Digital Data Flows Masterclass on Machine Learning and Speech

9 December 2020
# Future of Privacy Forum

## The Supporters

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>150+</td>
<td>Companies</td>
</tr>
<tr>
<td>25+</td>
<td>Leading Academics</td>
</tr>
<tr>
<td>15+</td>
<td>Advocates and Civil Society</td>
</tr>
<tr>
<td>5</td>
<td>Foundations</td>
</tr>
</tbody>
</table>

## The Mission

- Bridging the policymaker-industry-academic gap in privacy policy
- Developing privacy protections, ethical norms, & responsible business practices

## The Workstreams

- AI & Ethics
- Student Data
- Apps & Ad Tech
- Mobility & Location
- Privacy Enhancing Tech
- Smart Communities
Classes

1. Artificial Intelligence and Machine Learning  
   Date: 25 Oct 2018  
   Location: Brussels

2. Location Data: GPS, Wi-Fi, & Spatial Analytics  
   Date: 27 Nov 2018  
   Location: Brussels

3. De-Identification, Differential Privacy, and Homomorphic Encryption  
   Date: 30 Jan 2019  
   Location: Brussels

4. Online Advertising, Data Flows, Behavioral Targeting, and Cross-Device Tracking  
   Date: 1 May 2019  
   Location: Wash. DC

5. Mobile Apps: Operating Systems, Software Development Kits (SDKs), and User Controls  
   Date: 25 Jul 2019  
   Location: Virtual

6. Facial Recognition and Biometric Data  
   Date: 27 Feb 2020  
   Location: Wash. DC

7. Connected Cars and Autonomous Vehicles  
   Date: 25 Jun 2020  
   Location: Virtual

8. Blockchain Technologies  
   Date: 29 Oct 2020  
   Location: Virtual

9. Machine Learning and Speech  
   Date: 9 Dec 2020  
   Location: Virtual

Access recordings and materials for all previous classes at [www.fpf.org/classes](http://www.fpf.org/classes)
Speakers

**Professor Marine Carpuat**, Associate Professor in Computer Science at the University of Maryland

**Dr. Prem Natarajan**, VP, Alexa AI-NU
Neural Machine Translation for Communication Across Language Barriers

Marine Carpuat
marine@cs.umd.edu

Future of Privacy Forum
Digital Data Flows Masterclass 2020/12/09
7,111 languages spoken today

23 account for 50% of population

43% endangered

Source: Ethnologue
The Chinese capital, with its surprisingly high-speed Internet, sophisticated technology such as face-recognition software, has invested heavily in artificial intelligence and has unrivaled international energy, and is one of the most exciting cities for exploration-minded foreigners.
An English sentence $e$ is translated into the French sentence

$$f^* = \arg\max_f p(f|e; \theta)$$

“Learning” refers to finding good values for the model parameters

$$\theta^* = \arg\max_\theta \sum_i \log p(f_i | e_i; \theta)$$
Translation as Deep Learning

\[ p(f \mid e; \theta) = \prod_{t=1}^{T} p(f_t \mid f_{<t}, e; \theta) \]
requires millions of translation examples not available for many languages!

raises fundamental machine learning challenges

intractably large output space, infinitely many correct outputs...

makes errors that have real world impact

yet models are opaque, and developed independency from use cases
sequence-to-sequence models are general purpose tools to generate language used for dialog, text summarization, question answering, style transfer, etc.

Beyond Translation

share architecture design with language models and word representation models (aka word embeddings)
How can we train neural models with limited translation data?
Learning Paradigms to Exploit Diverse Data Sources

**Supervised learning** from parallel samples (translations) 🇺🇳TED 🇪🇺

**Unsupervised learning** from unpaired monolingual samples 🌏 Facebook Twitter 🐦
Supervised Learning

requires parallel samples: input sequence paired with a correct output sequence

\[ \theta^* = \arg\max_{\theta} \sum_i \log p(f_i | e_i; \theta) \]
Learn a **language model** per language which predicts the next word given previous words in a sentence.

### Unsupervised Learning

**fr**

\[ \theta_f^* = \arg\max_{\theta} \sum_f \log p(f_i | f_{<i}; \theta) \]

**en**

\[ \theta_e^* = \arg\max_{\theta} \sum_e \log p(e_i | e_{<i}; \theta) \]
Learn a **language model** per language which predicts the next word given previous words in a sentence

+ dictionaries to capture cross-lingual correspondences

Unsupervised Learning
learn from (few) paired + (many) unpaired samples

common approach: back translation
use auxiliary Machine Translation (MT) to translate target into source

simple and effective [Sennrich et al. 2016]
Alternative: reconstruction

More complex than backtranslation, but see Niu et al. [2019] for efficient variants

Provides a theoretical framework to analyze semi-supervised learning [Xu et al., 2020]
(Semi-)supervised learning is key to recent MT successes.
Better algorithms are still needed to scale to all languages.
How can we make neural models more responsive to user needs?
Neural MT still fails, sometimes catastrophically
Fluently inadequate translations

One type of catastrophic failure

An **adequate** translation accurately conveys the meaning of the source

A **fluent** translation is well formed in the output language

[Martindale et al. MT Summit 2019]
Fluently inadequate translations

<table>
<thead>
<tr>
<th>Kanlar içindeydi</th>
<th>Fluent</th>
<th>Disfluent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adequate</td>
<td>He was covered in blood.</td>
<td>There was blood on.</td>
</tr>
<tr>
<td>Inadequate</td>
<td>Drinking blood.</td>
<td>In the Kanlar.</td>
</tr>
</tbody>
</table>

Actual system outputs from WMT16 News TR-EN translation task. From left-to-right and top-to-bottom: online-G, online-B, dvorkanton, jhu
Can we detect them automatically?

How often do they occur?

For what types of systems are they most frequent?
We score fluency of a sentence using its language model probability

\[
p(e; \theta) = \prod_i p(e_i | e_{<i}; \theta)
\]

(modified to account for sentence length + outliers)
Quantifying Adequacy

Approach: compute semantic similarity between MT and human translation

“BLEU”: counts n-gram matches
“BVSS”: combines pairwise word similarity into sentence score

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU</td>
<td>94.33</td>
<td>99.08</td>
<td>96.65</td>
</tr>
<tr>
<td>Averaged Embeddings</td>
<td>84.56</td>
<td><strong>99.15</strong></td>
<td>91.28</td>
</tr>
<tr>
<td>BVSS</td>
<td><strong>99.39</strong></td>
<td>99.04</td>
<td><strong>99.22</strong></td>
</tr>
<tr>
<td>BVSS-Reference</td>
<td>99.00</td>
<td>99.03</td>
<td>99.01</td>
</tr>
<tr>
<td>BVSS-System</td>
<td>99.17</td>
<td>99.03</td>
<td>99.10</td>
</tr>
<tr>
<td>BLEU+BVSS</td>
<td><strong>99.61</strong></td>
<td><strong>99.81</strong></td>
<td><strong>99.71</strong></td>
</tr>
</tbody>
</table>
we study 36 MT systems spanning 3 training configurations

General (large, out-of-domain), TED (small, in-domain), Adapted (general model adapted to TED)

2 paradigms
Statistical MT (Joshua), Neural MT (Sockeye)

6 language pairs

<table>
<thead>
<tr>
<th></th>
<th>Arabic</th>
<th>Chinese</th>
<th>Farsi</th>
<th>German</th>
<th>Korean</th>
<th>Russian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joshua General</td>
<td>23.50</td>
<td>30.65</td>
<td>13.41</td>
<td>6.34</td>
<td>24.49</td>
<td>14.79</td>
</tr>
<tr>
<td>Joshua Adapted</td>
<td>27.11</td>
<td>31.35</td>
<td>17.71</td>
<td>10.24</td>
<td>25.23</td>
<td>15.70</td>
</tr>
<tr>
<td>Sockeye General</td>
<td>29.6</td>
<td>34.59</td>
<td>22.22</td>
<td>11.56</td>
<td>28.6</td>
<td>15.92</td>
</tr>
<tr>
<td>Sockeye TED Only</td>
<td>27.42</td>
<td>32.25</td>
<td>21.31</td>
<td>14.4</td>
<td>22.9</td>
<td>16.18</td>
</tr>
<tr>
<td>Sockeye Adapted</td>
<td>35.37</td>
<td>39.9</td>
<td>27.92</td>
<td>17.22</td>
<td>28.6</td>
<td>20.37</td>
</tr>
</tbody>
</table>

Table 6: BLEU scores for all systems
Fluently inadequate translations

For the same quality, neural models have more fluently inadequate than older statistical models.

Percent fluently inadequate for neural models improves rapidly as BLEU improves.
Neural models have consistently higher percent fluently inadequate than statistical models.

Domain adaptation improves score for both, but more dramatically for neural models.

Even highest percent fluently inadequate is <2%.
are infrequent but important
more likely to mislead users
millions of translation requests per day!

affect all models considered
happen more often for neural models, when translating out of domain, and low-resource languages
How can we make neural models more responsive to user needs?
Can We Make MT Sensitive to Formality?

Comment ça va?

Tone: formal

Formality sensitive MT

How are you doing?

Comment ça va?

Tone: informal

Formality sensitive MT

What’s up?

[Niu & Carpuat, EMNLP 2017]
Can We Control the Complexity Of MT?

Audience: fluent English speaker

El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.

The Mauritshuis museum is staging an exhibition focused solely on 17th century self-portraits.

Audience: 2nd language learner

El museo Mauritshuis abre una exposición dedicada a los autorretratos del siglo XVII.

The Mauritshuis museum is going to show self-portraits.

[Agrawal & Carpuat, EMNLP 2019]
Adapting translation output to different audiences via multi-task learning

Multi-task loss = \sum_{(s_i, g_e, e_o)} \log P(e_o | s_i, g_e; \theta) + \sum_{(e_i, g_e, e_o)} \log P(e_o | e_i, g_e; \theta) + \sum_{(s_i, e_o)} \log P(e_o | s_i; \theta)

- \( L_{CMT} \): Spanish sentences translated into simpler English
- \( L_{Simplify} \): Complex English sentences paired with simpler English
- \( L_{MT} \): Spanish-English translation examples
Tailoring MT output to different audiences

Possible with multi-task models
Outperform pipeline models
Powerful framework to incorporate diverse data sources

But open questions remain
How to better control the nature of style transfer operations? Can we relax data requirements with unsupervised methods?
Deep neural networks provide a powerful framework to model translation.

Semi-supervised learning is key to train high-performing models.

How can we make neural models more responsive to user needs?
Neural Machine Translation for Communication Across Language Barriers

Marine Carpuat
marine@cs.umd.edu

Future of Privacy Forum
Digital Data Flows Masterclass 2020/12/09
References


Thank you!

Contact us:
info@fpf.org
@FutureofPrivacy
www.fpf.org/classes