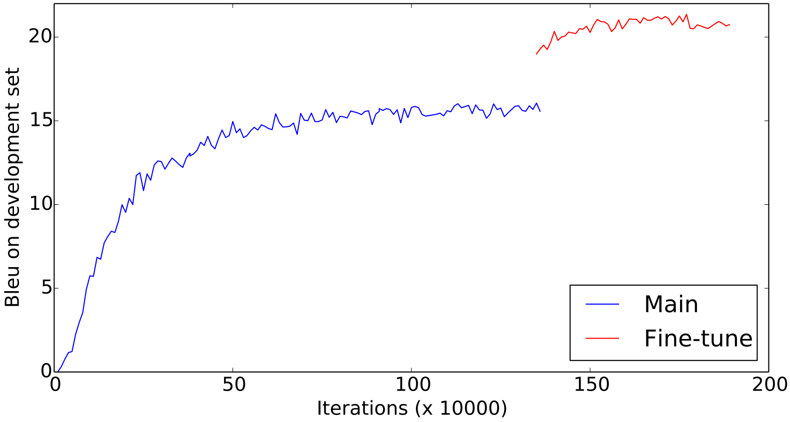
all these factors may affect translation of ambiguous source words

we can optimize performance on a specific text collection
$\rightarrow$ **domain adaptation**

Popular Domain Adaptation Techniques

- for phrase-based SMT:
  - weighting (or selection) of training data
  - weighted combination of in-domain and out-of-domain model(s)
- for neural MT:
  - *fine-tune* model with SGD on in-domain data
    (very effective)
  - domain indicator word (less effective)

Fine-Tuning for Domain Adaptation



Open Challenges

there are lots of open challenges...
...some of which we've already discussed

today: a small selection of challenges not discussed so far

Long Sentences

common claim: NMT performs poorly on long sentences

Long Sentences

attention helps

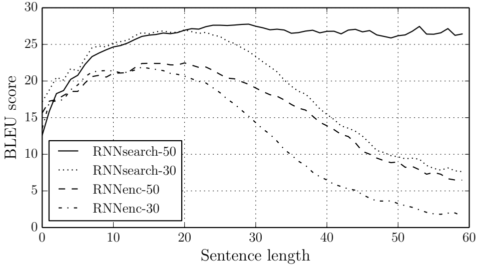


Figure 2: The BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models.

[Bahdanau et al., 2015]

Long Sentences

[Koehn and Knowles, 2017] find degradation on long sentences (system is not trained on long sentences)

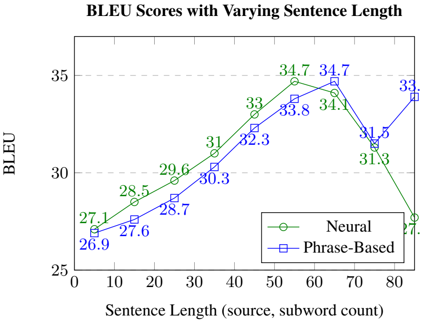


Figure 7: Quality of translations based on sentence length. SMT outperforms NMT for sentences longer than 60 subword tokens. For very long sentences (80+) quality is much worse due to too short output.

Long Sentences

we can avoid poor translations with reconstruction objective

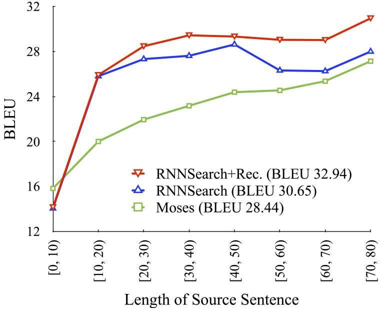
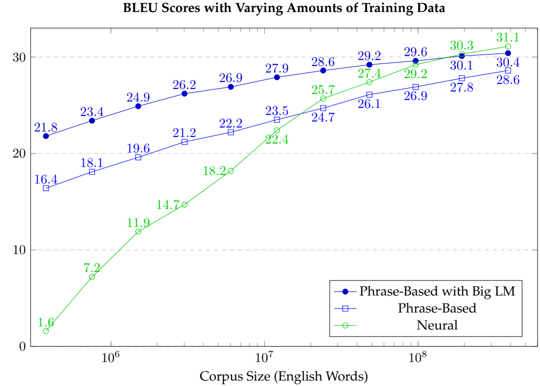


Figure 5: Performance of the generated translations with respect to the lengths of the input sentences on the test sets.

[Tu et al., 2016]

Low-Resource Neural MT



- learning curve is approximatively logarithmic
- phrase-based SMT performs better in low-data conditions
- even at 10^7 words (≈ 500 000 sentences), simple phrase-based system performs better than neural MT

discuss in pairs

which research that we discussed in previous lectures helps in low-resource settings?

Noisy Data

Ratio shuffled	0%	10%	20%	50%
SMT (BLEU)	32.7	32.7 (–0.0)	32.6 (–0.1)	32.0 (–0.7)
NMT (BLEU)	35.4	34.8 (–0.6)	32.1 (–3.3)	30.1 (–5.3)

Table 13.4: Impact of noise in the training data, with parts of the training corpus shuffled to contain mis-aligned sentence pairs. Neural machine translation degrades severely, while statistical machine translation holds up fairly well.

- effect of noise on phrase-based SMT:
  add some low-probability entries to translation model
- effect of noise on neural MT:
  change direction of parameter updates
  $\rightarrow$ model learns to rely more on target history than source text (?)

A Challenge Set for MT Evaluation [Isabelle et al., 2017]

Category	Subcategory	#	PBMT-1	NMT	Google NMT
Morpho-syntactic	Agreement across distractors	3	0%	100%	100%
	through control verbs	4	25%	25%	25%
	with coordinated target	3	0%	100%	100%
	with coordinated source	12	17%	92%	75%
	of past participles	4	25%	75%	75%
	Subjunctive mood	3	33%	33%	67%
Lexico-syntactic	Argument switch	3	0%	0%	0%
	Double-object verbs	3	33%	67%	100%
	Fail-to	3	67%	100%	67%
	Manner-of-movement verbs	4	0%	0%	0%
	Overlapping subcat frames	5	60%	100%	100%
	NP-to-VP	3	33%	67%	67%
	Factitives	3	0%	33%	67%
	Noun compounds	9	67%	67%	78%
	Common idioms	6	50%	0%	33%
	Syntactically flexible idioms	2	0%	0%	0%
Syntactic	Yes-no question syntax	3	33%	100%	100%
	Tag questions	3	0%	0%	100%
	Stranded preps	6	0%	0%	100%
	Adv-triggered inversion	3	0%	0%	33%
	Middle voice	3	0%	0%	0%
	Fronted should	3	67%	33%	33%
	Clitic pronouns	5	40%	80%	60%
	Ordinal placement	3	100%	100%	100%
	Inalienable possession	6	50%	17%	83%
	Zero REL PRO	3	0%	33%	100%

Table 3: Summary of scores by fine-grained categories. “#” reports number of questions in each category, while the reported score is the percentage of questions for which the divergence was correctly bridged. For each question, the three human judgments were transformed into a single judgment by taking system outputs with two positive judgments as positive, and all others as negative.

Idioms

from challenge set [Isabelle et al., 2017]

Source	His argument really **hit the nail on the head**.	
Ref	Son argument a vraiment **fait mouche**.	
PBMT-1	Son argument a vraiment **mis le doigt dessus**.	✓
NMT	Son argument a vraiment **frappé le clou sur la tête**.	✗
Google	Son argument a vraiment **frappé le clou sur la tête**.	✗

Discourse Phenomena

most MT systems operate on sentence level, but some translations require wider context.

example: most Romance languages mark gender in anaphoric pronouns

English	I made a decision.	Please respect it.
French	J'ai pris une décision.	Respectez-la s'il vous plaît.
French	J'ai fait un choix.	Respectez-le s'il vous plaît.

Further Reading

required reading

- Koehn, 13.8

Bibliography I

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural Machine Translation by Jointly Learning to Align and Translate.
In Proceedings of the International Conference on Learning Representations (ICLR).

Edunov, S., Ott, M., Auli, M., Grangier, D., and Ranzato, M. (2017).
Classical Structured Prediction Losses for Sequence to Sequence Learning.
CoRR, abs/1711.04956.

Isabelle, P., Cherry, C., and Foster, G. (2017).
A Challenge Set Approach to Evaluating Machine Translation.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017).
Six Challenges for Neural Machine Translation.
In Proceedings of the First Workshop on Neural Machine Translation, pages 28–39, Vancouver. Association for Computational Linguistics.

Shen, S., Cheng, Y., He, Z., He, W., Wu, H., Sun, M., and Liu, Y. (2016).
Minimum Risk Training for Neural Machine Translation.
In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany.

Tu, Z., Liu, Y., Shang, L., Liu, X., and Li, H. (2016).
Neural Machine Translation with Reconstruction.
CoRR, abs/1611.01874.