GPT를 이용한 소재문헌 자연어처리

chatGPT와 함께하는 미래 소재 개발의 시작! day 2

최재웅 박사

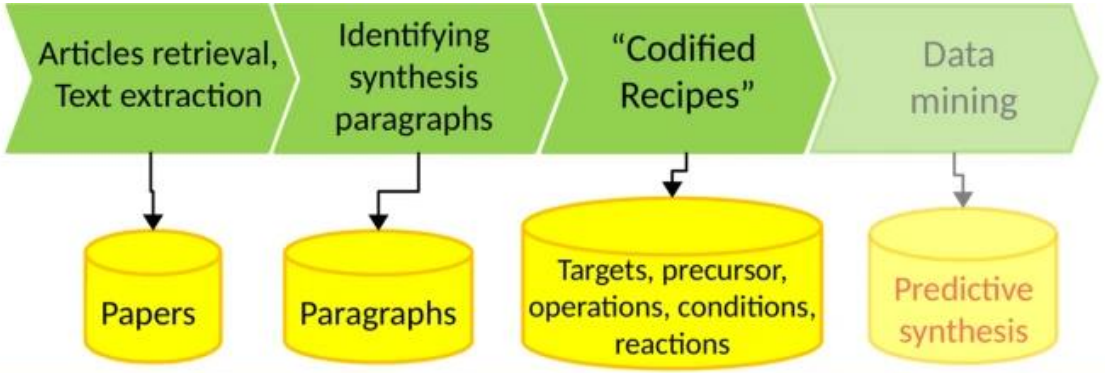
한국과학기술연구원 계산과학연구센터

2023.08.17

Today Contents

- ❖ 소재문헌 자연어처리
- ❖ Text classification
  - ✓ 실습: Zero-shot learning for text classification
- ❖ Information extraction

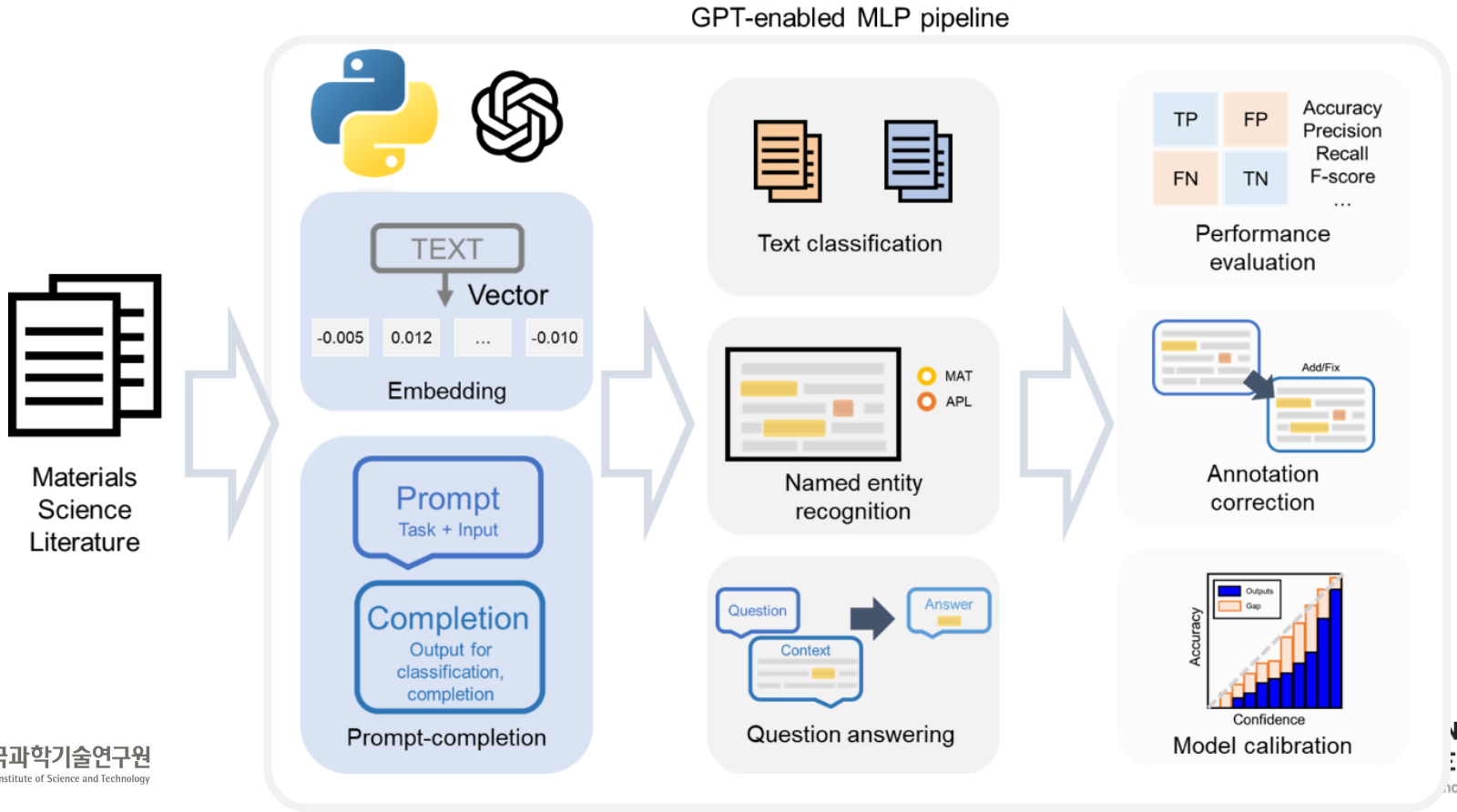
- ❖ **Why? 재료과학 분야, 소재 관련 문헌에는 실험에 대한 보고가 많음**
  - ✓ 이를 자동화된 방식으로 정리하면, 양질의 데이터베이스를 구축하고 새로운 소재 발굴 등의 시작점으로 활용할 수 있음
- ❖ **How? 사람이 하던 방식을 자연어처리와 기계학습으로 대체함**
  - ✓ 논문을 검색하고, 원하는 내용(예: 배터리 합성)을 다루는 논문을 선별하고, 논문 내에서 관련 파트(문단, 문장 수준)를 찾는 과정을 자연어처리를 통해 용이하게 함



Sci Data **6**, 203 (2019).

❖ Materials Language Processing (소재문헌 자연어처리),,,

- ✓ 소재문헌은 화학, 물리학 등의 지식이 반영된 텍스트 데이터
  - 소재문헌의 자연어 처리는 단순한 응용이 아닌 새로운 과학의 출발점이 될 수 있음



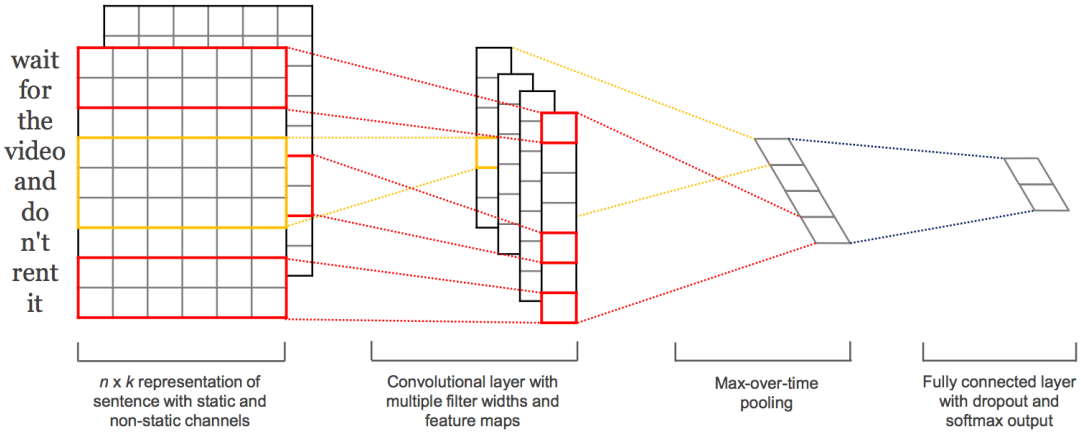
❖ 텍스트 분류

- ✓ 주어진 텍스트를 사전 정의된 카테고리로 분류하는 작업
- ✓ 일반적인 예시: 감성 분석, 스팸 필터링, 페이크 뉴스 탐지, 언어 감지, 주제 분류, 인용 추천 등
- ✓ 소재문헌 자연어처리에서의 예시
  - 논문 분류: 검색 결과에서 유효 집합 선별
    - 예: “battery”로 검색한 논문 집합(검색결과)에서, 배터리 실험 관련 논문을 초록(Abstract)을 기준으로 선별
    - Well-defined Category(분류 결과)와 대량의 Labelled dataset(분류 대상) 필요
  - 문단 분류: 유효 집합의 전체 본문에서 특정 문단 선별
    - 예: 위에서 선별한 배터리 실험 논문의 전체 문단 집합에서 Cell Synthesis 관련 문단을 선별
    - 이 또한, category와 그에 따른 labelled dataset 필요

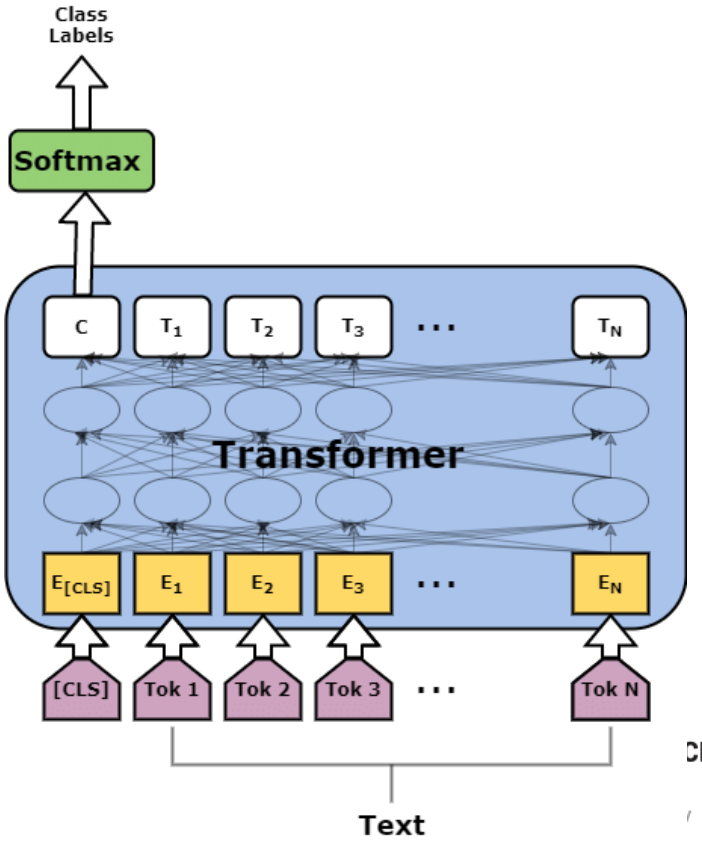
❖ 기존 텍스트 분류 모델

- ✓ (1) Text with Labels: 분류 목표(카테고리)가 있는 대량의 데이터
- ✓ (2) Model: hyperplane에서의 nonlinear function 추정
  - document representation
  - classification layer

Text CNN



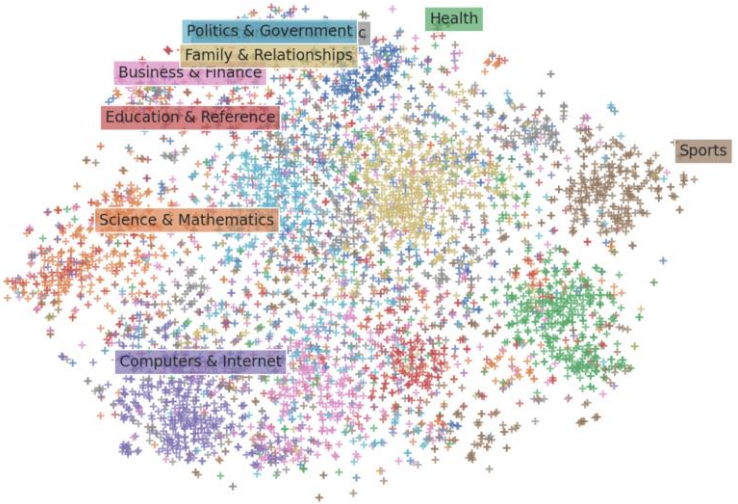
BERT classifier



https://paperswithcode.com/sota/text-classification-on-glue

❖ GPT를 활용한 텍스트 분류 모델

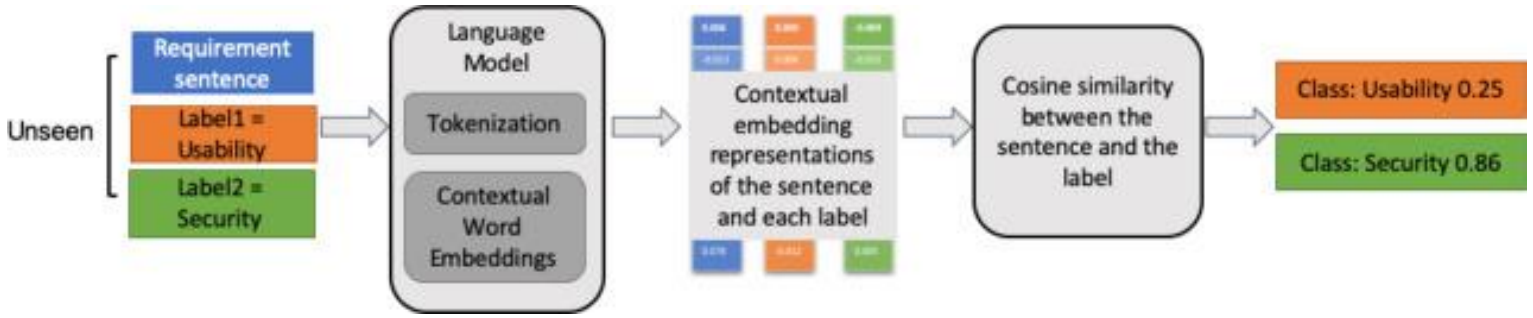
- ✓ 라벨링한 데이터 없이는 분류를 못할까? → Zero-shot learning
  - 학습 과정 중에 보지 못한 클래스에 대해 예측하는 작업
- ✓ How?
  - Attribute: In CV, provide more attribute of image
  - Descriptor: In NLP, embed the labels in the space of document embedding.



https://huggingface.co/tasks/zero-shot-classification

❖ GPT를 활용한 텍스트 분류 모델 (ZSL)

- ✓ 라벨이 있는 데이터(train/test)
  - 라벨: original label, expert-curated label, word embedding label 등 사용가능
- ✓ Model: GPT의 경우, text-embedding-ada-002
  - BERT, SentenceBERT 등의 다른 PLM을 통해, representation 가능



Zero shot learning 기반 문서 분류 과정

Information and Software Technology 159 (2023): 107202. ESANN (2016), pp. 423-428

❖ GPT를 활용한 텍스트 분류 모델 (ZSL)

- ✓ 실습: 배터리 관련 논문 검색 결과로부터 배터리 소재 관련 논문 찾기
  - BatteryBERT: A Pretrained Language Model for Battery Database Enhancement

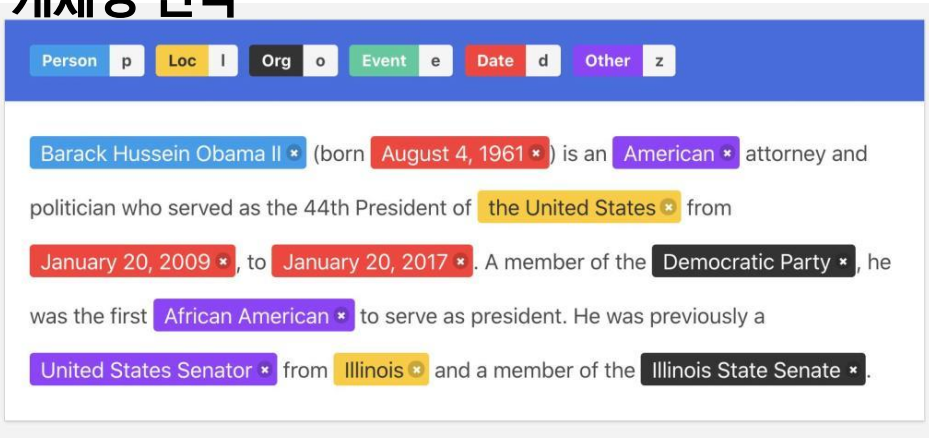
Jupyter Notebook 참고 (tutorials_GPT_MLP.ipynb)

Journal of Chemical Information and Modeling 62.24 (2022): 6365-6377.

❖ 정보 추출 (Information Extraction)

- ✓ 텍스트(비정형화된 데이터)로부터 구조화된 정보를 추출하는 작업
- ✓ 대표적으로, 개체명 인식 (Named Entity Recognition), 관계 추출 (Relation Extraction), 이벤트 추출 (Event extraction) 등이 있음
  - 개체명 인식: 텍스트에서 이름이 있는 중요한 개체를 식별 (어떤 부분이 어떤 유형인지를 탐지)
  - 관계 추출: 두 개체(Entity) 간의 관계를 식별하고 추출
  - 이벤트 추출: 개체명 인식과 구문 분석을 통해 얻은 정보를 활용하여 이벤트(사건이나 동작)를 추출 ~ Subject Action Object

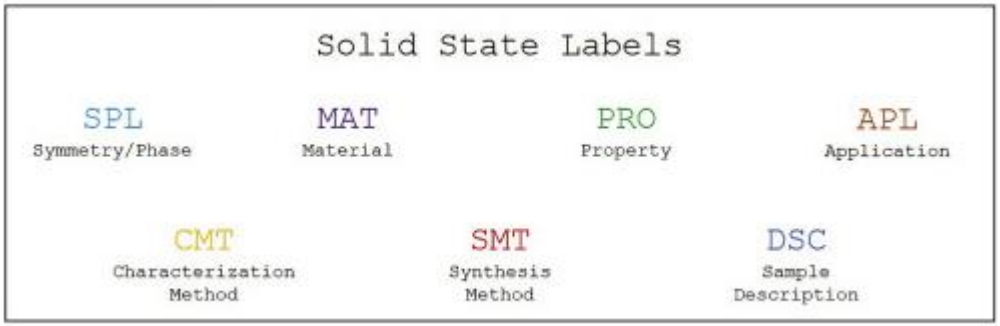
개체명 인식



관계 추출

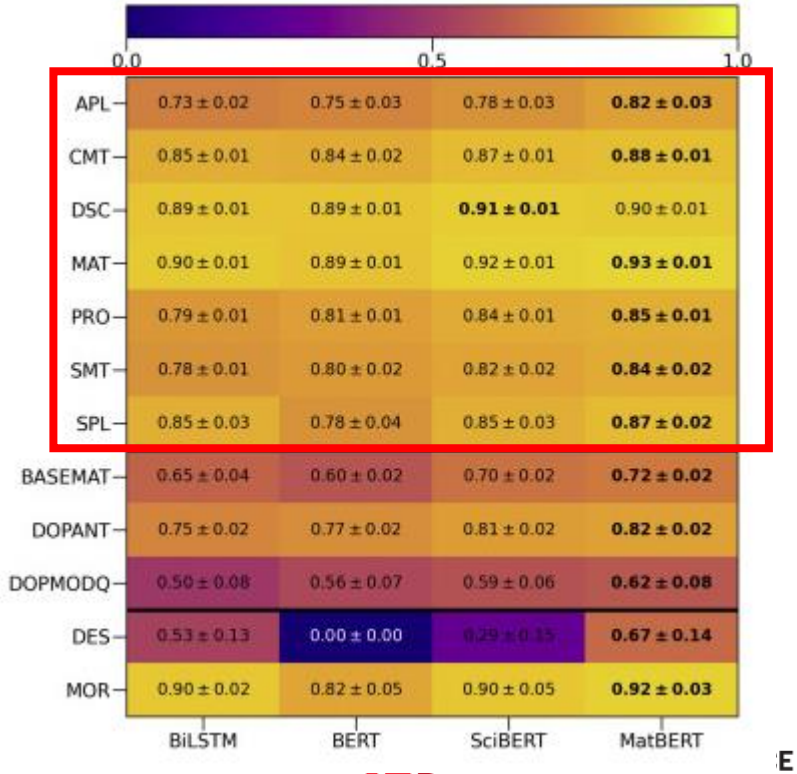
❖ 소재 문헌에서의 개체명 인식 예시 (1)

- ✓ 주어진 텍스트(문단)에서 자동으로 소재(MAT; Materials), 물성(PRO; Property), 응용분야(APL; Applications) 을 추출할 수 없을까?



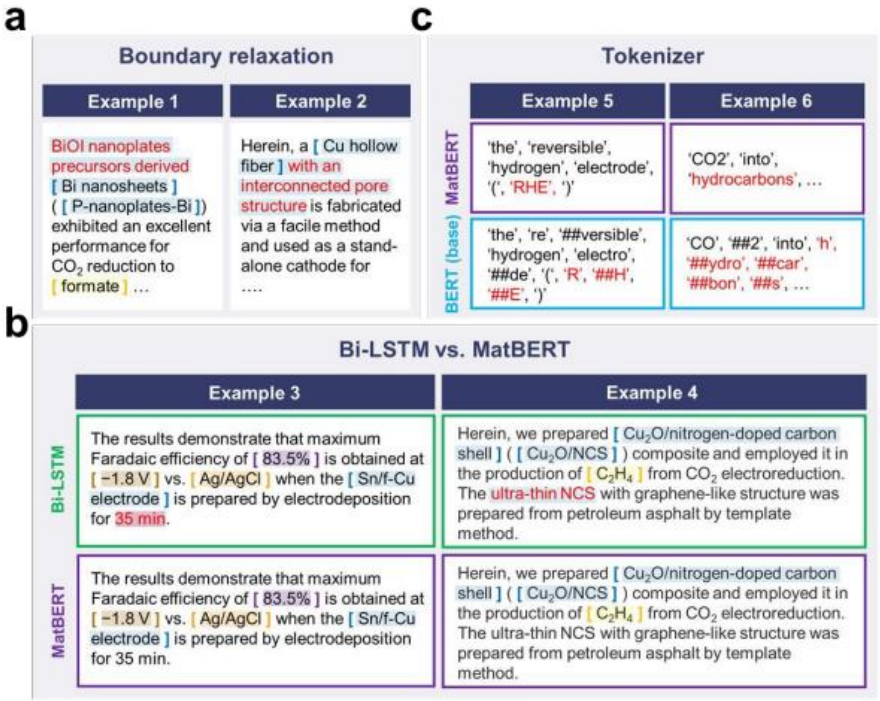
Monoclinic zirconia nanowires were synthesized by chemical vapor decomposition using ZrCl4 powder as a starting material. Based on x-ray diffraction and Raman spectroscopy data, the resulting crystal structure was found to be single crystalline monoclinic zirconia. An Au thin film was pre-depositing on a graphite substrate as a catalyst.

Patterns 3.4 (2022)

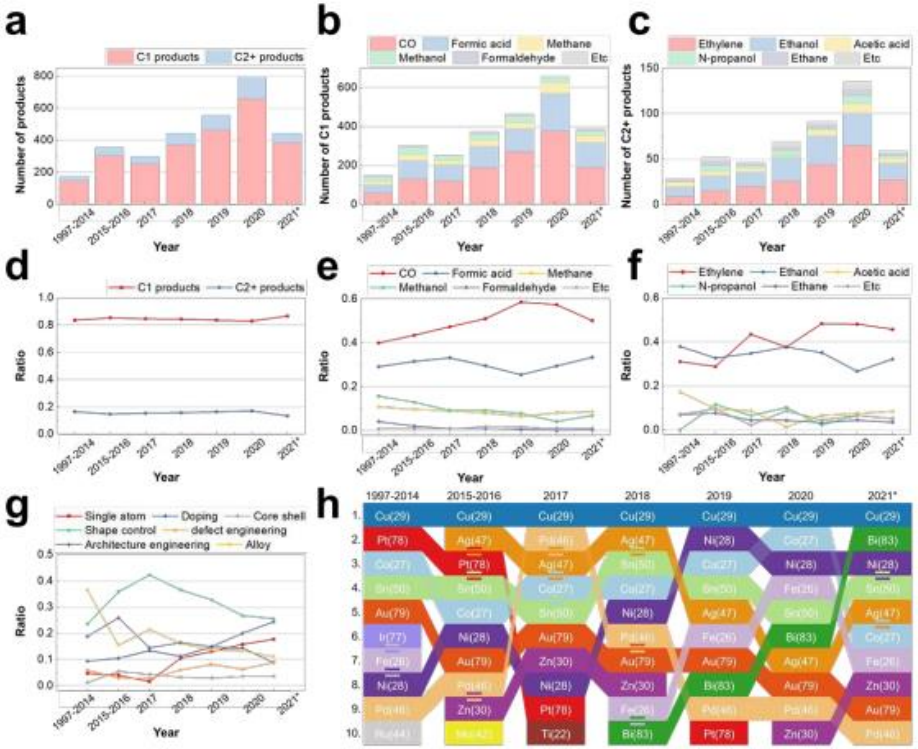


❖ 소재 문헌에서의 개체명 인식 예시 (2)

- ✓ 촉매 문헌에서 catalyst, product, electrolyte, reference electrode 등의 물질을 추출하고, current density, faradaic efficiency, stability hour 등의 물성을 추출

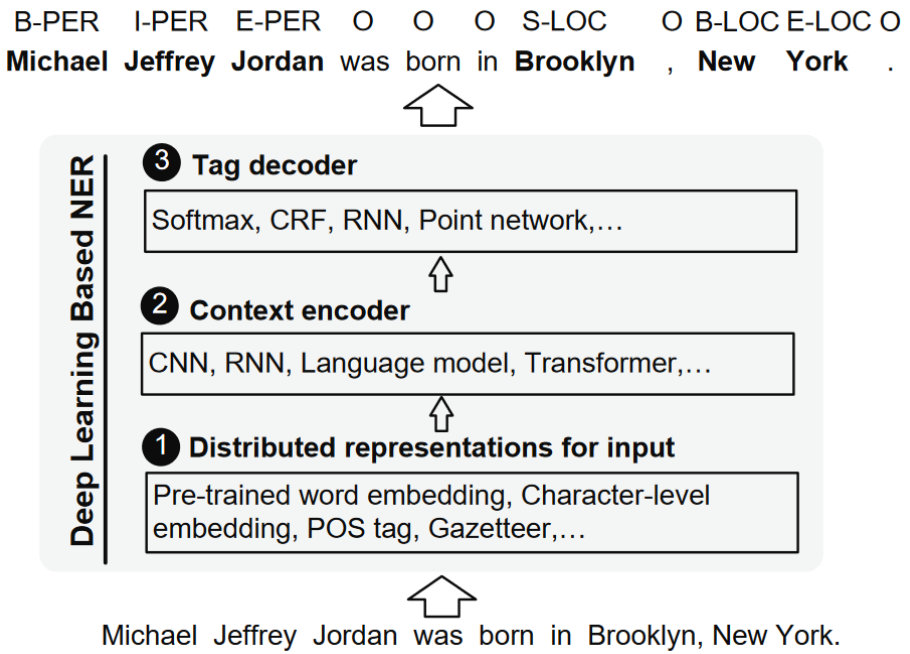


JMCA (2023)



❖ 기존 개체명 인식 모델

- ✓ Sequence 인코딩: 단어의 의미
- ✓ Context 인코딩: Sequential 특성 반영
- ✓ Tag 디코딩: Tag 전후관계 반영



원본 문장:
"Yesterday afternoon, John J. Smith travelled to Washington."

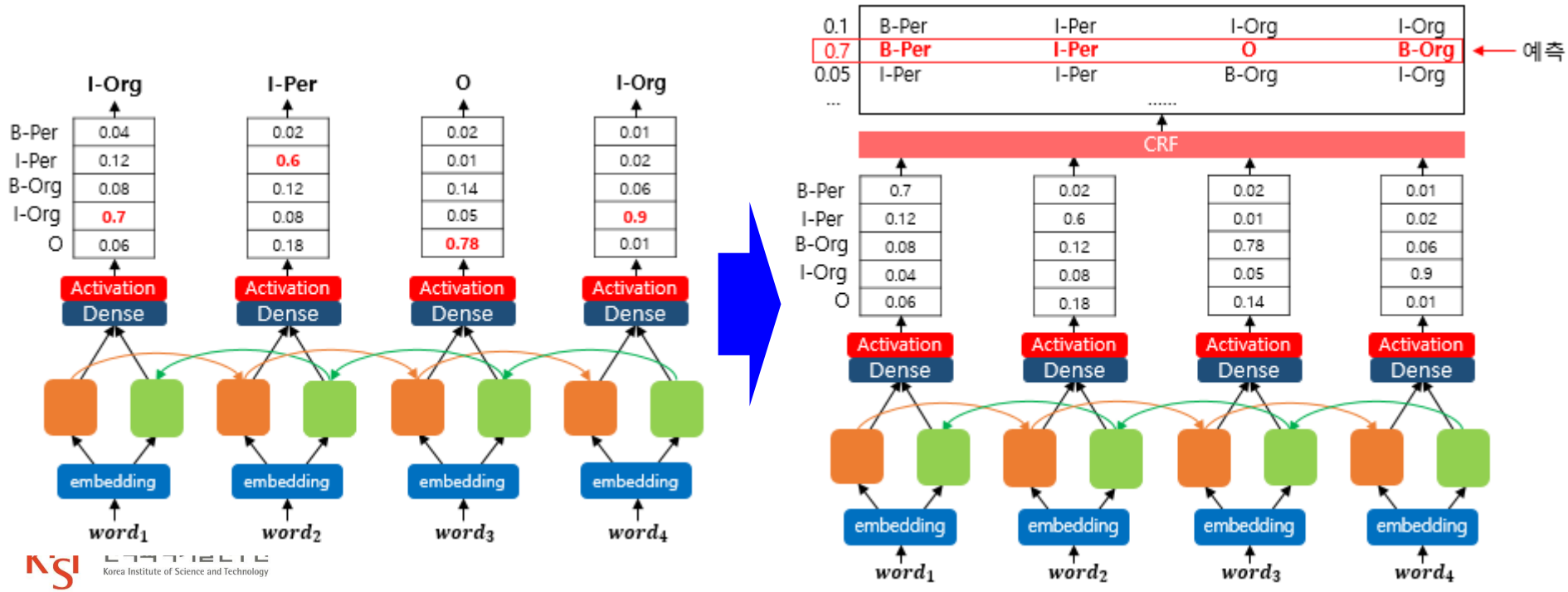
태깅 예시

Tokens	IO	BIO	BMEWO	BMEWO+
Yesterday	O	O	O	BOS_O
afternoon	0	O	O	O
,	0	O	O	O_PER
John	I_PER	B_PER	B_PER	B_PER
J	I_PER	I_PER	M_PER	M_PER
.	I_PER	I_PER	M_PER	M_PER
Smith	I_PER	I_PER	E_PER	E_PER
traveled	O	0	O	PER_O
to	O	O	O	O_LOC
Washington	I_LOC	B_LOC	W_LOC	W_LOC
.	O	O	O	O_EOS

❖ 기존 개체명 인식 모델

✓ Tag 디코딩

- CRF: 시퀀스 내 단어의 레이블을 결정할 때, 이전 단어의 레이블과의 의존관계를 고려함
- 첫번째 단어의 BiLSTM 셀과 활성화함수를 거친 출력값은 CRF 층의 입력이 되고, 최종적으로 CRF layer는 레이블 시퀀스 중에서 가장 높은 점수를 가지는 시퀀스를 예측함



❖ GPT 기반의 모델을 어떻게 평가해야 할까 (recent works)

- ✓ Performance
  - Top N Accuracy, Precision, F-measure
- ✓ Explainability
  - 추론 과정에 대한 설명을 제공할 수 있는가 ~ Self-check and human-check 모두 요구됨
- ✓ Calibration
  - ECE score 등을 통해 모델의 불확실성 체크해야 함
- ✓ Faithfulness
  - 답변이 실제 input에서 비롯되었는지에 대한 체크가 필요함 (Manually)

❖ GPT 기반의 모델 평가 결과 (General Task)

- ✓ Performance: 어려운 task일수록 fine-tuned model이 ChatGPT의 prompt engineering 모델보다 성능이 높다!
  - 쉬운 Task에서는 ChatGPT가 높은 성능을 보였다!
- ✓ Calibration: BERT 기반 모델들이 ChatGPT 보다 신뢰성이 높다

Task	Dataset	BERT	RoBERTa	SOTA	ChatGPT
Entity Typing(ET)	BBN	80.3	79.8	82.2 (Zuo et al., 2022)	85.6
	OntoNotes 5.0	69.1	68.8	72.1 (Zuo et al., 2022)	73.4
Named Entity Recognition(NER)	CoNLL2003	92.8	92.4	94.6 (Wang et al., 2021)	67.2
	OntoNotes 5.0	89.2	90.9	91.9 (Ye et al., 2022)	51.1
Relation Classification(RC)	TACRED	72.7	74.6	75.6 (Li et al., 2022a)	20.3
	SemEval2010	89.1	89.8	91.3 (Zhao et al., 2021)	42.5
Relation Extraction(RE)	ACE05-R	87.5 | 63.7	88.2 | 65.1	91.1 | 73.0 (Ye et al., 2022)	40.5 | 4.5
	SciERC	65.4 | 43.0	63.6 | 42.0	69.9 | 53.2 (Ye et al., 2022)	25.9 | 5.5
Event Detection(ED)	ACE05-E	71.8	72.9	75.8 (Liu et al., 2022a)	17.1
	ACE05-E+	72.4	72.1	72.8 (Lin et al., 2020)	15.5
Event Argument Extraction(EAE)	ACE05-E	65.3	68.0	73.5 (Hsu et al., 2022)	28.9
	ACE05-E+	64.0	66.5	73.0 (Hsu et al., 2022)	30.9
Event Extraction(EE)	ACE05-E	71.8 | 51.0	72.9 | 51.9	74.7 | 56.8 (Lin et al., 2020)	17.0 | 7.3
	ACE05-E+	72.4 | 52.7	72.1 | 53.4	71.7 | 56.8 (Hsu et al., 2022)	16.6 | 7.8

Table 2: The performances of ChatGPT and several baseline models on 14 IE datasets on the Standard-IE setting. We report the performance on the whole test set. All results are directly cited from public papers or re-implemented using official open-source code.

	BERT	RoBERTa	ChatGPT
BBN*(ET)*	0.012	0.012	0.026
CoNLL*(NER)*	0.052	0.044	0.204
SemEval*(RC)*	0.023	0.031	0.460
ACE05-R*(RE)*	0.020	0.014	0.745
ACE05-E*(ED)*	0.161	0.226	0.656
ACE05-E*(EAE)*	0.154	0.168	0.699
ACE05-E*(EE)*	0.211	0.288	0.699

Table 7: The expected calibration error (ECE) is used to measure the calibration of a given model, and the lower, the better. Results are calculated on the whole test set.

감사합니다.
Q&A

최재웅 박사

(jwchoi95@kist.re.kr)

한국과학기술연구원 계산과학연구센터