단어에서 분자로, 제1부 (Word to Molecule, Part I)

성예닮 (Yedarm Seong) 서울대학교 인공지능 협동과정 (Interdisciplinary Program in Artificial Intelligence, Seoul National University)

2025년 11월 7일 (Nov 07, 2025)

목차 (Table of Contents)

1. 언어 모델(LMs)의 짧은 역사 (Brief History of Language Models (LMs))
2. 토큰화 (Tokenization)
3. 디코더 전용 및 인코더 전용 언어 모델 (Decoder-only & Encoder-only LMs)
4. 분자 표현법 (Molecule Representations)
5. 화학 언어 모델 (Chemical LMs)
6. 참고문헌 (References)

1. 언어 모델(LMs)의 짧은 역사 (Brief History of Language Models (LMs))

언어 모델(LMs)을 위해 단어를 어떻게 표현할 것인가? (How to represent words for LMs?)

- 원-핫 인코딩 및 단어 가방 (One-hot Encoding & Bag-of-Words)
- 분산 표현 (Distributed Representation) (Word2Vec, GloVe 등)
- 문맥화된 표현 (Contextualized Representation) (Transformer, BERT, GPT 등)

원-핫 인코딩 (One-hot Encoding)

정의 (Definition)

- 어휘 사전(Vocabulary)의 크기가 V일 때, 각 단어는 V차원의 벡터(Vector)로 표현됩니다. 이 벡터는 해당 단어에 대응하는 인덱스(Index)만 '1'이고 나머지는 모두 '0'인 형태입니다.

예시 (Example)

- "the cat sat on the mat"
  1. 어휘 사전(Vocabulary) 구축: {the, cat, sat, on, mat} (크기 $V = 5$)
  2. 인덱스(Index) 할당: the(0), cat(1), sat(2), on(3), mat(4)
  3. 벡터(Vector) 변환: $cat = [0, 1, 0, 0, 0]$, $mat = [0, 0, 0, 0, 1]$

단어 가방 (Bag-of-Words, BoW)

정의 (Definition)

- 문서(문장이나 기사 등) 전체를 하나의 수치 벡터(Numerical vector)로 표현하는 간단한 방법입니다. 문법(Grammar)이나 단어 순서(Word order)는 무시하고 어휘 사전에 포함된 각 단어의 출현 빈도(Frequency)만 계산합니다. 비유적으로 문서의 모든 단어를 "가방(Bag)"에 넣고 빈도만 고려하는 방식입니다.

2. 토큰화 (Tokenization)

토큰화란 무엇인가? (What is Tokenization?)

- 텍스트(Text)를 컴퓨터가 이해할 수 있는 숫자 시퀀스(Sequence of numbers), 즉 토큰(Tokens)으로 변환하는 과정입니다.
- 언어 모델(BERT, GPT 등)은 수치 입력(Numerical inputs)을 필요로 하며, 텍스트는 처리를 위해 의미 있는 단위(Meaningful units)로 분할되어야 합니다.

고전적인 문제: 어휘 사전 외 단어 (Out-of-Vocabulary, OOV)

- 어휘 사전이 "단어(Word)" 단위로 구축되면, 사전에 없는 단어(신조어, 오타, 이름 등)는 모두 `<UNK>` (Unknown)로 처리되어 상당한 정보 손실이 발생합니다.

서브워드(Subwords)의 등장

- 단어를 더 작고 의미 있는 단위인 서브워드(Subwords)로 분해합니다.
- **장점 (Advantages)**:
  - 어휘 사전 외 단어(OOV) 문제 해결: 모르는 단어도 아는 서브워드의 조합으로 표현 가능합니다.
  - 형태학적 정보(Morphological information) 유지: "playing"과 "plays"가 공통 어근 "play"를 공유함을 학습 가능합니다.
  - 어휘 사전 크기 조절: 거대한 단어 단위 사전과 너무 작은 문자 단위 사전 사이의 균형을 맞춥니다.

3. 디코더 전용 및 인코더 전용 언어 모델 (Decoder-only & Encoder-only LMs)

디코더 전용 언어 모델 (Decoder-only LM): 인과적 언어 모델링 (Causal Language Modeling)

- 자기회귀(Autoregressive) 모델링으로도 알려져 있습니다.
- **정의 (Definition)**: 현재 시점 i까지의 모든 이전 토큰들($w_1, \ldots, w_{i-1}$)이 주어졌을 때, 다음 토큰(w_i)을 예측하는 작업입니다.

인코더 전용 언어 모델 (Encoder-only LM): 마스크 언어 모델링 (Masked Language Modeling, MLM)

- **정의 (Definition)**: 입력 시퀀스(Input sequence)의 특정 토큰을 무작위로 마스킹(Masking)하고, 주변 문맥(Context)을 바탕으로 마스킹된 토큰을 예측하도록 모델을 학습시키는 방식입니다. 양방향(Bidirectional) 문맥을 모두 고려합니다.

4. 분자 표현법 (Molecule Representations)

컴퓨터 작업을 위해 분자를 표현하는 일반적인 방법들:

- **분자 그래프 (Molecular Graphs)**: 노드(Nodes)는 원자, 에지(Edges)는 결합을 나타냅니다.
- **SMILES (Simplified Molecular Input Line Entry System)**: 분자를 선형 문자열(Linear strings)로 표현하는 널리 사용되는 방식입니다.
- **SELFIES (Self-Referencing Embedded Strings)**: 항상 유효한 분자 구조를 생성하도록 설계된 견고한 문자열 표현법(Robust string representation)입니다.

- **InChI (International Chemical Identifier)**: 화학 물질의 구조를 인코딩하는 표준화된 텍스트 식별자(Standardized textual identifier)입니다.

5. 화학 언어 모델 (Chemical LMs)

화학 언어 모델(Chemical LMs)은 화학 구조와 특성을 이해하고 생성하기 위해 설계된 전문 언어 모델입니다. 자연어 처리(NLP) 기술을 활용하여 SMILES, SELFIES, InChI와 같은 화학 표현법(Chemical representations)을 처리합니다. 신약 개발(Drug discovery), 재료 과학(Materials science), 분자 특성 예측(Molecular property prediction) 등에 응용됩니다.

InChI 기반 화학 언어 모델(Chemical LM)의 특징

- **다중 토큰화 (Multi-tokenization)**: / (계층, Layers) 별로 분할하여 화학식(Formula), 연결성(Connectivity), 수소(Hydrogens), 입체화학(Stereochemistry) 계층을 각각 토큰화합니다.
- **계층 내/계층 간 문맥 (Intra-/Inter-layer Context)**: 로컬 어텐션(Local attention, 계층 내 문맥 파악)과 글로벌 어텐션(Global attention, 전체 시퀀스의 장거리 의존성 파악)을 조합하여 InChI 문자열을 효과적으로 학습합니다.

6. 참고문헌 (References)

Firth, John (1957). “A synopsis of linguistic theory, 1930-1955”. In: Studies in linguistic analysis, pp. 10–32.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: Advances in neural information processing systems 30.

Radford, Alec et al. (2018). “Improving language understanding by generative pre-training”. In.

Devlin, Jacob et al. (2019). “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186.

Gage, Philip (1994). “A new algorithm for data compression”. In: C Users Journal 12.2, pp. 23–38.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). “Edinburgh neural machine translation systems for WMT 16”. In: arXiv preprint arXiv:1606.02891.

Li, Juncai and Xiaofei Jiang (2021). “Mol-BERT: An Effective Molecular Representation with BERT for Molecular Property Prediction”. In: Wireless Communications and Mobile Computing 2021.1, p. 7181815.

Wang, Sheng et al. (2019). “Smiles-bert: large scale unsupervised pre-training for molecular property prediction”. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics, pp. 429–436.