NLP for Low-resource or Endangered Languages and Cross-lingual transfer

Antonios Anastasopoulos
MTMA

May 31, 2019
There are about 7000 languages in the world
There are about **7000** languages in the world

According to UNESCO, **43%** of the world’s languages are endangered or vulnerable.
from The Endangered Languages Project

- : dormant
- : critically endangered
- : endangered
- : at risk

[from The Endangered Languages Project]
[from The Endangered Languages Project]

- ●: dormant
- ○: critically endangered
- ▲: endangered
- ▫: at risk
[from The Endangered Languages Project]
Language Documentation

1. Collect (record) data
2. Transcribe and translate
3. Perform analysis
4. Elicit further paradigms
5. Prepare a grammar
"We collect and archive language recordings now while the speakers are still alive. That’s all. We have the whole of the future to transcribe and process the recordings..."

Steven Bird
My work

Develop methods that will automate and speed up the language documentation process:

Alignment (segmentation)

Transcription

Translation

Analysis

‘The cat sat on the mat’
Speech Transcription using Translations

el gato se sentò en la alfombra
the cat sat on the mat

Leveraging Translations for Speech Transcription in Low-Resource Settings
Antonios Anastasopoulos and David Chiang, Interspeech 2018
multi-source models: encoder-decoder model
multi-source models: results
el gato se sentò en la alfombra

\[ P(y_1 \cdots y_M) \]

softmax

\[ s_1^{1} \cdots s_M^{1} \quad \uparrow \text{decoder} \]

\[ c_1^{1} \cdots c_M^{1} \quad \uparrow \text{attention} \]

\[ h_1^{1} \cdots h_{N_1}^{1} \quad \uparrow \text{encoder} \]

\[ x_1^{1} \cdots x_{N_1}^{1} \]

\[ s_1^{2} \cdots s_M^{2} \quad \uparrow \text{decoder} \]

\[ c_1^{2} \cdots c_M^{2} \quad \uparrow \text{attention} \]

\[ h_1^{2} \cdots h_{N_2}^{2} \quad \uparrow \text{encoder} \]

\[ x_1^{2} \cdots x_{N_2}^{2} \]

the cat sat on the mat

[Dutt et al. 2017]
Character Error Rate

- **Ainu (2k)**: speech 40.7, translation 74.9, ensemble 40.6
- **Mboshi (5k)**: speech 29.8, translation 68.2, ensemble 36.8
- **Spanish (17k)**: speech 52.0, translation 44.6, ensemble 42.2

Multi-source models: results
multi-source models: multisource
multi-source models: results
multi-source models: attention parameter sharing
Character Error Rate

- **Ainu (2k)**: 40.7, 40.6, 46.0
- **Mboshi (5k)**: 29.8, 36.8, 37.5, 28.6
- **Spanish (17k)**: 52.0, 44.6, 42.2, 41.6, 38.7

multi-source models: results
Speech Transcription and Translation

el gato se sentò en la alfombra
the cat sat on the mat

Tied Multitask Models for Speech Transcription and Translation

Antonios Anastasopoulos and David Chiang. NAACL 2018.
multi-task models: pivot
Multi-task models: end-to-end

An Attentional Model for *Speech Translation without Transcription*
Long Duong, Antonios Anastasopoulos, Trevor Cohn, Steven Bird, and David Chiang.
NAACL 2016.
**Multi-task models: simple multitask**

- **Encoder**: $h_1 \cdots h_N$
- **Decoder**: $s_1 \cdots s_{M_1}^1 \cdots s_{M_2}^2$
- **Attention**: $c_1^1 \cdots c_{M_1}^1 \cdots c_1^2 \cdots c_{M_2}^2$
- **Softmax**: $P(y_1^1 \cdots y_{M_1}^1)$, $P(y_1^2 \cdots y_{M_2}^2)$

**Input**: $x_1 \cdots x_N$

**Examples**:
- **Spanish**: el gato se sentó en la alfombra
- **English**: the cat sat on the mat
Transcription Character Error Rate

Transcription character BLEU

multi-task models: results
multi-task models: triangle

```
\text{the cat sat on the mat}

\text{el gato se sentó en la alfombra}

P(y_1 \cdots y_{M^2})

\uparrow \text{softmax}

s_{1} \cdots s_{M^2}

\uparrow \text{decoder}

\text{attentions}

\text{P(y}_1^1 \cdots y_{M^1}^1)\text{)}

\uparrow \text{softmax}

s_{1}^1 \cdots s_{M^1}^1

\uparrow \text{decoder}

\text{c}_1^1 \cdots c_{M^2}^1

\uparrow \text{encoder}

x_1 \cdots x_N
```

\text{multi-task models: triangle}
multi-task models: results
If A attends over B... and B attends over C...
this should be similar to A attending directly over C.

\[ \mathcal{R}_{\text{trans}} = -\lambda_{\text{trans}} \| \mathbf{A}^{12} \mathbf{A}^1 - \mathbf{A}^2 \|_2^2. \]
multi-task models: results
We can improve translation and transcription accuracy by jointly performing the two tasks.

Translation can be further improved by using intermediate representations and transitivity.
Other (relevant and ongoing) work

Build a tool for linguists that uses ML in its backend to aid annotation:

work with Graham Neubig
Data Augmentation, Cross-Lingual Transfer, and other nice things
Using related languages for MT

"Generalized Data Augmentation for Low-Resource Translation"

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig
ACL 2019
Transfer for MT

Typical scenario: continued training

Machine Translation

The current best approach is a semi-supervised one:

- Back-translation of target-side monolingual data

What if we don’t have tons of monolingual data for a language?

Does the quality of the back-translated data matter?
Generalized Back-Translation

For low-resource languages, there maybe exist a related high-resource one e.g.

1. Azerbaijani (Turkish)
2. Belarusian (Russian)
3. Galician (Portuguese)
4. Slovak (Czech)

We should use them!
Generalized Back-Translation
Generalized Back-Translation

Typical:

- only use [1] for data augmentation

  OR

- add [b] to [c] and train.

But HRL to LRL might be easier!
From HRL to LRL

Assuming a parallel dataset is probably too much.

If the languages are related enough:

1. Get monolingual embeddings
2. Align the embedding space [Lample et al, 2018]
3. Learn a dictionary
4. Word substitution in HRL to create pseudo-LRL
From ENG to LRL through HRL

ENG to LRL system would be bad (duh!)

ENG to HRL system would be better…

… and HRL to LRL might be easy-ish (cause related)
## Results

<table>
<thead>
<tr>
<th>Training Data</th>
<th>BLEU for X→ENG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AZE (TUR)</td>
</tr>
<tr>
<td>1 Base Supervised NMT</td>
<td>11.83</td>
</tr>
<tr>
<td>2 Base Unsupervised NMT</td>
<td>0.47</td>
</tr>
<tr>
<td><strong>Standard Supervised Back-translation</strong></td>
<td></td>
</tr>
<tr>
<td>3 ${\hat{S}^s_{E\rightarrow L}, M_E}$</td>
<td>11.84</td>
</tr>
<tr>
<td>4 ${\hat{S}^s_{E\rightarrow H}, M_E}$</td>
<td>12.46</td>
</tr>
<tr>
<td><strong>Augmentation from HRL-ENG</strong></td>
<td></td>
</tr>
<tr>
<td>5 ${\hat{S}^s_{H\rightarrow L}, \tau_{HE}}$</td>
<td>11.92</td>
</tr>
<tr>
<td>6 ${\hat{S}^m_{H\rightarrow L}, \tau_{HE}}$</td>
<td>(unsupervised MT)</td>
</tr>
<tr>
<td>7 ${S^w_{H\rightarrow L}, \tau_{HE}}$</td>
<td>(word subst.)</td>
</tr>
<tr>
<td>8 ${S^m_{H\rightarrow L}, \tau_{HE}}$</td>
<td>(modified UMT)</td>
</tr>
<tr>
<td>9 ${\hat{S}^w_{H\rightarrow L}, \hat{S}^m_{H\rightarrow L}, \tau_{HE}}$</td>
<td>15.24</td>
</tr>
<tr>
<td><strong>Augmentation from ENG by pivoting</strong></td>
<td></td>
</tr>
<tr>
<td>10 ${\hat{S}^w_{E\rightarrow H\rightarrow L}, M_E}$</td>
<td>(word subst.)</td>
</tr>
<tr>
<td>11 ${\hat{S}^m_{E\rightarrow H\rightarrow L}, M_E}$</td>
<td>(modified UMT)</td>
</tr>
<tr>
<td><strong>Combinations</strong></td>
<td></td>
</tr>
<tr>
<td>12 ${\hat{S}^w_{H\rightarrow L}, \hat{S}^w_{E\rightarrow H\rightarrow L}, \tau_{HE}, M_E}$</td>
<td>(word subst.)</td>
</tr>
<tr>
<td>13 ${\hat{S}^w_{H\rightarrow L}, \hat{S}^m_{H\rightarrow L}, \tau_{HE}, \tau_{HE}}$</td>
<td>15.91</td>
</tr>
<tr>
<td>${\hat{S}^w_{E\rightarrow H\rightarrow L}, \hat{S}^m_{E\rightarrow H\rightarrow L}, M_E, M_E}$</td>
<td></td>
</tr>
</tbody>
</table>
Takeaways

Translating from HRL to LRL:

- it is better to use word substitution than simple NMT or standard UMT (cf lines 5,6 to 7,8,9)

Pivoting from ENG though HRL, improvements but not as much.

Best of both worlds works best (line 12)

- More ENG data, as good as possible LRL data
Using Related Languages for Morphological Inflection
Inflection task and SIGMORPHON

Low-resource target training data (Asturian)

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Asturian</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>facer</td>
<td>fechu</td>
<td>V;V.PTCP;PST</td>
</tr>
<tr>
<td>aguar</td>
<td>aguà</td>
<td>V;PRS;2;PL;IND</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

High-resource source language training data (Spanish)

<table>
<thead>
<tr>
<th>Spanish</th>
<th>Spanish Form</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>tocar</td>
<td>tocando</td>
<td>V;V.PTCP;PRS</td>
</tr>
<tr>
<td>bailar</td>
<td>bailaba</td>
<td>V;PST;IPFV;3;SG;IND</td>
</tr>
<tr>
<td>mentir</td>
<td>mintió</td>
<td>V;PST;PFV;3;SG;IND</td>
</tr>
</tbody>
</table>

SIGMORPHON challenge:
100 language pairs (43 test languages)
Previous Work

Concatenate tags and lemma, single encoder-decoder

- Issue: inherently different (order, function)

Half task is identifying stem/root and copy characters, so other works focus on copying

- explicit copy mechanism, or
- hard monotonic attention, or
- learn to output the string transduction steps
Augmentation approach: hallucinating data

Most low-resource languages have just 50 or 100 examples.

You can hallucinate more data:

```
bar
|   |   |
baiba
```

replacing the red parts with random characters
Results on transfer from single language

If languages are genetically distant, transfer does NOT help.

Same alphabet crucial: see Kurmanji-Sorani

<table>
<thead>
<tr>
<th>L1</th>
<th>L2</th>
<th>L1+L2</th>
<th>+H</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>latin</td>
<td>czech</td>
<td>15</td>
<td>71.4</td>
<td>77.4</td>
</tr>
<tr>
<td>bengali</td>
<td>greek</td>
<td>12.4</td>
<td>70.5</td>
<td>71.6</td>
</tr>
<tr>
<td>sorani</td>
<td>irish</td>
<td>10.3</td>
<td>66.3</td>
<td>65.6</td>
</tr>
<tr>
<td>italian</td>
<td>ladin</td>
<td>48</td>
<td>74</td>
<td>74</td>
</tr>
<tr>
<td>latvian</td>
<td>lithuanian</td>
<td>7.1</td>
<td>48.4</td>
<td>50.5</td>
</tr>
<tr>
<td>english</td>
<td>murrinhpatha</td>
<td>36</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td>italian</td>
<td>neapolitan</td>
<td>70</td>
<td>83</td>
<td>84</td>
</tr>
<tr>
<td>urdu</td>
<td>old english</td>
<td>13.8</td>
<td>43.4</td>
<td>44.3</td>
</tr>
<tr>
<td>slovene</td>
<td>old saxon</td>
<td>10.7</td>
<td>52.3</td>
<td>50.5</td>
</tr>
<tr>
<td>russian</td>
<td>portuguese</td>
<td>34.5</td>
<td>88.8</td>
<td>87.7</td>
</tr>
<tr>
<td>swahili</td>
<td>quechua</td>
<td>4.2</td>
<td>92.1</td>
<td>91.6</td>
</tr>
<tr>
<td>portuguese</td>
<td>russian</td>
<td>25.6</td>
<td>76.3</td>
<td>74.3</td>
</tr>
<tr>
<td>kurmanji</td>
<td>sorani</td>
<td>6.2</td>
<td>69</td>
<td>66.7</td>
</tr>
<tr>
<td>zulu</td>
<td>swahili</td>
<td>46</td>
<td>81</td>
<td>76</td>
</tr>
<tr>
<td>kannada</td>
<td>telugu</td>
<td>76</td>
<td>94</td>
<td>94</td>
</tr>
</tbody>
</table>

Average 27.72 67.77 68.55
But we can do better if transferring from multiple (related) languages e.g.

<table>
<thead>
<tr>
<th>L1</th>
<th>L2</th>
<th>L1+L2</th>
<th>+H</th>
<th>+L1 + H</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>turkish</td>
<td>81</td>
<td>80</td>
<td>81</td>
<td></td>
<td></td>
</tr>
<tr>
<td>persian</td>
<td>35</td>
<td>74</td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bashkir</td>
<td>azeri</td>
<td>37</td>
<td>66</td>
<td>67</td>
<td>66.7±0.9</td>
</tr>
<tr>
<td>uzbek</td>
<td>27</td>
<td>74</td>
<td>70</td>
<td></td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>84</td>
<td>83</td>
<td>87</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Interpreting the model
Takeaways

1. Monolingual data hallucination can take you a long way…

2. … and it’s preferable to cross-lingual transfer from distant languages

3. If close enough languages, both data hallucination and cross-lingual transfer should help

4. The closer the languages, the larger the improvements

Main Issues:

• Data Hallucination is language-agnostic. A more informed sampling could probably do better

• Different alphabets really hurt performance (Dutch-Yiddish, Kurmanji-Sorani). Need to find either an a priori mapping between the two, or map the to a common space (IPA?)
What language should you use for cross-lingual transfer?

"Choosing Transfer Languages for Cross-Lingual Learning"

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell and Graham Neubig
ACL 2019
Setting

Cross-lingual transfer on 4 tasks:

1. MT: 54x54 pairs (X-Eng TED)
2. POS-tagging: 60x26
3. Entity Linking: 53x9
4. DEP parsing: 30x30
Learning to Rank

For each language pair, extract features:

1. dataset dependent:
   - dataset size
   - type-token ratio
   - word/subword overlap

2. dataset-independent:
   1. typological features (from URIEL)

[plug: check out the lang2vec python library, now with pre-computed distances!]
Learning to Rank

For each test language and a list of potential transfer languages (each pair represented by the features), train a model to rank the candidate languages.

Model: tree-based LambdaRank (good in limited feature/data settings)

<table>
<thead>
<tr>
<th>Method</th>
<th>MT</th>
<th>EL</th>
<th>POS</th>
<th>DEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>word overlap ( o_w )</td>
<td>28.6</td>
<td>30.7</td>
<td>13.4</td>
<td>52.3</td>
</tr>
<tr>
<td>subword overlap ( o_{sw} )</td>
<td>29.2</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>size ratio ( s_{tf}/s_{tk} )</td>
<td>3.7</td>
<td>0.3</td>
<td>9.5</td>
<td>24.8</td>
</tr>
<tr>
<td>type-token ratio ( d_{ttt} )</td>
<td>2.5</td>
<td>–</td>
<td>7.4</td>
<td>6.4</td>
</tr>
<tr>
<td>genetic ( d_{gen} )</td>
<td>24.2</td>
<td>50.9</td>
<td>14.8</td>
<td>32.0</td>
</tr>
<tr>
<td>syntactic ( d_{syn} )</td>
<td>14.8</td>
<td>46.4</td>
<td>4.1</td>
<td>22.9</td>
</tr>
<tr>
<td>featural ( d_{fco} )</td>
<td>10.1</td>
<td>47.5</td>
<td>5.7</td>
<td>13.9</td>
</tr>
<tr>
<td>phonological ( d_{pho} )</td>
<td>3.0</td>
<td>4.0</td>
<td>9.8</td>
<td>43.4</td>
</tr>
<tr>
<td>inventory ( d_{inv} )</td>
<td>8.5</td>
<td>41.3</td>
<td>2.4</td>
<td>23.5</td>
</tr>
<tr>
<td>geographic ( d_{geo} )</td>
<td>15.1</td>
<td>49.5</td>
<td>15.7</td>
<td>46.4</td>
</tr>
<tr>
<td><strong>LANGRANK (all)</strong></td>
<td>51.1</td>
<td>63.0</td>
<td>28.9</td>
<td>65.0</td>
</tr>
<tr>
<td><strong>LANGRANK (dataset)</strong></td>
<td>53.7</td>
<td>17.0</td>
<td>26.5</td>
<td>65.0</td>
</tr>
<tr>
<td><strong>LANGRANK (URIEL)</strong></td>
<td>32.6</td>
<td>58.1</td>
<td>16.6</td>
<td>59.6</td>
</tr>
</tbody>
</table>

[ Available as a python package too: https://github.com/neulab/langrank ]
Other Cool Things
Language Technology for Language Documentation and Revitalization

Hackathon-type workshop at CMU, Aug 12-16, 2019
- Language community members
- Documentary linguists
- Computational linguists
- Computer scientists and developers

Example projects:
- Building and using tools for rapid dictionary creation
- Building and using tools for development of speech recognition systems
- Building and using tools to analyze the syntax of language, and extract example sentences for educational materials
- Creating a plugin that incorporates language technology into language documentation software such as ELAN/Praat