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
  - version “3.0“

#	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4	By the 1750s , the suite had come to be seen as old-fashioned , superseded by the symphony and concerto , and few composers were still writing suites during that time .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

dataset stats

	normal	simple
#word	7,400,555	5,634,887
#char	40,242,640	29,680,984
#word-type	212,292	165,170
#hapax	109,988	82,487
#mean word per sent	26.00	19.79
#mean char per sent	141.36	104.26
#mean char per word	5.43	5.26

- simple is ~25% shorter
- mostly due to shortened sentences, partly shorter words
- many words occur only once and have to be treated as UNKNOWN

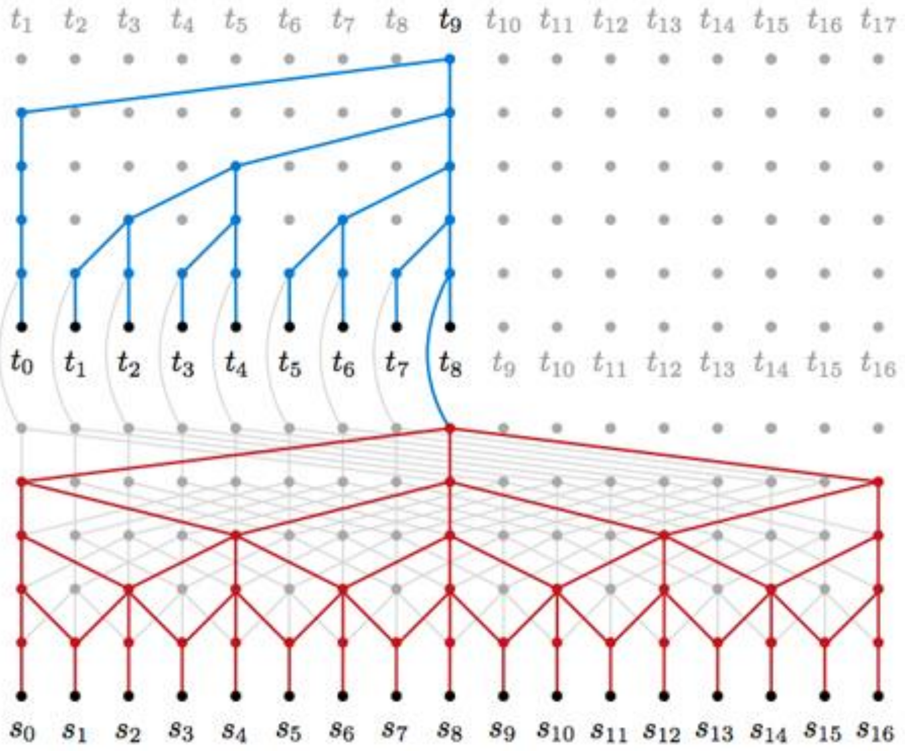
character-level modelling

	normal	simple
#word	7,400,555	5,634,887
#char	40,242,640	29,680,984
#word-type	212,292	165,170
#hapax	109,988	82,487
#char-type	2,880	2,359
%top 255 chars	99.97	99.97

- more robust
  - UNKNOWN chars only make up 0.03% of text
  - no special treatment for numbers
  - may learn morphology
- no tokenizer required
- easier applicable to other languages

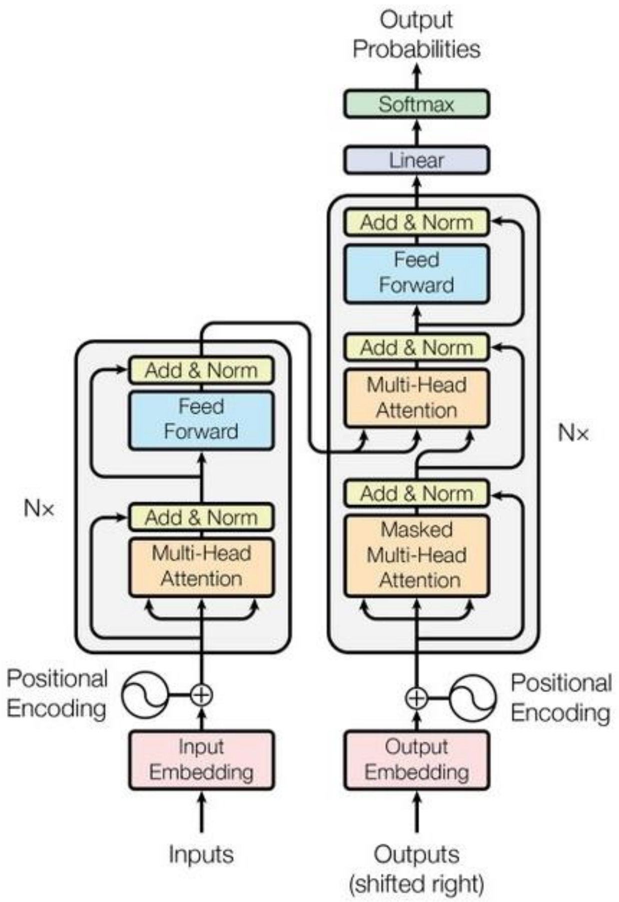
model choice

- ByteNet
  - as a baseline
  - already character-level


t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13} t_{14} t_{15} t_{16} t_{17}
t_0 t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13} t_{14} t_{15} t_{16}
s_0 s_1 s_2 s_3 s_4 s_5 s_6 s_7 s_8 s_9 s_{10} s_{11} s_{12} s_{13} s_{14} s_{15} s_{16}

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

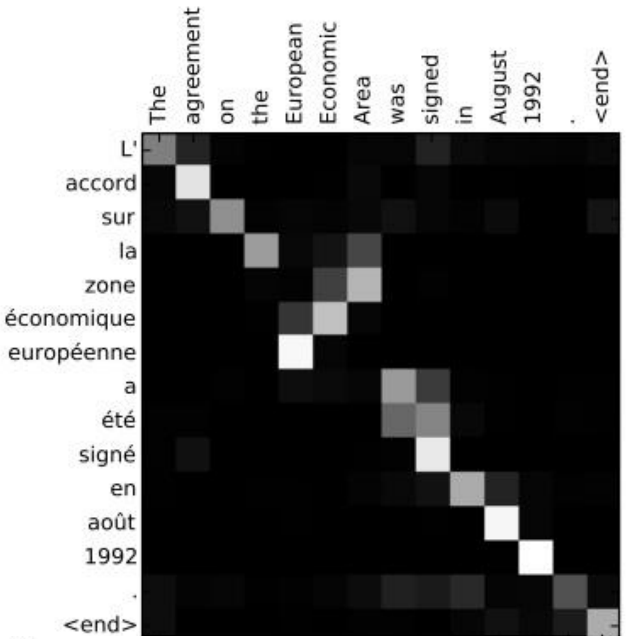


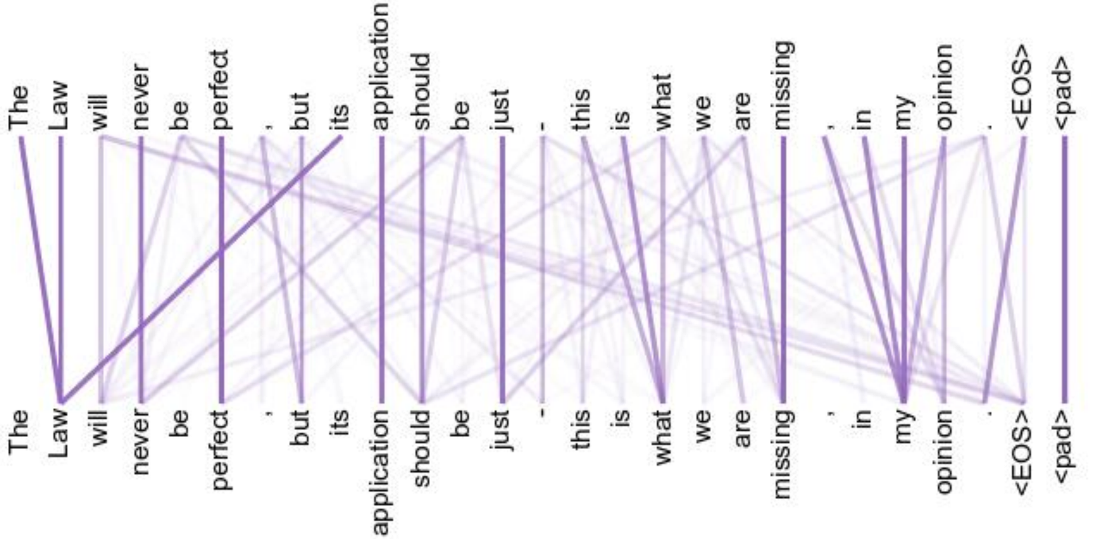
exploration(1)

- 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 - <u>Maya</u>
  - data / evaluation - <u>Sonu</u>
  - pre-training - <u>Philipp</u>
  - word-level or subword-level modelling / beam search - <u>Kuan</u>

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