

Project Presentation

Measure Text Fluency

By:

- Aditya Rathi (2020201041)
- Prashant Raj (2020201057)
- Jayant Ingle (2020201019)



Outline

- Measure Text Fluency
- Timeline
- Generating Dataset
- Metrics Implemented
 - BLEU Metric
 - ROUGE Metric
 - METEOR Metric
- Input and Output
- Analysis and Results
- Conclusions

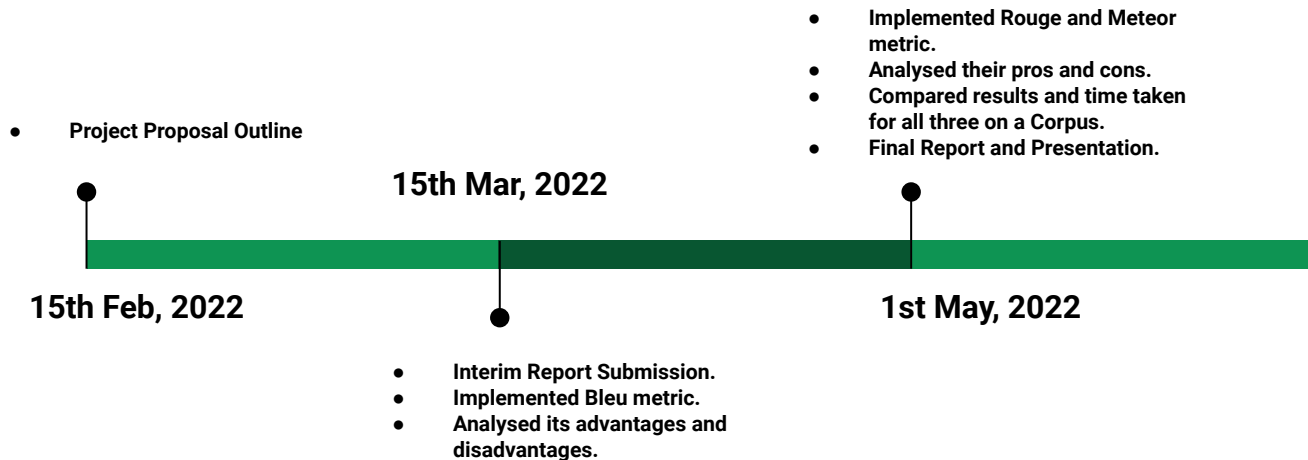


Measure Text Fluency

- The aim of this project is to implement metrics to evaluate text fluency of machine translations.
- Fluency is commonly considered as one of the dimensions of text quality of MT. Fluency measures the quality of the generated text, without taking the source sentence into account. It accounts for criteria such as grammar, spelling, choice of words, and style.
- Given a candidate sentence (a translation obtained from MT) and a reference translation, we try to implement different metrics to compute their fluency.




Timeline





Generating Dataset

- Used French - English corpus provided to us for Assignment 3.
- Used Google Translate to generate machine translations for French to English sentences.
- Used generated output sentences as Candidate Sentences and original English corpus given to us as Reference Sentences.

B1	 =GOOGLETRANSLATE(A1, "fr", "en")	
	A	B
1	Quand j'avais la vingtaine, j'ai vu mes tout premiers clients com	When I had my twenties, I saw my very first customers as a psychotherapist.
2	J'étais étudiante en thèse en psychologie clinique à Berkeley.	I was a thesis student in clinical psychology in Berkeley.
3	Elle, c'était une femme de 26 ans appelée Alex.	Elle, c'était une femme de 26 ans appelée Alex.
4	Lorsqu'Alex est entrée pour sa première séance, elle portait un	When Alex entered her first session, she wore jeans and a great top too wide, she
5	Lorsque j'ai entendu ça, j'ai été si soulagée.	When I heard that, I was so relieved.
6	Ma camarade de classe avait eu un pyromane comme premier	My classmate had had a pyromaniac as the first patient.
7	Et moi, j'avais une fille de 20 ans et quelques qui voulait parler c	And I had a 20 year old girl and a few who wanted to talk about boys.
8	Je pensais pouvoir gérer ça.	I thought I could manage this.
9	Mais je ne l'ai pas géré.	But I didn't manage it.




Metrics Implemented

- BLEU
- ROUGE
- METEOR



BLEU Metric

- BLEU stands for Bi-Lingual Evaluation Understudy. It is a segment level algorithm that judges translations on a per-word basis.
- BLEU was the first widely adopted metric used for machine translation.



$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} .$$

Then,

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) .$$

The “BLEU score” is the geometric mean of all four n-gram precisions

$$\text{BLEU-4} \sim \sqrt[4]{p_1 \cdot p_2 \cdot p_3 \cdot p_4}$$



ROUGE Metric

- ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation.
- It is essentially a set of metrics for evaluating automatic summarization of texts as well as machine translations.
- It works by comparing an automatically produced summary or translation against a set of reference summaries (typically human-produced).



Recall

number of n-grams found in model and reference

number of n-grams in reference

$$\text{count}_{\text{match}}(\text{gram}_n)$$

$$\text{count}(\text{gram}_n)$$



METEOR Metric

- METEOR (Metric for Evaluation of Translation with Explicit ORdering) (Banerjee 2005) is an evaluation specifically designed to address several observed weaknesses in BLEU.
- METEOR is a recall-oriented metric, whereas BLEU is generally precision-oriented metric.
- Unlike BLEU which only calculates precision, METEOR calculates both precision and recall, and combines the two with a large bias towards recall, to calculate the harmonic mean.



Input and Output

```
NLP_Project$ time python3 main.py -t cmd -em all
```

```
Candidate sentence :I was a Ph.D. student in clinical psychology at Berkeley.
```

```
Reference sentence :I was the thesis student in clinical psychology in Berkeley.
```

```
bleu score : 0.32466791547509893
```

```
rouge score : 0.8333333333333334
```

```
meteor score : 0.6776
```

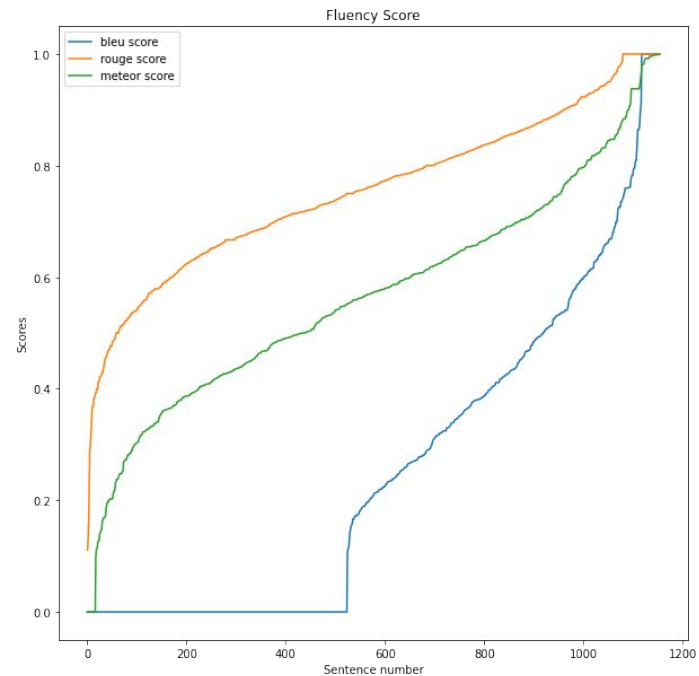
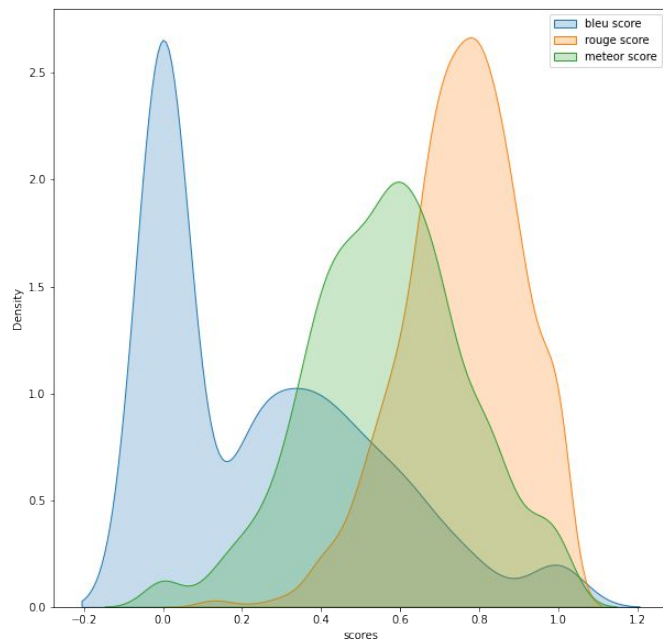
```
real 0m12.605s
```

```
user 0m1.271s
```

```
sys 0m0.921s
```



Analysis and Results





Analysis and Results

	bleu score	rouge score	meteor score
count	1154.000000	1154.000000	1154.000000
mean	0.252634	0.754477	0.567195
std	0.281222	0.148792	0.202155
min	0.000000	0.111111	0.000000
25%	0.000000	0.666667	0.430090
50%	0.208717	0.763636	0.572785
75%	0.438972	0.859155	0.698414
max	1.000000	1.000000	0.999500



Analysis and Results

- Sentence count : 1154

Criterion	Bleu Score	Rouge Score	Meteor Score
Average Score	0.208717	0.763636	0.572785
Total time taken (in seconds)	0.528282165527 3438	0.094687700271 60645	151.7340083122 2534



Conclusions

- When the sentence length increases, meteor takes more time to compute the score due to the computation of all possible combinations.
- Time comparison
 - Meteor Metric > Bleu Metric > Rouge Metric
- Text Fluency comparison
 - Rouge Metric > Meteor Metric > Bleu Metric
- BLEU is a precision score based metric, and doesn't account for recall.
- Rouge and Meteor accounts for both precision and recall.
- Rouge also accounts for F1 score.



Thank You