

KOREAN BIO-MEDICAL CORPUS (KBMC) FOR MEDICAL NAMED ENTITY RECOGNITION

Sungjoo Byun¹, Jiseung Hong², Sumin Park¹, Dongjun Jang¹,
Jean Seo¹, Minseok Kim¹, Chaeyoung Oh¹, Hyopil Shin¹

¹Seoul National University

{byunsj, mam3b, qwer4107, seemdog, snumin44, nyong10,
hpshin}@snu.ac.kr

²KAIST

jiseung.hong@kaist.ac.kr



BIO-MEDICAL NAMED ENTITY RECOGNITION

1. NER contributes to processing medical terminology. Medical NER enables language models to identify and process medical terminologies and jargon.
2. NER facilitates information extraction from unstructured data.

PROBLEM?

Medical NER datasets are insufficient;

- This problem becomes even more challenging as domain-specific NER tasks require extensive labeling, particularly for specific entity categories like Disease, Body, and Treatment.
- The difficulty is further amplified due to the necessity of expert-level knowledge in medical domains.

There is no open-source medical NER dataset for Korean.

KBMC (KOREAN BIO-MEDICAL CORPUS)

The first open-source medical NER dataset for Korean.

Index	Token	Translation	Label
1	간질	Interstitial	B-Disease
2	폐렴	pneumonia	I-Disease
3	은	(particle) is	O
4	간	liver	B-Body
5	과	and	O
6	폐	lung	B-Body
7	의	of	O
8	역할	function	O
9	이	(particle) is	O
10	저하	deteriorated	O
11	되어	has	O
12

12
1	치료	treatments	O
2	는	(particle) are	O
3	항암제	anticancer drug	B-Treatment
4	치료	treatment	I-Treatment
5	,	,	O
6	방사선	radiation	B-Treatment
7	치료	therapy	I-Treatment
8	,	,	O
9	골수	bone marrow	B-Treatment
10	이식	transplantation	I-Treatment
11	등	etc	O
12	이	(particle) are	O
13	있으며	There (are)	O
14

Named Entity (NE)	Scheme	# of NE
Disease	B (Begin)	10,595
	I (Inside)	10,089
Body	B (Begin)	5,215
	I (Inside)	1,158
Treatment	B (Begin)	1,193
	I (Inside)	839

Label Distribution of KBMC

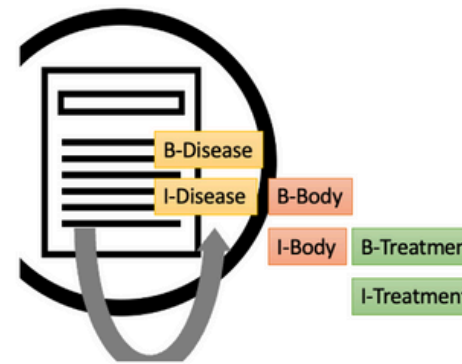
6,150 sentences, 153,971 tokens in total

KBMC (KOREAN BIO-MEDICAL CORPUS) CONSTRUCTION PROCESS



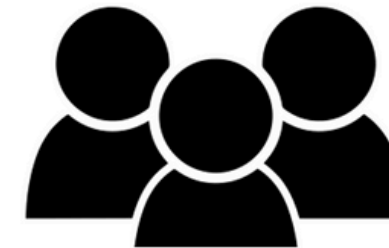
Text Source

- Disease names from Korean Standard Terminology Of Medicine (KOSTOM)
- Create sentences using ChatGPT API.



Automatic Pre-annotation

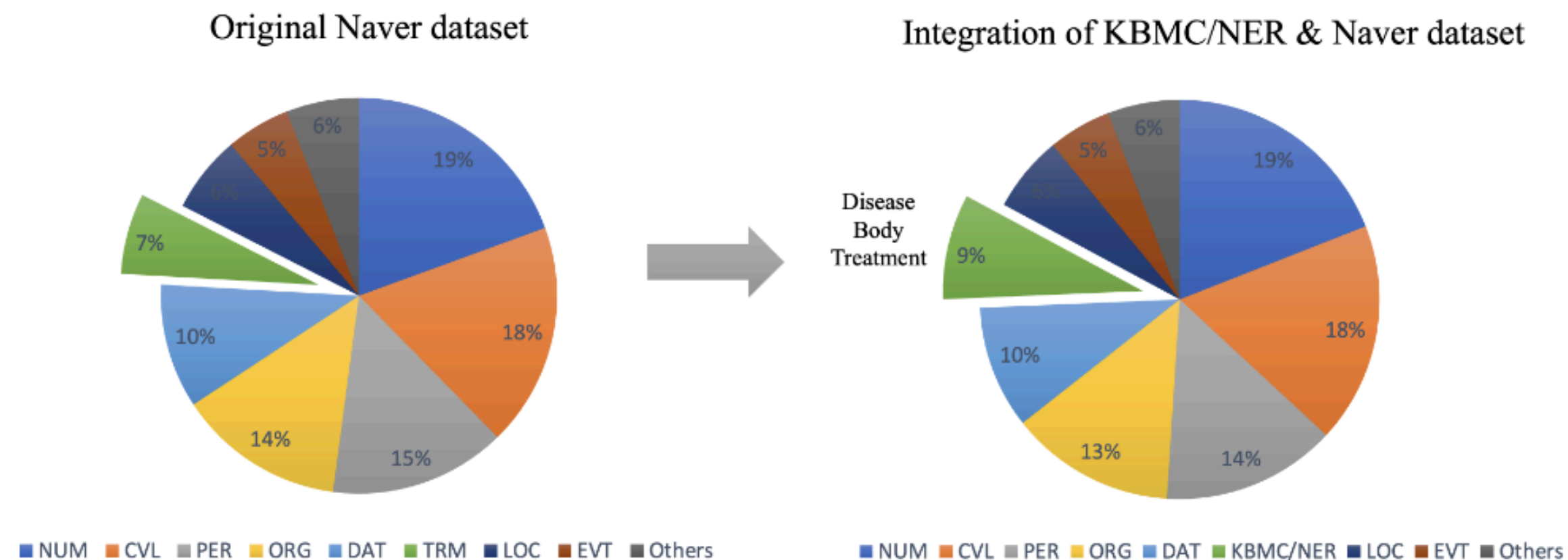
- Facilitates human annotation process



Human annotation

- Human annotation and accuracy review

DATA APPLICATION



The distribution of Named Entity labels in two datasets: the original Naver NER dataset (left), and a combined version of the Naver NER dataset (partial) and KBMC (right).

For data augmentation and comparison of NER in general and domain-specific text, the Naver NER dataset is concatenated with KBMC. The concatenated version includes 13 general Named Entities and 3 medical Named Entities.

EXPERIMENT

Model	Avg.F1(General)	medical NE	F1 of medical NER
KM-BERT (Kim et al., 2022)	87.08	TRM	75.35
KR-BERT (Lee et al., 2020b)	86.51	TRM	75.26
Ko-BERT	88.01	TRM	78.21
KR-ELECTRA (Lee and Shin, 2022)	87.62	TRM	76.25
Ko-ELECTRA	88.00	TRM	76.58
BiLSTM-CRF (Huang et al., 2015)	55.23	TRM	42.23

Medical Named Entities and NER Performance:
General NER dataset (The Naver Dataset) solely used.

Model	Avg.F1(General)	Medical NEs	F1 of Medical NER
KM-BERT	88.53 (+1.45)	Disease	98.04 (+22.69)
		Body	98.13 (+22.78)
		Treatment	98.53 (+23.18)
KR-BERT	87.48 (+0.97)	Disease	98.04 (+22.78)
		Body	98.32 (+23.06)
		Treatment	97.82 (+22.56)
KoBERT	88.70 (+0.69)	Disease	98.25 (+20.04)
		Body	98.22 (+20.01)
		Treatment	98.18 (+19.97)
KR-ELECTRA	88.63 (+1.01)	Disease	98.21 (+21.96)
		Body	98.31 (+22.06)
		Treatment	98.53 (+22.28)
KoELECTRA	88.86 (+0.86)	Disease	98.05 (+21.47)
		Body	97.72 (+21.14)
		Treatment	96.56 (+19.98)
BiLSTM-CRF	56.68 (+1.45)	Disease	88.18 (+45.95)
		Body	81.44 (+39.21)
		Treatment	61.14 (+18.91)

Medical Named Entities and Performance: KBMC applied.

The dataset is divided into 90% for training and 10% for testing.

To avoid data imbalance, we maintain consistent proportions of general and medical data in both training and evaluation.

KBMC APPLICABILITY ASSESSMENT

	Avg.F1	Precision	Recall
MedSpaCy	95.69	97.02	95.52

In order to test the utility of KBMC, we also test our dataset using MedSpaCy. KBMC dataset demonstrates remarkable performance on a clinical text processing toolkit in Python as well.

CONCLUSION

KBMC sentences	Translation	NER Tags
전신 적 다한증 은 신체 전체 에 힘이 빠져서 일상 생활 이 어려워지는 질환 으로 , 근육 통증 과 무기 력 감 이 동반 됩니다 .	Systemic myasthenia is a condition in which the whole body loses strength, making daily life difficult, accompanied by muscle pain and a sense of lethargy.	Disease-B Disease-I Disease-I O O O O O O O O O O O O O O Disease-B Disease-I O Disease-B Disease-I Disease-I O O O O
췌장암 이란 췌장 에 생긴 암세포 로 이루어진 종괴 (종양 덩어리) 이다 .	Pancreatic cancer refers to a tumor (a lump of tumor) made up of cancer cells that form in the pancreas.	Disease-B O Body-B O O O O O Disease-B O Disease-B O O O O
이러한 병명 은 폐 기능 저하 로 인한 호흡 곤란 기침 천식 발작 등의 증상 을 유발 하여 일상 생활 에 큰 영향을 미칩니다 .	Such diseases lead to symptoms such as respiratory distress, coughing, asthma attacks, etc., caused by decreased lung function, greatly affecting daily life.	O O O Disease-B Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I Disease-I O O O O O O O O O O O O O O
버킷 림프종 은 림프절 에서 발생하는 악성 종양 으로 , 조기 발견 과 치료 가 중요하며 항암 치료 나 방사선 치료 등 다양한 치료법 이 존재 합니다 .	Burkitt lymphoma is a malignant tumor that originates in the lymph nodes. Early detection and treatment are crucial, and various treatment methods, such as chemotherapy and radiation therapy, exist.	Disease-B Disease-I O Body-B O O O Disease-B Disease-I O O O O O O O O Treatment-B Treatment-I O Treatment-B Treatment-I O O O O O O O

In our research, we introduce KBMC, the first open-source biomedical NER dataset tailored for the Korean language. KBMC provides a training ground for language models to detect and categorize medical Named Entities, addressing the issue of data scarcity in this domain.

With KBMC, models can recognize a broader spectrum of medical terms.

We anticipate that our KBMC dataset will contribute substantially to ongoing research in the field of medical NLP.



THANK YOU



byunsj@snu.ac.kr