Challenges and research directions in Neural Machine Translation

multilingual, unsupervision, fairness

Marta R. Costa-jussà
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona
Words Embeddings
Contextual Words Embeddings

S1 I play football

play (S1)

compete

practise
Contextual Words Embeddings

S1 I play football
S2 I attended a play in the big theatre
Contextual Word Embeddings use Transformers
Neural Machine Translation

Birds can fly

Los pájaros pueden volar

Encoder

Decoder
Three different architectures in 2 years

RNN WITH ATTENTION (2015)
Three different architectures in 2 years

RNN WITH ATTENTION (2015)

CNN (2017)
Three different architectures in 2 years

RNN WITH ATTENTION (2015)

TRANSFORMER (2017)

CNN (2017)
Big Successes
Google translation Evolution
Microsoft claims human parity

Achieving Human Parity on Automatic Chinese to English News Translation

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou

Microsoft AI & Research
Is Machine Translation Solved?
Quality in Machine Translation depends on training data

<table>
<thead>
<tr>
<th>Parallel Text</th>
<th>Amount of Parallel Data</th>
<th>Translation Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Spanish</td>
<td>ENGLISH</td>
</tr>
<tr>
<td>Birds can fly</td>
<td>Los pájaros pueden volar</td>
<td>Birds sleep</td>
</tr>
<tr>
<td>Dog sleep</td>
<td>Los perros duermen</td>
<td>XXL</td>
</tr>
</tbody>
</table>
Quality in Machine Translation depends on training data

<table>
<thead>
<tr>
<th>Parallel Text</th>
<th>Amount of Parallel Data</th>
<th>Translation Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Spanish</td>
<td></td>
</tr>
<tr>
<td>Birds can fly</td>
<td>Los pájaros pueden volar</td>
<td>XXL</td>
</tr>
<tr>
<td>Dog sleep</td>
<td>Los perros duermen</td>
<td></td>
</tr>
<tr>
<td>Turkish</td>
<td>Basque</td>
<td></td>
</tr>
<tr>
<td>Kuşlar uçabilir</td>
<td>hegaztiak hegan</td>
<td>XXS</td>
</tr>
</tbody>
</table>

ENGLISH
Birds sleep

SPANISH
Los pájaros duermen

Turkish
Kuşlar uyku

BASQUE
hegaztiak
Research Directions

Multilingual systems & Unsupervised systems
Low-resourced languages can benefit from high-resourced Universal Encoder-Decoder
Low-resourced languages can benefit from high-resourced

Universal Encoder-Decoder

Language-Specific Encoder-Decoders
Low-resourced languages can benefit from high-resourced

**Universal Encoder-Decoder**

- Shared Vocabulary
- ✓ Zero-shot
- ✓ Transfer learning from high-resourced languages to low-resourced (with the same script)
- x Detrimental for high resourced languages

**Language-Specific Encoder-Decoders**

- ✓ Independent vocabulary
- ✓ Zero-shot
- ✓ Incremental training of new languages and domains
- x No transfer learning from high-resourced to low-resourced (with the same script)
Machine Translation can be trained on unlabelled data
Questioning our data...
Biased Translations

She is a doctor

He is a doctor
Bad Translations can generate Automated Bias

Palestinian man is arrested by police after posting ‘Good morning’ in Arabic on Facebook which was wrongly translated as ‘attack them’
Towards Fairer Systems

Data augmentation & Debiasing algorithms
## Data augmentation

<table>
<thead>
<tr>
<th>Biased Text</th>
<th>Debias Text</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>&lt;Malik:he&gt;</code> is an aspiring singer who works as a salesman in a car showroom. One day he meets <code>&lt;Sonia:he&gt;</code> Saxena daughter of Mr. Saxena when goes to deliver a car to home as birthday present</td>
<td><code>&lt;Malik:he&gt;</code> is an aspiring singer who works as a salesman in a car showroom. One day he meets <code>&lt;Sonia:he&gt;</code> Saxena daughter of Mr. Saxena when goes to deliver a car to home as birthday present</td>
</tr>
</tbody>
</table>
Debiased algorithms
Debiased algorithms
Inspiration from other areas... healthcare?

Under-representation of racial and ethnic minorities in clinical trials

**SRC:** Underrepresentation of Hispanics and Other Minorities in Clinical Trials: Recruiters’ Perspectives, Occa, A, Morgan, SE, Potter JE, 2017
More about our research

www.talp.upc.edu

www.costa-jussa.com