# Pronoun-Targeted Fine-tuning for NMT with Hybrid Losses

Prathyusha Jwalapuram, Shafiq Joty and Youlin Shen

School of Computer Science and Engineering
Nanyang Technological University

# Outline

# Machine Translation

Source (French) translated to Target (English).

**French-English translation example**

**French**: Il était créatif, généreux, drôle, affectueux et talentueux, et il va beaucoup me manquer.

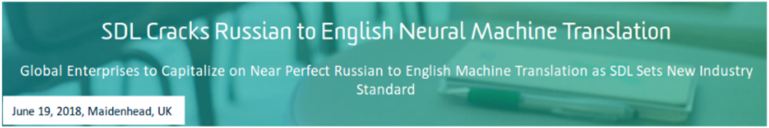
**English**: He was creative, generous, funny, loving and talented, and I will miss him dearly.

# Claims of Human Parity

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Microsoft reaches a historic milestone, using AI to match human performance in translating news from Chinese to English

March 14, 2018 | Allison Linn




[taken from Sennrich [2018a]]

# But...

- Läubli et al. [2018] conduct studies to show this is not true.
  $\Rightarrow$ Evaluation is not robust!
- NMT models still poor in translating discourse phenomena
  $\Rightarrow$ *e.g.*, pronouns, connectives, coherence

# But...

**French-English translation example**

**French:** **Il** était créatif, généreux, drôle ...

**Human:** He was creative, generous, funny ...

**MT:** It was creative, generous, funny ...

- Predominantly used MT metric: **BLEU**
- Measures n-gram word overlap with reference translation

# Discourse Phenomena

In typical text (discourse), sentences are related:

- John lives near the park.
  **He** often goes **there**. **(pronouns)**
- Eva walked into town to buy ice-cream.
  **But** the shop was closed. **(connective)**

[taken from Hardmeier [2018]]

# Context Aware Machine Translation

- MT and MT evaluation traditionally at sentence level.
- Recent systems now try to model **extra-sentential context**.
- But BLEU cannot reflect any improvements.

# Backtranslation

- Training strategies typically used to improve MT are:
  1. **Backtranslation** [Sennrich et al., 2015]
     - Additional pseudo-parallel data

# Backtranslation

- Training strategies typically used to improve MT are:
  1. **Backtranslation** [Sennrich et al., 2015]
     - Additional pseudo-parallel data

**Backtranslation Example**

**Monolingual:** she was eating biscuits afterwards.

**En-De MT:** **sie aß anschließend kekse.**

**Reference:** sie hat anschließend ein paar hundekuchen gefressen.

**De-En**

sie aß anschließend kekse.

$\Downarrow$

she was eating biscuits afterwards.

## Motivation

- NMT models poor in translating discourse phenomena like pronouns [Sennrich, 2018b]
- Elaborate contextual models are not consistent in performance across languages [Jwalapuram et al., 2020]

A typical pronoun translation error is mistranslation of the gender:

### Pronoun Translation Error

*S:* Mir wurdediese Wohnungin Earls Court gezeigt, und **sie** hatte ...
*T:* I was shown this apartment in Earls Court , and **she** had ...
*Correct:* I was shown this apartment in Earls Court , and **it** had ...

# Motivation

- Traditional conditional language model used for MT objective may be proving inadequate

# Motivation

- Traditional conditional language model used for MT objective may be proving inadequate
- Propose **hybrid conditional generative-discriminative** losses $\Rightarrow$ improve the learning power of the model

# Motivation

- Traditional conditional language model used for MT objective may be proving inadequate
- Propose **hybrid conditional generative-discriminative** losses $\Rightarrow$ improve the learning power of the model
  1. Target improvement of pronoun translations through **fine-tuning**
  2. Without using additional data
  3. Leverage existing training data the model has failed to learn from

# Conditional Language Model loss

For a source-target sentence pair $(x, y)$, a CLM predicts a conditional probability distribution $P_\theta(y_{1:n}|x)$, where $n$ = number of tokens in the target text and $\boldsymbol{c}$ = context vector that summarizes the relevant input.

$$P_\theta(y_{1:n}|x) = \prod_{t=1}^{n} P_\theta(y_t|y_{<t}, \boldsymbol{c}) \tag{1}$$

$$\mathcal{L}_g = -\frac{1}{n}\sum_{t=1}^{n} \log P_\theta(y_t|y_{<t}, \boldsymbol{c}) \tag{2}$$

# Fine-tuning Framework - Intuition for Additional Loss

- Characterised as
  1. Incorrect output produced by the model - "negative" class
  2. Target token is from "positive" class

# Fine-tuning Framework - Intuition for Additional Loss

- Characterised as
  1. Incorrect output produced by the model - "negative" class
  2. Target token is from "positive" class
- Main intuition: promote positive sample over negative sample rather than over entire vocabulary.

## Fine-tuning Framework - Intuition for Additional Loss

- Characterised as
  1. Incorrect output produced by the model - “negative” class
  2. Target token is from “positive” class
- Main intuition: promote positive sample over negative sample rather than over entire vocabulary.
- Two variants:
  - Log-likelihood loss
  - Max-margin loss

# Log-likelihood loss

Maximize the probability of the reference token by minimizing:

$$\mathcal{L}_{nll} = -\frac{1}{n}\sum_{t=1}^{n} \log \frac{\exp(\hat{y}_t^{+})}{\left(\exp(\hat{y}_t^{+}) + \exp(\hat{y}_t^{-})\right)} \tag{3}$$

- $y^{+}$ is the reference (positive) translation.
- $y^{-}$ is the model (negative) output.

# Max-Margin Loss

Pairwise ranking loss [Collobert et al., 2011] that maximizes the distance between positive and negative samples.

$$\mathcal{L}_{mm} = \frac{1}{n}\sum_{t=1}^{n} \max\{0, \mu - \hat{y}_t^{+} + \hat{y}_t^{-}\} \tag{4}$$

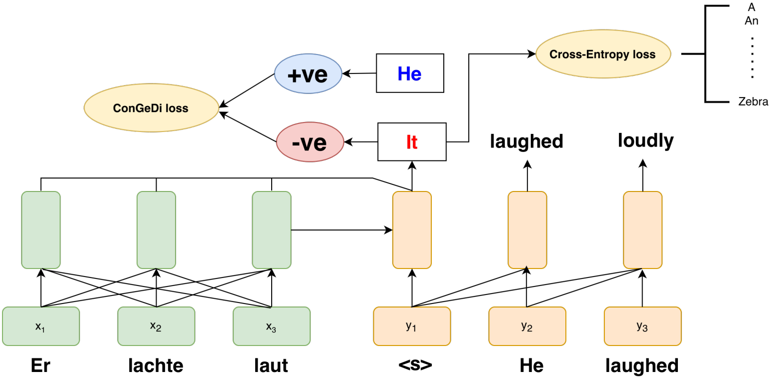
- $\mu$ is the margin.

# Training

- Losses can be applied on all tokens or targeted towards pronouns.
- Final fine-tuning loss combines
  - discriminative loss $\mathcal{L}_d$ aimed at correcting the mistakes
  - generative loss $\mathcal{L}_g$ needed to preserve the translation adequacy and fluency
  - weighted by $\lambda$

$$\mathcal{L}_{gd} = \lambda \mathcal{L}_g + (1 - \lambda)\mathcal{L}_d \tag{5}$$

# Training


A
An
Zebra
Cross-Entropy loss
+ve
He
ConGeDi loss
-ve
It
laughed
loudly
x1
x2
x3
y1
y2
y3
Er
lachte
laut
<s>
He
laughed

# Fine-tuning Data

Given a training corpus $\mathcal{D} = (\mathcal{S}, \mathcal{R})$, where $\mathcal{S}$ is the source and $\mathcal{R}$ is the target/reference text

- Translate $\mathcal{D}$ using a baseline model $\mathcal{M}$ to obtain source to target translations $\mathcal{T}_{\mathcal{M}}$.
- Align $\mathcal{T}_{\mathcal{M}}$ with reference $\mathcal{R}$.
- Find pronoun translations in $\mathcal{T}_{\mathcal{M}}$ that do not match reference $\mathcal{R}$.

# Fine-tuning Data

**Pronoun Translation Error**

*S:* Mir wurdediese Wohnungin Earls Court gezeigt, und **sie** hatte ...
*T:* I was shown this apartment in Earls Court , and **she** had ...
*Correct:* I was shown this apartment in Earls Court , and **it** had ...

- For each sentence with a mistranslated pronoun, extract the source sentences from $\mathcal{S}$.
- The corresponding source and reference sentences form the pronoun-targeted fine-tuning subset, referred to as $\mathcal{D}_{\text{prn}} = (\mathcal{S}', \mathcal{T}')$.

# Baseline Models

SEN2SEN: 6-layer base Transformer model; translates each sentence independently.

CONCAT: 6-layer base Transformer, translates sentence given one previous sentence as context.

- German-English (De-En) translation task
- 2.5M pairs of parallel training data (IWSLT, Europarl, Newscommentary)
- 300K pairs of fine-tuning subset data
- Tested on WMT14 test data and targeted pronoun testset [Jwalapuram et al., 2019]

# Baseline Results

| | | WMT14 | Pronoun Testset | | | |
|---|---|---|---|---|---|---|
| **Model** | **Train** | **BLEU** | **BLEU** | **P** | **R** | **F1** |
| SEN2SEN | $\mathcal{D}$ | 31.64 | 35.56 | 77.92 | 66.01 | 69.55 |
| CONCAT | $\mathcal{D}$ | **31.81** | **36.16** | **80.39** | **68.49** | **72.03** |

- Simple context model outperforms the sentence-level model.

# Training on Targeted Data

| Fine-tuning data for SEN2SEN | WMT14 BLEU | Pronoun Testset BLEU | P | R | F1 |
|---|---|---|---|---|---|
| $\mathcal{D}$ (baseline) | **31.64** | 35.56 | 77.92 | 66.01 | 69.55 |
| $\mathcal{D}_{\text{prn}}$ | 30.43 | 34.72 | 79.49 | 67.55 | 71.02 |
| $\mathcal{D} + \mathcal{D}_{\text{prn}}$ (shuffled) | 31.31 | 35.48 | 78.35 | 67.02 | 70.35 |
| $\mathcal{D} + \mathcal{D}_{\text{prn}}$ | 31.23 | 35.39 | 79.61 | 67.99 | **71.40** |
| $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | 31.56 | 35.57 | 79.25 | 68.01 | **71.35** |
| $\mathcal{D}$ (Increased training) | 31.53 | 35.60 | 78.14 | 66.15 | 69.77 |
| CONCAT | | | | | |
| $\mathcal{D}$ (baseline) | 31.81 | 36.16 | 80.39 | 68.49 | 72.03 |
| $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | 31.31 | 36.12 | 81.20 | 69.35 | 72.84 |

- BLEU scores drop with only fine-tuning data, but improvement in pronoun translations.
- Increased training does not improve pronoun translations $\rightarrow$ improvement from the targeted dataset.

# Max-margin Loss + Targeted Data

| Model | Fine-tuning data | WMT14 BLEU | Pronoun Testset BLEU | P | R | F1 |
|---|---|---|---|---|---|---|
| Baseline SEN2SEN | - | 31.64 | 35.56 | 77.92 | 66.01 | 69.55 |
| Baseline CONCAT | - | 31.81 | 36.16 | 80.39 | 68.49 | 72.03 |
| **All tokens** | | | | | | |
| SEN2SEN | $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | **32.14*** | 36.16 | 78.83 | 66.15 | 69.77* |
| SEN2SEN | $2\mathcal{D} + \mathcal{D}_{\text{rand}}$ | 31.86 | 35.88 | 78.07 | 66.00 | 69.65 |
| SEN2SEN | $\mathcal{D}$ | 31.75 | 36.34 | 78.27 | 66.36 | 69.91 |
| CONCAT | $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | 31.75 | **36.70** | **81.25** | **69.27** | **72.88** |
| **Only Pronouns** | | | | | | |
| SEN2SEN | $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | 31.81* | 36.43 | 78.62 | 66.82 | 70.37* |
| SEN2SEN | $2\mathcal{D} + \mathcal{D}_{\text{rand}}$ | 31.71 | 36.12 | 78.65 | 66.72 | 70.32 |
| SEN2SEN | $\mathcal{D}$ | 31.89 | 36.20 | 78.31 | 66.32 | 69.98 |
| CONCAT | $2\mathcal{D} + \mathcal{D}_{\text{prn}}$ | **31.99*** | **36.64** | **80.87** | **69.07** | **72.64** |

- * statistically significant; CONCAT best performing model.
- Fine-tuning with random subset does not lead to similar improvements.
- **All tokens** vs. **Only pronouns** ⇒ BLEU vs. F1.

# Log-likelihood loss + Targeted data

| Model | Fine-tuning data | WMT14 BLEU | Pronoun Testset BLEU | P | R | F1 |
|---|---|---|---|---|---|---|
| Baseline SEN2SEN | - | 31.64 | 35.56 | 77.92 | 66.01 | 69.55 |
| Baseline CONCAT | - | 31.81 | 36.16 | 80.39 | 68.49 | 72.03 |
| **All tokens** | | | | | | |
| SEN2SEN | $2\mathcal{D}+\mathcal{D}_{\text{prn}}$ | 31.83* | 36.50 | 79.18 | 67.16 | 70.78* |
| SEN2SEN | $2\mathcal{D}+\mathcal{D}_{\text{rand}}$ | 31.73 | 36.16 | 78.32 | 66.62 | 70.15 |
| SEN2SEN | $\mathcal{D}$ | 31.77 | 36.24 | 78.35 | 66.17 | 69.86 |
| CONCAT | $2\mathcal{D}+\mathcal{D}_{\text{prn}}$ | **31.85** | **36.61** | **80.91** | **68.91** | **72.57** |
| **Only Pronouns** | | | | | | |
| SEN2SEN | $2\mathcal{D}+\mathcal{D}_{\text{prn}}$ | 31.73 | 36.30 | 79.01 | 66.80 | 70.50* |
| SEN2SEN | $2\mathcal{D}+\mathcal{D}_{\text{rand}}$ | **32.05** | 36.43 | 78.35 | 66.25 | 69.87 |
| SEN2SEN | $\mathcal{D}$ | **32.05** | 35.81 | 78.58 | 66.52 | 70.22 |
| CONCAT | $2\mathcal{D}+\mathcal{D}_{\text{prn}}$ | **32.00*** | **36.57** | **80.89** | **68.66** | **72.39** |

- Comparable results for log-loss

# Examples

| WMT14 Testset | |
|---|---|
| Source | 14 stunden kämpften die ärzte um das überleben des opfers , jedoch vergeblich . |
| Reference | for 14 hours, doctors battled to save the life of the victim , ultimately in vain . |
| Baseline | 14 hours of doctors fought for the victim's survival , but in vain . |
| Our best model | the doctors fought 14 hours for the survival of the victim , but in vain . |
| **Pronoun Testset** | |
| Context | ... die die amerikanische flamme in die umnachtete welt bringe : lady liberty geht voran . |
| Source | sie soll die fackel der freiheit von den vereinigten staaten in den rest der welt tragen . |
| Context | ... taking the american flame out to the benighted world : **lady liberty** is stepping forward . |
| Reference | she is meant to be carrying the torch of liberty from the united states to the rest of the world . |
| Baseline | it is meant to carry the torch of freedom from the united states to the rest of the world . |
| Our best model | she is supposed to carry the torch of freedom from the united states to the rest of the world . |

# Comparison with Backtranslation

| | WMT14 | Pronoun Testset | | | |
|---|---|---|---|---|---|
| **Model** | **BLEU** | **BLEU** | **P** | **R** | **F1** |
| Baseline SEN2SEN | 31.64 | 35.56 | 77.92 | 66.01 | 69.55 |
| Backtranslation | **32.57** | **38.54** | 80.61 | 67.14 | 71.37 |
| Best fine-tuned SEN2SEN | 32.14 | 36.16 | 78.83 | 66.15 | 69.77 |
| Best fine-tuned CONCAT | 32.00 | 36.57 | **80.89** | **68.66** | **72.39** |

- Backtranslation (+76M) has best BLEU.
- But CONCAT outperforms for pronoun translations.

# IWSLT13 testset

| | SEN2SEN | | CONCAT | |
|---|---|---|---|---|
| **Model** | **BLEU** | **Prn. F1** | **BLEU** | **Prn. F1** |
| Baseline | 31.64 | 60.47 | 32.10 | 62.01 |
| Backtranslation | 30.30 | 58.02 | - | - |
| **All tokens** | | | | |
| Max-margin | 31.88 | 60.87 | **32.95** | 61.90 |
| Log-likelihood | 32.02 | 60.64 | 32.78 | **62.10** |
| **Only Pronouns** | | | | |
| Max-margin | 32.13 | 60.61 | **33.13** | **62.20** |
| Log-likelihood | 32.16 | 60.83 | 32.78 | 61.97 |

- Backtranslation fails to generalize.
- Fine-tuning improves results here as well.

# French-English

| Model | Fine-tuning loss | WMT14 BLEU | Pronoun Testset BLEU | P | R | F1 |
|---|---|---|---|---|---|---|
| Baseline Sen2Sen | - | 35.61 | 34.53 | 90.64 | 64.00 | 73.73 |
| Baseline Concat | - | 36.06 | 35.18 | 84.86 | 72.07 | 75.86 |
| **All tokens** | | | | | | |
| Sen2Sen | max-margin | **36.12***| 35.31 | 93.61 | 64.26 | 74.56* |
| Sen2Sen | log-likelihood | 36.04* | 35.39 | 96.39 | 66.95 | **77.38*** |
| Concat | max-margin | 35.98 | 35.41 | 85.93 | 72.48 | 76.48 |
| Concat | log-likelihood | 35.98 | 35.09 | 85.07 | 71.43 | 75.51 |
| **Only Pronouns** | | | | | | |
| Sen2Sen | max-margin | 36.05* | 35.34 | 93.48 | 67.24 | 76.96 |
| Sen2Sen | log-likelihood | 35.86* | 35.09 | 93.62 | 63.74 | 73.88 |
| Concat | max-margin | 35.97 | 35.26 | 85.71 | 71.97 | 76.07 |
| Concat | log-likelihood | **36.09** | 35.55 | 85.85 | 72.38 | **76.50** |

- 2.53M pairs training data, 500K pairs fine-tuning subset.
- Consistent improvements with fine-tuning.

# Czech-English

| Model | Fine-tuning loss | WMT14 BLEU | Pronoun Testset BLEU | P | R | F1 |
|---|---|---|---|---|---|---|
| Baseline Sen2Sen | - | 25.23 | 21.88 | 82.65 | 48.78 | 60.40 |
| Baseline Concat | - | 28.27 | 24.19 | 71.94 | 55.57 | 60.37 |
| **All tokens** | | | | | | |
| Sen2Sen | max-margin | **26.13***| 22.49 | 84.18 | 50.71 | **62.16*** |
| Sen2Sen | log-likehood | 26.08* | 22.65 | 83.02 | 49.02 | 60.53 |
| Concat | max-margin | 27.56 | 23.69 | 73.82 | 57.81 | 62.45* |
| Concat | log-likelihood | 27.50 | 23.85 | 74.43 | 58.17 | **62.89*** |
| **Only Pronouns** | | | | | | |
| Sen2Sen | max-margin | 26.10* | 22.56 | 83.02 | 49.96 | 61.03 |
| Sen2Sen | log-likelihood | 26.01* | 22.62 | 83.90 | 49.17 | 60.88 |
| Concat | max-margin | 27.48 | 23.76 | 74.20 | 57.72 | 62.53* |
| Concat | log-likelihood | 27.59 | 23.72 | 74.18 | 57.77 | 62.54 |

- 992K pairs training data, 100K pairs fine-tuning subset.
- Consistent improvements with fine-tuning.

# Analysis

- Max-margin and Log-likelihood loss perform comparably.

# Analysis

- Max-margin and Log-likelihood loss perform comparably.
- Sen2Sen model BLEU improvements but no pronoun translation improvements $\rightarrow$ lack of context.

# Analysis

- Max-margin and Log-likelihood loss perform comparably.
- SEN2SEN model BLEU improvements but no pronoun translation improvements $\rightarrow$ lack of context.
- General BLEU improvements $\rightarrow$ targeted data subset that model failed to learn from.

# Conclusions and Future Work

- Fine-tuning framework is generic; *e.g.*, can be applied to NEs.

# Conclusions and Future Work

- Fine-tuning framework is generic; *e.g.*, can be applied to NEs.
- Adapt to other directed generation tasks; *e.g.*, coherence/factual correctness in abstractive summarization or controlled text generation.

# Conclusions and Future Work

- Fine-tuning framework is generic; *e.g.*, can be applied to NEs.
- Adapt to other directed generation tasks; *e.g.*, coherence/factual correctness in abstractive summarization or controlled text generation.
- Address training issues from datasets; *e.g.*, correct biases (such as gender) in data or improve system robustness.

# Conclusions and Future Work

- Fine-tuning framework is generic; *e.g.*, can be applied to NEs.
- Adapt to other directed generation tasks; *e.g.*, coherence/factual correctness in abstractive summarization or controlled text generation.
- Address training issues from datasets; *e.g.*, correct biases (such as gender) in data or improve system robustness.
- End-to-end system that automatically filters targeted data.

# Thank you

**Thank you!**

Link to full paper (EMNLP 2020):

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Christian Hardmeier. Discourse in machine translation. urlhttps://ufal.mff.cuni.cz/mtm18/files/03-discourse-in-mt-christian-hardmeier.pdf,
2018.

Prathyusha Jwalapuram, Shafiq Joty, Irina Temnikova, and Preslav Nakov. Evaluating pronominal anaphora in machine translation: An evaluation measure and a test suite. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2964–2975, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1294. URL https://www.aclweb.org/anthology/D19-1294.

Prathyusha Jwalapuram, Barbara Rychalska, Shafiq R. Joty, and Dominika Basaj. Can your context-aware MT system pass the DiP benchmark tests? : Evaluation benchmarks for discourse phenomena in machine translation. *ArXiv*, abs/2004.14607, 2020.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has machine translation achieved human parity? a case for document-level evaluation. In *EMNLP*, 2018.

Rico Sennrich. Why the time is ripe for discourse in machine translation. http://homepages.inf.ed.ac.uk/rsennric/wnmt2018.pdf, 2018a.

Rico Sennrich. Why the time is ripe for discourse in machine translation. 2018b. URL http://homepages.inf.ed.ac.uk/rsennric/wnmt2018.pdf.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. *ArXiv*, abs/1511.06709, 2015.