Tuning Neural MT

Guido Zarrella
MITRE Corporation
Outline

- Tuning MT: when the system you have isn’t the system you need

- Neural MT tuning methods differ from those for Statistical MT

- Genre or Domain matters (a lot):
  - In-genre test: BLEU = 25.6
  - Out-of-genre test: BLEU = 7.5 (-18.1)

- You care about NMT tuning because…
  - Tuned w/ monolingual data only: BLEU = 10.3 (+2.8)
  - Trained on a small parallel set: BLEU = 13.5 (+6.0)
  - Tuned (transfer learning): BLEU = 15.0 (+7.5) to 16.9 (+9.4)
Tuning a system you have, to get the system you need
Tuning a system you have, to get the system you need
Tuning Machine Translation

In SMT, tuning involves learning a **weighted combination** of scoring features output by trained components: translation tables, language models, reordering models, …

For example: **Minimum Error Rate Training (MERT)**

or **Margin-infused Relaxed Algorithm (MIRA)**
### Tuning Statistical Machine Translation

**SampleRank Training for Phrase-Based Machine Translation**  
Barry Haddow, Abhishek Arun, Philipp Koehn 2011

<table>
<thead>
<tr>
<th>Pair</th>
<th>System</th>
<th>untuned</th>
<th>MERT-tuned</th>
</tr>
</thead>
<tbody>
<tr>
<td>fr-en</td>
<td>WMT-SMALL</td>
<td>28.0</td>
<td>29.2 (0.2)</td>
</tr>
<tr>
<td></td>
<td>WMT-LARGE</td>
<td>29.4</td>
<td>32.5 (0.1)</td>
</tr>
<tr>
<td>de-en</td>
<td>WMT-SMALL</td>
<td>25.0</td>
<td>25.3 (0.1)</td>
</tr>
<tr>
<td></td>
<td>WMT-LARGE</td>
<td>26.6</td>
<td>26.8 (0.2)</td>
</tr>
</tbody>
</table>

Image Credit: NLP Group at Northeastern University, China
Need for Domain Adaptation

**Newswire source**

目前日本有关方面已经派出三只巡逻艇，协同韩国方面在出事水域开展搜寻遇难者的工作。

**Semiconductor source**

利用在线应力测试技术表征了掺入Pt后对镍硅化物薄膜应力性质的影响。

**Human translation**

Currently, Japanese authorities have three dispatched patrol boats to coordinate with the South Koreans in searching for the victims in the area of the incident.

**Machine translation**

Japan has dispatched three patrol boats to the area, in coordination with the South Koreans to search for the victims in the area of the incident work.

The effect of Pt doping on the stress in the nickel silicide film has been characterized using an in-situ stress measurement.

Stress tests use online technology characterized by incorporation of Pt on nickel silicide films nature of the stress
## Need for Domain Adaptation

<table>
<thead>
<tr>
<th>System</th>
<th>Description</th>
<th>Score (BLEU)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Semi-</td>
<td>Chem-</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>conductor</td>
<td>bio</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>Stand-alone product, statistical</td>
<td>9.4</td>
<td>9.7</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>Stand-alone product, rule-based</td>
<td>11.2</td>
<td>11.9</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>Web-based, statistical</td>
<td>15.1</td>
<td>22.8*</td>
<td></td>
</tr>
<tr>
<td>MITRE</td>
<td>Statistical</td>
<td>16.1</td>
<td>17.9</td>
<td></td>
</tr>
</tbody>
</table>
“With the exception of fr-es and ru-en the neural system is always comparable or better than the phrase-based system.”

*Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions*
Marcin Junczys-Dowmunt, Tomasz Dwojak, Hieu Hoang
Neural Machine Translation

\[ f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .}) \]

Subtitle Corpus for Discourse


<table>
<thead>
<tr>
<th>language</th>
<th>files</th>
<th>tokens</th>
<th>sentences</th>
<th>af</th>
<th>ar</th>
<th>bg</th>
<th>bn</th>
<th>br</th>
<th>bs</th>
<th>ca</th>
<th>cs</th>
<th>da</th>
<th>de</th>
<th>el</th>
<th>en</th>
<th>eo</th>
<th>es</th>
<th>et</th>
<th>eu</th>
<th>fa</th>
<th>fi</th>
<th>fr</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>af</td>
<td>32</td>
<td>0.2M</td>
<td>27.4k</td>
<td>6.2k</td>
<td>7.6k</td>
<td></td>
<td></td>
<td></td>
<td>1.8k</td>
<td></td>
<td>10.5k</td>
<td></td>
<td></td>
<td>7.9k</td>
<td>11.6k</td>
<td>16.2k</td>
<td>12.6k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.1k</td>
<td>2.8k</td>
</tr>
<tr>
<td>ar</td>
<td>67,608</td>
<td>329.8M</td>
<td>60.8M</td>
<td>6.2k</td>
<td></td>
<td>16.2M</td>
<td>62.2k</td>
<td>13.0k</td>
<td>6.1M</td>
<td>0.3M</td>
<td>16.5M</td>
<td>7.5M</td>
<td>7.1M</td>
<td>15.3M</td>
<td>19.4M</td>
<td>19.4k</td>
<td>18.3M</td>
<td>6.9M</td>
<td>0.1M</td>
<td>3.0M</td>
<td></td>
<td></td>
<td>10.8M</td>
</tr>
<tr>
<td>bg</td>
<td>90,376</td>
<td>523.4M</td>
<td>80.2M</td>
<td>7.7k</td>
<td>17.8M</td>
<td>60.7k</td>
<td>13.8k</td>
<td>7.5M</td>
<td>0.3M</td>
<td>21.1M</td>
<td>8.2M</td>
<td>8.9M</td>
<td>19.3M</td>
<td>26.4M</td>
<td>23.4k</td>
<td>24.8M</td>
<td>7.7M</td>
<td>0.1M</td>
<td>2.8M</td>
<td>13.2M</td>
<td>18.5M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bn</td>
<td>76</td>
<td>0.6M</td>
<td>0.1M</td>
<td></td>
<td></td>
<td></td>
<td>64.1k</td>
<td>62.7k</td>
<td>36.6k</td>
<td>3.1k</td>
<td>61.1k</td>
<td>58.2k</td>
<td>54.7k</td>
<td>58.5k</td>
<td>69.3k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>br</td>
<td>32</td>
<td>0.2M</td>
<td>23.1k</td>
<td></td>
<td></td>
<td></td>
<td>13.3k</td>
<td>14.1k</td>
<td>2.7k</td>
<td>5.3k</td>
<td>14.5k</td>
<td>10.0k</td>
<td>7.5k</td>
<td>14.4k</td>
<td>17.7k</td>
<td>1.1k</td>
<td>15.6k</td>
<td>15.0k</td>
<td>0.7k</td>
<td>4.4k</td>
<td>8.1k</td>
<td>15.4k</td>
<td></td>
</tr>
<tr>
<td>bs</td>
<td>30,511</td>
<td>179.5M</td>
<td>28.4M</td>
<td>1.8k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ca</td>
<td>711</td>
<td>4.0M</td>
<td>0.5M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>96.2k</td>
<td>0.2M</td>
<td>0.3M</td>
</tr>
<tr>
<td>cs</td>
<td>125,126</td>
<td>715.3M</td>
<td>112.8M</td>
<td>10.7k</td>
<td></td>
<td>18.1M</td>
<td>24.7M</td>
<td>63.3k</td>
<td>14.8k</td>
<td>8.5M</td>
<td>0.4M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>da</td>
<td>24,079</td>
<td>162.4M</td>
<td>23.6M</td>
<td>6.1k</td>
<td></td>
<td></td>
<td>8.0M</td>
<td>9.3M</td>
<td>60.9k</td>
<td>10.1k</td>
<td>4.0M</td>
<td>0.2M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>27,742</td>
<td>186.3M</td>
<td>26.9M</td>
<td>8.0k</td>
<td></td>
<td>7.6M</td>
<td>10.0M</td>
<td>57.2k</td>
<td>7.7k</td>
<td>4.0M</td>
<td>0.2M</td>
<td>10.6M</td>
<td>5.4M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>el</td>
<td>114,230</td>
<td>683.1M</td>
<td>101.6M</td>
<td>11.8k</td>
<td></td>
<td>16.8M</td>
<td>22.3M</td>
<td>60.5k</td>
<td>14.6k</td>
<td>8.1M</td>
<td>0.3M</td>
<td>23.0M</td>
<td>9.1M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>en</td>
<td>322,294</td>
<td>2.5G</td>
<td>336.6M</td>
<td>16.7k</td>
<td></td>
<td>21.9M</td>
<td>31.6M</td>
<td>75.0k</td>
<td>18.5k</td>
<td>11.1M</td>
<td>0.4M</td>
<td>33.8M</td>
<td>11.0M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eo</td>
<td>89</td>
<td>0.5M</td>
<td>79.3k</td>
<td></td>
<td></td>
<td></td>
<td>19.9k</td>
<td>24.3k</td>
<td>1.1k</td>
<td>7.6k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>es</td>
<td>191,987</td>
<td>1.3G</td>
<td>179.2M</td>
<td>12.9k</td>
<td></td>
<td>20.3M</td>
<td>29.2M</td>
<td>69.1k</td>
<td>16.0k</td>
<td>10.2M</td>
<td>0.4M</td>
<td>30.7M</td>
<td>10.4M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>et</td>
<td>23,515</td>
<td>140.7M</td>
<td>22.9M</td>
<td>2.2k</td>
<td></td>
<td>7.5M</td>
<td>8.9M</td>
<td>58.6k</td>
<td>15.4k</td>
<td>4.0M</td>
<td>0.2M</td>
<td>9.2M</td>
<td>5.7M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eu</td>
<td>188</td>
<td>1.4M</td>
<td>0.2M</td>
<td></td>
<td></td>
<td></td>
<td>0.1M</td>
<td>0.1M</td>
<td>3.3k</td>
<td>0.7k</td>
<td>80.9k</td>
<td></td>
<td>0.1M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fa</td>
<td>6,469</td>
<td>44.3M</td>
<td>7.4M</td>
<td>2.1k</td>
<td></td>
<td>3.1M</td>
<td>2.9M</td>
<td>46.3k</td>
<td>4.4k</td>
<td>1.4M</td>
<td>0.1M</td>
<td>3.1M</td>
<td>2.2M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fi</td>
<td>44,594</td>
<td>208.5M</td>
<td>38.7M</td>
<td>2.8k</td>
<td></td>
<td>11.5M</td>
<td>14.8M</td>
<td>57.9k</td>
<td>8.3k</td>
<td>5.7M</td>
<td>0.2M</td>
<td>15.3M</td>
<td>9.0M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>105,070</td>
<td>672.8M</td>
<td>90.9M</td>
<td>7.5k</td>
<td></td>
<td>15.5M</td>
<td>21.3M</td>
<td>61.4k</td>
<td>16.3k</td>
<td>7.4M</td>
<td>0.3M</td>
<td>21.8M</td>
<td>8.5M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>de</td>
<td>370</td>
<td>1.0M</td>
<td>0.2M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.5k</td>
<td>28.0k</td>
<td>11.8k</td>
<td>71.3k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Arabic to English

- Trained on 21 million conversational segments from movie subtitles
  - 256 million training steps (sentences)
  - 19 days on K40 GPU

- NMT BLEU = 25.6
  - SMT BLEU = 25.3

- Serialized as 536 MB model
  - Deployable to laptops
people would think that he was the terrorist. right.

- there's a boy in the cage.

we're just here to see our friend rigby, sir.

- glass is all over the floor.

broke the stereo.

like nathan.

let's get ice cream.

he's down checking a buoy in the channel.

oh, my god, please.

cervical lymph node has black flecks.

you came for your uncle's wedding.

yeah, and doctors say i should get more

and more each day.

people say he was a terrorist. right.

- there's a boy in the cage.

we're just here to see our friend, sir.

the glass is all around someone.

like a . .

let's go get ice cream.

he's out there asking for a consult.

oh, god, please.

the black is a black.

i came for your uncle's wedding.

yeah, the doctors said i would remember more every day.
In a new domain

“tourism accounts for almost N % of the austrian gross domestic product .”

“the industry are nearly N , of the most common population .”

On Wikipedia: **BLEU = 7.4**
Tuning NMT?

$f = (\text{La, croissance, économique, s'est, ralentie, ces, dernières, années, .})$

Black Box NMT

$e = (\text{Economic, growth, has, slowed, down, in, recent, years, .})$
Transfer Learning

- Our core strategy is to employ transfer learning between deep neural networks pre-trained on massive datasets.

- Knowledge gained in one context can be re-used to solve different but related problems.

https://edpsychexperience.wordpress.com/2013/03/25/013112-learning-learning-transfer
Wikipedia Adaptation Experiments

- Incremental training: we pick up where OpenSubtitles left off
  - With tiny parallel tuning set (n=1024)
  - With small parallel training set (n=32768)
  - With full parallel training set (n=148136)
  - With varying amounts of in-domain monolingual data
  - With expanded vocabularies

- About 22 minutes per 100k training updates

Incrementally Adapting OpenSubtitles to Wikipedia

![Graph showing BLEU score over increasing tuning updates, with a trend line labeled "train from scratch, n=33k".]
Incrementally Adapting OpenSubtitles to Wikipedia

- **Graph Description:**
  - **Y-axis:** BLEU score ranging from 0 to 20.
  - **X-axis:** Number of tuning updates from 0 to 20,000,000.
  - **Lines:**
    - Blue line: training from scratch, n=33k.
    - Red line: tuning, n=33k.
  - **Legend:**
    - **Train from scratch, n=33k**
    - **Tuning, n=33k**

- **Graph Analysis:**
  - The graph illustrates the improvement in BLEU score as the number of tuning updates increases for both training from scratch and tuning.
  - Training from scratch shows a steady increase in BLEU score.
  - Tuning shows a more erratic but overall increasing trend in BLEU score.
  - The performances are compared under the same conditions, demonstrating the effectiveness of incremental adaptation.
Incrementally Adapting OpenSubtitles to Wikipedia

![Graph showing BLEU scores over tuning updates for different training methods.

- **Blue line**: train from scratch, n=33k
- **Red line**: tuning, n=33k
- **Orange line**: tuning, n=148k

The graph plots BLEU scores against the number of tuning updates, illustrating the performance improvement over time for each method.
Incrementally Adapting OpenSubtitles to Wikipedia

![Graph showing Bleu scores over tuning updates]

- **tuning, n=1k**
- **monolingual English only, n=33k**

# Tuning Updates
Incrementally Adapting OpenSubtitles to Wikipedia

![Graph showing BLEU scores over tuning updates for different scenarios.]

- Tuning, n=1k
- Monolingual English only, n=33k
- Tuning n=1k + Monolingual n=147k
Reference: tourism accounts for almost N % of the austrian gross domestic product.

Train from scratch, 33k: world is up for N % of the total reserves.

Untuned: the industry are nearly N, of the most common population.

1k tuning: tourism costs nearly N (of the most common population).

33k tuning: tourism often manifests approximately N % of the gdp.

... ensembling?
Results

- **Genre & domain matter (a lot)**
  - In-genre test: BLEU = **25.6**
  - Out-of-genre test: BLEU = **7.5** (**-18.1**)

- **Incremental training helps**
  - Trained, parallel in-domain: BLEU = **13.5** (**+6.0**)
  - Tuned, parallel in-domain: BLEU = **15.0** (**+7.5**) to **16.9** (**+9.4**)

- **Monolingual data helps when parallel data is scarce**
  - Tuned, 33k monolingual in-domain: BLEU = **10.3** (**+2.8**)
  - Tuned, 1k parallel in-domain: BLEU = **10.6** (**+3.1**)

- **Expanding vocabulary doesn’t increase BLEU (yet)**
Conclusions

- All parameters in a NMT system are tunable
  - can create great diversity from one “well trained” seed system
  - … in minutes or hours, with little or no additional parallel data

- Government use cases poised to benefit most
  - Collect many partially trained systems on the shelf?

- Still open question how to best create systems optimized for tuning

- Sharing models? Share training code too.
Thank You

Guido Zarrella
jzarrella@mitre.org
@gzco