Final Project 06
Final Project Presentation

NSGC - Neural Spell & Grammar Checker (en/pt)

Rafael Ito

25/06/2020

Presentation

- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

Presentation

- **Introduction**
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

Introduction

GEC shared tasks: (Grammatical Error Correction)

- HOO (Helping Our Own)
  - HOO-2011
  - HOO-2012
- CoNLL (The SIGNLL Conference on Computational Natural Language Learning)
  - CoNLL-2013
  - CoNLL-2014
- NLPCC 2018 (Chinese)
- BEA 2019 (Building Educational Applications)

Presentation

- Introduction
- **Methodology**
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

Presentation

- Introduction
- Methodology
- **Datasets**
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

- **Datasets**

- CoNLL-2013
- CoNLL-2014
- JFLEG
- BEA
- ReGRA

- **Datasets**

- **CoNLL-2013**
- CoNLL-2014
- JFLEG
- BEA
- ReGRA

CoNLL-2013

- **language**: English
- **corpus**: NUCLE
- **data format**: SGML (Standard Generalized Markup Language)
- **annotation format**: M2
- **metric**: M^2 (MaxMatch), F_1 score
- **test set available**: yes

- **Datasets**

- CoNLL-2013
- **CoNLL-2014**
- JFLEG
- BEA
- ReGRA

CoNLL-2014

Datasets

- **language**: English
- **corpus**: NUCLE
- **data format**: SGML (Standard Generalized Markup Language)
- **annotation format**: M2
- **metric**: M^2 (MaxMatch), $F_{0.5}$ score
- **test set available**: yes

- **Datasets**

- CoNLL-2013
- CoNLL-2014
- **JFLEG**
- BEA
- ReGRA

JFLEG

Datasets

- **language**: English
- **corpus**: GUG
- **data format**: text
- **annotation format**: text
- **metric**: GLEU
- **test set available**: yes

- **Datasets**

- CoNLL-2013
- CoNLL-2014
- JFLEG
- **BEA**
- ReGRA

Datasets

BEA

- **language**: English
- **corpus**: W&I, LOCNESS
- **data format**: JSON, M2
- **annotation format**: JSON, M2
- **metric**: ERRANT
- **test set available**: no

- **Datasets**

- CoNLL-2013
- CoNLL-2014
- JFLEG
- BEA
- **ReGRA**

ReGRA

Datasets

- **language**: Portuguese
- **corpus**: Own (?)
- **data format**: text (adjusted by hand)
- **annotation format**: text (adjusted by hand)
- **metric**: GLEU
- **provided by**: Osvaldo Novais de Oliveira Junior

Presentation

- Introduction
- Methodology
- Datasets
- **Metrics**
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

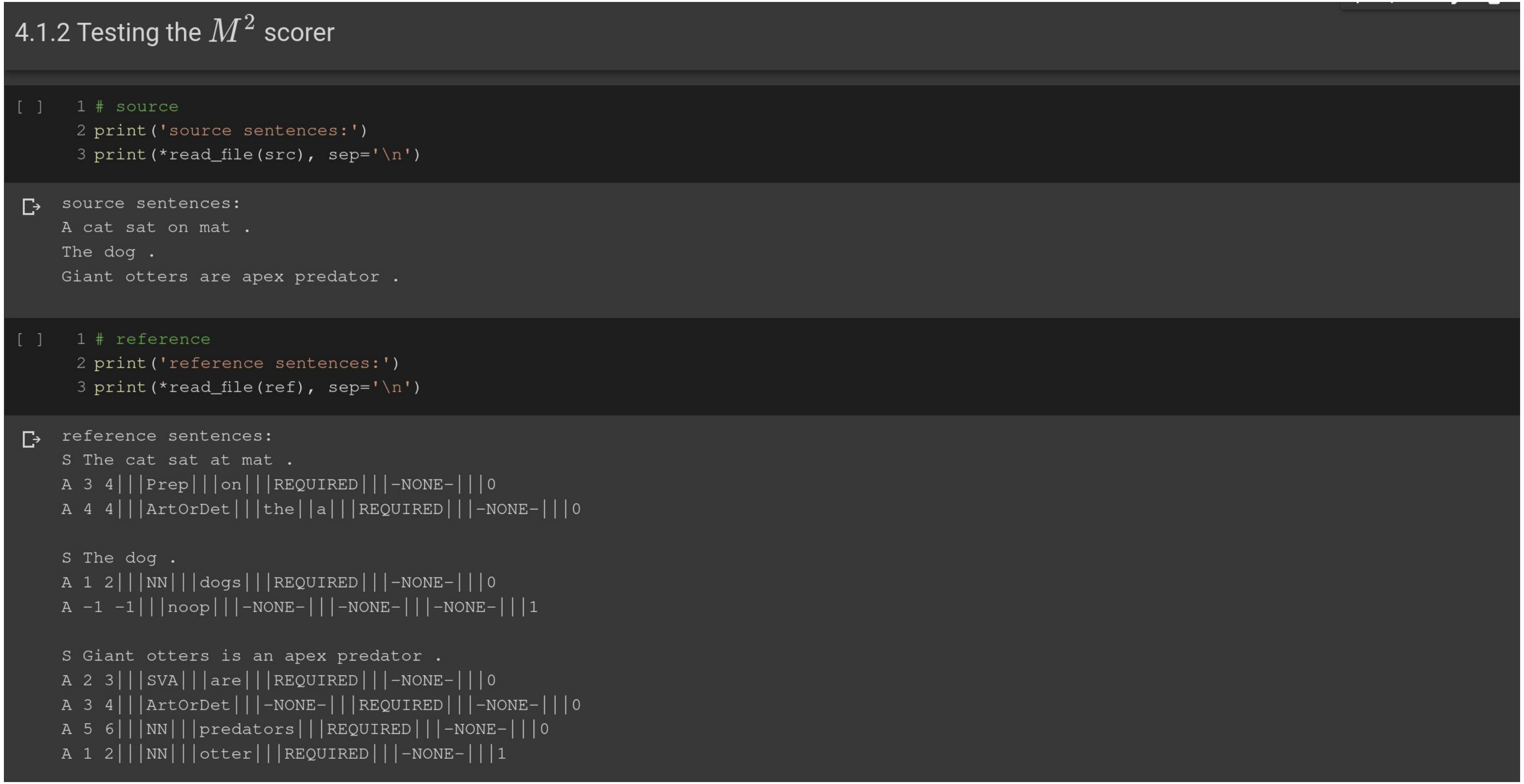
- **Metrics**

- M^2 (MaxMatch)
- GLEU
- Edit distance

- **Metrics**

- **M^2 (MaxMatch)**
- GLEU
- Edit distance

M^2 (MaxMatch)



4.1.2 Testing the M^2 scorer

```
# source
print('source sentences:')
print(*read_file(src), sep='\n')
```

```
source sentences:
A cat sat on mat .
The dog .
Giant otters are apex predator .
```

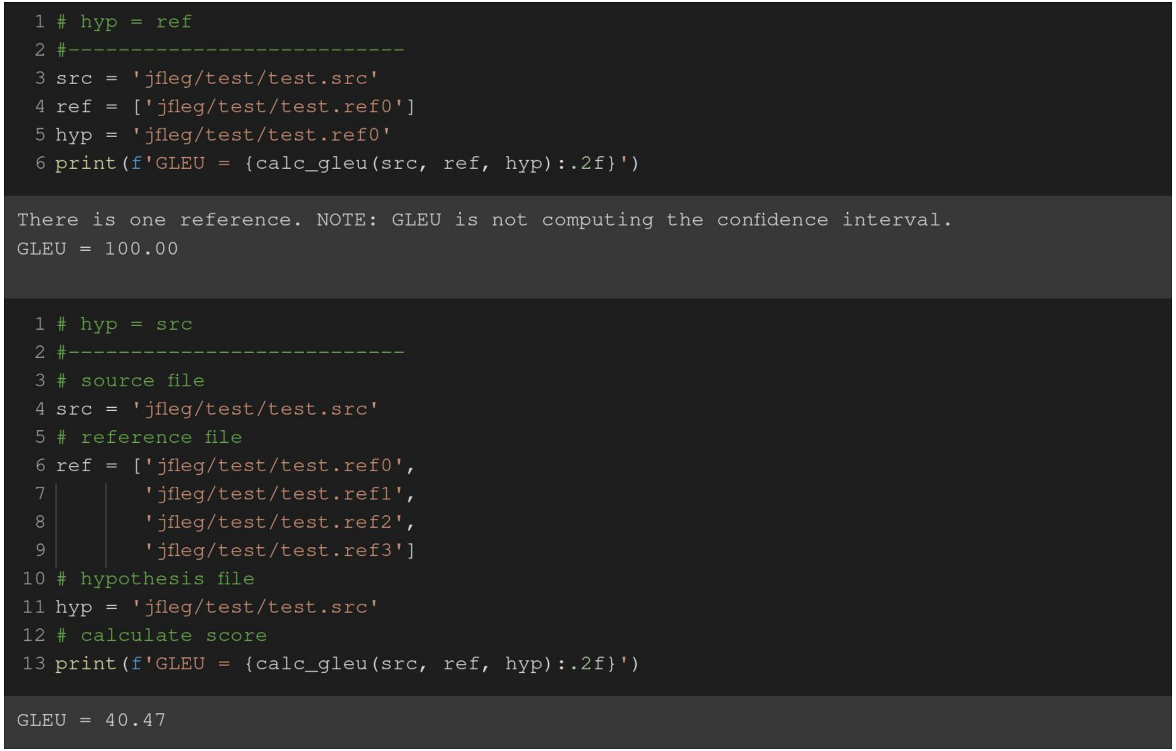
```
# reference
print('reference sentences:')
print(*read_file(ref), sep='\n')
```

```
reference sentences:
S The cat sat at mat .
A 3 4|||Prep|||on|||REQUIRED|||-NONE-|||0
A 4 4|||ArtOrDet|||the||a|||REQUIRED|||-NONE-|||0

S The dog .
A 1 2|||NN|||dogs|||REQUIRED|||-NONE-|||0
A -1 -1|||noop|||-NONE-|||-NONE-|||-NONE-|||1

S Giant otters is an apex predator .
A 2 3|||SVA|||are|||REQUIRED|||-NONE-|||0
A 3 4|||ArtOrDet|||-NONE-|||REQUIRED|||-NONE-|||0
A 5 6|||NN|||predators|||REQUIRED|||-NONE-|||0
A 1 2|||NN|||otter|||REQUIRED|||-NONE-|||1
```

- **Metrics**

- M^2 (MaxMatch)
- **GLEU**
- Edit distance

GLEU



```python
# hyp = ref
#---------------------------
src = 'jfleg/test/test.src'
ref = ['jfleg/test/test.ref0']
hyp = 'jfleg/test/test.ref0'
print(f'GLEU = {calc_gleu(src, ref, hyp):.2f}')
```

```
There is one reference. NOTE: GLEU is not computing the confidence interval.
GLEU = 100.00
```

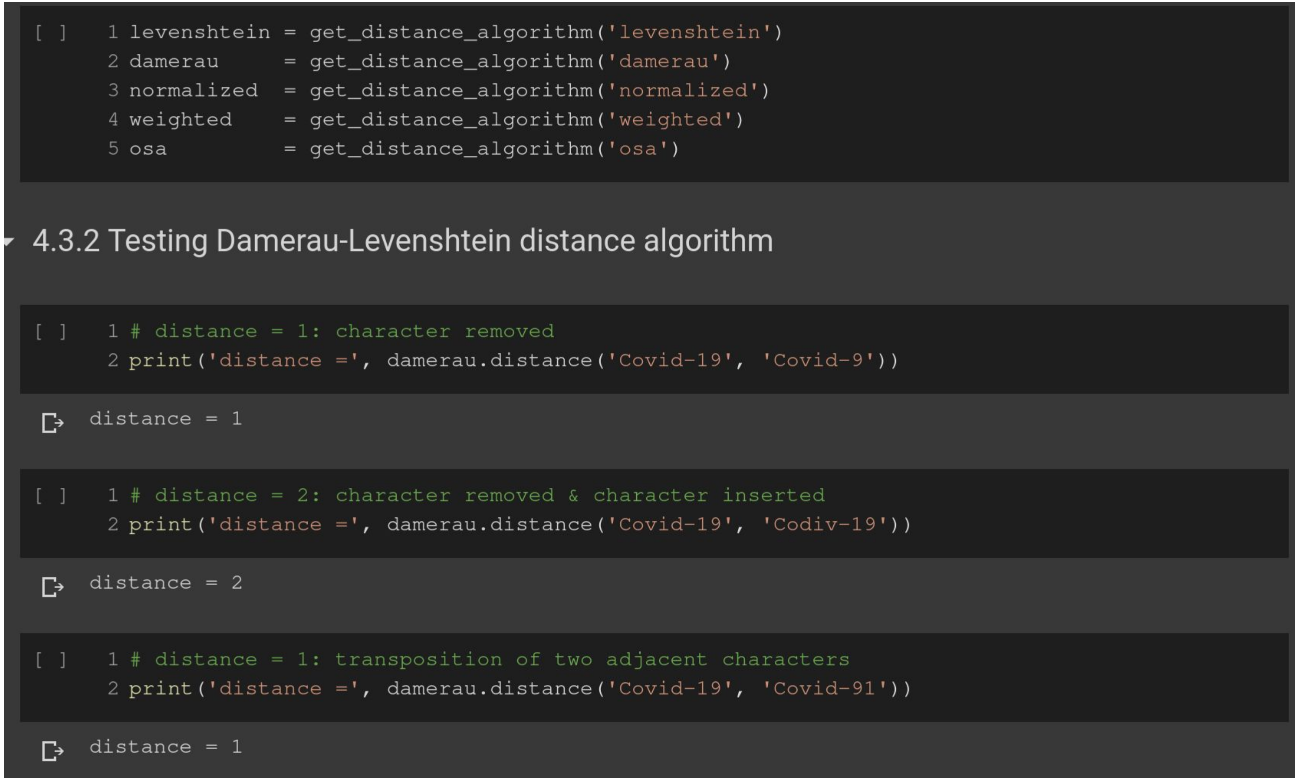
```python
# hyp = src
#---------------------------
# source file
src = 'jfleg/test/test.src'
# reference file
ref = ['jfleg/test/test.ref0',
       'jfleg/test/test.ref1',
       'jfleg/test/test.ref2',
       'jfleg/test/test.ref3']
# hypothesis file
hyp = 'jfleg/test/test.src'
# calculate score
print(f'GLEU = {calc_gleu(src, ref, hyp):.2f}')
```

```
GLEU = 40.47
```

- **Metrics**

- M^2 (MaxMatch)
- GLEU
- **Edit distance**

Edit distance



```python
levenshtein = get_distance_algorithm('levenshtein')
damerau     = get_distance_algorithm('damerau')
normalized  = get_distance_algorithm('normalized')
weighted    = get_distance_algorithm('weighted')
osa         = get_distance_algorithm('osa')
```

4.3.2 Testing Damerau-Levenshtein distance algorithm

```python
# distance = 1: character removed
print('distance =', damerau.distance('Covid-19', 'Covid-9'))
```

```
distance = 1
```

```python
# distance = 2: character removed & character inserted
print('distance =', damerau.distance('Covid-19', 'Codiv-19'))
```

```
distance = 2
```

```python
# distance = 1: transposition of two adjacent characters
print('distance =', damerau.distance('Covid-19', 'Covid-91'))
```

```
distance = 1
```

Presentation

- Introduction
- **Methodology**
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- Future Work

Methodology

Steps:

1. get source sentence
2. tokenize it
3. mask each token one at a time
4. input masked sentences to the model
5. get predictions
6. compare with the original masked token (edit distance)
7. decide whether change or keep original token

Methodology

1. get source sentence
2. tokenize it

source:

Ele comprou este carro **à** prazo.

reference (gold):

Ele comprou este carro a prazo.

tokenize:

['Ele', 'comprou', 'este', 'carro', 'à', 'prazo', '.']

BERT’s special tokens:

[CLS] → ID: 101

[SEP] → ID: 102

[MASK] → ID: 103

Methodology

3. mask each token one at a time
4. input masked sentences to the model

convert to IDs:

[101, 787, 10107, 860, 3883, 353, 6620, 119, 102]

mask:

[CLS, **MASK**, 10107, 860, 3883, 353, 6620, 119, SEP]
[CLS, 787, **MASK**, 860, 3883, 353, 6620, 119, SEP]

...

[CLS, 787, 10107, 860, 3883, 353, 6620, **MASK**, SEP]

Methodology

5. get predictions
6. compare with the original masked token (edit distance)

hyperparameters:

topk = 2

threshold = 2

- **Case 1:** masked token in prediction

Ele comprou este carro à prazo .

[CLS, **MASK**, 10107, 860, 3883, 353, 6620, 119, SEP]

predictions: [‘Você’, ‘Ele’] ⇒ keep original!

7. decide whether change or keep original token

Methodology

5. get predictions
6. compare with the original masked token (edit distance)

hyperparameters:

topk = 2

threshold = 2

- **Case 2:** masked token not in prediction

Ele comprou **este** carro à prazo .

[CLS, 787, 10107, **MASK**, 3883, 353, 6620, 119, SEP]

predictions: [‘o’, ‘um’], edit distance > 2 ⇒ keep original!

7. decide whether change or keep original token

Methodology

5. get predictions
6. compare with the original masked token (edit distance)

hyperparameters:

topk = 2

threshold = 2

- **Case 3:** masked token not in prediction

Ele comprou este carro **à** prazo .

[CLS, 787, 10107, 860, 3883, **MASK**, 6620, 119, SEP]

predictions: [‘a’, ‘no’], edit distance < 2 ⇒ change!

7. decide whether change or keep original token

Methodology

hyperparameters:

topk = 2

threshold = 2

source:

Ele comprou este carro **à** prazo.

reference (gold):

Ele comprou este carro a prazo.

system output:

Ele	comprou	este	carro	a	prazo	.
keep	keep	keep	keep	**change**	keep	keep

Hyperparameters

source: Ele comprou este carro **à** prazo .

- **k = 2, threshold = 3:**

  Ele comprou este carro a prazo . ✓

- **k = 3, threshold = 3:**

  Ele comprou seu carro à prazo . ✗

- **k = 2, threshold = 1:**

  Ele comprou este carro a prazo . ✓

- **k = 5, threshold = 5:**

  Ele comprou esse carro à parte . ✗

Doing the same with T5

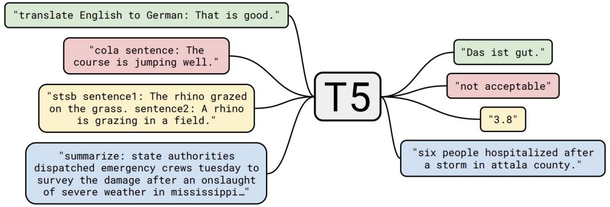
- **source:**
  - **People also do not do nothing .**

- **references:**
  - **People also do not do anything .**
  - **People also do not do nothing .**
  - **People also do something .**
  - **People also do not do nothing .**

T5's special tokens:

</s> → ID: 1

<extra_id_0> → ID: 32099



"translate English to German: That is good."

"cola sentence: The course is jumping well."

"stsb sentence1: The rhino grazed on the grass. sentence2: A rhino is grazing in a field."

"summarize: state authorities dispatched emergency crews tuesday to survey the damage after an onslaught of severe weather in mississippi…"

T5

"Das ist gut."

"not acceptable"

"3.8"

"six people hospitalized after a storm in attala county."

Doing the same with T5

mask:

<extra_id_0> also do not do nothing . **</s>**
People **<extra_id_0>** do not do nothing . **</s>**
People also **<extra_id_0>** not do nothing . **</s>**
People also do **<extra_id_0>** do nothing . **</s>**
People also do not **<extra_id_0>** nothing . **</s>**
People also do not do **<extra_id_0>** . **</s>**
People also do not do nothing **<extra_id_0>** **</s>**

- model's output:

  People also do not do anything .

hyperparameters:

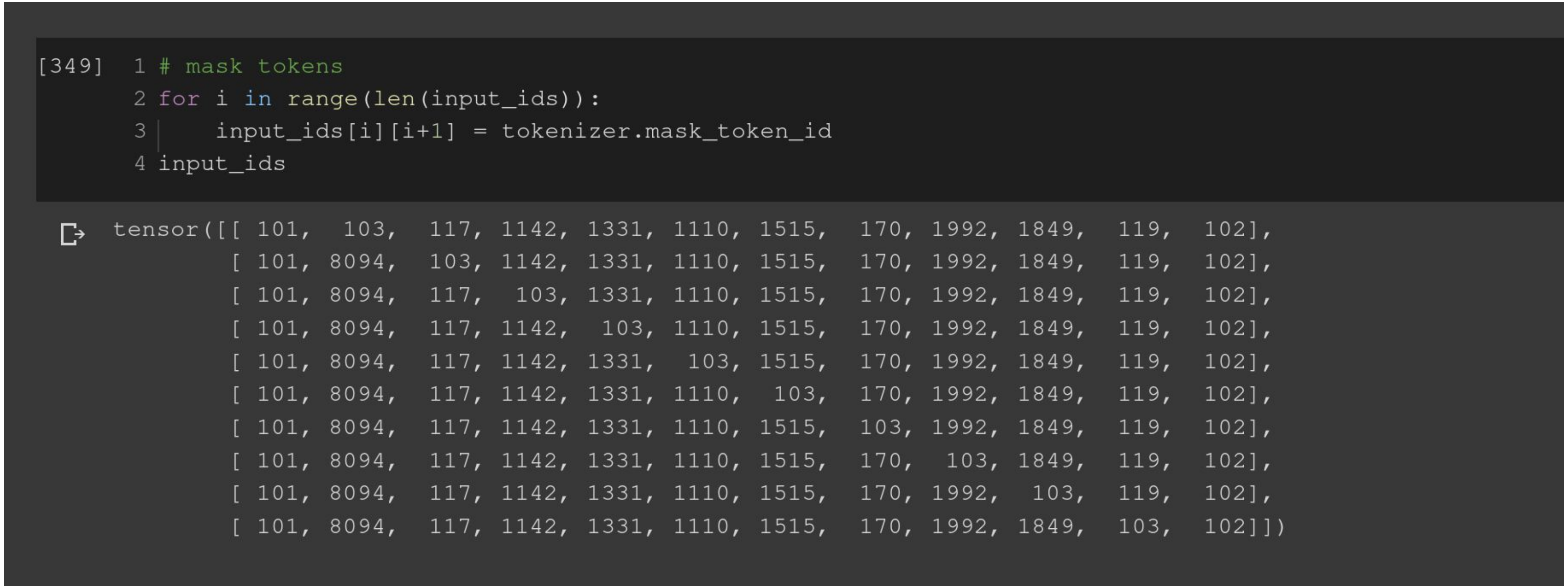
topk = 5

threshold = 2

Presentation

- Introduction
- Methodology
- Datasets
- Metrics
- **Notebooks**
- Experiments
- Results
- Problems
- Conclusion
- Future Work

Notebooks



```
[349]
# mask tokens
for i in range(len(input_ids)):
    input_ids[i][i+1] = tokenizer.mask_token_id
input_ids
```

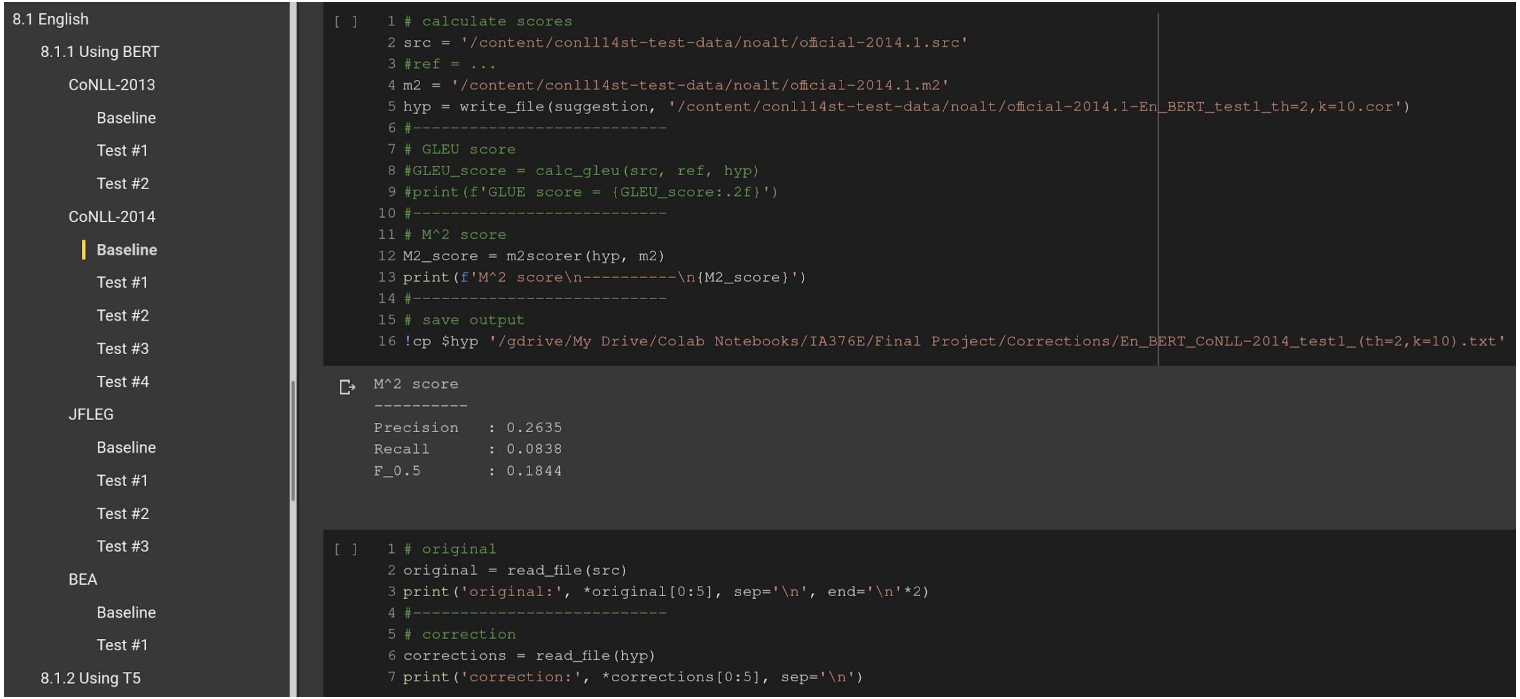
```
tensor([[ 101,  103,  117, 1142, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  103, 1142, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117,  103, 1331, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142,  103, 1110, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331,  103, 1515,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110,  103,  170, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  103, 1992, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170,  103, 1849,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170, 1992,  103,  119,  102],
        [ 101, 8094,  117, 1142, 1331, 1110, 1515,  170, 1992, 1849,  103,  102]])
```

https://colab.research.google.com/drive/194LQ5UyrmFJOKUPL7qyAcDFkcfWF3qV1?authuser=1#scrollTo=gTGvw969QXqO

Presentation

- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- **Experiments**
- Results
- Problems
- Conclusion
- Future Work

Experiments



```
[ ]
# calculate scores
src = '/content/conll14st-test-data/noalt/official-2014.1.src'
#ref = ...
m2 = '/content/conll14st-test-data/noalt/official-2014.1.m2'
hyp = write_file(suggestion, '/content/conll14st-test-data/noalt/official-2014.1-En_BERT_test1_th=2,k=10.cor')
#---------------------------
# GLEU score
#GLEU_score = calc_gleu(src, ref, hyp)
#print(f'GLUE score = {GLEU_score:.2f}')
#---------------------------
# M^2 score
M2_score = m2scorer(hyp, m2)
print(f'M^2 score\n----------\n{M2_score}')
#---------------------------
# save output
!cp $hyp '/gdrive/My Drive/Colab Notebooks/IA376E/Final Project/Corrections/En_BERT_CoNLL-2014_test1_(th=2,k=10).txt'
```

```
M^2 score
----------
Precision   : 0.2635
Recall      : 0.0838
F_0.5       : 0.1844
```

```
[ ]
# original
original = read_file(src)
print('original:', *original[0:5], sep='\n', end='\n'*2)
#---------------------------
# correction
corrections = read_file(hyp)
print('correction:', *corrections[0:5], sep='\n')
```

Presentation

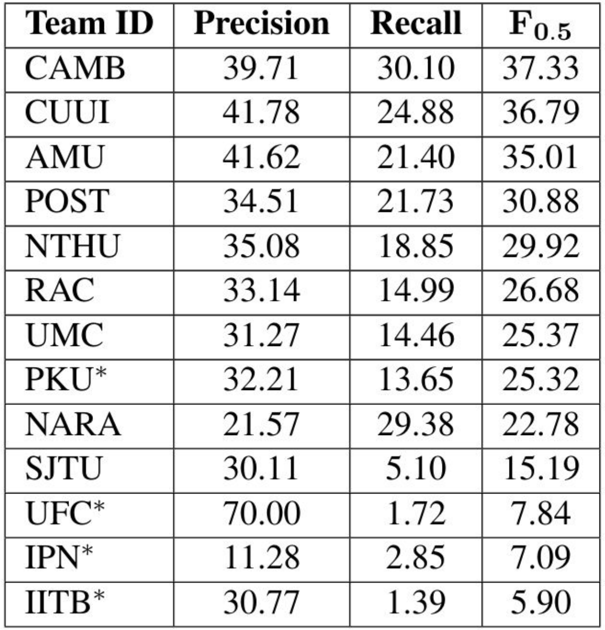
- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- **Results**
- Problems
- Conclusion
- Future Work

Results

	baseline (0~100)	number of tests	best result	metric
CoNLL-2013	0	3	15.3	M^2
CoNLL-2014	0	9	18.44	M^2
JFLEG	40.47	4	44.56	GLEU
BEA	0	2	16.82	M^2
ReGRA	36.99	8	38.29	GLEU

Results

CoNLL-2014 results



Team ID	Precision	Recall	$F_{0.5}$
CAMB	39.71	30.10	37.33
CUUI	41.78	24.88	36.79
AMU	41.62	21.40	35.01
POST	34.51	21.73	30.88
NTHU	35.08	18.85	29.92
RAC	33.14	14.99	26.68
UMC	31.27	14.46	25.37
PKU*	32.21	13.65	25.32
NARA	21.57	29.38	22.78
SJTU	30.11	5.10	15.19
UFC*	70.00	1.72	7.84
IPN*	11.28	2.85	7.09
IITB*	30.77	1.39	5.90

Table 7: Scores (in %) *without* alternative answers. The teams that submitted their system output after the deadline have an asterisk affixed after their team names.

My results	
Precision	26.35
Recall	8.38
$F_{0.5}$	18.44

Presentation

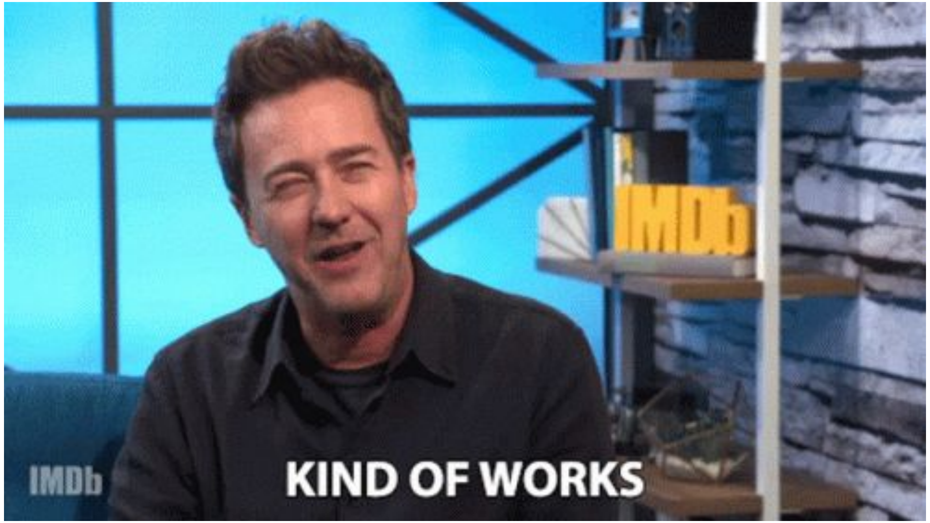
- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- **Problems**
- Conclusion
- Future Work

Problems

- Subwords when tokenizing
- Insert tokens
    - source: Forexample , My cousin is 12years old .
    - reference: For example , my cousin is 12 years old .
- Remove tokens
    - Example: I often look at TV → I often watch TV
- Portuguese dataset

Presentation

- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- **Conclusion**
- Future Work

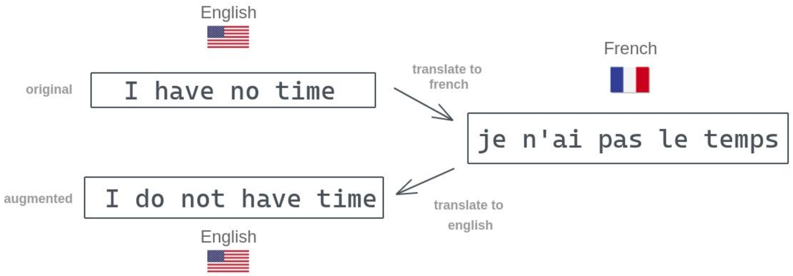


Presentation

- Introduction
- Methodology
- Datasets
- Metrics
- Notebooks
- Experiments
- Results
- Problems
- Conclusion
- **Future Work**

Future Work

- **soft check:**
  - dictionary-based
  - back translation



English

original I have no time

translate to french

French

je n'ai pas le temps

translate to english

augmented I do not have time

English

MARIANNMT

Fast Neural Machine Translation in C++

- **committee machines (one for each “rule”)**
- **test Portuguese GEC with T5**
- **use commercial GEC system to get a superior reference for pt-br**
- **package to be installed with pip and/or run on regular terminal**

Future Work

- Insert token:
  - Ex: I like pizza
    - [MASK] I like pizza
    - I [MASK] like pizza
    - I like [MASK] pizza
- 2-gram substitution:
  - Ex: I like pizza
    - [MASK] pizza
    - I [MASK]
- edit distance threshold based on the length of the word masked