

Narzędzie do tłumaczenia tekstu wykorzystujące transformery 'translation' z portalu Hugging Face

Strona tytułowa:

Tytuł: Narzędzie do tłumaczenia tekstu wykorzystujące transformery 'translation' z portalu Hugging Face

Autorzy: Przemysław Danak, Bruno Blajda

1. Abstrakt

W tym projekcie opracowano narzędzie do tłumaczenia tekstu wykorzystujące transformery 'translation' z portalu *Hugging Face*. Skupiono się na tłumaczeniu z języka angielskiego na język francuski i wykorzystano do tego model transformatorowy T5. Projekt obejmował przygotowanie danych treningowych, wytrenowanie modelu i jego przetestowanie.

2. Wstęp

2.1 Cel

Celem projektu było zbudowanie systemu tłumaczenia tekstów z języka angielskiego na francuski z wykorzystaniem modelu T5. System miał być w stanie generować tłumaczenia o jakości zbliżonej do ludzkiej, uwzględniając kontekst i poprawność gramatyczną zdań.

2.2 Zakres

Zakres projektu obejmował: przygotowanie środowiska i instalację wymaganych bibliotek (*transformers*, *datasets*, *evaluate*, *sacrebleu*), pobranie i podział zbioru danych *OPUS Book*, przygotowanie tekstu dla modelu, trenowanie modelu i testowanie modelu.

2.3 Metodyka

Projekt realizowano w oparciu o następujące kroki:

- wybór modelu - wykorzystano model *google-t5/T5-small* z biblioteki *Hugging Face Transformer*,
- przygotowanie danych - przetworzono zbiór *OPUS Books*, podzielony na zbiory treningowy (80%) i testowy (20%),
- dopasowanie modelu - model dopasowano do specyficznego zadania tłumaczenia z angielskiego na francuski,
- ocena - wyniki oceniano za pomocą miary BLEU, porównując tłumaczenia modelu z referencyjnymi tłumaczeniami.

3. Część Teoretyczna

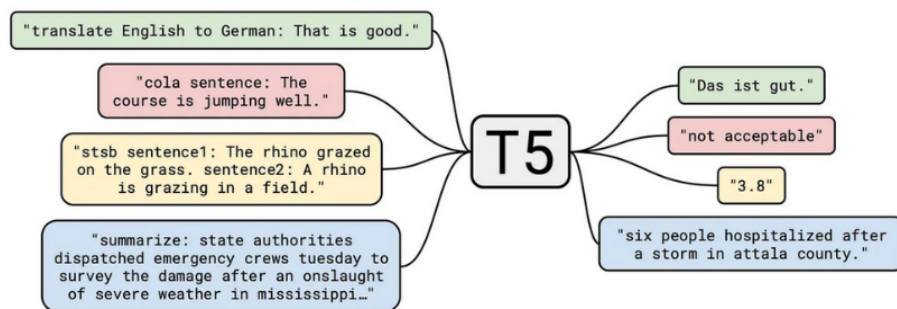
Omówiono różne podejścia do tłumaczenia maszynowego [2, 3]:

- tłumaczenie maszynowe oparte na regułach - opiera się na zestawie reguł językowych, w tym regułach gramatycznych,
- tłumaczenie maszynowe statystyczne - opiera się na dużych dwujęzycznych korpusach i modelach statystycznych do generowania tłumaczeń,

- tłumaczenie maszynowe analizujące składnię tekstu - dzieli tekst na frazy zamiast pojedynczych słów (większa dokładność kontekstowa),
- tłumaczenie maszynowe oparte na sieciach neuronowych - wykorzystuje modele głębokiego uczenia,
- hybrydowe tłumaczenie maszynowe - łączy dwie lub więcej metod, które zostały wymienione powyżej.

Model T5 - uniwersalny transformator, który zamienia każde zadanie językowe na problem tłumaczenia tekstu na tekst. Kluczowe cechy:

- mechanizm uwagi - skupia się na istotnych częściach tekstu,
- dwukierunkowe przetwarzanie - analizuje kontekst całego zdania,
- elastyczność - obsługuje wiele zadań językowych takie jak: tłumaczenia, podsumowywanie tekstu, analiza sentymentu.



Działanie modelu T5. [4]

Miara BLEU ocenia jakość tłumaczeń maszynowych, porównując je z tłumaczeniami referencyjnymi. BLEU mierzy, jak bardzo tłumaczenie przypomina tłumaczenia wykonane przez ludzi.

4. Część Praktyczna

4.1 Przygotowanie środowiska

Zainstalowano wymagane biblioteki:

- *Transformers* - umożliwia łatwe korzystanie z najnowszych modeli NLP,
- *Datasets* - udostępnia ogromną liczbę zbiorów danych do nauki maszynowej,
- *Evaluate* - ocenia modele NLP, udostępnia miarę BLEU (Bilingual Evaluation Understudy) - porównanie tłumaczenia z tłumaczeniem przygotowanym przez ludzi,
- *SacreBLEU* - zapewnia standaryzację w ocenie tłumaczeń metryką BLEU.

```
pip install transformers datasets evaluate sacrebleu
```

4.2 Pobranie i podział danych

Za pomocą biblioteki Datasets pobrano zbiór danych *OPUS Books*. Dane podzielono na zbiory:

- Treningowy (80%) - do nauki modelu.
- Testowy (20%) - do oceny jakości tłumaczeń.

Poniższy kod opracowano na podstawie *Hugging Face* [1].

```
from datasets import load_dataset

books = load_dataset("opus_books", "en-fr")
books = books["train"].train_test_split(test_size=0.2)
books["train"][0]
```

Poniżej skorzystano z *JSON Formatter* [5].

```
{
  "id": "126191",
  "translation": {
    "en": "Michaud and Grivet prided themselves on their correct attitude."
    "fr": "Michaud et Grivet s'applaudirent de leur excellente tenue."
  }
}
```

4.3 Tokenizacja i przetwarzanie danych

Pobrano tokenizator. Został użyty model *google-t5/t5-small*. Model ten jest szybszy w trenowaniu i generowaniu tłumaczeń w porównaniu z większymi wersjami oraz mniej wymagający pod względem zasobów sprzętowych.

Przygotowano funkcję, która dodaje prefiks „translate English to French:” do zdań angielskich i przetwarza je na tokeny numeryczne.

Poniższy kod opracowano na podstawie *Hugging Face* [1].

```
from transformers import AutoTokenizer

checkpoint = "google-t5/t5-small"
tokenizer = AutoTokenizer.from_pretrained(checkpoint)

source_lang = "en"
target_lang = "fr"

prefix = "translate English to French: "

def preprocess_function(examples):
    inputs = [prefix + example[source_lang] for example in examples["translation"]]
    targets = [example[target_lang] for example in examples["translation"]]
    model_inputs = tokenizer(inputs, text_target=targets, max_length=128, truncation=True)
    return model_inputs

tokenized_books = books.map(preprocess_function, batched=True)
print(tokenized_books['train'][0])
```

Poniżej skorzystano z *JSON Formatter* [5].

```
{
  "id": "124887",
  "translation": {
    "en": "He went to find old Michaud, and told him he had just recognized Camille"
    "fr": "Il alla chercher le vieux Michaud et lui dit qu'il venait de reconnaître"
  },
  "input_ids": [ ],
  "attention_mask": [ ],
  "labels": [ ]
}
```

4.4 Trenowanie modelu

Przetrenowano model i zapisano go w celu późniejszego przetestowania. Zostały ustawione hiperparametry - kluczowe ustawienia, które kontrolują sposób, w jaki model się uczy. Wykorzystano *Seq2SeqTrainer*, który korzysta z funkcji oceny - pozwala lepiej dostosować model oraz ograniczono się do liczby 5000 zdań.

Poniższy kod opracowano na podstawie *Hugging Face* [1].

```
from transformers import AutoModelForSeq2SeqLM, Seq2SeqTrainingArguments, Seq2SeqTrainer

model = AutoModelForSeq2SeqLM.from_pretrained(checkpoint)

small_train_dataset = tokenized_books["train"].select(range(4000))
small_test_dataset = tokenized_books["test"].select(range(1000))

training_args = Seq2SeqTrainingArguments(
    output_dir="translation_model",
    save_strategy="epoch",
    save_steps=500,
    eval_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    weight_decay=0.01,
    save_total_limit=2,
    num_train_epochs=3,
    predict_with_generate=True,
    fp16=True,
    report_to=["none"]
)

trainer = Seq2SeqTrainer(
    model=model,
    args=training_args,
    train_dataset=small_train_dataset,
    eval_dataset=small_test_dataset,
    tokenizer=tokenizer,
    data_collator=data_collator,
    compute_metrics=compute_metrics,
)

trainer.train()

trainer.save_model("translation_model")
```

4.5 Testowanie modelu

Do testowania przygotowano kilka zdań wraz z tłumaczeniami. Jakość tłumaczeń z modelu oceniono za pomocą metryki BLEU. Na podstawie wyników można wywnioskować, że model generuje akceptowalne tłumaczenia.

Poniższy kod opracowano na podstawie *Hugging Face* [1]. Do tłumaczenia zdań referencyjnych użyto *Tłumacza DeepL* [6].

```

from transformers import AutoModelForSeq2SeqLM, AutoTokenizer
from evaluate import load

model = AutoModelForSeq2SeqLM.from_pretrained("translation_model")
tokenizer = AutoTokenizer.from_pretrained("translation_model")

sentences = [
    "translate English to French: The cat is sleeping.",
    "translate English to French: The sky is blue.",
    "translate English to French: I like apples.",
    "translate English to French: She is my friend.",
    "translate English to French: We are happy.",
    "translate English to French: Learning new languages can be a challenging but rewarding experience.",
    "translate English to French: I love programming and artificial intelligence.",
    "translate English to French: The weather is beautiful today, don't you think?",
    "translate English to French: Although the weather forecast predicted sunshine, it started raining heavily in the afternoon.",
    "translate English to French: The implications of artificial intelligence on society are both fascinating and deeply complex.",
    "translate English to French: The orchestra played a symphony that captivated the audience with its intricate melodies and harmonies."
]

references = [
    ["Le chat dort."],
    ["Le ciel est bleu."],
    ["J'aime les pommes."],
    ["Elle est mon amie."],
    ["Nous sommes heureux."],
    ["Apprendre de nouvelles langues peut être une expérience difficile mais enrichissante."],
    ["J'aime la programmation et l'intelligence artificielle."],
    ["Il fait beau aujourd'hui, vous ne trouvez pas ?"],
    ["Alors que les prévisions météorologiques annonçaient du soleil, il s'est mis à pleuvoir abondamment dans l'après-midi."],
    ["Les implications de l'intelligence artificielle sur la société sont à la fois fascinantes et profondément complexes."],
    ["L'orchestre a joué une symphonie qui a captivé le public par ses mélodies et ses harmonies complexes."]
]

def translate_and_evaluate(sentences, references, model, tokenizer, metric):
    predictions = []
    scores = []

    print("Translation Results:\n")
    for i, (text, reference) in enumerate(zip(sentences, references), 1):
        inputs = tokenizer(text, return_tensors="pt")
        outputs = model.generate(inputs.input_ids, max_new_tokens=50)
        prediction = tokenizer.decode(outputs[0], skip_special_tokens=True)
        predictions.append(prediction)

        individual_score = metric.compute(predictions=[prediction], references=[reference])["score"]
        scores.append(individual_score / 100)

        print(f"{i}. English Sentence: {text}")
        print(f"   Reference Translation: {reference[0]}")
        print(f"   Model Translation: {prediction}")
        print(f"   BLEU Score: {individual_score / 100:.2f}\n")

    average_bleu = sum(scores) / len(scores)
    return average_bleu

metric = load("sacrebleu")

average_bleu = translate_and_evaluate(sentences, references, model, tokenizer, metric)

print(f"Overall Average BLEU Score: {average_bleu:.4f}")

```

5. Wyniki

Wyniki w większości przypadków są poprawne. W celu poprawy modelu można skorzystać z większej ilości zdań z tłumaczeniami (treningowych i testowych), odpowiednio zmienić hiperparametry (w tym *num_train_epoch* - liczba epok) lub zmienić model na *google-t5/t5-base* lub *google-t5/t5-large*.

```

Translation Results:

1. English Sentence: translate English to French: The cat is sleeping.
   Reference Translation: Le chat dort.
   Model Translation: Le chat dorme.
   BLEU Score: 0.35

2. English Sentence: translate English to French: The sky is blue.
   Reference Translation: Le ciel est bleu.
   Model Translation: Le ciel est bleu.
   BLEU Score: 1.00

3. English Sentence: translate English to French: I like apples.
   Reference Translation: J'aime les pommes.
   Model Translation: Je m'aime les pommes.
   BLEU Score: 0.40

4. English Sentence: translate English to French: She is my friend.
   Reference Translation: Elle est mon amie.
   Model Translation: Elle est mon ami.
   BLEU Score: 0.43

5. English Sentence: translate English to French: We are happy.
   Reference Translation: Nous sommes heureux.
   Model Translation: Nous sommes heureux.
   BLEU Score: 1.00

6. English Sentence: translate English to French: Learning new languages can be a challenging but rewarding experience.
   Reference Translation: Apprendre de nouvelles langues peut être une expérience difficile mais enrichissante.
   Model Translation: L'apprentissage de nouvelles langues peut être une expérience difficile mais enrichissante.
   BLEU Score: 0.90

7. English Sentence: translate English to French: I love programming and artificial intelligence.
   Reference Translation: J'aime la programmation et l'intelligence artificielle.
   Model Translation: Je m'aime la programmation et l'intelligence artificielle.
   BLEU Score: 0.68

8. English Sentence: translate English to French: The weather is beautiful today, don't you think?
   Reference Translation: Il fait beau aujourd'hui, vous ne trouvez pas ?
   Model Translation: Le temps est beau aujourd'hui, ne pensez-vous pas?
   BLEU Score: 0.21

9. English Sentence: translate English to French: Although the weather forecast predicted sunshine, it started raining heavily in the afternoon.
   Reference Translation: Alors que les prévisions météorologiques annonçaient du soleil, il s'est mis à pleuvoir abondamment dans l'après-midi.
   Model Translation: Bien que les prévisions météorologiques prévoient le soleil, elle a commencé à pâtir en plein vent l'après-midi.
   BLEU Score: 0.18

10. English Sentence: translate English to French: The implications of artificial intelligence on society are both fascinating and deeply complex.
   Reference Translation: Les implications de l'intelligence artificielle sur la société sont à la fois fascinantes et profondément complexes.
   Model Translation: Les implications de l'intelligence artificielle sur la société sont à la fois fascinantes et profondément complexes.
   BLEU Score: 1.00

```

Wyniki

6. Bibliografia

1. <https://huggingface.co/docs/transformers/en/tasks/translation>
2. <https://summalinguae.com/pl/technologie-jezykowe/5-rodzajow-tlumaczenia-maszynowego/>
3. <https://gtelocalize.com/types-of-machine-translation/#3-Neural-Machine-Translation-NMT>
4. <https://cameronrwolfe.substack.com/p/t5-text-to-text-transformers-part>
5. <https://jsonformatter.curiousconcept.com/>
6. <https://www.deepl.com>