



explicharr

automatic text simplification with deep learning

team 3 - explicharr

Kuan Yu – Maya Angelova – Philipp Schoneville - Sonu Rauniyar

- train (translation) model for text simplification task
- Wikipedia dataset & character-level Transformer

dataset

- Wikipedia dataset
 - 285.000 aligned sentence pairs
 - normal to simplified text

normal

simple

-
- | | | |
|---|--|--|
| 1 | She then went on to study at the Central School of Speech and Drama in 1977 , where she met her future comedy partner , Jennifer Saunders . | French studied acting at the London Central School of Speech and Drama , and met Jennifer Saunders there in 1977 . |
| 2 | It has been depicted with brownish-grey plumage , yellow feet , a tuft of tail feathers , a gray , naked head , and a black , yellow , and green beak . | They had gray feathers and yellow feet . |
| 3 | Paramedics provide advanced levels of care for medical emergencies and trauma . | They provide care for medical emergencies and trauma . |
| 4 | By the 1750s , the suite had come to be seen as old-fashioned , superseded by the symphony and concerto , and few composers were still writing suites during that time . | By the 1750s composers had stopped writing suites . |

normal

simple

-
- | | | |
|---|--|--|
| 1 | The losing team gets zero points . | A team gets 3 points for a win . |
| 2 | The Trump Building is a 70-story skyscraper in New York City . | The Trump Building is a skyscraper in New York City , United States . |
| 3 | The tower is the tallest mid-block building in New York City . | Before it was built , the tallest building in the world was the Woolworth Building . |
| 4 | The second season of Bad Girls Club premiered on December 4 , 2007 , on Oxygen . | The Bad Girls Club season 2 is the second season of The Bad Girls Club . |

- many of the sentences in the dataset are unchanged

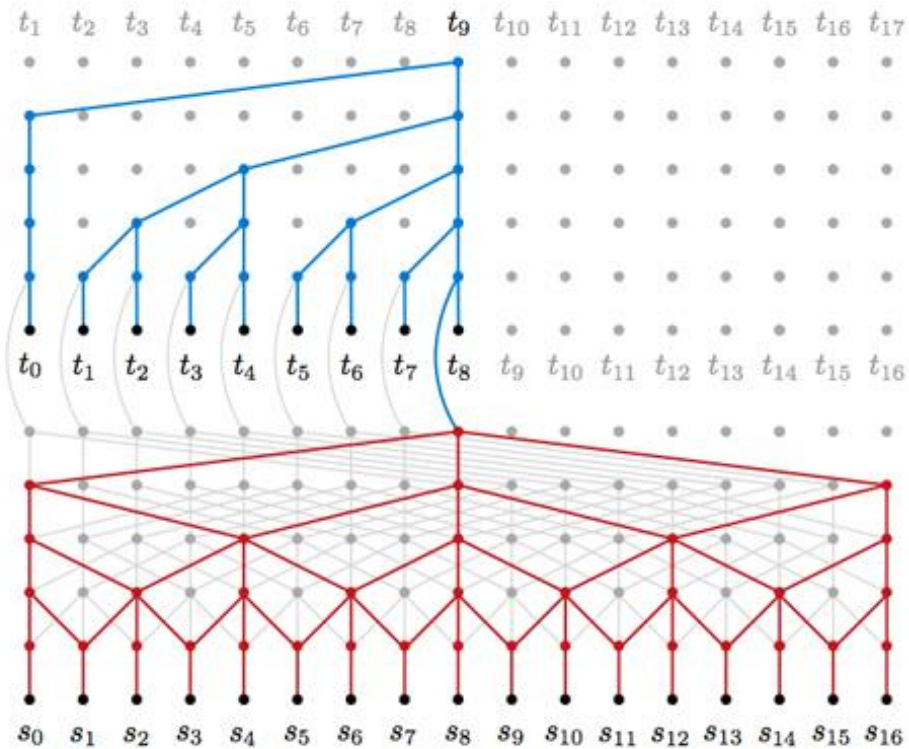
character-level modelling

	normal	simple
#word	4.284.135	3.832.824
#word-type	162.491	147.078
#hapax	83.997	76.049
#char-type	133	134

- more robust
 - many words don't occur in simple data -> translate to UNKNOWN
 - no special treatment for numbers
 - may learn morphology
- no tokenizer required
- easier applicable to other languages

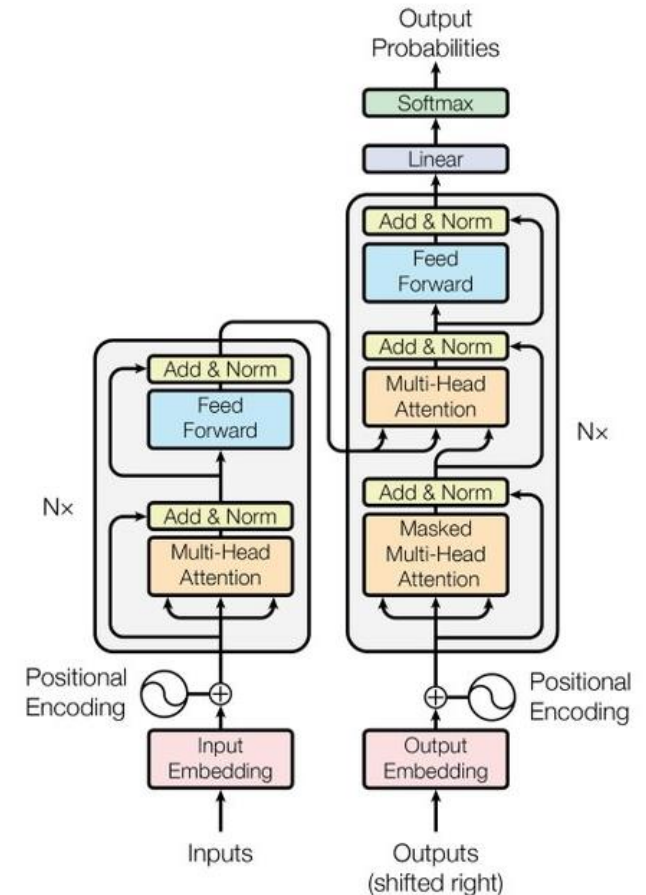
model choice

- ByteNet
 - as a baseline
 - already character-level



model choice

- Transformer
 - state of the art for word-level translation model
 - faster training than RNN architectures
 - no information bottleneck
 - compared to CNNs there is no limited input field

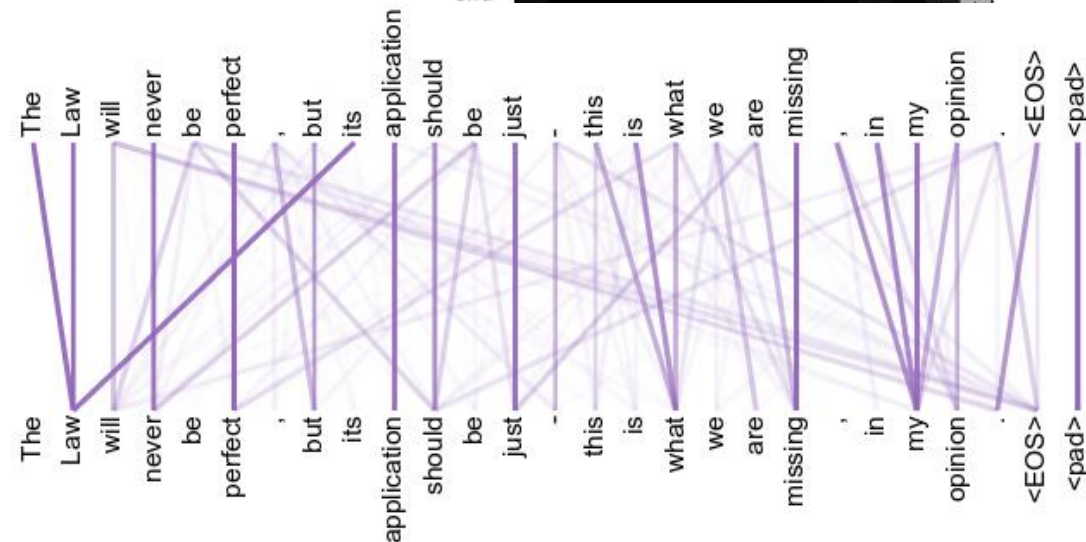
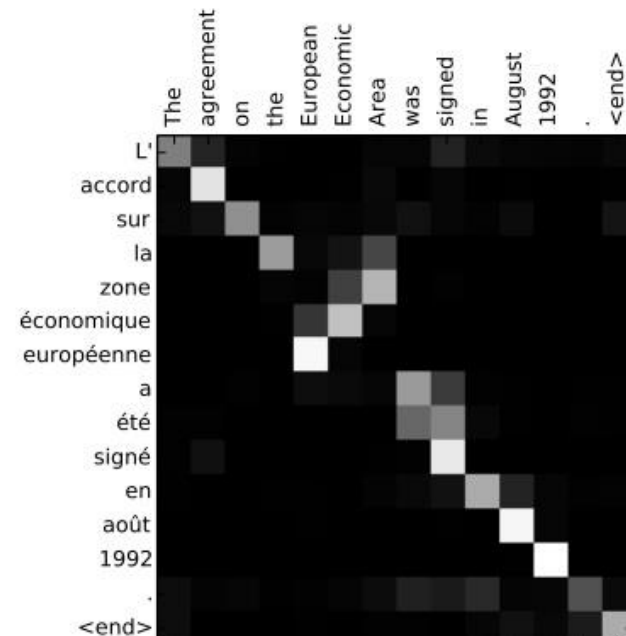


exploration(1)

- data
 - find more (and better) data
 - make adjustments to existing data set (e.g. unchanged sentences)
- evaluation
 - alternatives to BLEU score
 - other team's solution
 - character level evaluation: e.g. chrF, characTER

exploration(2)

- attention mechanisms
 - look at different forms of attention mechanisms
- visualization
 - visualize the model's attention
 - use for error analysis



exploration(3)

- pre-training
 - different architectures (e.g. autoencoders, generative models)
 - different data (e.g. encoder: normal wikipedia – decoder: simple)
 - different parts (e.g. only embeddings)

exploration(4)

- word-level and subword-level modelling
- beam search
- (conditional random field)

time plan

1

end of May

- get baseline on ByteNet and Transformer

2

June

- diverge and explore:
 - attention mechanism / visualization - Maya
 - data / evaluation - Sonu
 - pre-training - Philipp
 - word-level or subword-level modelling / beam search - Kuan

time plan

3

July

- combine design choices
- train and evaluate final model(s)

4

end of July

- presentation and writing paper

THANK YOU FOR YOUR
ATTENTION