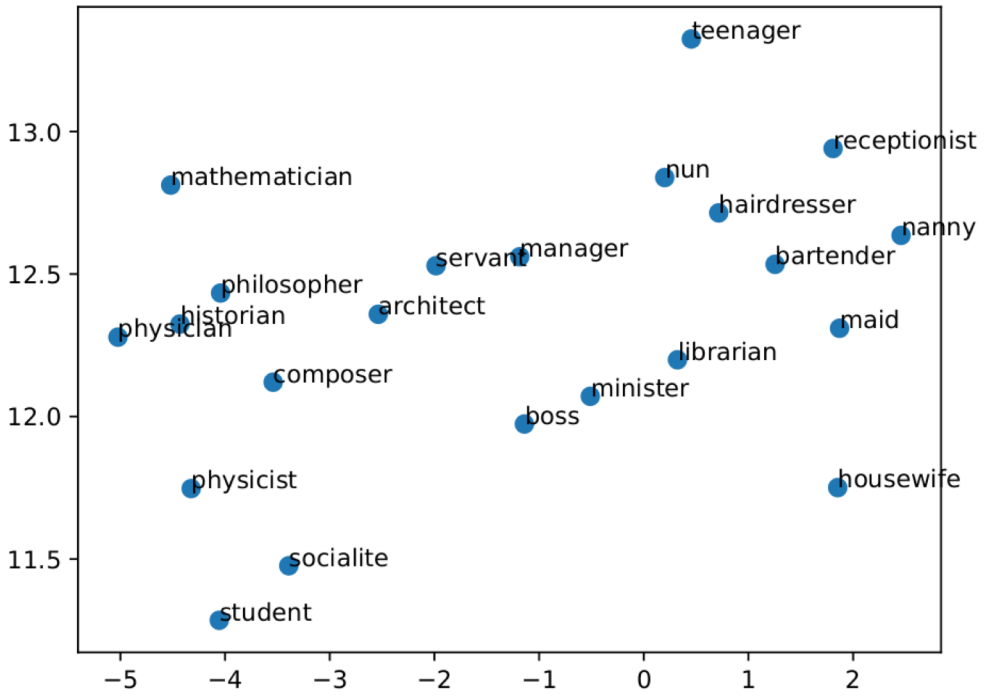
Challenges and research directions in Neural Machine Translation

multilingual, unsupervision, fairness

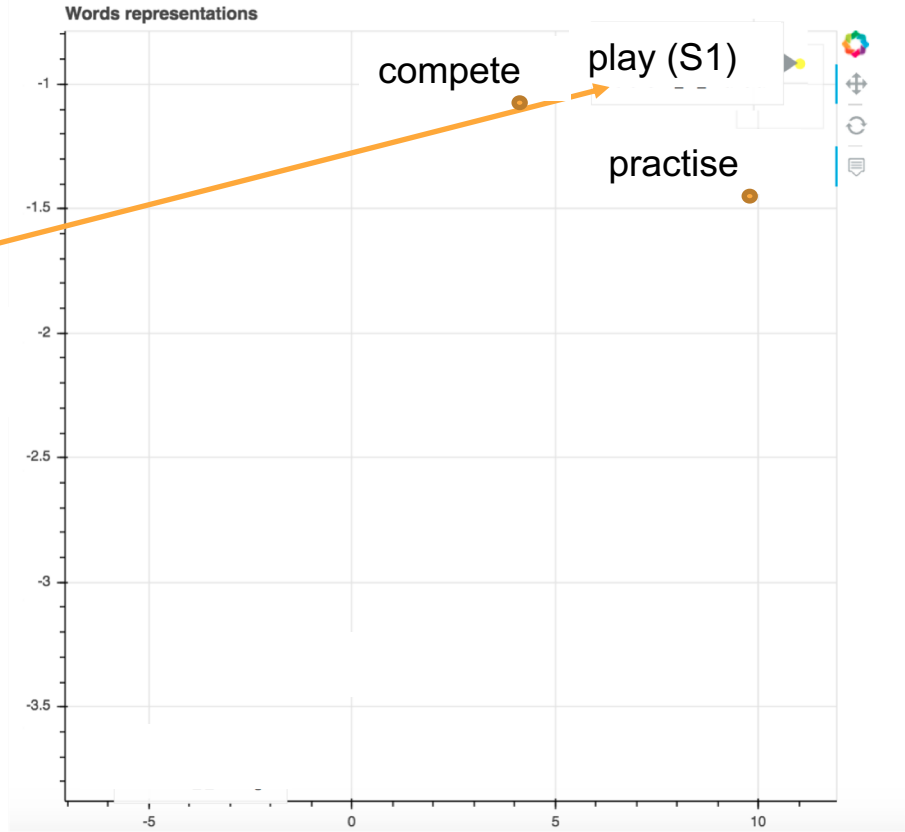
Marta R. Costa-jussà
TALP Research Center, Universitat Politècnica de Catalunya, Barcelona

Words Embeddings



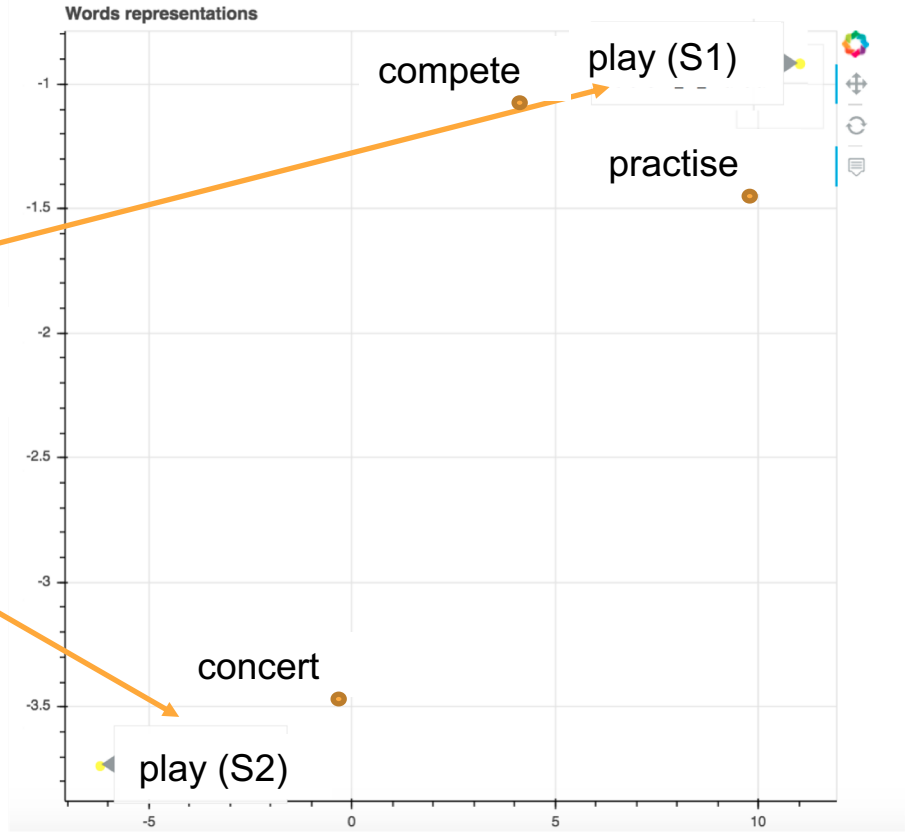
Contextual Words Embeddings

S1 I play football

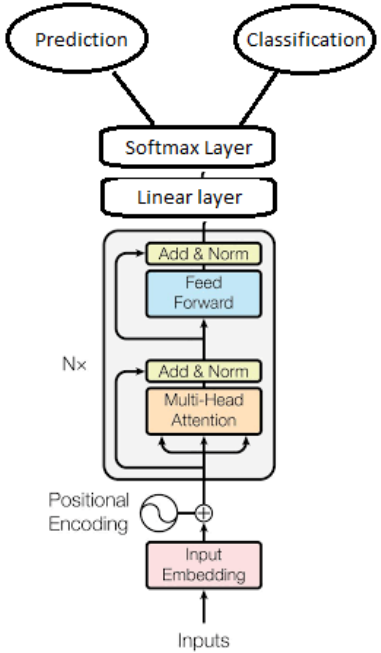


Contextual Words Embeddings

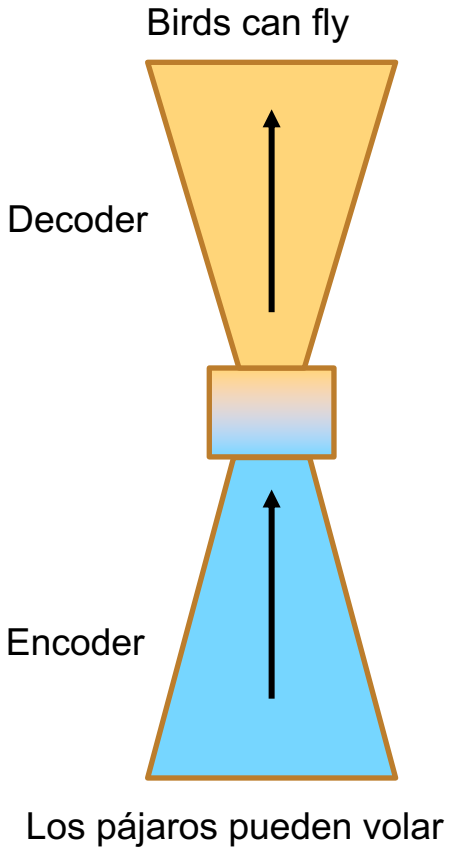
S1 I play football
S2 I attended a play in the big theatre



Contextual Word Embeddings use Transformers

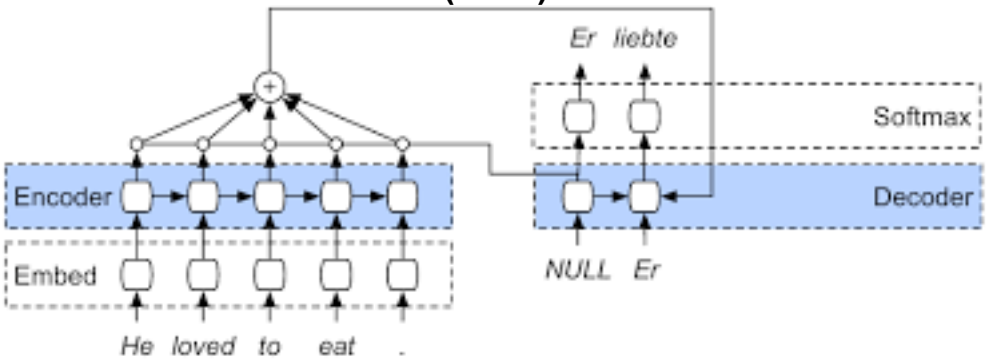


Neural Machine Translation



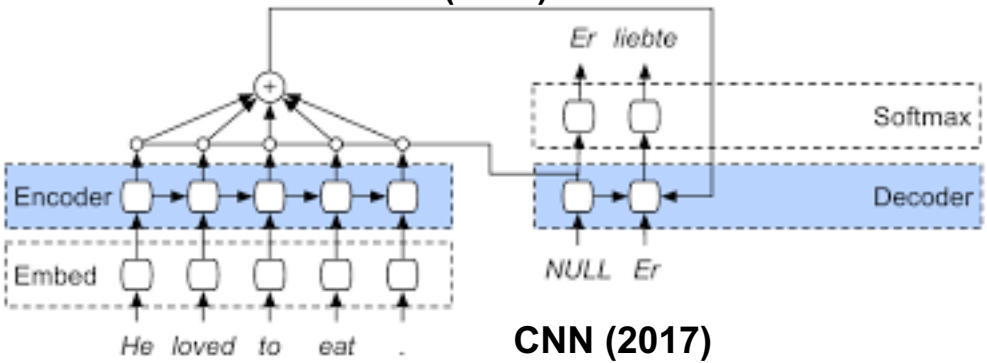
Three different architectures in 2 years

RNN WITH ATTENTION (2015)

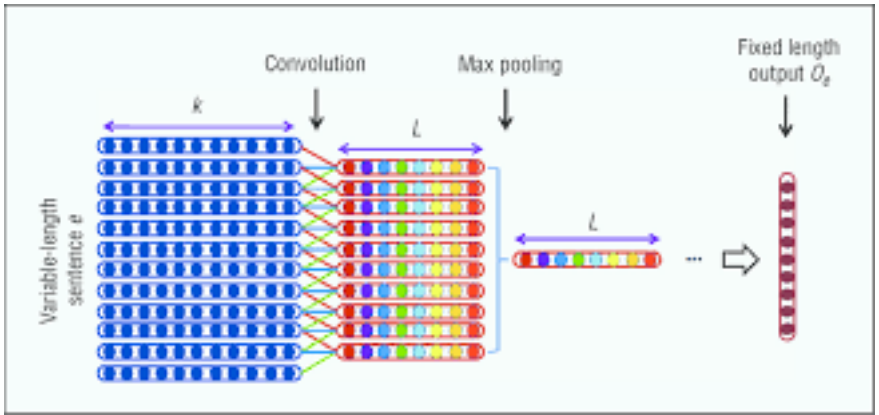


Three different architectures in 2 years

RNN WITH ATTENTION (2015)

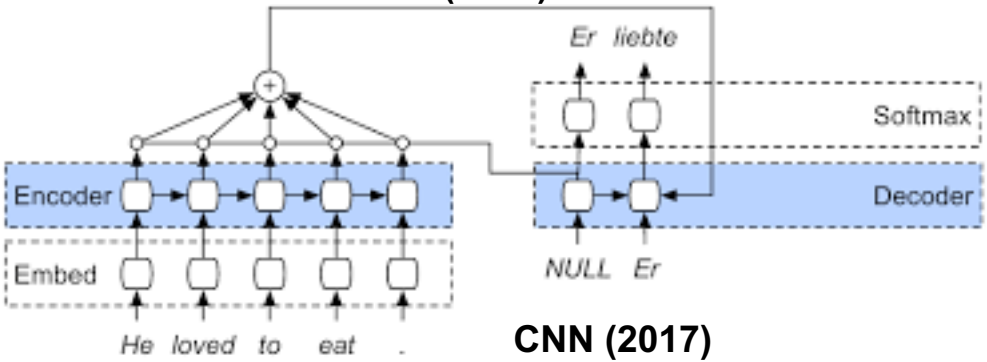


CNN (2017)

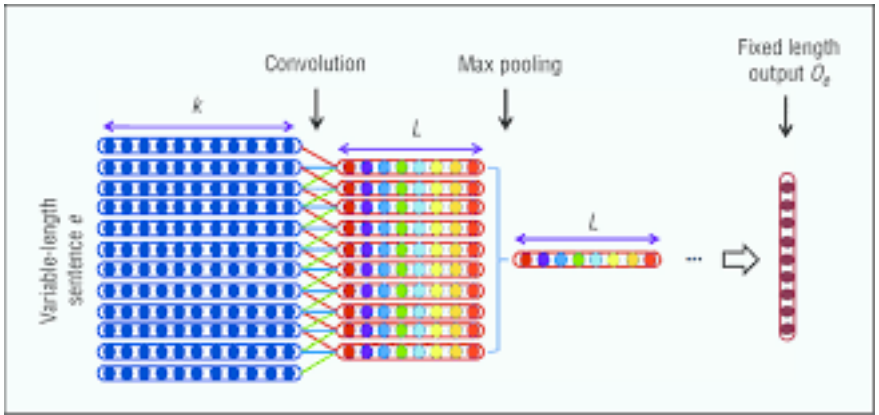


Three different architectures in 2 years

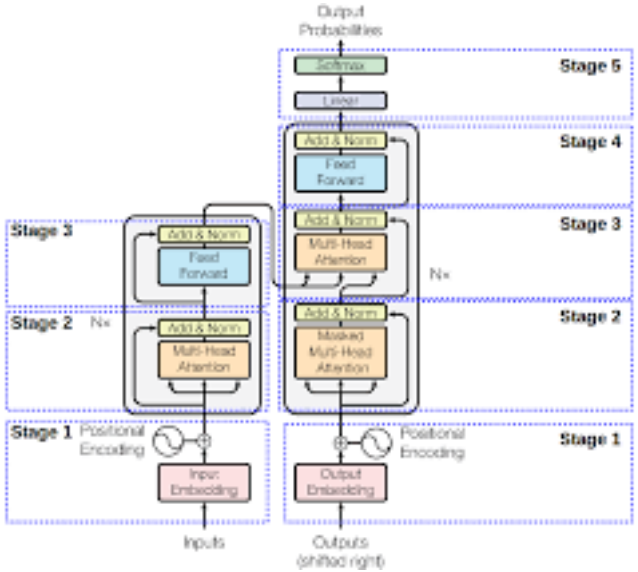
RNN WITH ATTENTION (2015)



CNN (2017)

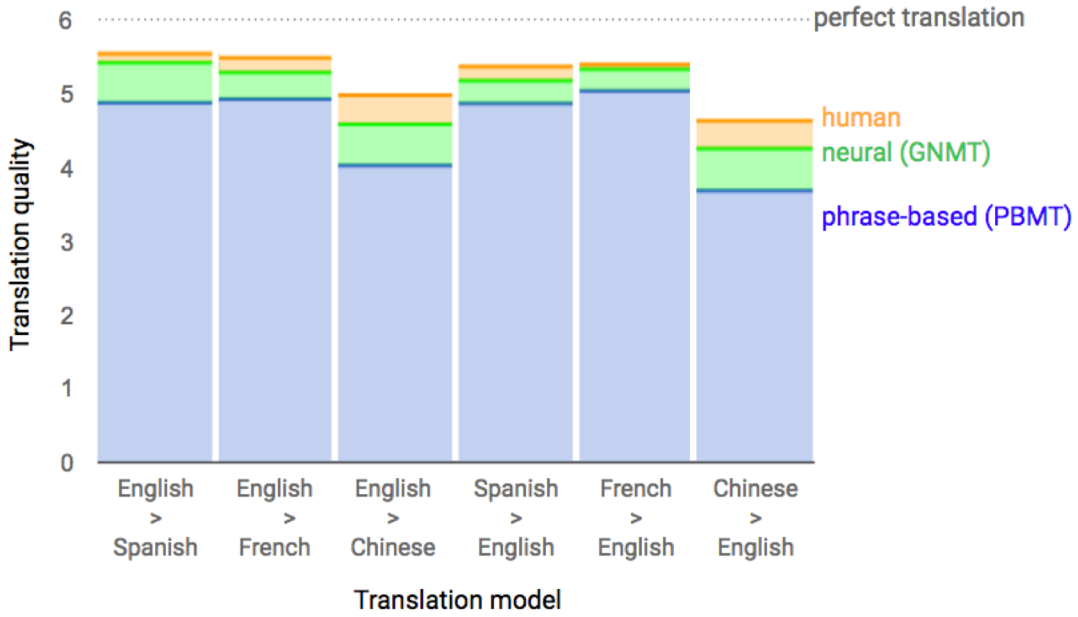


TRANSFORMER (2017)



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

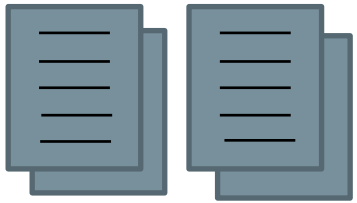
Is Machine Translation Solved?

Quality in Machine Translation depends on training data

Parallel Text

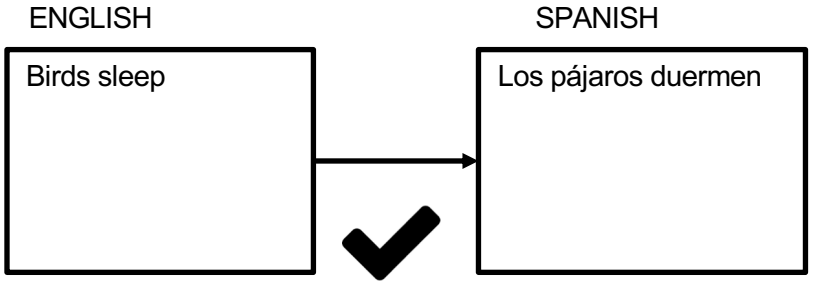
English	Spanish
Birds can fly	Los pájaros pueden volar
Dog sleep	Los perros duermen

Amount of Parallel Data



XXL

Translation Quality



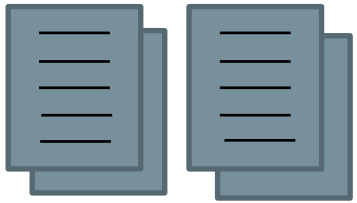
Quality in Machine Translation depends on training data

Parallel Text

English	Spanish
Birds can fly	Los pájaros pueden volar
Dog sleep	Los perros duermen

Turkish	Basque
Kuşlar uçabilir	hegaztiak hegan

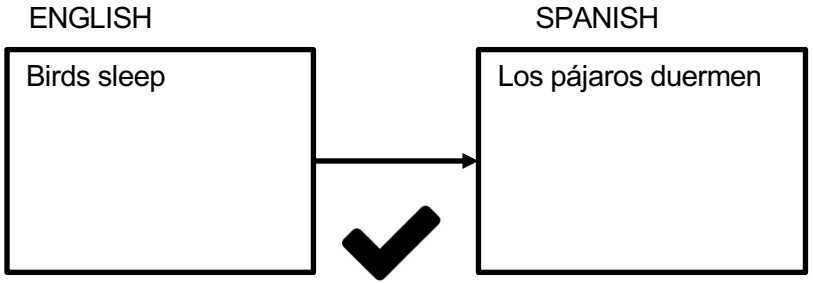
Amount of Parallel Data

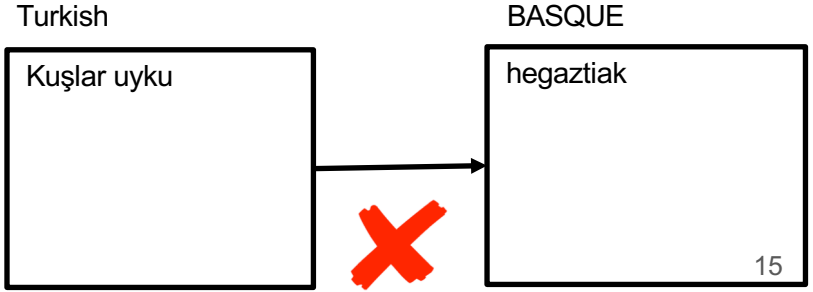


XXL

XXS

Translation Quality



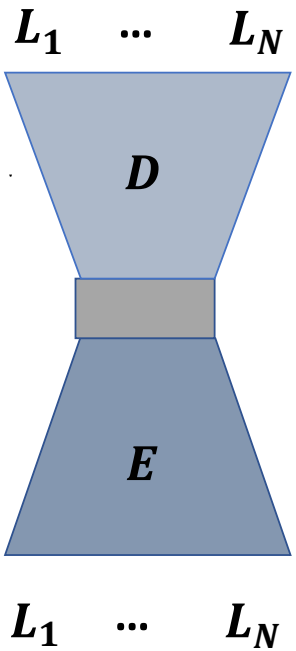


Research Directions

Multilingual systems & Unsupervised systems

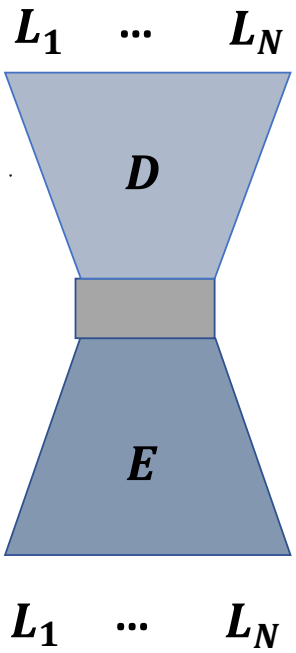
Low-resourced languages can benefit from high-resourced

Universal Encoder-Decoder

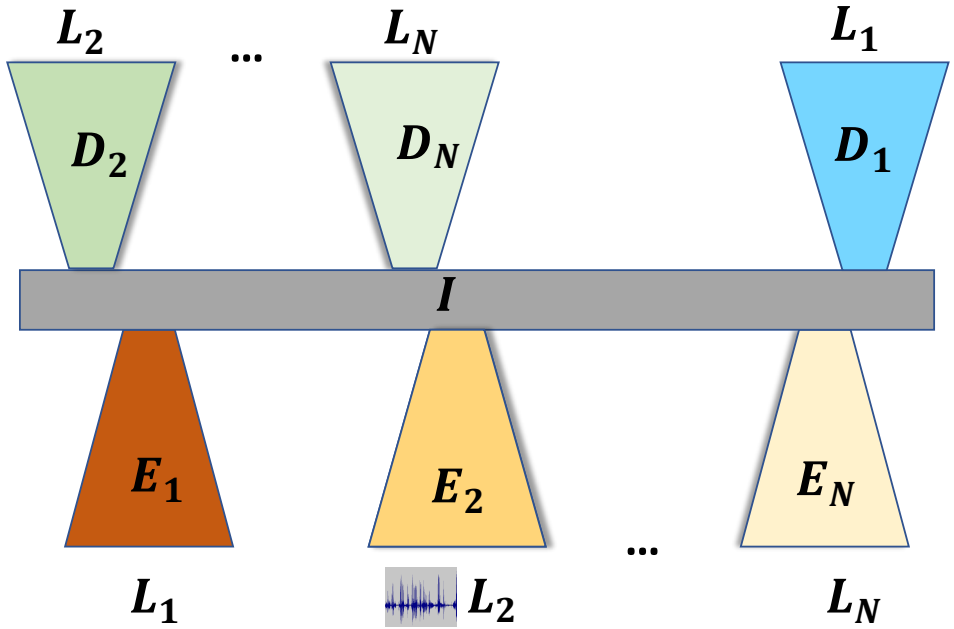


Low-resourced languages can benefit from high-resourced

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

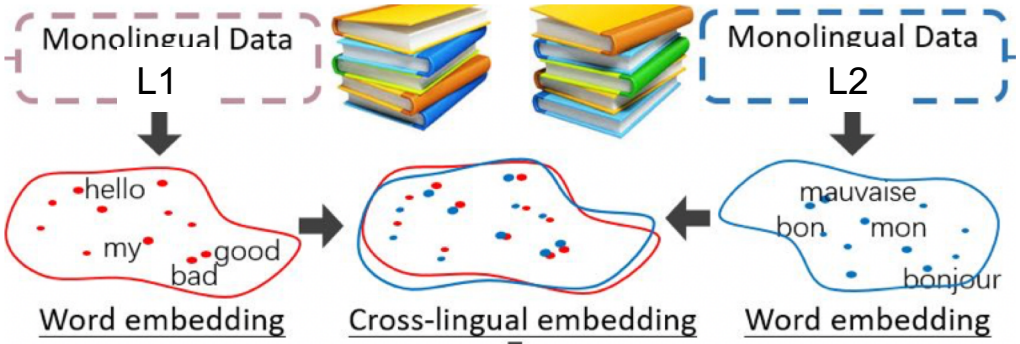
Language-Specific Encoder-Decoders

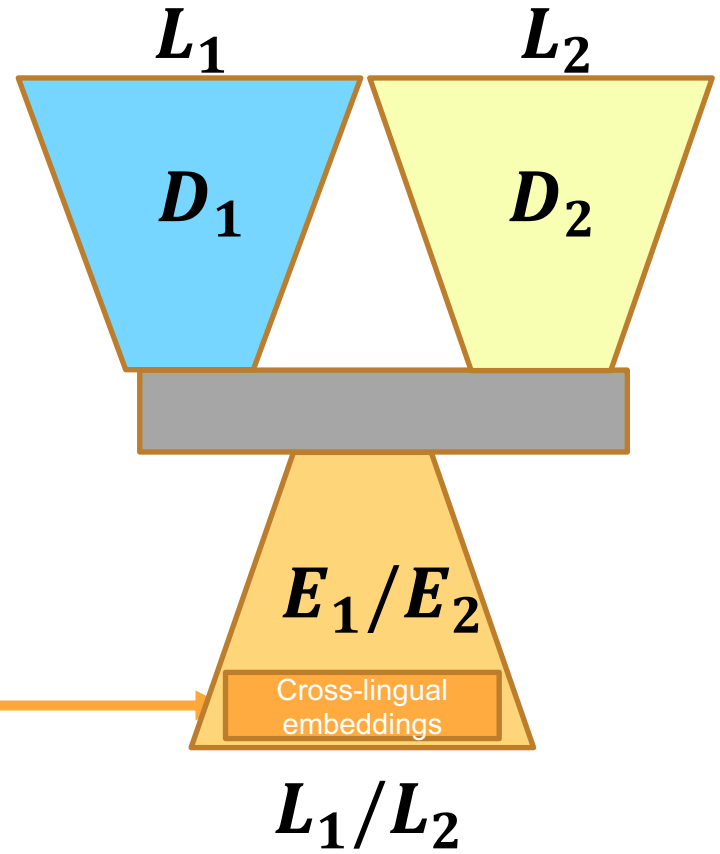
- ➢ Independent vocabulary

- ✓ Zero-shot
- ✓ Incremental training of new languages and domains

- x No transfer learning from high-resourced to low-resourced (with the same script)

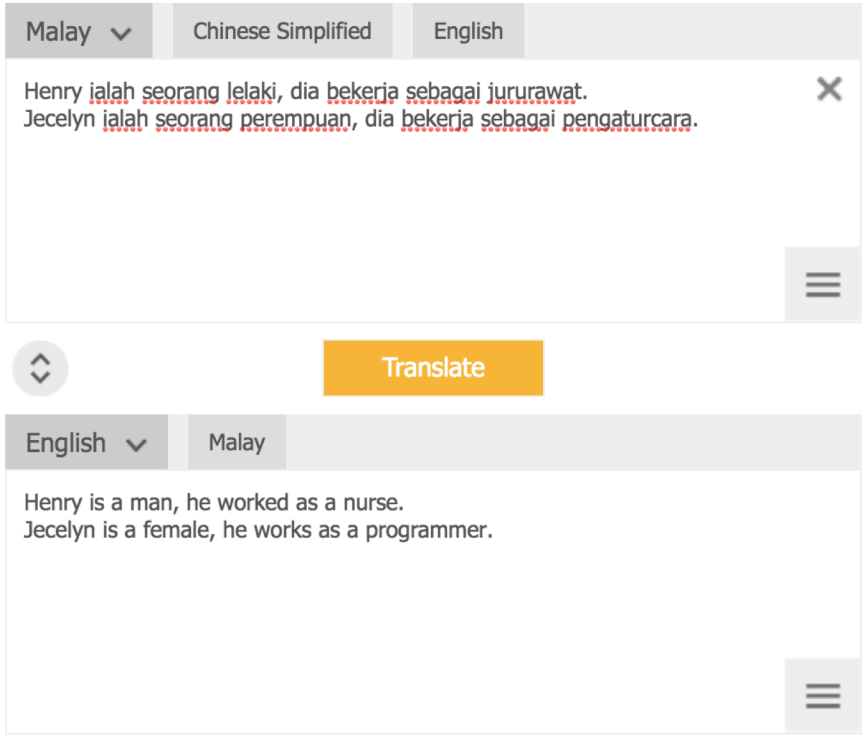
Machine Translation can be trained on unlabelled data

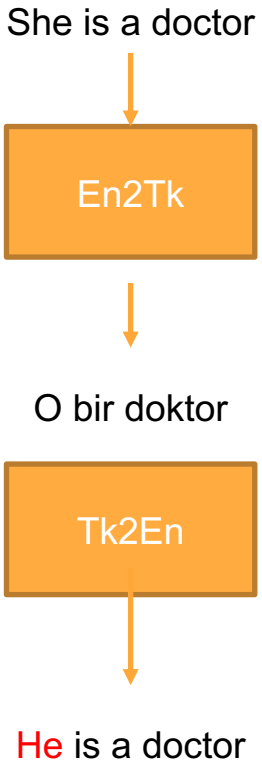




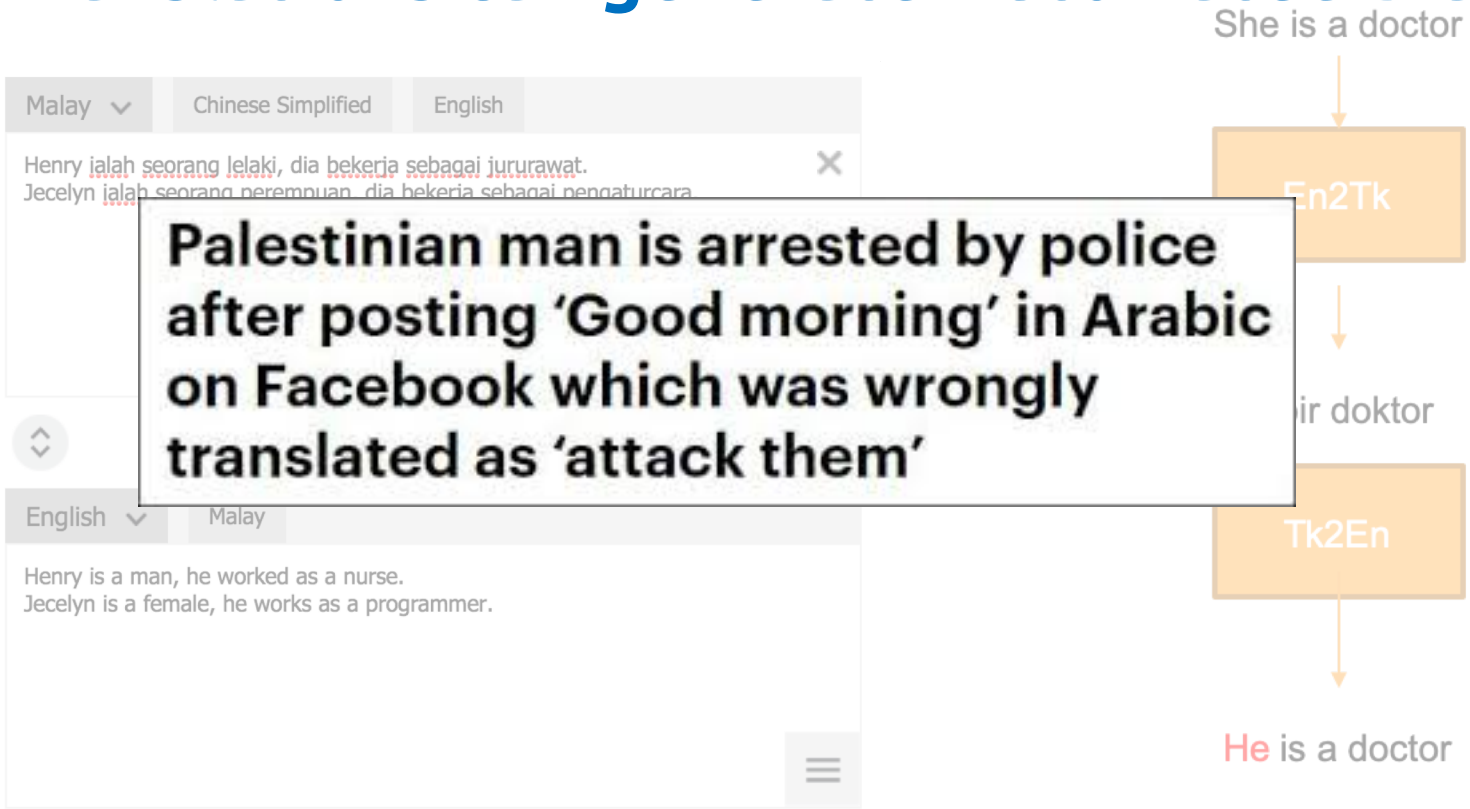
Questioning our data…

Biased Translations





Bad Translations can generate Automated Bias

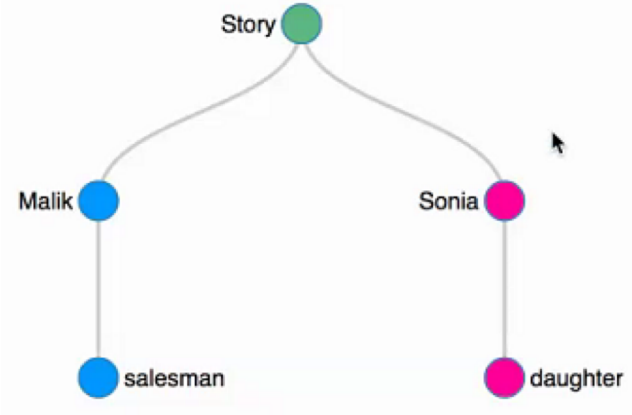


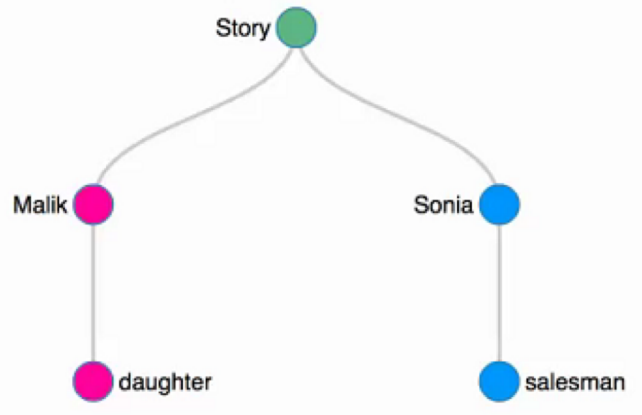
Towards Fairer Systems

Data augmentation & Debiasing algorithms

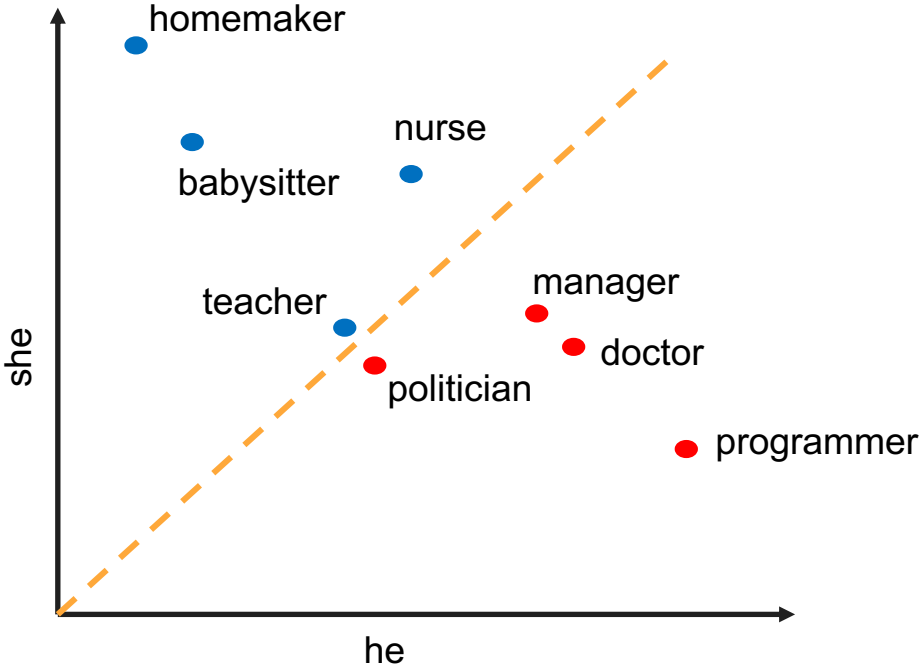
Data augmentation

Biased Text	Debiased Text
<Malik:he> is an aspiring singer who works as a salesman in a car showroom. One day he meets <Sonia:she> Saxena daughter of Mr. Saxena when goes to deliver a car to home as birthday present	<Malik:she> is an aspiring singer who works as a salesman in a car showroom. One day he meets <Sonia:he> Saxena daughter of Mr. Saxena when goes to deliver a car to home as birthday present

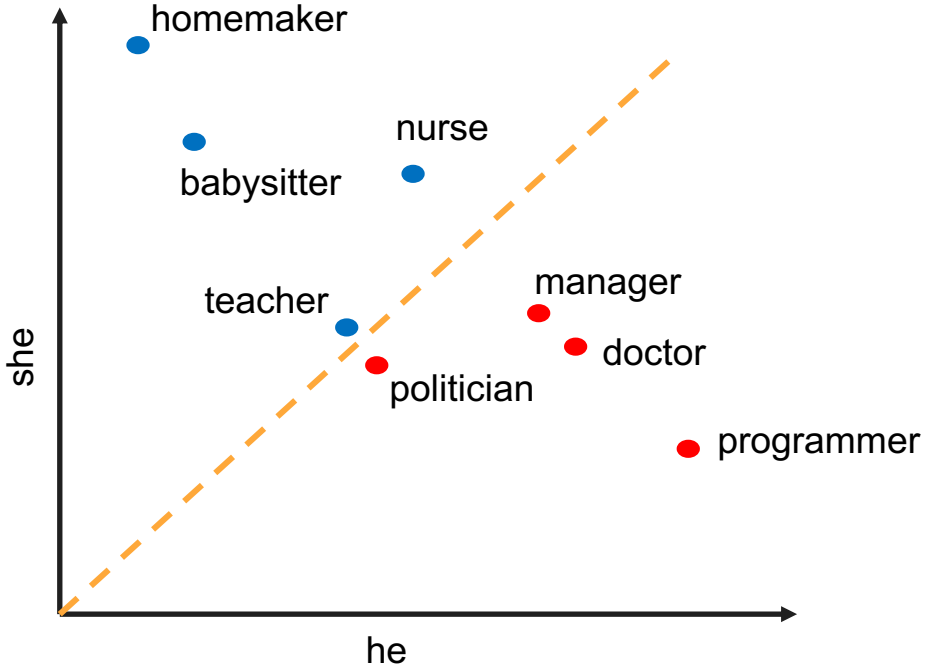


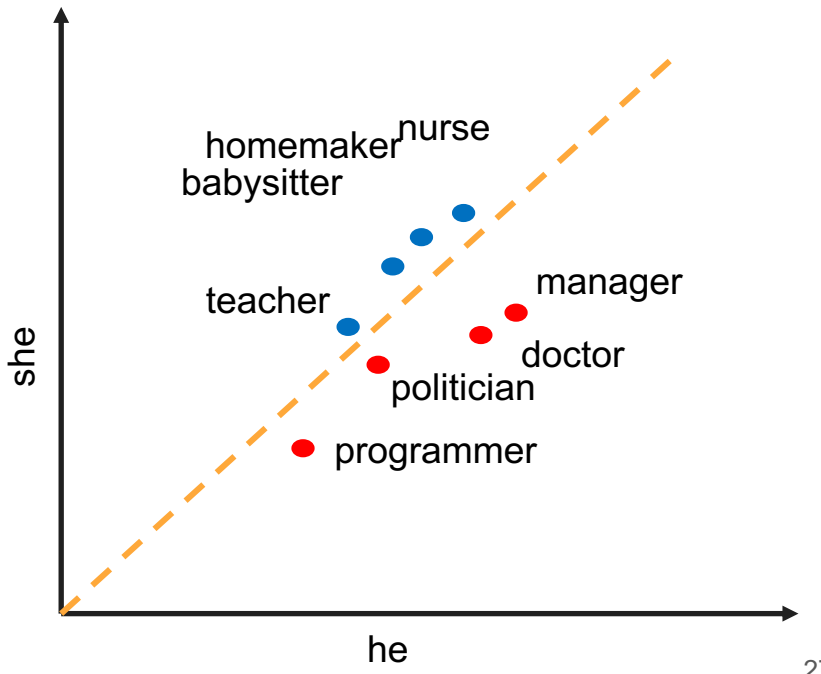


Debiased algorithms



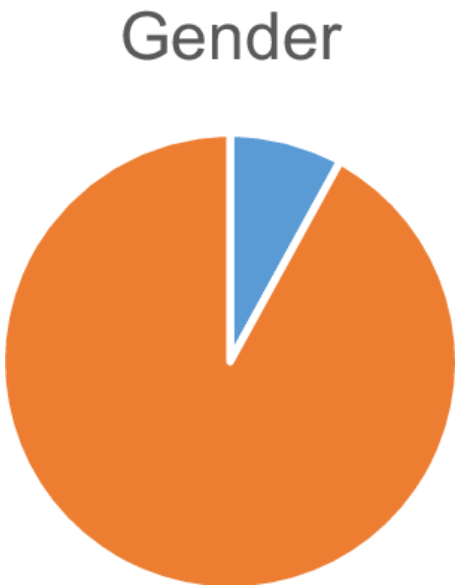
Debiased algorithms

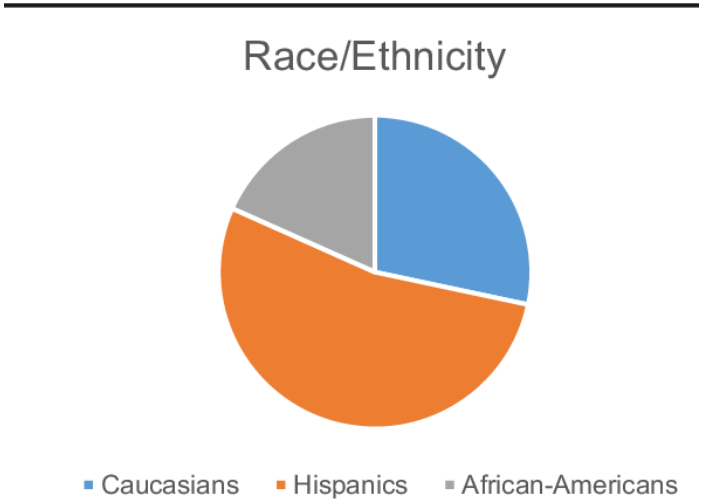




Inspiration from other areas… healthcare?

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