Artficial Intelligence Laboratory

Ch8. Matching Tokenizers and Datasets

Transformers for NLP

정보융합공학과 AI전공
정주경

부산대학교
PUSAN NATIONAL UNIVERSITY

목 차

Matching datasets and tokenizers

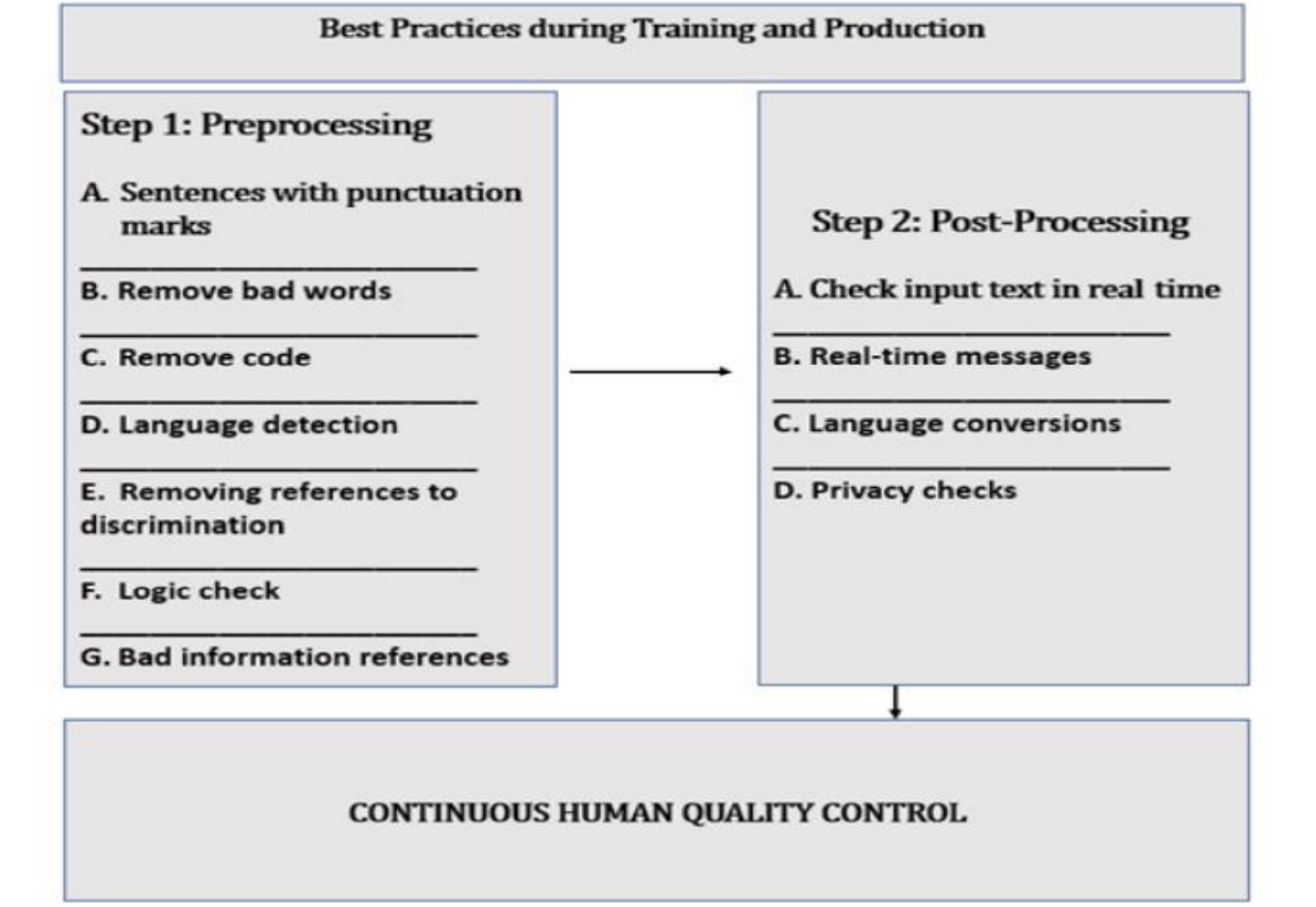
1. Matching datasets and tokenizers

1. Preprocessing
2. Post-processing
3. Continuous human quality control

Matching datasets and tokenizers

데이터 수집 및 관리

데이터셋 품질 관리 프로세스



Matching datasets and tokenizers

Step 1: Prepocessing

Standard heuristics to datasets

- Sentences with punctuation marks
  - 구두점(마침표, 물음표 등)으로 끝나는 문장 권장
- Remove bad words
  - 욕설, 혐오 표현
- Remove code
  - 일반적으로 nlp task의 경우 코드 제거
- Language detection
  - 데이터셋의 컨텐츠가 원하는 언어로 되어 있는지 확인
- Removing references to discrimination
  - 어떠한 형태의 차별도 막는게 좋음
- Logic check
  - 자연어 추론(NLI)을 수행하는 Transformer 모델을 실행
- Bad information references
  - 작동하지 않는 링크, 비윤리적인 웹 사이트 참조하는 text 제거

Matching datasets and tokenizers

Step 2: Post-processing

- **Check input text in real time**
  - 나쁜 정보 받아들이지 않기
  - 실시간으로 입력 구문 분석, 허용되지 않는 데이터 필터링(Preprocessing)

- **Real-time messages**
  - 거부된 데이터를 필터링 된 이유와 함께 저장하여 사용자가 로그를 참조할 수 있도록

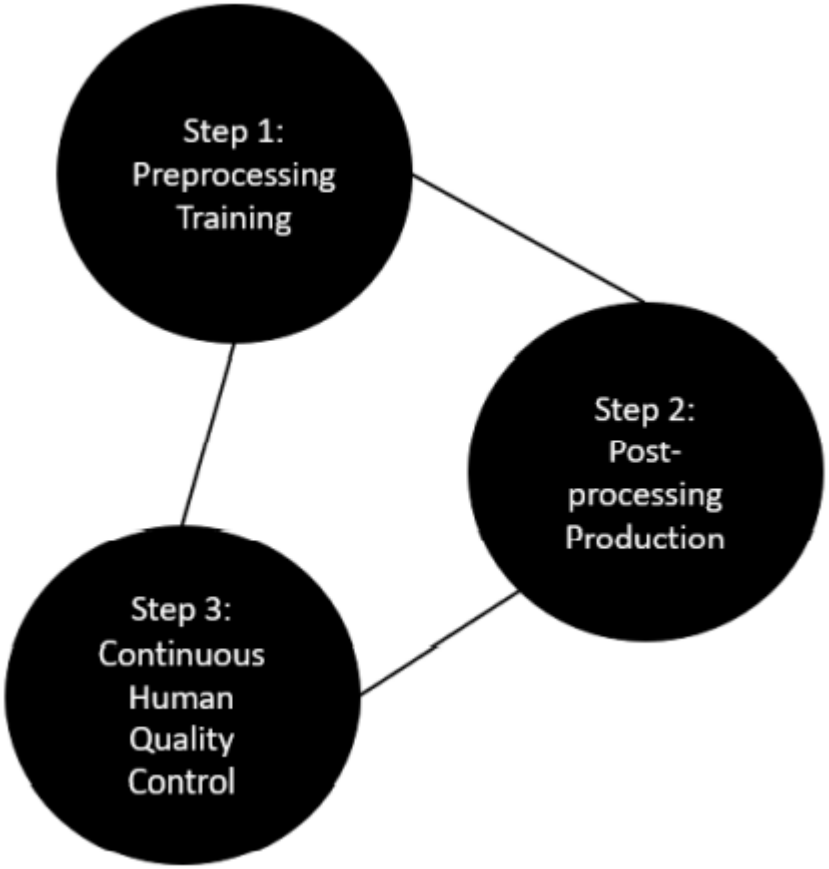
- **Language conversions**
  - rare vocabulary -> standard vocabulary

- **Privacy checks**
  - 데이터셋 및 작업에서 개인 정보 제외

Matching datasets and tokenizers

Step 3: Continuous human quality control

- Transformer는 복잡한 NLP작업 수행 => 인간의 개입은 여전히 의무적
- 접근 방법: Transformer training, 출력 제어, 결과를 다시 training set에 제공
- 지속적인 품질 관리 통해 Transformer의 training dataset을 개선, 성능 향상



Word2Vec tokenization

1. Word2Vec tokenization

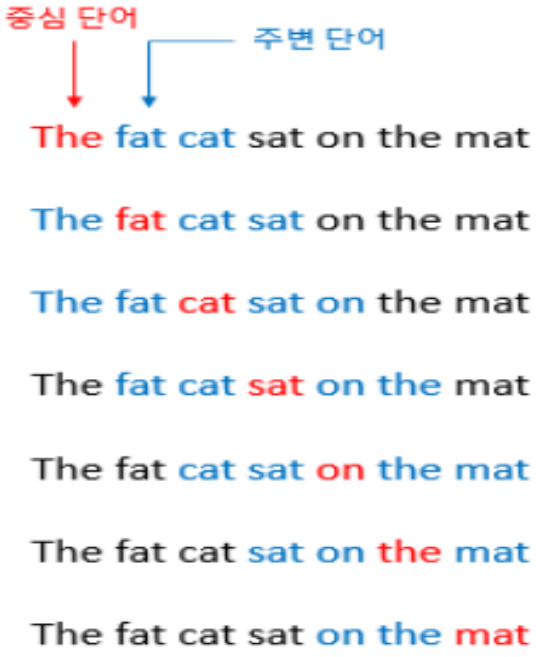
    1. Word2Vec
        - CBOW(Continuous Bag of Words)
        - Skip-gram

    2. Word pairs that tokenizers miscalculated

Word2Vec tokenization: CBOW

- You shall know a word by the company it keeps
- CBOW(Continuous Bag of Words): 주변 단어들을 입력으로 중간에 있는 단어 예측

  - 예측해야 하는 단어를 중심 단어, 예측에 사용되는 단어들을 주변 단어
  - Window: 중심 단어를 예측하기 위해 앞, 뒤로 몇 개의 단어를 볼지 결정
  - Sliding Window: window를 옆으로 움직여 주변 단어와 중심 단어 변경



중심 단어	주변 단어
[1, 0, 0, 0, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0]
[0, 1, 0, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0]
[0, 0, 1, 0, 0, 0, 0]	[1, 0, 0, 0, 0, 0, 0], [0, 1, 0, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0]
[0, 0, 0, 1, 0, 0, 0]	[0, 1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]
[0, 0, 0, 0, 1, 0, 0]	[0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 1, 0]	[0, 0, 0, 1, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 0, 1]
[0, 0, 0, 0, 0, 0, 1]	[0, 0, 0, 0, 1, 0, 0], [0, 0, 0, 0, 0, 1, 0]

Word2Vec tokenization: Skip-gram

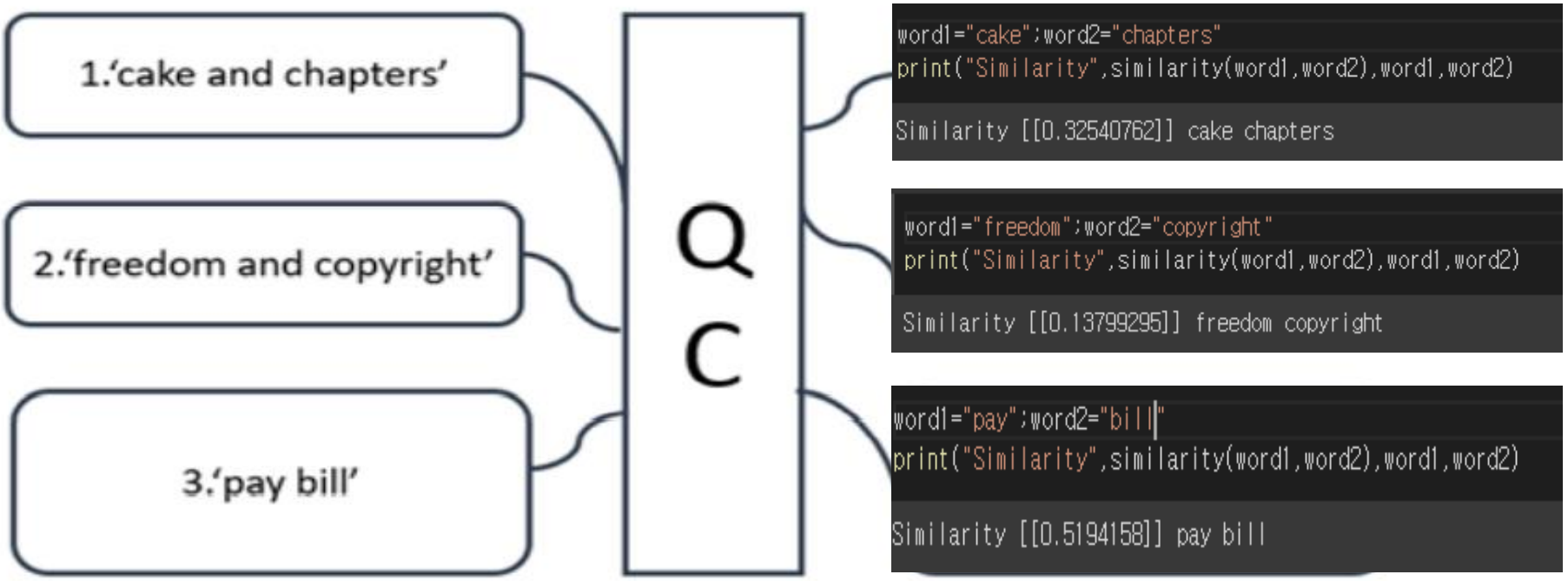
- Skip- Gram: 중간 단어들을 입력으로 주변 단어 예측

중심 단어	주변 단어
cat	The
cat	Fat
cat	sat
cat	on
sat	fat
sat	cat
sat	on
sat	the

- 전반적으로 Skip-gram이 CBOW보다 성능이 좋다고 알려져 있음

Word2Vec tokenization

- Word pairs that tokenizers miscalculated
  - 1. Cake와 Chapter는 서로 어울리지 않지만 토크나이저는 cosine similarity가 높음
  - 2. Freedom은 언론의 자유를 의미, copyright은 무료 전자책의 편집자가 적은 note 의미
  - 3. bill은 지불할 금액, 권리 청구서의 의미를 가지는 Polysemy(다의어)



QC = Quality Control

Examples of some of the limits encountered with tokenizers

1. Examples of some of the limits encountered with tokenizers
    1. Case 0: Words in the dataset and the dictionary
    2. Case 1: Words not in the dataset or the dictionary
    3. Case 2: Noisy relationships
    4. Case 3: Rare words
    5. Case 4: Replacing rare words
    6. Case 5: Entailment

Examples of some of the limits encountered with tokenizers

Case 0: Words in the dataset and the dictionary

- “freedom” and “liberty” are in the dataset
- Freedom: 아무런 규제가 없는 상태
- Liberty: 부당한 규제로부터 해방, 합당한 규제는 있어도 된다는 뜻을 지님

=> 단어의 정의는 비슷하지만 역사적, 문화적으로 갖는 가치가 다름

```
word1="freedom";word2="liberty"
print("Similarity",similarity(word1,word2),word1,word2)
```

Output:

book

```
Similarity [[0.79085565]] freedom liberty
```

current

```
Similarity [[0.38110557]] freedom liberty
```

Examples of some of the limits encountered with tokenizers

Case 1: Words not in the dataset or the dictionary

- "corporations(회사)" and "rights(권리)"

- The dictionary does not contain the word 'corporations'

```
word1="corporations";word2="rights"
print("Similarity",similarity(word1,word2),word1,word2)
```

```
corporations :[unk] key not found in dictionary
Similarity 0 corporations rights
```

⇒ 누락된 단어는 문제를 유발, 트랜스포머 모델의 output을 왜곡

- Several possibilities need to be checked:

- Unk는 데이터셋에 있지만 tokenized dictionary에는 선택X
- Unk가 데이터셋에 포함X => 'corporations'가 왜 dictionary에 없는지 설명
- 사용자가 Transformer에 입력을 보내고 Transformer가 토큰화 X

=> BPE를 쓰면 일부 해결

Examples of some of the limits encountered with tokenizers: BPE

Case 1: Words not in the dataset or the dictionary : BPE

- **If a team only uses byte-level BPE to fix the problem:**

- “corporations” => “corp” + “o” + “ra” + “tion” + “s”

- **unk는 word pieces로 분해**
- **원래 토큰의 의미를 전달X**

- **Transformer의 성능 : 0.8 -> 0.9**

- **좋아 보일 수 있지만 영어에서 corp는 corporation 또는 corporal를 의미**

=> Corp와 다른 단어 사이에 혼란과 나쁜 연관성

Examples of some of the limits encountered with tokenizers

Case 2: Noisy relationships

- **The dataset contained the words “etext” and “declaration”**
  - **“etext” : Project Gutenberg’s 사이트 내 각 전자책 서문을 의미**
  - **“declaration” : Declaration of Independence(독립선언서)의 실제 내용과 관련된 단어**

```
word1="etext";word2="declaration"
print("Similarity",similarity(word1,word2),word1,word2)
```

Output: book

```
Similarity [[0.880751]] etext declaration
```

current

```
Similarity [[0.56551445]] etext declaration
```

⇒ **트랜스포머가 텍스트를 생성할 때 “etext is a declaration”같은 잘못된 자연어 추론 생성**

Examples of some of the limits encountered with tokenizers

Case 3: Rare words

- 데이터셋에 있지만 눈에 띄지 않았거나 모델이 제대로 학습X
- 의학, 법률, 공학 용어 같은 전문 용어
- Slang(속어)
- 미국, 영국, 호주 등 다른 영어권 나라들마다의 변형
- 수세기전에 쓰여져 잊혀지거나 전문가들만이 사용

⇒ Rare words는 특정 작업에 대한 트랜스포머 출력을 파괴할 수 있음

- "justiciar" : (중세)법무 장관, 고등 법원 판사

```
word1="justiciar";word2="judgement"
print("Similarity",similarity(word1,word2),word1,word2)

Similarity [[0.2268471]] justiciar judgement
```

- 'justiciar': 13세기 초 Magna Carta에서 추출

=> Magna Carta의 몇몇 조항들은 21세기 영국에서 여전히 유효

Magna Carta: 1215년에 영국 국민의 법적 및 정치적 권리 확인서. 흔히 영국 현대법의 기초로 여겨짐

Examples of some of the limits encountered with tokenizers

Case 4: Replacing rare words

- ‘justiciar’의 어원은 French Latin-like “judicaire(법, 판단력, 사법)”
- ‘justiciar’를 같은 meta-concept을 전달하는 ‘judge’로 대체
- 현대적 의미인 ‘judge’로 ‘judgement’유사도 비교

```
word1="justiciar";word2="judge"
print("Similarity",similarity(word1,word2),word1,word2)

word1="judge";word2="judgement"
print("Similarity",similarity(word1,word2),word1,word2)
```

```
Similarity [[0.33639437]] justiciar judge
Similarity [[0.15304697]] judge judgement
```

=> 0.9를 초과하는 상관 관계를 찾을 때까지 대체 단어 찾기

Examples of some of the limits encountered with tokenizers

Case 5: Entailment

- Entailment verification을 위해 cosine similarity 계산
- “Pay” + “debt” (빚을 갚다) 고정된 순서로 유사도 계산

```
word1="pay";word2="debt"
print("Similarity",similarity(word1,word2),word1,word2)
```

```
Similarity [[0.5163679]] pay debt
```

- 여러 단어 쌍으로 데이터 집합을 확인하고 의미가 있는지 확인
- Cosine Similarity가 0.9이상이면, 데이터셋에 추가할 수 있음

Standard NLP tasks with specific vocabulary

1. Standard NLP tasks with specific vocabulary
    1. Generating unconditional samples with GPT-2

    2. Controlling tokenized data

Standard NLP tasks with specific vocabulary

- **Generating unconditional samples with GPT-2**

- GPT-2가 의료 텍스트에 어떻게 대처하는지 살펴보기
- 데이터셋에는 Matrina Conte, Nadia Loy(2020)의 *"Multi-cue kinetic model with non-local sensing for cell migration on a fibers network with chemotaxis"* 논문 포함

output

```
======== SAMPLE 1 ========
 Therefore, we shall consider the quantities
S(x + λvˆ), q(x + λvˆ, vˆ), ∀ x ∈ Ω, ∀ vˆ ∈ S
d-1
, λ ≤ R.
Of course, next to the border of the domain Ω, we shall always consider λ such that x + λvˆ ∈ Ω.
In order to analyze qualitatively the impact of the non-locality at the macroscopic level, we
study, as previously done in [36, 37], the impact of the directional cues S and q with respect to
the size of the cell, that is related to its sensing radius R. Thus, we introduce the characteristic
length of variation of S as
|S :=
1
max
x∈Ω
|∇S|
S
```

- 생성된 문장의 구조는 비교적 허용 가능
- 출력의 문법은 나쁘지 않다
- 비전문가에게는 결과가 사람이 생성한 것처럼 보일 수 있다.

Standard NLP tasks with specific vocabulary

Generating unconditional samples with GPT-2

Therefore, we shall consider the quantities
(그러므로 우리는 수량을 고려할 것이다.)
Of course, next to the border of the domain Ω, we shall always consider λ such that x + λvˆ ∈ Ω.
(물론 도메인 Ω의 경계 옆에서, 우리는 항상 x + λvˆ ∈ Ω가 되도록 λ을 고려해야 한다.)
In order to analyze qualitatively the impact of the non-locality at the macroscopic level, we
study, as previously done in [36, 37], the impact of the directional cues S and q with respect to
the size of the cell, that is related to its sensing radius R.
(거시적 수준에서 비국소성의 영향을 정성적으로 분석하기 위해, 이전 [36, 37]에서 수행한 것처럼, 감지 반경 R과 관련된 셀의 크기에 대한 방향 신호 S와 q의 영향을 연구한다.)
Thus, we introduce the characteristic length of variation of S as
(따라서, 우리는 S의 특징적인 변화 길이를 도입한다.)

- 문법상 오류는 없어 보이지만 내용은 말이 안 되는 경우가 있음
- 의료데이터를 넣었음에도 원하는 내용의 샘플을 찾기가 힘듬

=> GPT 모델의 사전학습 데이터를 기준으로는 rare words이기 때문

Standard NLP tasks with specific vocabulary

- **Generating unconditional samples with GPT-2**

- 데이터셋의 크기를 늘리기

⇒ 우리가 찾고 있는 데이터가 들어 있는지 확신X, 더 많은 데이터로 나쁜 상관관계

EX) 코로나를 포함하는 데이터셋 사용

- 코로나는 위험한 바이러스는 아니지만, 일반적인 독감 같습니다.
- 코로나는 매우 위험한 바이러스입니다.
- 코로나는 바이러스가 아니라 실험실에서 만든 것입니다.
- 코로나는 연구소에서 만든 것이 아닙니다.
- 백신은 위험합니다.
- 백신은 구세주입니다.
- 정부가 팬더믹을 제대로 관리하지 못했습니다.
- 정부는 필요한 일을 했습니다.

⇒ 데이터셋이 더 작아야 하고 과학 논문의 콘텐츠로 제한되어야 한다는 것을 의미

⇒ 신뢰할 수 있는 결과를 내기 위해서는 많은 노력이 필요

Standard NLP tasks with specific vocabulary

Controlling tokenized data

- 사전훈련된 토크나이저로 인코딩된 GPT-2 모델에 첫 단어 입력

This suggests that →

```
This : 1212
Ġsuggests : 5644
Ġthat : 326
```

- Rare word: “amoeboid(아메바 같은)”

amoeboid →

```
Ġam : 716
o : 78
eb : 1765
oid : 1868
```

- “amoeboid” => “am” + “o” + “eb” + “oid”
- “am”은 “I am”과 같은 다른 순서로 인식 가능성 있음
- “oid”는 “tabloid”에서 분해된 것으로 학습될 수 있음

- Tokenizer 한계

Word2Vec : amoeboid같은 rare word가 없을 가능성이 높음

BPE: Polysemy(다의어)문제 ex) am => I + am, am + bush (매복)

T5 Bill of Rights Sample

1. T5 Bill of Rights Sample
    1. Summarizing the Bill of Rights, version1

    2. Summarizing the Bill of Rights, version2 (modern everyday English)

T5 Bill of Rights Sample

Summarizing the Bill of Rights, version1

- **text**

No person shall be held to answer for a capital, or otherwise infamous crime, unless on a presentment or indictment of a Grand Jury, exceptin cases arising in the land or naval forces, or in the Militia, when in actual service in time of War or public danger; nor shall any person be subject for the same offense to be twice put in jeopardy of life or limb; nor shall be compelled in any criminal case to be a witness against himself,nor be deprived of life, liberty, or property, without due process of law;nor shall private property be taken for public use without just compensation.

- **Summarized text**

no person shall be held to answer for a capital, or otherwise infamous crime. except in cases arisingin the land or naval forces or in the militia, when in actual service in time of war or public danger

T5 Bill of Rights Sample

Summarizing the Bill of Rights, version2 (modern everyday English)

- **text**

A person must be indicted by a Grand Jury for a capital or infamous crime. There are excpetions in time of war for a person in the army, navy, or national guard. A person can not be judged twice for the same offense or put in a situation of double jeopardy of life. A person can not be asked to be a witness against herself or himself. A person cannot be deprived of life, liberty or property without due process of law. A person must be compensated for property taken for public use.

- **Summarized text**

there are exceptions in time of war for a person in the army, navy, or national guard. no person can be deprived of life, liberty or property without due process of law. there must be compensation for property taken for

⇒ **T5모델은 현대 일상 영어로 많이 학습되었기 때문**

Conclusion

- 특정 프로젝트의 경우 특정 데이터셋에 대해 트랜스포머 훈련을 계속 해야함

Summary

1. Transformer 훈련시 데이터셋 품질 관리 프로세스로 Preprocessing, Post-Processing, 지속적인 인간의 품질 관리

2. Word2Vec은 희귀한 단어가 나오면 [unk]가 나오는 문제

3. BPE를 통해 일부 해결할 수 있지만 다의어 문제

4. 특정 프로젝트의 경우 특정 데이터셋에 대하여 Transformer 훈련을 계속 해야함