

# TEAM - 3

## SENTENCE SIMPLIFICATION WITH CHARACTER-LEVEL TRANSFORMER

Kuan Yu • Sonu Rauniyar • Maya  
Angelova • Philipp Schoneville

Faculty:  
Dr. Stober & Prof.Dr.  
Manfred Stede

## TEAM 3 - “EXPLICHARR”

- Text simplification can be regarded as translation from a language into a sublanguage with reduced linguistic complexities
- System must know distributional properties of source and target languages & preserve mapping of semantic and discourse structures

# DATASET : WE USED !

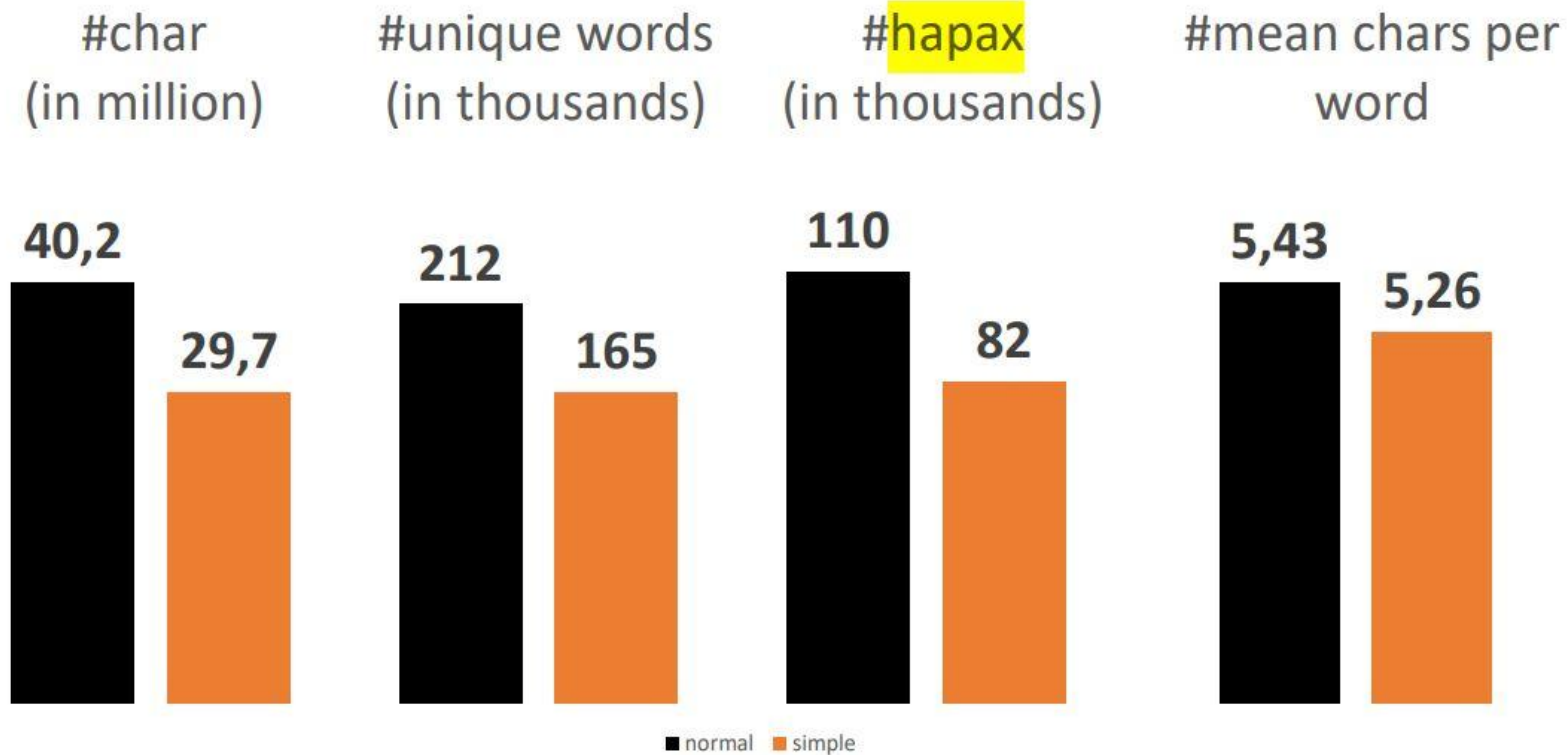
## ◉ Wikipedia dataset

- version “3.0”
- 285.000 aligned sentence pairs from Standard to Simple Wikipedia dataset
- Used classified ones as “good” and “good partial” by Hwang et al.

## ◉ Source :

Hwang, William et al. (2015). “Aligning sentences from standard wikipedia to simple wikipedia”.

# STATS OF THE DATASET



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

# CHAR LEVEL MODEL

- ◉ Zipfian distribution in Natural Language makes it hard to model vocabularies
- ◉ Sentence was treated on char level rather than word
- ◉ Top 256 most frequent ones account for 99.97% of the character count
- ◉ Removed aligned sentences which are identical or longer than 256 characters
- ◉ Finally left with 226,208 instances, used random 1% as validation set & rest training set

# FEW FACTS ABOUT USED “WIKI” DATASET

- ◉ Many target sentences don't make sense at all, it confuses the model :P
- ◉ what model could learn was “**limited**” by this dataset!
- ◉ Argued that wiki dataset isn't ideal for Text simplification (*Xu, Wei, Chris Callison-Burch, and Courtney Napoles (2015). “Problems in current text simplification research: New data can help”*)

# EVALUATION

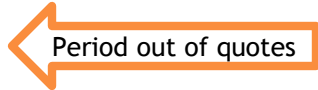
- ◉ Matching the source sentences against the target sentences gave a high BLEU score of 29.53
- ◉ Judged the goodness of learning by monitoring the validation loss and accuracy.
- ◉ Our Baseline model, word-level transformer had a BLEU score of 25.44

# TEST RESULTS OF THE “MODEL EXPLICHARR”

- All tests results are from validation set
- Model never saw these sentences while training
- about 80% of results are exact copies of source Sentence

Text normalization:

Testing -> “period out of quotes”

- s: Punch 's wife was originally called `` Joan . "
  - p: Punch 's wife was originally called `` Joan " . 
  - t: Punch 's wife was originally `` Joan " .
- s: source sentences
  - t: target sentences
  - p: predicted sentences



# TEXT NORMALIZATION

## ❖ Testing -> “Bracketing”

- s: Buddha-Bhagavan ) .

- p: Buddha-Bhagavan .



Perfectly removes the brackets

- t: Buddhists believe there have been many Buddhas who were alone , called pacceka-buddhas .

- s: Edo ( 江戸 ?

- p: Edo ( ? )



Perfectly adds the brackets

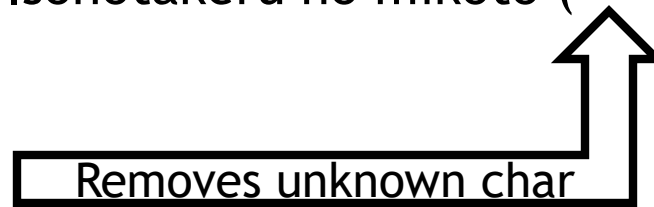
- t: Edo ( 江戸 ? )

- s: source sentences
- t: target sentences
- p: predicted sentences

# TEXT NORMALIZATION

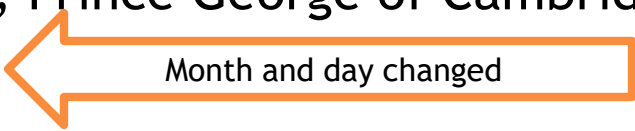
- ❖ Testing -> “unknown characters”
- ❖ Ignores the unknown characters

- s: The enshrined kami is Isonotakeru no mikoto ( 五十猛命 ? )
- p: The enshrined kami is Isonotakeru no mikoto (   ? )



- t: This place is special to the kami named Isonotakeru no mikoto ( 五十猛命 ? )
- s: source sentences
  - t: target sentences
  - p: predicted sentences



# TEXT NORMALIZATION

- ❖ Testing -> “dates”
  - ❖ Perfectly aligns date with the “US date format”
  - s: Their first child , Prince George of Cambridge , was born on **22 July** 2013 .
  - p: Their first child , Prince George of Cambridge , was born on **July 22** , 2013 .

Month and day changed
  - t: On July 24 , it was announced that the baby would be called Prince George of Cambridge , with the full given name George Alexander Louis .
- s: source sentences
  - t: target sentences
  - p: predicted sentences

# TEXT NORMALIZATION

○ Its tries to be smart & produce normal sentences.

- s: **is a** 1982 **arcade**-style platform video game by Nintendo .
- p: **In** 1982 , the style platform video game by Nintendo . 
- t: Donkey Kong Jr. is an arcade video game made by Nintendo in 1982 .
- s: **reflects** the influence of the sexagenary cycle as a count of years .
- p: The influence of the sexagenary cycle as a count of years . 
- t: Ancient records show that the sexagenary cycle was used to count years in the 3rd century B.C. .

# TEXT NORMALIZATION

- s: ( It is used in all of the Star Wars and Indiana Jones movies . )
- p: It is used in all of the Star Wars and Indiana Jones movies .
- t: It is used in all of the Indiana Jones movies .

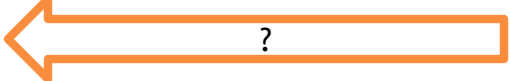
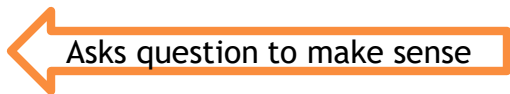
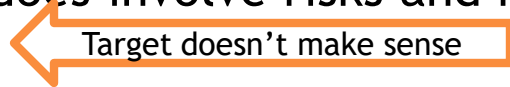
Brackets off

❖ But when it doesn't know how, it produces a trivial sentence.

- s: and Mario 's Time Machine .
- p: The machine is a machine .
- t: In the Dr. Mario games first seen in 1990 , Mario is a doctor .

Trivial sentence

# TEXT NORMALIZATION

- s: **and Airplane II : The Sequel .**
  - p: The Sequel is a sequel . 
  - t: ( 1980 ) and Airplane II : The Sequel ( 1982 ) .
- ❖ Or it tries to make sense of the parts it doesn't understand.
- s: **and it is because** they **don't understand what Parkour is ;`**  
**Who is** the best ? '
  - p: They are the best ? 
  - t: Star Jumping Parkour **does** involve risks and many people get injured every day . 

# TEXT NORMALIZATION

- s: It is officially known as Avtomat Kalashnikova ( Russian : **АВТОМАТ Калашникова** ) .
- p: It is known as Avtomat Kalashnikova ( Russian : : **1950 : 1998** ) .
- t: The letters AK stand for Avtomat Kalashnikova , which is Russian for Kalashnikov 's Automatic Rifle .

Tries make sense of “АВТОМАТ Калашникова “

❖ It tends to remove the beginning of a sentence.

- s: **With one huge blow from his olive-wood club** , Hercules killed the watchdog .
- p: Hercules killed the watchdog .
- t: Herakles killed her .

First part of sentence omitted

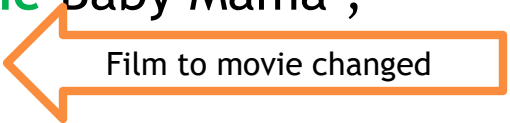
# TEXT NORMALIZATION

- s: **For example** , the speed of sound in gases depends on temperature .
- p: The speed of sound in gases depends on temperature .
- t: Loudness depends on sound intensity , sound frequency , and the person 's hearing .
- s: **In it** , Goldilocks is re-imagined as a girl in her 20s .
- p: Goldilocks is re-imagined as a girl in her 20s .
- t: She finally became Goldilocks sometime in the early 20th century .

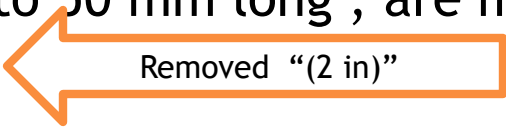


# TEXT NORMALIZATION


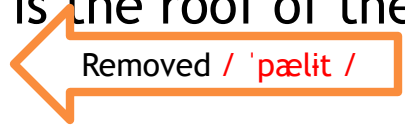
## ❖ BUT not always:

- s: In 2008 , she starred in the comedy **film** Baby Mama , alongside former SNL co-star Amy Poehler .
- p: In 2008 , she starred in the comedy **movie** Baby Mama , alongside former SNL co-star Amy Poehler .  

- t: In 2008 , she starred in the comedy movie Baby Mama , with Amy Poehler .

# TEXT NORMALIZATION

- ❖ It may also remove parts in the middle.
- s: The spines , which may be up to 50 mm ( 2 in ) long , are modified hairs , mostly made of keratin .
- p: The spines , which may be up to 50 mm long , are modified hairs , mostly made of keratin .  

- t: The spines , which may be up to 50 mm long , are modified hairs , mostly made of keratin .
-

# TEXT NORMALIZATION

- s: Sodium iodate ( **NaIO3** ) is the sodium salt of iodic acid .
  - p: Sodium iodate is the sodium salt of iodic acid . 
  - t: Sodium iodate is a chemical compound .
- 
- s: The palate / **'pælit** / is the roof of the mouth in humans and other mammals .
  - p: The palate is the roof of the mouth in humans and other mammals . 
  - t: The palate is the roof of the mouth in humans and other mammals .

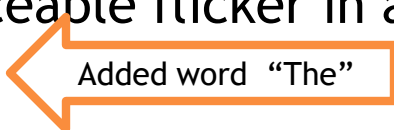
# TEXT NORMALIZATION

## ❖ AND it may remove some words:

- s: It is **located** in the Market Square , Aylesbury , Buckinghamshire .
  - p: It is in the Market Square , Aylesbury , Buckinghamshire .
  - t: The King 's Head is a National Trust property in Aylesbury , Buckinghamshire , England , UK .
- 
- s: Home Alone received mixed reviews from **film** critics .
  - p: Home Alone received mixed reviews from critics .
  - t: Home Alone 3 ( 1997 ) and Home Alone 5 : The Holiday Heist ( 2012 ) have completely different characters and actors .

# TEXT NORMALIZATION

## ❖ OR add some words.

- s: **However** , low frequency also causes noticeable flicker in arc lamps and incandescent light bulbs .
- p: **The** low frequency also causes noticeable flicker in arc lamps and incandescent light bulbs .  

- t: This happens so fast that a light bulb does not stop glowing .

# TEXT NORMALIZATION

❖ OR replace them.

- s: In the **film** 's story the location was named as Footscray station .
- p: In the **movie** 's story the location was named as Footscray station .
- t: In the movie it was called Footscray station .

# TEXT NORMALIZATION

I am  
NOT  
for  
maths!

## ❖ It sucks at math :P

- s: For example ,  $2 + 3 = 3 + 2 = 5$  and  $2 \cdot 3 = 3 \cdot 2 = 6$  , but  $23 = 8$  , whereas  $32 = 9$  .
- p: For example ,  $2 + 3 = 3 = 3 = 3 = 3 = 3 = 3 = 3$  , whereas  $32 = 9$  .
- t: Then people write powers using the ^ sign :  $2^3$  means .

# TEXT NORMALIZATION

- ❖ It can get really confused but this is very rare !!
  - s: 1 and  $-1$  divide ( are divisors of ) every integer .
  - p: In the 19th century , the state of the state of the state of the state of the state of the state of the state .
  - t: For example , if the last digit is even ( 0 , 2 , 4 , 6 or 8 ) , then 2 is a divisor .



THANK YOU!