Semantic, Stylistic & Other Data Divergences in Neural Machine Translation

Marine Carpuat
marine@cs.umd.edu
$e^* = \arg\max_e p(e|f; \theta)$

Nature of data matters more in Neural MT
This Talk: Data Divergences in NMT

Examine implicit equivalence assumptions about bitext and MT

Show that divergences from these assumptions occur and matter for neural MT
Translation Divergences

“the same information is conveyed in the source and target text, but the structure of the sentences are different”

[Dorr 1994]

en: Maria did not slap the green witch

es: Maria no daba una botefada a la bruja verde
Divergence (according to WordNet)

• S: (n) divergence, divergency
  (the act of moving away in different direction from a common point)

• S: (n) deviation, divergence, departure, difference
  (a variation that deviates from the standard or norm)
Semantic Divergences

Assumption:
source and target side in bitext have the same meaning

Our hypothesis:
bitext sides are not always semantically equivalent and this matters for NMT
Assumption: References can substitute for predicted translations during training

Our hypothesis: Modeling divergences between references and predictions improves NMT
Assumption:
MT output should preserve all properties of input

Our hypothesis:
We can tailor NMT style while preserving input meaning
Assumption: source and target side in bitext have the same meaning

Yet: parallel documents ≠ parallel segments “traduttore, traditore”: translators can alter source meaning
Divergence Examples

En: i don't know what i'm gonna do.
Fr: j'en sais rien.

En: you help me with zander and i helped you with joe.
Fr: tu m'as aidee avec zander, je t'ai aidee avec joe.

En: - has the sake chilled? - no, it's fine.
Fr: - c'est assez chaud?
How Frequent are Divergent Examples? A Crowdsourcing Experiment

<table>
<thead>
<tr>
<th></th>
<th>Equivalent</th>
<th>Divergent</th>
</tr>
</thead>
<tbody>
<tr>
<td>CommonCrawl</td>
<td>56</td>
<td>38</td>
</tr>
<tr>
<td>OpenSubs</td>
<td>44</td>
<td>56</td>
</tr>
</tbody>
</table>
Approach: cross-lingual semantic similarity model

Predict semantic similarity with the “Very Deep Pairwise Similarity Model” [He & Lin 2016]

Initialize with bilingual word embeddings
Approach: Generate (Noisy) Synthetic Training Examples

Sentence aligned bitext

“Equivalent” examples

Divergent examples

[Munteanu & Marcu 2006]
Intrinsic Evaluation: ConvNet trained on synthetic examples performs best.

F-score for divergent pair detection

- **Our approach**
- **Parallel vs. non-parallel**
- **Bilingual embeddings**
- **MT scores**

OpenSubtitles

- Our approach: 78
- Parallel vs. non-parallel: 65
- Bilingual embeddings: 55
- MT scores: 50

CommonCrawl

- Our approach: 80
- Parallel vs. non-parallel: 60
- Bilingual embeddings: 50
- MT scores: 40
**Intrinsic Evaluation:** ConvNet trained on synthetic examples performs best

F-score for divergent pair detection

<table>
<thead>
<tr>
<th></th>
<th>OpenSubtitles</th>
<th>CommonCrawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our approach</td>
<td>80</td>
<td>75</td>
</tr>
<tr>
<td>Parallel vs. non-parallel</td>
<td>65</td>
<td>60</td>
</tr>
<tr>
<td>Bilingual embeddings</td>
<td>60</td>
<td>55</td>
</tr>
<tr>
<td>MT scores</td>
<td>55</td>
<td>45</td>
</tr>
</tbody>
</table>

Worse F-score when using same synthetic examples with non-neural classifier [Munteanu & Marcu 2006]
Intrinsic Evaluation: ConvNet trained on synthetic examples performs best

F-score for divergent pair detection

Worse F-score when using only bilingual word embeddings
Intrinsic Evaluation: ConvNet trained on synthetic examples performs best

Worse F-score when using NMT scores
Intrinsic Evaluation: ConvNet trained on synthetic examples performs best.

F-score for divergent pair detection

- Our approach
- Parallel vs. non-parallel
- Bilingual embeddings
- MT scores
- Supervised cross-lingual entailment

Worse F-score when using a supervised cross-lingual entailment classifier [Carpuat et al. 2017]
Do semantic divergences impact MT?

English > French tasks from IWSLT

<table>
<thead>
<tr>
<th>Training Set</th>
<th>OpenSubtitles</th>
<th>33.5M segment pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>In domain Test Set</td>
<td>MSLT: Microsoft Speech Language Translation (IWSLT16)</td>
<td>5000 segment pairs</td>
</tr>
<tr>
<td>Out of domain Test Set</td>
<td>TED talks (IWLST15)</td>
<td>1300 segment pairs</td>
</tr>
</tbody>
</table>
Downsampling via cross-lingual semantic similarity helps NMT training

[Vyas, Niu & Carpuat, NAACL 2018]
Downsampling via cross-lingual semantic similarity doesn’t hurt BLEU at test time

<table>
<thead>
<tr>
<th>Model</th>
<th>MSLT BLEU</th>
<th>TED BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Avg.</td>
<td>Ensemble</td>
</tr>
<tr>
<td>Random</td>
<td>43.49</td>
<td>45.64</td>
</tr>
<tr>
<td>Parallel</td>
<td>40.65</td>
<td>42.12</td>
</tr>
<tr>
<td>Entailment</td>
<td>39.64</td>
<td>41.86</td>
</tr>
<tr>
<td>Semantic Sim.</td>
<td><strong>45.53</strong></td>
<td><strong>47.23</strong>*</td>
</tr>
<tr>
<td>All</td>
<td>44.64</td>
<td>46.26</td>
</tr>
</tbody>
</table>

[Vyas, Niu & Carpuat, NAACL 2018]
Beyond filtering divergent examples

Fixing divergences by deleting extra info
[Pham et al. EMNLP 2018]

Curriculum learning with noise & domain criteria
[Wang et al. NAACL 2019]
A Probabilistic Curriculum for Sampling Training Data

[Zhang et al. NAACL 2019]
Preview: Divergence-based Curriculum improves BLEU

[BLEU on fr-en MSLT]

- All data
- Rand half
- Rand half + length curriculum
- Rand half + divergence curriculum

[Richburg & Carpuat, unpublished]
All bitexts contain semantically divergent examples

We can detect them with deep semantic similarity models trained on synthetic examples

Neural machine translation is sensitive to such divergences

Filtering out divergent examples helps

Open questions

What kind of divergences? How do they differ from noise?
Curriculum Learning for Domain Adaptation in Neural Machine Translation. Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat and Kevin Duh. NAACL 2019

Identifying Semantic Divergences in Parallel Text without Annotations. Yogarshi Vyas, Xing Niu and Marine Carpuat. NAACL 2018


github.com/yogarshi/SemDiverge
github.com/kevinduh/sockeye-recipes
Assumption:
References can substitute for predicted translations during training

Our hypothesis:
Modeling divergences between references and predictions improves NMT
Exposure Bias: Gap Between Training and Inference

Model Translation

Inference

Maximum Likelihood Training

Reference

\[
P(y|x) = \prod_{t=1}^{T} p(y_t|y_{<t}, x)
\]

\[
\text{Loss} = \sum_{t=1}^{T} \log p(y_t|y_{<t}, x)
\]
How to Address Exposure Bias?

Expose models to their own predictions during training

But how to compute the loss when the partial translation diverges from the reference?

Our method: learn to align the reference words with partial translations during training.
Existing Methods

Search-based Methods
Computationally expensive

Reinforcement Learning with Sentence-Level Reward
[Ranzato et al., 2015, Bahdanau et al., 2016]
Inefficient and unstable

Scheduled Sampling
[Venkatraman et al. 2015, Bengio et al. 2015, Goyal et al. 2017]
Simple and efficient, but ...
Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>  

\[
P = \text{choose randomly}
\]

[Bengio et al., NeurIPS 2015]
Existing Method: Scheduled Sampling

Reference: <s> We made dinner </s>

\[ P \text{ = choose randomly} \]

[Bengio et al., NeurIPS 2015]
Existing Method: Scheduled Sampling

Reference: `<s> We made dinner </s>`

Incorrect synthetic reference: “We will dinner”

\[
J = \log p(\text{“dinner”} | \text{“<s> We will”, source})
\]

[Bengio et al., NeurIPS 2015]
Our Solution: Learning How To Align Reference with Partial Translations

Reference: <s> We made dinner </s>

\[ a_1 \log p(\text{“dinner”} | \text{“<s>”}, \text{source}) + a_2 \log p(\text{“dinner”} | \text{“<s> We”}, \text{source}) + \]
\[ a_3 \log p(\text{“dinner”} | \text{“<s> We will”}, \text{source}) + a_4 \log p(\text{“dinner”} | \text{“<s> We will make”}, \text{source}) \]
Our Solution: Learning How To Align Reference with Partial Translations

Reference: <s> We made dinner </s>

We make dinner

Soft Alignment

\[ a_i \propto \exp(\text{Embed}_{\text{dinner}} \cdot h_i) \]

\[ a_1 \log p(\text{“dinner”} \mid \text{“<s>”}, \text{source}) + a_2 \log p(\text{“dinner”} \mid \text{“<s> We”}, \text{source}) + \\
 a_3 \log p(\text{“dinner”} \mid \text{“<s> We will”}, \text{source}) + a_4 \log p(\text{“dinner”} \mid \text{“<s> We will make”}, \text{source}) \]
Our Solution: Learning How To Align Reference with Partial Translations

Reference: <s> We made dinner </s>

We make dinner

Soft Alignment

\[ a_i \propto \exp(\text{Embed}_{\text{dinner}} \cdot h_i) \]

\[
\begin{align*}
 a_1 \log p(\text{“dinner”} | \text{“<s>”, source}) + & a_2 \log p(\text{“dinner”} | \text{“<s> We”, source}) + \\
 a_3 \log p(\text{“dinner”} | \text{“<s> We will”, source}) + & a_4 \log p(\text{“dinner”} | \text{“<s> We will make”, source})
\end{align*}
\]
Training Objective

Ours:

Soft alignment between $y_t$ and $\tilde{y}_j$

\[
J_{SA} = \sum_{(x, y) \in D} \sum_{t=1}^{T} \sum_{j=1}^{T'} \log a_{tj} p(y_t | \tilde{y}_j, x)
\]

Scheduled Sampling:

Hard alignment by time index $t$

\[
J_{SS} = \sum_{(x, y) \in D} \sum_{t=1}^{T} \log p(y_t | \tilde{y}_t, x)
\]
Training Objective

**Ours:**
Soft alignment between $y_t$ and $\tilde{y}_j$

$$J_{SA} = \sum_{(x,y)\in D} \sum_{t=1}^{T} \sum_{j=1}^{T'} \log a_{tj} p(y_t | \tilde{y}_j, x)$$

**Scheduled Sampling:**
Hard alignment by time index $t$

$$J_{SS} = \sum_{(x,y)\in D} \sum_{t=1}^{T} \log p(y_t | \tilde{y}_t, x)$$
Training Objective

Ours:

Soft alignment between $y_t$ and $\tilde{y}_j$

$$J_{SA} = \sum_{(x,y) \in D} \sum_{t=1}^{T} \sum_{j=1}^{T'} \log \sum_{j=1}^{T'} a_{tj} p(y_t | \tilde{y}_j, x)$$

Combined with maximum likelihood:

$$J = J_{SA} + J_{ML}$$

Scheduled Sampling:

Hard alignment by time index $t$

$$J_{SS} = \sum_{(x,y) \in D} \sum_{t=1}^{T} \log p(y_t | \tilde{y}_t, x)$$
Experiments

Data
- IWSLT14 de-en
- IWSLT15 vi-en

<table>
<thead>
<tr>
<th>Task</th>
<th>sentences (K)</th>
<th>vocab (K)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>train</td>
<td>dev</td>
</tr>
<tr>
<td>de-en</td>
<td>153.3</td>
<td>7.0</td>
</tr>
<tr>
<td>vi-en</td>
<td>121.3</td>
<td>1.5</td>
</tr>
</tbody>
</table>

Model
- Bi-LSTM encoder, LSTM decoder, multilayer perceptron attention
- Differentiable sampling with Straight-Through Gumbel Softmax
- Based on AWS sockeye
Our Method Outperforms Maximum Likelihood and Scheduled Sampling

![Bar Chart](chart.png)

- **de-en**: Baseline, Scheduled Sampling, Differentiable Scheduled Sampling, Our Method
- **en-de**: Baseline, Scheduled Sampling, Differentiable Scheduled Sampling, Our Method
- **vi-en**: Baseline, Scheduled Sampling, Differentiable Scheduled Sampling, Our Method

**BLEU Scores**

- **en-de**: Baseline: 22, Scheduled Sampling: 22, Differentiable Scheduled Sampling: 22, Our Method: 22
- **vi-en**: Baseline: 22, Scheduled Sampling: 24, Differentiable Scheduled Sampling: 23, Our Method: 24
Our Method Needs No Annealing

Scheduled sampling: BLEU drops when used without annealing!
A new training objective

1. Generate translation prefixes via differentiable sampling
2. Learn to align the reference words with sampled prefixes

Better BLEU than the maximum likelihood and scheduled sampling (de-en, en-de, vi-en)

Simple to train, no annealing schedule required

Reference Divergences
Flexible Reference Word Order for Neural Machine Translation
Weijia Xu, Xing Niu, Marine Carpuat.
NAACL 2019

github.com/Izecson/saml-nmt
Assumption:
MT output should preserve all properties of input

Our hypothesis:
We can tailor NMT style while preserving input meaning
Style Matters for Translation

Is it more "Hey Dude" or "Dear Sir"?
Improve translation accuracy by telling us the tone of the content.

- Informal
- Friendly
- Business
- Formal
- Other

Possible instructions:
- Voice
- Links
- Purpose & Audience
- Casual, romantic, funny, serious etc.
- To your website, screen shots or other docs.
- This is going to my most important client etc.

www.gengo.com
Does Style Matter for Machine Translation?

We focus on **formality**

Goal: Can we produce MT output with varying formality?

Prior work: other aspects of style
- conversational language [Lewis et al. 2015]
- politeness (du vs. Sie) [Sennrich et al. 2016]
- personalization (gender) [Rabinovich et al. 2017]
Formality-Sensitive Machine Translation (FSMT)

Comment ça va?  Source \( f \)  \( \rightarrow \)  FSMT \( (\theta) \)  \( \rightarrow \)  Translation-1 \( (e_1) \)  How are you doing?

Desired formality level \( \ell \)  Translation-2 \( (e_2) \)  What's up?

How to train?

\( f \)  \( \ell_1 \)  \( e_1 \)  Ideal training data doesn’t occur naturally!

\( f \)  \( \ell_2 \)  \( e_2 \)

[Niu, Martindale & Carpuat, EMNLP 2017]
Formality in MT Corpora

delegates are kindly requested to bring their copies of documents to meetings.

in these centers, the children were fed, medically treated and rehabilitated on both a physical and mental level.

there can be no turning back the clock.

I just wanted to introduce myself.

-yeah, bro, up top.
Formality Transfer (FT)

Given a large parallel formal-informal corpus (e.g., Grammarly’s Yahoo Answers Formality Corpus) these are sequence-to-sequence tasks

[Rao and Tetreault, 2018]
Formality Sensitive MT as Multitask Formality Transfer + MT

How are you doing?  What's up?
Comment ça va?

To formal or informal?

Bi-FT + FSMT

Formal-Target or Informal-Target

EN How are you doing?
EN What's up?
Multitask Formality Transfer + MT

Model: shared encoder, shared decoder as in multilingual NMT [Johnson et al. 2017]

Training objective:

\[ \mathcal{L}_{MT} + \mathcal{L}_{FT} \]

\[ \mathcal{L}_{MT} = \sum_{(X,Y)} \log P(Y|X; \theta) \quad \text{MT pairs} \]

\[ \mathcal{L}_{FT} = \sum_{(Y_{\ell}, Y_{\ell})} \log P(Y_{\ell}|Y_{\ell}, \ell; \theta) \quad \text{FT pairs} \]
Multitask Formality Transfer + MT Training Data

FT

Side constraint [Sennrich et al. 2016]

50k sentence pairs from Grammarly’s Yahoo Answers Formality Corpus
Multitask Formality Transfer + MT
Training Data

<table>
<thead>
<tr>
<th>FT</th>
<th>Informal-EN</th>
<th>Formal-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;I&gt;</td>
<td>Formal-EN</td>
<td>Informal-EN</td>
</tr>
<tr>
<td>&lt;I&gt;</td>
<td>FR</td>
<td>Formal-EN</td>
</tr>
<tr>
<td>&lt;I&gt;</td>
<td>FR</td>
<td>Informal-EN</td>
</tr>
</tbody>
</table>

Data selected from OpenSubtitles

[Moore & Lewis, 2010]
Evaluation – Formality Transfer

Test set
Grammarly’s Yahoo Answers Formality Corpus
1K sent pairs per direction
4 references
Automatic metric: BLEU

[Rao & Tetreault, 2018]
Multitask Model

Model
1 layer LSTM encoder decoder
MLP attention

Shared 30k BPE vocab
Tied src emb, trg emb, output layer
512 embeddings, hidden layers

Toolkit: AWS Sockeye
## Results – Formality Transfer (BLEU)

<table>
<thead>
<tr>
<th>Model</th>
<th>Informal$\rightarrow$Formal E&amp;M</th>
<th>Informal$\rightarrow$Formal F&amp;R</th>
<th>Formal$\rightarrow$Informal E&amp;M</th>
<th>Formal$\rightarrow$Informal F&amp;R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Source</td>
<td>49.09</td>
<td>51.03</td>
<td>29.85</td>
<td>29.85</td>
</tr>
<tr>
<td>PBMT (Rao and Tetreault, 2018)</td>
<td>68.22</td>
<td>72.94</td>
<td>33.54</td>
<td>32.64</td>
</tr>
<tr>
<td>NMT Baseline (Rao and Tetreault, 2018)</td>
<td>58.80</td>
<td>68.28</td>
<td>30.57</td>
<td>36.71</td>
</tr>
<tr>
<td>NMT Combined (Rao and Tetreault, 2018)</td>
<td>68.41</td>
<td>74.22</td>
<td>33.56</td>
<td>35.03</td>
</tr>
<tr>
<td>NMT Baseline</td>
<td>65.34</td>
<td>71.28</td>
<td>32.36</td>
<td>36.23</td>
</tr>
<tr>
<td>Bi-directional FT</td>
<td>66.30</td>
<td>71.97</td>
<td>34.00</td>
<td>36.33</td>
</tr>
<tr>
<td>+ training on E&amp;M + F&amp;R</td>
<td>69.20</td>
<td>73.52</td>
<td>35.44</td>
<td>37.72</td>
</tr>
<tr>
<td>+ ensemble decoding ($\times$4)</td>
<td>71.36</td>
<td>74.49</td>
<td>36.18</td>
<td>38.34</td>
</tr>
<tr>
<td>+ multi-task learning</td>
<td><strong>72.13</strong></td>
<td><strong>75.37</strong></td>
<td><strong>38.04</strong></td>
<td><strong>39.09</strong></td>
</tr>
</tbody>
</table>
## Results – Formality Transfer (BLEU)

<table>
<thead>
<tr>
<th>Model</th>
<th>Informal→Formal</th>
<th>Formal→Informal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E&amp;M</td>
<td>F&amp;R</td>
</tr>
<tr>
<td>Original Source</td>
<td>49.09</td>
<td>51.03</td>
</tr>
<tr>
<td>PBMT (Rao and Tetreault, 2018)</td>
<td>68.22</td>
<td>72.94</td>
</tr>
<tr>
<td>NMT Baseline (Rao and Tetreault, 2018)</td>
<td>58.80</td>
<td>68.28</td>
</tr>
<tr>
<td>NMT Combined (Rao and Tetreault, 2018)</td>
<td>68.41</td>
<td>74.22</td>
</tr>
<tr>
<td>NMT Baseline</td>
<td>65.34</td>
<td>71.28</td>
</tr>
<tr>
<td>Bi-directional FT</td>
<td>66.30</td>
<td>71.97</td>
</tr>
<tr>
<td>+ training on E&amp;M + F&amp;R</td>
<td>69.20</td>
<td>73.52</td>
</tr>
<tr>
<td>+ ensemble decoding (×4)</td>
<td>71.36</td>
<td>74.49</td>
</tr>
<tr>
<td>+ multi-task learning</td>
<td><strong>72.13</strong></td>
<td><strong>75.37</strong></td>
</tr>
</tbody>
</table>
## Results – Formality Transfer

### Human Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Formality Difference I-F</th>
<th>Formality Difference F-I</th>
<th>Meaning Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rao&amp;Tetreault baseline</td>
<td>0.54</td>
<td>0.45</td>
<td>2.94</td>
</tr>
<tr>
<td>Multitask FT+MT</td>
<td>0.59</td>
<td>0.64</td>
<td>2.92</td>
</tr>
</tbody>
</table>

- **Range** for Formality Difference I-F = [0,2]
- **Range** for Formality Difference F-I = [0,2]
- **Range** for Meaning Preservation = [0,3]

300 samples per model
3 judgments per sample
Protocol based on Rao & Tetreault
Multitask Formality Transfer + MT
Training Data

Data selected [Moore & Lewis, 2010] from OpenSubtitles

Selected bilingual data is similar to GYAFC (FT😊)
GYAFC ≠ domain of translation data (FSMT☹️)
Multitask Formality Transfer + MT Training Data Variants

MultiTask Select:

- <F> Informal-EN   Formal-EN
- <I> Formal-EN    Informal-EN
- <F> FR           FR
- <I> FR           Informal-EN

MultiTask Rand:

- <F> Informal-EN   Formal-EN
- <I> Formal-EN    Informal-EN
- FR             EN

Side constraint:

- <F> FR           Formal-EN
- <I> FR           Formal-EN
Evaluation – Formality Sensitive MT

French-English

Training Data
50K pairs from GYAFC
2.5M pairs selected from OpenSubtitles 2016

Test
Microsoft Spoken Language Corpus
1 reference of unknown formality
Formality Sensitive MT
BLEU Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>FR to formal EN</th>
<th>FR to informal EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTask Select</td>
<td>25.02</td>
<td>25.20</td>
</tr>
<tr>
<td>MultiTask Rand</td>
<td>25.24</td>
<td>25.14</td>
</tr>
<tr>
<td>Side constraint</td>
<td>27.15</td>
<td>26.70</td>
</tr>
<tr>
<td>Phrase-based MT + formality reranking</td>
<td>29.12</td>
<td>29.02</td>
</tr>
</tbody>
</table>

[Niu & Carpuat 2017]
Formality Transfer MT
Human Evaluation

<table>
<thead>
<tr>
<th>Model</th>
<th>Formality Difference</th>
<th>Meaning Preservation</th>
</tr>
</thead>
<tbody>
<tr>
<td>MultiTask Rand</td>
<td>0.35</td>
<td>2.95</td>
</tr>
<tr>
<td>Side constraint</td>
<td>0.32</td>
<td>2.90</td>
</tr>
<tr>
<td>Phrase-based MT + formality reranking</td>
<td>0.05</td>
<td>2.97</td>
</tr>
<tr>
<td>[Niu &amp; Carpuat 2017]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

300 samples per model
3 judgments per sample
Protocol based on Rao & Tetreault
<table>
<thead>
<tr>
<th>Analysis: Multitask model makes more formality changes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
</tr>
<tr>
<td>Refrain from the commentary and respond to the question, Chief Toohey.</td>
</tr>
<tr>
<td><strong>Formal</strong></td>
</tr>
<tr>
<td><strong>MultiTask</strong></td>
</tr>
<tr>
<td>You need to be quiet and answer the question, Chief Toohey.</td>
</tr>
<tr>
<td><strong>Side constraint</strong></td>
</tr>
<tr>
<td>Please refrain from any comment and answer the question, Chief Toohey.</td>
</tr>
<tr>
<td><strong>PBMT</strong></td>
</tr>
<tr>
<td>Please refrain from comment and just answer the question, the Tooheys’s boss.</td>
</tr>
<tr>
<td><strong>Informal</strong></td>
</tr>
<tr>
<td><strong>MultiTask</strong></td>
</tr>
<tr>
<td>Shut up and answer the question, Chief Toohey.</td>
</tr>
<tr>
<td><strong>Side constraint</strong></td>
</tr>
<tr>
<td>Please refrain from comment and answer the question, chief Toohey.</td>
</tr>
<tr>
<td><strong>PBMT</strong></td>
</tr>
<tr>
<td>Please refrain from comment and answer my question, Tooheys’s boss.</td>
</tr>
</tbody>
</table>
Analysis: Multitask model introduces more meaning errors

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
<th>Formal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Try to file any additional motions as soon as you can.</td>
<td>Try to introduce the <strong>sharks</strong> as soon as you can.</td>
</tr>
<tr>
<td>MultiTask</td>
<td>You should try to introduce the <strong>sharks</strong> as soon as you can.</td>
<td></td>
</tr>
<tr>
<td>Side constraint</td>
<td>Try to present additional requests as soon as you can.</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Try to introduce any additional requests as soon as you can.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Informal</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Try to introduce <strong>sharks</strong> as soon as you can.</td>
<td></td>
</tr>
<tr>
<td>MultiTask</td>
<td>Try to introduce extra requests as soon as you can.</td>
<td></td>
</tr>
<tr>
<td>Side constraint</td>
<td>Try to introduce extra requests as soon as you can.</td>
<td></td>
</tr>
<tr>
<td>PBMT</td>
<td>Try to introduce any additional requests as soon as you can.</td>
<td></td>
</tr>
</tbody>
</table>
Multi Task Loss so far:

\[ \mathcal{L}_{MT} + \mathcal{L}_{FT} \]

\[ \mathcal{L}_{MT} = \sum_{(X, Y) \text{ MT pairs}} \log P(Y \mid X; \theta) \]

\[ \mathcal{L}_{FT} = \sum_{(Y_{\ell}, Y_{\bar{\ell}}) \text{ FT pairs}} \log P(Y_{\ell} \mid Y_{\bar{\ell}}, \ell; \theta) \]

Hypothesis:
Training with complete FSMT examples can improve formality control while preserving meaning.

\[ (X, \ell, Y_{\ell}) \text{ FSMT triplets} \]
Improving Multitask Training with Synthetic Supervision

1. Online Style Inference (OSI): predict formality of MT samples on the fly

2. Replace MT loss by OSI loss

\[ \mathcal{L}_{OSI} = \sum_{(X, \ell_Y, Y)} \log P(Y | X, \ell_Y; \theta) \]

\[ \mathcal{L} = \mathcal{L}_{FT} + \mathcal{L}_{OSI} \]
Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source
Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source

$\ell_F = <F>$

FSMT

Formal ($Y_F$) EN

How are you doing?

FR Source ($X$)

Comment ça va?

$\ell_I = <I>$

FSMT

Informal ($Y_I$) EN

What's up?
Synthetic Supervision: Predict formality of MT samples on the fly

By comparing reference to formal vs. informal translations of source

\[(X, Y) \leftrightarrow (X, \ell_F, Y) \text{ if } CED(Y_I, Y_F) = H_Y(Y_I) - H_Y(Y_F) > \tau\]
Formality is marked more strongly in Online Source Inference outputs than in MultiTask outputs.
Human Evaluation: Meaning Preservation

Online Style Inference preserves the meaning of references better than Multitask
Our new multitask formality transfer + MT model

Improves English formality transfer

Can produce distinct formal/informal translations of same input

Introduces more formality rewrites while preserving meaning, esp. with synthetic supervision


github.com/xingniu/multitask-ft-fsmt
Semantic Divergences

Reference Divergences

Style Divergences
From Parallel Text to Machine Translation

Modeling divergences between reference & predictions improves NMT

\[
\{(f_1, e_1), (f_2, e_2), \ldots (f_N, e_N)\}
\]

\[
e^* = \arg\max_e p(e|f; \theta)
\]

Detecting semantic divergence helps NMT training

NMT can tailor output style while preserving input meaning
From Parallel Text to Machine Translation

How can we design training to best exploit available data?

What properties of training samples matter for training?

Can we recast MT as a language generation task?

\[
\{(f_1, e_1), (f_2, e_2), \ldots (f_N, e_N)\}
\]

\[
e^* = \arg \max_e p(e|f; \theta)
\]
Semantic, Stylistic & Other Data Divergences in Neural Machine Translation

Marine Carpuat
marine@cs.umd.edu
<table>
<thead>
<tr>
<th>Type</th>
<th>Informal translation</th>
<th>Formal translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Filler</td>
<td>And I think his wife has family there.</td>
<td>I think his wife has family there.</td>
</tr>
<tr>
<td>Completeness</td>
<td>▼</td>
<td></td>
</tr>
<tr>
<td>Quotation</td>
<td>The gas tax is simply not sustainable, said Lee.</td>
<td>“The gas tax is simply not sustainable,” said Lee.</td>
</tr>
<tr>
<td>Yes-No</td>
<td>You like shopping?</td>
<td>Do you like shopping?</td>
</tr>
<tr>
<td>Subject</td>
<td>Sorry it’s my fault.</td>
<td>I’m sorry it’s my fault.</td>
</tr>
<tr>
<td>Article</td>
<td>Cookies where I work.</td>
<td>The cookies where I work.</td>
</tr>
<tr>
<td>Relativizer</td>
<td>Other stores you can’t buy.</td>
<td>The other stores where you can’t buy.</td>
</tr>
<tr>
<td>Paraphrasing</td>
<td>▼</td>
<td></td>
</tr>
<tr>
<td>Contraction</td>
<td>I think he’d like that, but we’ll see.</td>
<td>I think he would like that, but we will see.</td>
</tr>
<tr>
<td>Possessive</td>
<td>Fay’s innovation perpetuated over the years.</td>
<td>The innovation of Fay has perpetuated over the years.</td>
</tr>
<tr>
<td>Adverb</td>
<td>I told you already.</td>
<td>I already told you.</td>
</tr>
<tr>
<td>Idiom</td>
<td>Hi, how’s it going?</td>
<td>Hi, how are you?</td>
</tr>
<tr>
<td>Slang</td>
<td>You gotta let him digest.</td>
<td>You have to let him digest.</td>
</tr>
<tr>
<td>Word-1</td>
<td>Actually my dad’s some kind of technician so he understands, but my mom’s very old.</td>
<td>In fact, my father is some kind of technician so he understands, but my mother is very old.</td>
</tr>
<tr>
<td>Word-2</td>
<td>Maybe a little more in some areas.</td>
<td>Perhaps a little more in certain areas.</td>
</tr>
<tr>
<td>Word-3</td>
<td>It’s really necessary for our nation.</td>
<td>This is essential for our nation.</td>
</tr>
<tr>
<td>Phrase-1</td>
<td>Yeah, me neither.</td>
<td>Yeah, neither do I.</td>
</tr>
<tr>
<td>Phrase-2</td>
<td>I think he’s moving to California now.</td>
<td>I think he is moving to California at the moment.</td>
</tr>
<tr>
<td>Phrase-3</td>
<td>It could be a Midwest thing.</td>
<td>This could be one thing from the Midwest.</td>
</tr>
</tbody>
</table>
Intrinsic Evaluation: ConvNet trained on synthetic examples performs best

<table>
<thead>
<tr>
<th>Divergence Detection Approach</th>
<th>OpenSubtitles</th>
<th>Common Crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+P</td>
<td>+R</td>
</tr>
<tr>
<td>Sentence Embeddings</td>
<td>65</td>
<td>60</td>
</tr>
<tr>
<td>MT Scores (1 epoch)</td>
<td>67</td>
<td>53</td>
</tr>
<tr>
<td>Non-entailment</td>
<td>58</td>
<td>78</td>
</tr>
<tr>
<td>Non-parallel</td>
<td>70</td>
<td>83</td>
</tr>
<tr>
<td>Semantic Dissimilarity</td>
<td><strong>76</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>