explicharr

Kuan Yu •Maya Angelova • Philipp Schoneville • Sonu Rauniyar

06.06.2018

team 3 - explicharr

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

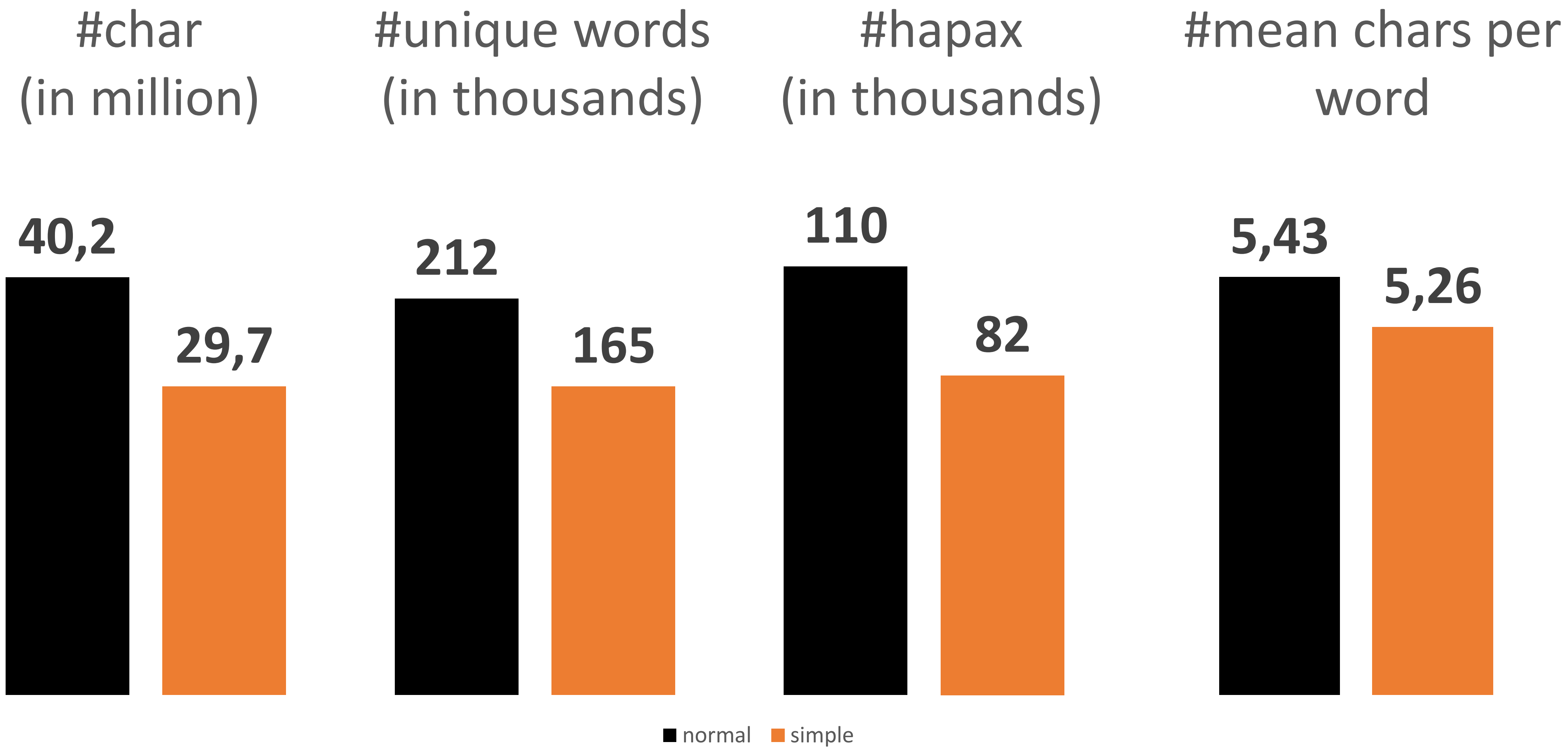
dataset

- Wikipedia dataset
  - version “3.0“
  - 285.000 aligned sentence pairs
  - normal to simplified text

#	normal	simple
1	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 .
2	Paramedics provide advanced levels of care for medical emergencies and trauma .	They provide care for medical emergencies and trauma .
3	By the 1750s , the suite had come to be seen as old-fashioned , superseded by the symphony and concerto , and few composers were still writing suites during that time .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 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 .
3	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



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

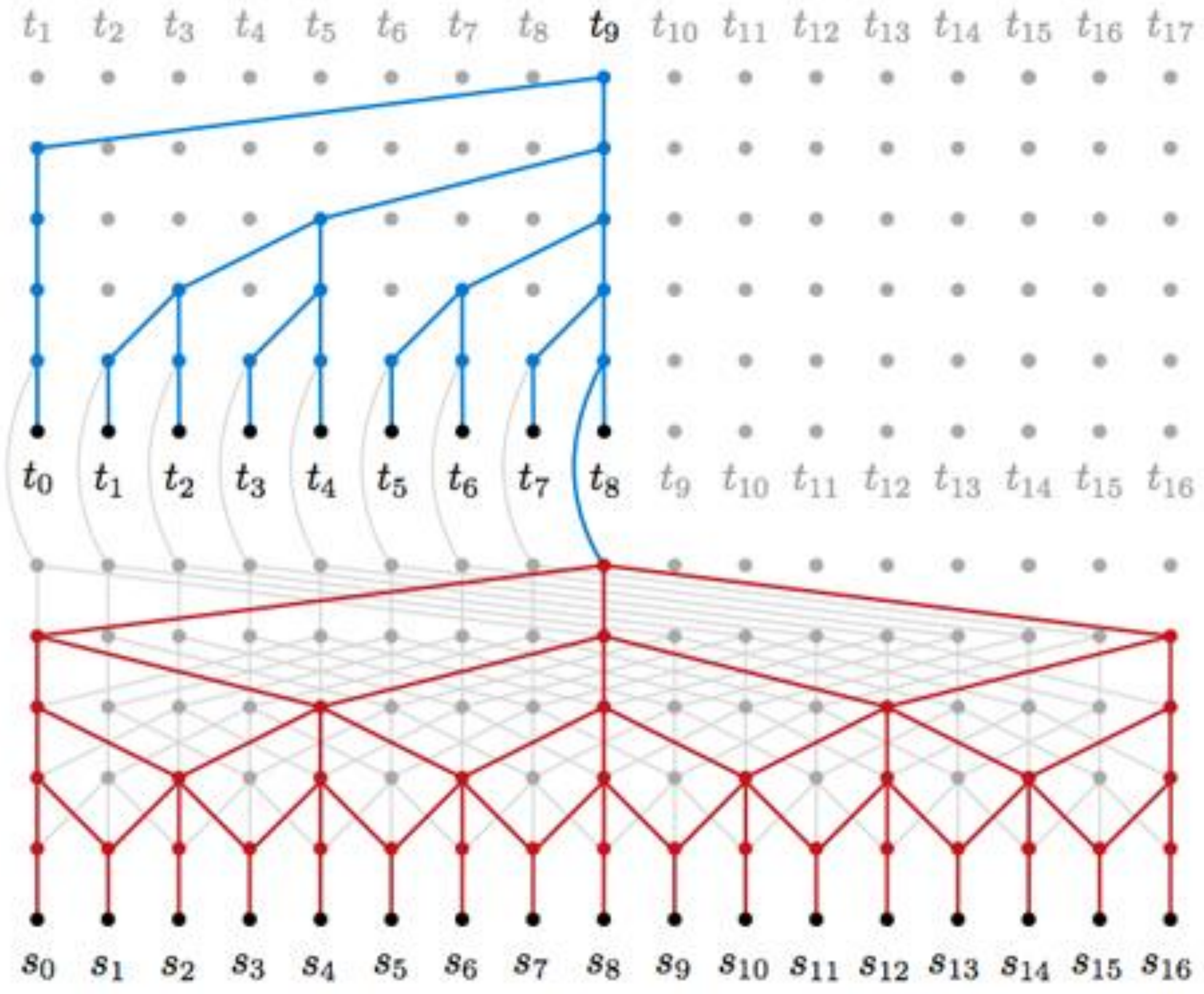
character-level modelling

	normal	simple
#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

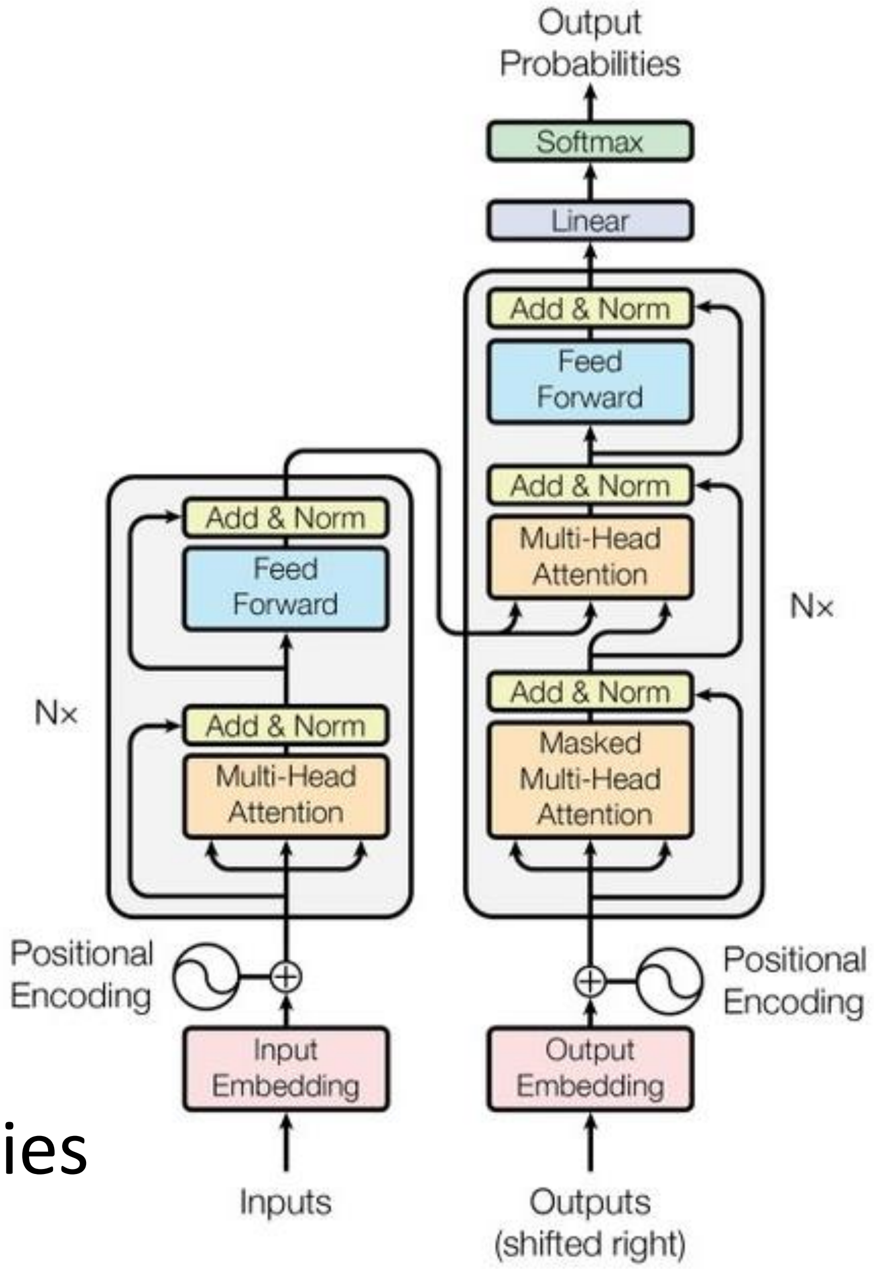
- ByteNet
  - Convolutional architecture
  - as a baseline
  - already character-level


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 word-level translation model
  - faster training than RNN architectures
  - no information bottleneck

  - compared to CNNs: no limited receptive field
  - shorter path length for long range dependencies



#	input	output
1	Some of the largest reservoirs in the world can be found along the Volga .	Some of the largest that is found in the world , including the United States , Australia , Russia , Italy , Italy , Italy , Italy , Italy , Italy , Italy , Italy , Italy , Italy , Russia , Russia , Russia , Russia , Russia , Russia , Russia , Russia , Russia , Russia , Russia ,
2	Furthermore , spectroscopic studies have shown evidence of hydrated minerals and silicates , which indicate rather a stony surface composition .	Some scientists think that that when they are moving around the surface of the Earth , they can also be seen when they are moving around the surface of the Earth , or that is seen is a star that is about a person who finds that is .

Transformer
Bleu score: 0.1530

#	input	output
1	Jeddah is the principal gateway to Mecca , Islam 's holiest city , which able-bodied Muslims are required to visit at least once in their lifetime .	1825 is the main gateway to mecca , islam 's holiest city .
2	Convinced that the grounds were haunted , they decided to publish their findings in a book An Adventure (1911) , under the pseudonyms of Elizabeth Morison and Frances Lamont .	1825 decided to publish their findings in a book an adventure (1911) .

Transformer (Pytorch)
Bleu score: 0.2544

1825 was born in london .

1825 was put on christmas eve in 1890 .

1825 is a band from the united states .

1825 was a law enforcement agency in new york city .

1825 's clouds are made of ice .

1825 are very rare .

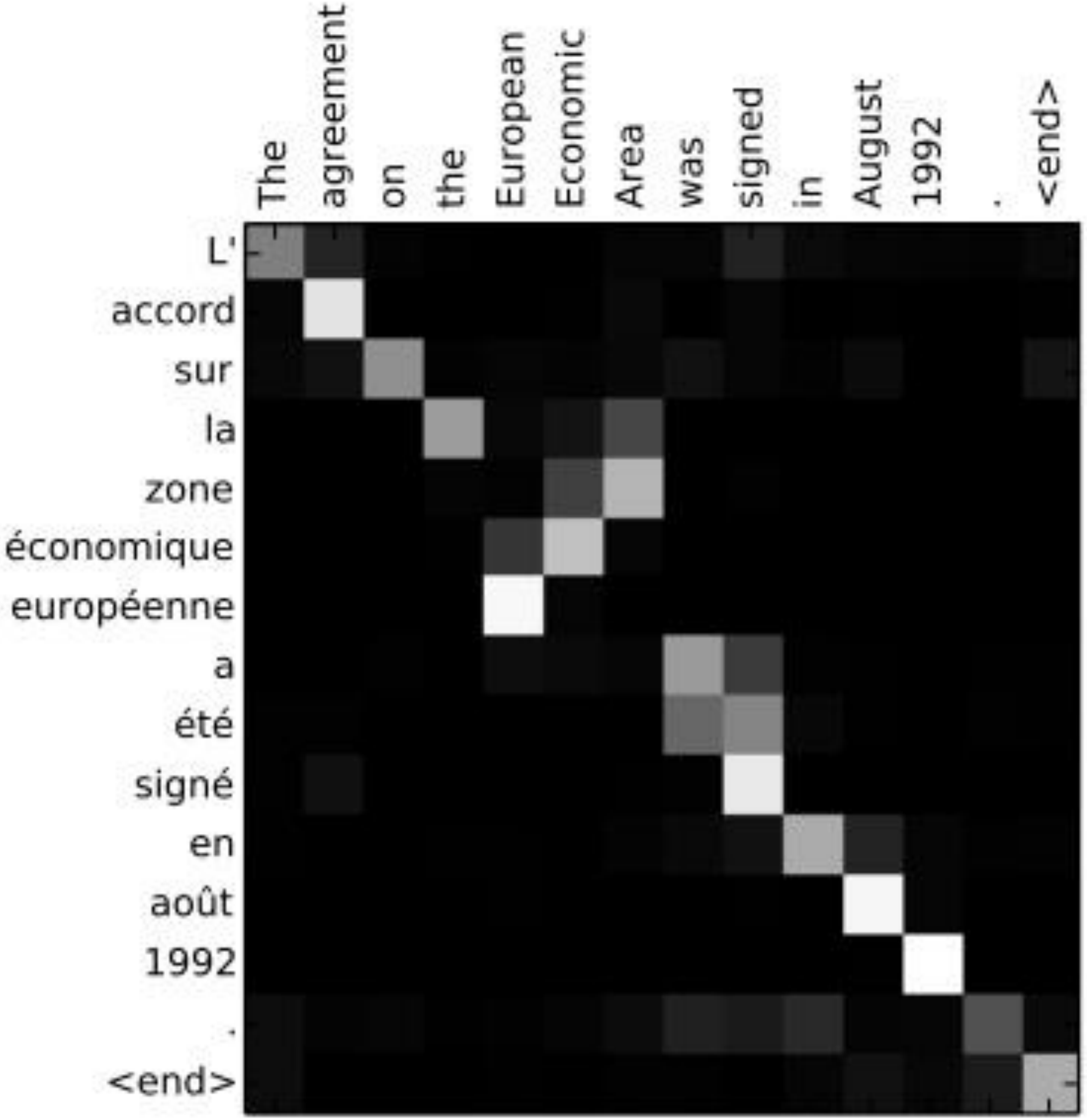
1825 left the newly conquered region .

exploration(1)

- data
  - find more (and better) data (e.g. newsela)
  - 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 and introspection



exploration(3)

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