질병분류기호 추천 딥러닝 모델

2021 KU 메디컬 해커톤

정근영, 윤동근

Introduction

- 2021 KU 메디컬 해커톤
  - 진행 기간
    - 9.13 ~ 9.29
  - 주제
    - 의료관련 자유주제
    - 원하는 팀의 경우 건국대병원 초진 데이터 제공
  - 참가대상
    - 학생부 -> 학부생, 일반부 -> 대학원생 및 일반인
    - 학생부 12팀, 일반부 2팀 참가

Introduction

- 심사 결과
  - 학생부 대상 수상

- 시상식
  - 10.15 (금), 더 클래식500
  - 1등 팀 데모 시연 및 발표

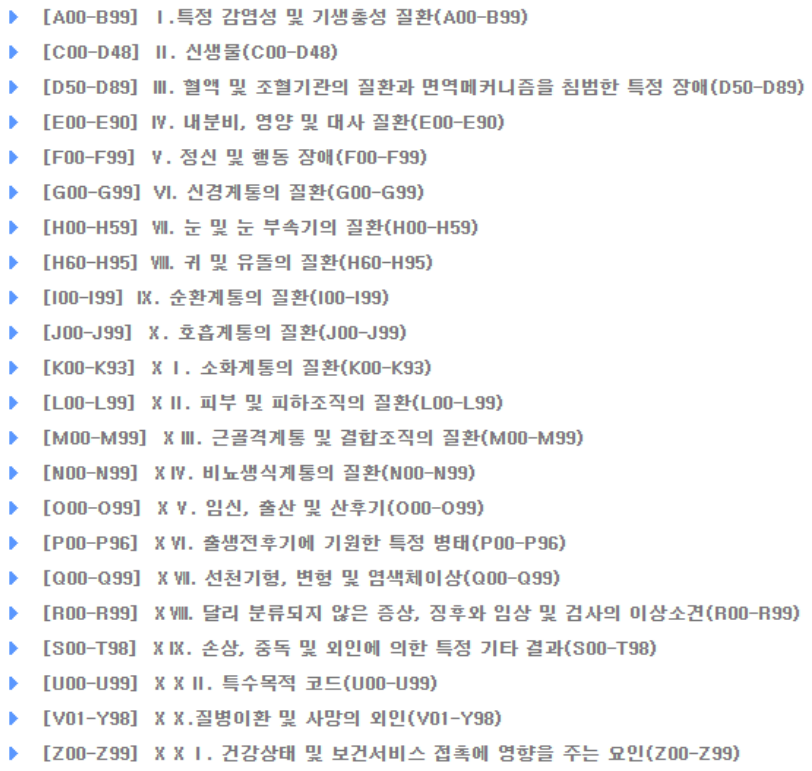
Topic

주호소 및 현병력 (Input)	진단코드 (Target)	진단명	진단분류	진단소분류	진료과
GB polyp	K828-001	GB polyp	담낭의 기타 명시된 질환	담낭의 기타 질환	소화기내과
개인의원에서 시행한 검진 복부 초음파 검사에서 담낭 내 용종이 관찰되어 내원함	K828-001	GB polyp	담낭의 기타 명시된 질환	담낭의 기타 질환	소화기내과
AF	I639-001	Cerebral Infarction	상세불명의 뇌경색증	뇌경색(증)	심장혈관내과
10/31 AF 으로 의뢰 현재 요양병원 거주	I639-001	Cerebral Infarction	상세불명의 뇌경색증	뇌경색(증)	심장혈관내과
consult f/u	H2692	Cataract, unspecified, Bilateral	상세불명의 백내장, 양쪽	기타 백내장	안과
DM- HT- 2019년 4월부터 좌안 시력저하 ** 양안 산동 good	H2692	Cataract, unspecified, Bilateral	상세불명의 백내장, 양쪽	기타 백내장	안과

- 의사가 초진시 입력하는 항목은 '주호소 및 현병력'과 '진단코드'
- 진단코드를 분류/추천하는 모델을 만드는 것이 진료에 가장 직접적인 도움

KCD Code

- 한국표준질병사인분류(Korean Standard Classification of Diseases, KCD)
  - 병원, 약국, 보험 등 범용적으로 사용
  - WHO의 ICD10 코드 차용



Issue 1: Domain Specific Words

주호소 및 현병력 (Input)	진단코드 (Target)
GB polyp	K828-001
개인의원에서 시행한 검진 복부 초음파 검사에서 담낭 내 용종이 관찰되어 내원함	K828-001
AF	I639-001
10/31 AF 으로 의뢰 현재 요양병원 거주	I639-001
consult f/u	H2692
DM- HT- 2019년 4월부터 좌안 시력저하 ** 양안 산동 good	H2692

- 일반적인 언어모델로 최적의 성능을 기대할 수 없음
  - 언어모델이 자주 마주하지 못한 전문용어
  - '폐암'과 '간암'의 유사한 Vector Representation
- 의학 데이터로 사전학습된 모델 필요

Issue 2: Absence of Korean Pre-Trained

주호소 및 현병력 (Input)	진단코드 (Target)
GB polyp	K828-001
개인의원에서 시행한 검진 복부 초음파 검사에서 담낭 내 용종이 관찰되어 내원함	K828-001
AF	I639-001
10/31 AF 으로 의뢰 현재 요양병원 거주	I639-001
consult f/u	H2692
DM- HT- 2019년 4월부터 좌안 시력저하 ** 양안 산동 good	H2692

- 영문 의학 데이터로 사전학습된 모델 사용
- 한국어 전문용어?
  - 번역

Translation

1. Seq2Seq

- High-Resource
- 전문용어 오역

2. Bilingual Lexicon Extraction (BLE)

- Low-Resource
- 전문용어는 동음이의어가 적음
- Task 특성상 Sequence의 의미보다는 특정 Token의 등장여부가 중요

word2word

- **word2word: A Collection of Bilingual Lexicons for 3,564 Language Pairs**
  - Choe et al., LREC 2020
  - OpenSubtitles2018 데이터
  - 62 Languages, 3,564 Directed Language Pairs

- **Bilingual Lexicon Extraction (BLE)**
  - Bilingual Parallel Corpus에서 단어 수준의 관련성을 추출
  - Low-Resource Machine Translation, Cross-Lingual Word Embeddings
  - **Methods**

    1) Co-occurances

$$p(y|x) = \frac{p(x,y)}{p(x)}$$

    1) Pointwise Mutual Information

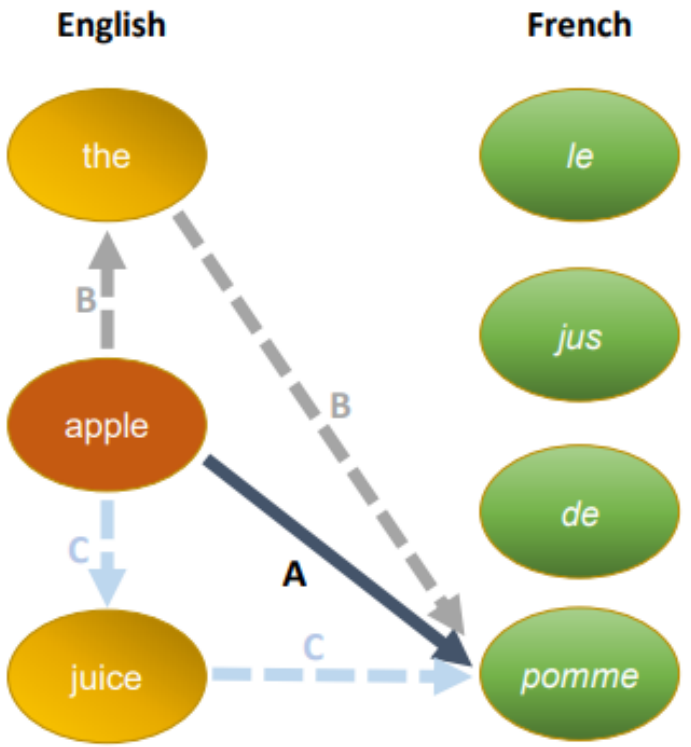
$$\text{PMI}(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

    1) Controlled Predictive Effects (Proposed Method)

word2word: CPE

- Source Token만 Target Token에 영향을 끼치는 것이 아님
- Source Language의 다른 Token들도 고려 필요

$$\text{CPE}(y \mid x) = p(y \mid x) - \sum_{x' \in \mathcal{X}} p(y \mid x')p(x' \mid x)$$



word2word: Performance

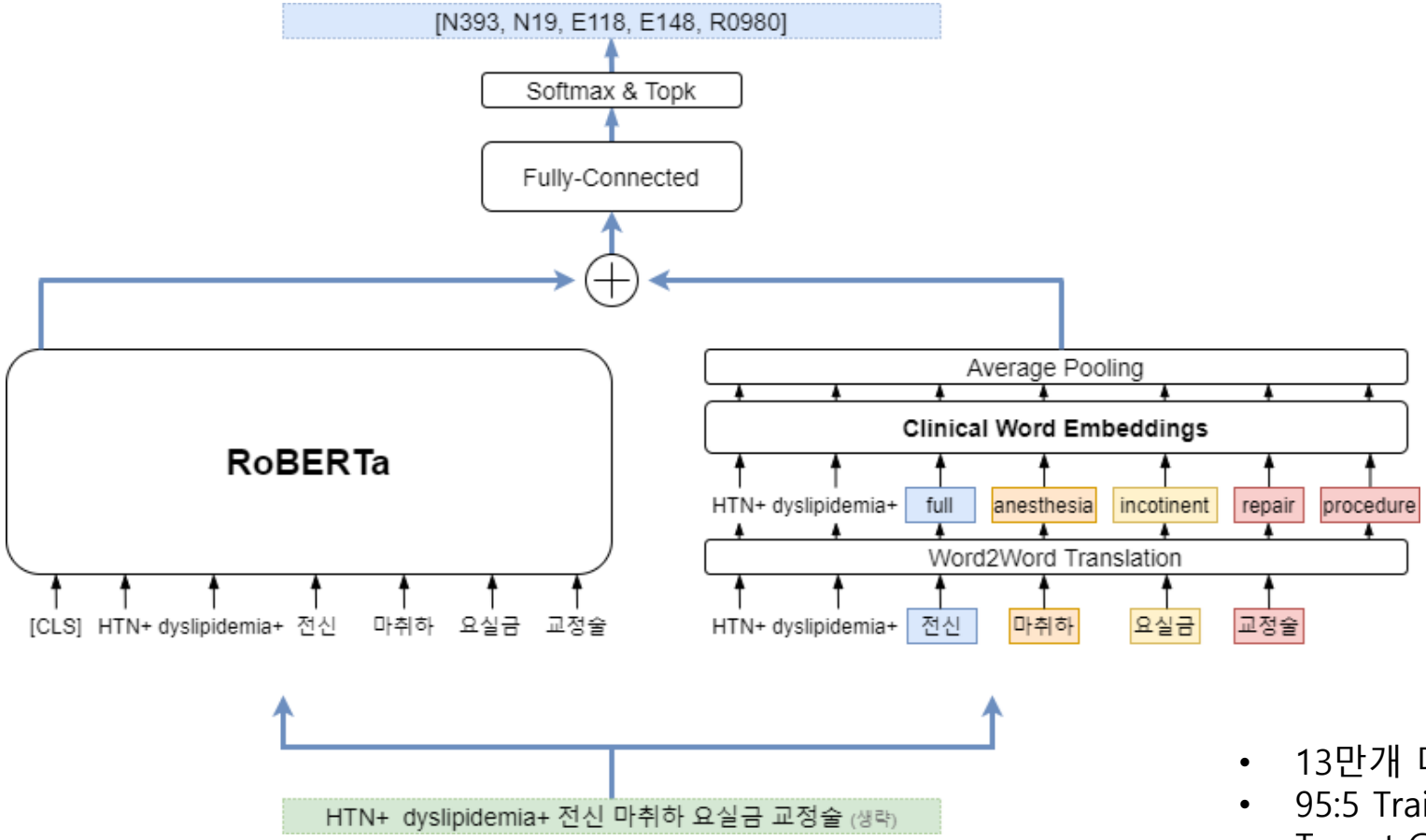
Metric (%)	Method	en-es	es-en	en-fr	fr-en	en-de	de-en	en-ru	ru-en	en-zh	zh-en	en-it	it-en
# Sentence Pairs		61.4M		41.8M		22.5M		25.9M		4.8M		35.2M	
P@1	Co-occurrence	22.3	25.5	18.7	21.9	10.5	23.5	3.3	11.4	5.4	3.8	24.9	24.1
	PMI	72.7	72.3	73.9	72.1	62.1	71.9	32.8	55.0	24.8	33.1	68.1	69.5
	MUSE	81.7	**83.3**	82.3	**82.4**	74.0	72.4	51.7	63.7	42.7	37.5	66.2	58.7
	CPE	**82.4**	79.5	**83.6**	80.7	**82.4**	**81.1**	**66.7**	**68.9**	**56.0**	**58.7**	**80.9**	**82.1**
P@5	Co-occurrence	67.8	71.4	63.1	66.3	63.7	65.5	52.3	51.8	46.0	36.3	61.9	68.5
	PMI	**92.3**	**90.4**	**92.5**	**90.1**	90.5	**88.1**	74.1	79.5	58.7	66.1	**90.3**	**91.1**
	MUSE	-	-	-	-	-	-	-	-	-	-	80.4	76.5
	CPE	90.1	88.4	91.7	89.3	**90.7**	87.7	**79.5**	**80.0**	**73.5**	**72.8**	89.8	89.9

Metric (%)	Method	en-ar	ar-en	en-zh	zh-en	en-ja	ja-en	en-ko	ko-en	en-th	th-en	en-vi	vi-en
# Sentence Pairs		29.8M		11.2M		2.1M		1.4M		3.3M		3.5M	
P@1	Co-occurrence	23.3	1.1	2.1	0.4	5.0	0.3	22.9	0.4	0.6	0.5	4.0	2.1
	PMI	13.3	20.7	8.5	20.6	33.5	16.7	14.0	14.9	18.3	13.4	20.5	16.5
	CPE	**30.3**	**27.9**	**48.3**	**34.3**	**49.3**	**40.4**	**39.1**	**38.1**	**48.1**	**31.0**	**30.0**	**37.7**
P@5	Co-occurrence	46.9	35.2	50.5	27.1	30.7	29.1	36.6	26.9	55.6	24.4	39.3	28.3
	PMI	57.0	**61.6**	78.7	**65.3**	64.0	60.5	48.8	57.7	64.5	52.7	**50.1**	60.4
	CPE	**58.1**	50.5	**80.9**	60.1	**66.8**	**66.4**	**54.9**	**60.0**	**69.3**	**53.1**	48.9	**62.2**

Clinical Word Embeddings

- Gary Weissman - github.com/gweissman/clinical_embeddings
  - PMC 데이터 (의학 논문 모음)
  - word2vec, fastText, GloVe를 다양한 차원으로 제공
  - 300차원 fastText 채택

Our Model



- 13만개 데이터
- 95:5 Train-Valid Split
- Target Class 1084개

Performance

Top-k Accuracy : 입력된 증상에 대해 모델이 결과로 내놓은 k개의 질병 코드 중,실제 정답 질병 코드가 포함되어 있을 확률

K	1	3	5
Baseline	42%	67%	-
Ours	50%	72%	80%

Future Improvements

- Multi-Stage Learning
  - 진단중분류(155) -> 진단소분류(474) -> 진단분류(883) -> 진단코드(1084)

- 한국어 의학 데이터 사전학습 언어 모델
  - 네이버 지식인 의료 전문가 답변, 하이닥, 논문 초록 등 한국어 데이터 수집
  - PMC 영문 데이터