Korean Bio-Medical Corpus (KBMC) for Medical Named Entity Recognition

Sungjoo Byun[1], Jiseung Hong[2], Sumin Park[1], Dongjun Jang[1], Jean Seo[1], Minseok Kim[1], Chaeyoung Oh[1], Hyopil Shin[1]
[1]Seoul National University
{byunsj, mam3b, qwer4107, seemdog, snumin44, nyong10, hpshin}@snu.ac.kr
[2]KAIST
jiseung.hong@kaist.ac.kr

Bio-medical Named Entity Recognition

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

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	…	…	…

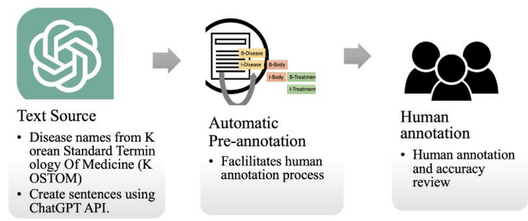
6,150 sentences, 153,971 tokens in total

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

KBMC (Korean Bio-Medical Corpus)

- The first open-source medical NER dataset for Korean.



Construction Process of KBMC

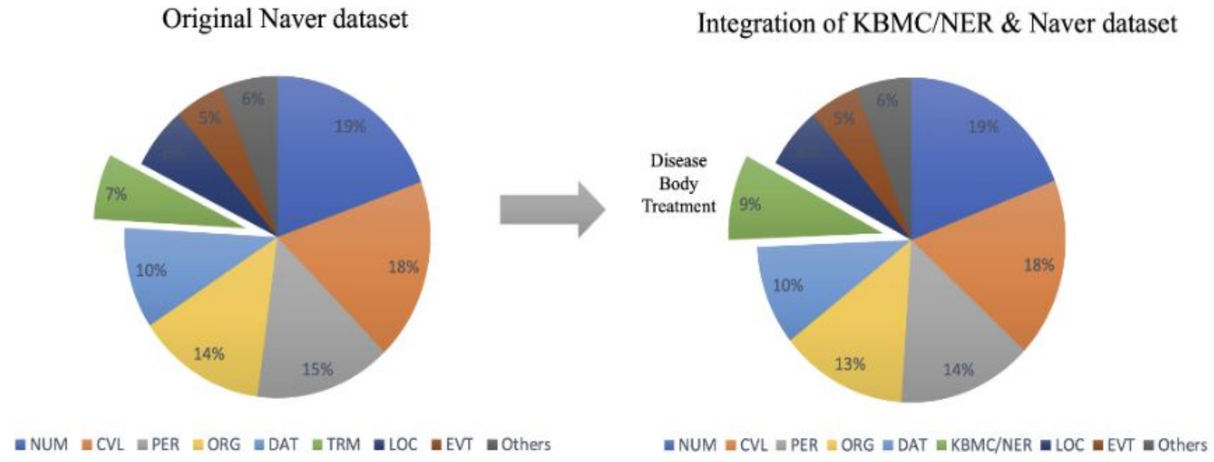
KBMC Applicability Assessment

	Avg.F1	Precision	Recall
MedSpaCy	95.69	97.02	95.52

- KBMC demonstrates remarkable performance on a clinical text processing toolkit in Python, MedSpaCy as well.

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.

Conclusion

- KBMC enables language to recognize a broader spectrum of medical terms, enhancing their understanding and processing of clinical texts.

Contributions:

- We describe and publicly release Korean Bio-Medical Named Entity Recognition Corpus (KBMC), the first open-source Korean medical NER dataset. This contributes to solving the data scarcity problem.

- Crucial role in medical data processing. Medical NER would facilitate the sensitive data anonymization process and contribute to the reconstruction of medical data that lack standardized formats.

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. F1 scores for medical entities are nearly 20 points higher than the TRM label.

- The Naver dataset contains the label TRM (Term) representing both medical and IT-related terms.
- In the concatenated dataset, sentences that include TRM from the original dataset have been replaced with KBMC for more accurate classification of medical terms into refined categories.

Selected References

Cole Pearson, Naeem Seliya, and Rushit Dave.2021. Named entity recognition in unstructured medical text documents.
Veysel Kocaman and David Talby. 2022. Accurate clinical and biomedical named entity recognition at scale. Software Impacts, 13:100373.
Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python.
Sam Henry, Kevin Buchan, Michele Filannino, Am- ber Stubbs, and Ozlem Uzuner. 2019. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. Journal of the American Medical Informatics Association, 27(1):3–12.
Maximilian Hofer, Andrey Kormilitzin, Paul Goldberg, and Alejo J. Nevado-Holgado. 2018. Few-shot learning for named entity recognition in medical text.CoRR, abs/1811.05468