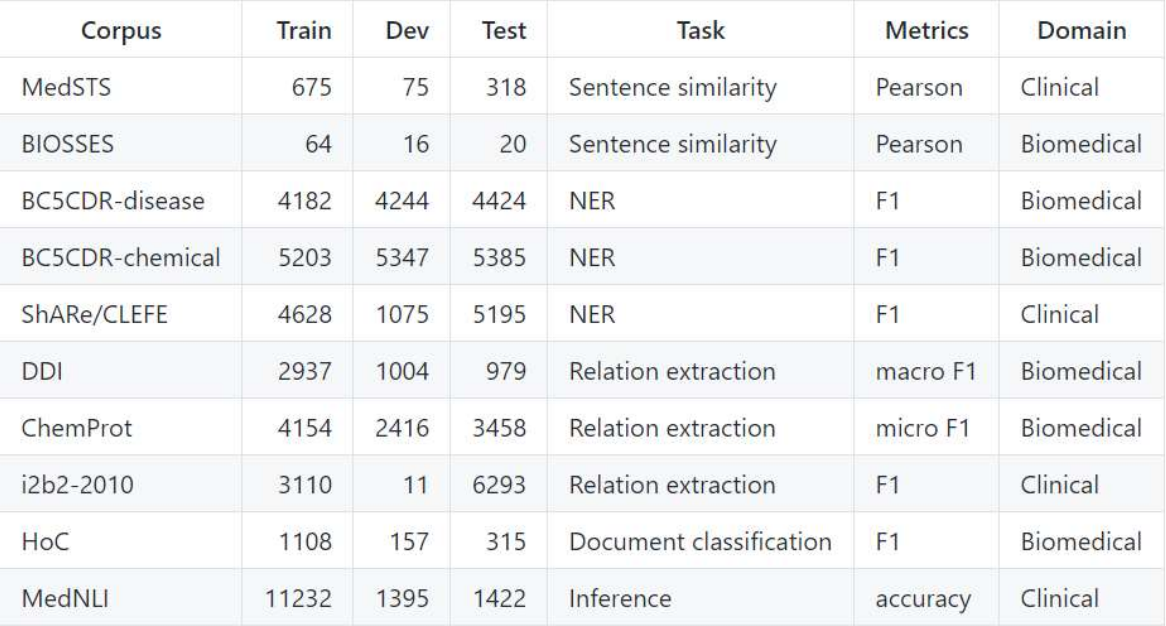
Bio-Bert-Data

Blue-Data

Blue dataset은 5개의 서로 다른 Bio-NLP tasks에서 빈번하게 사용되는 dataset을 지칭함.

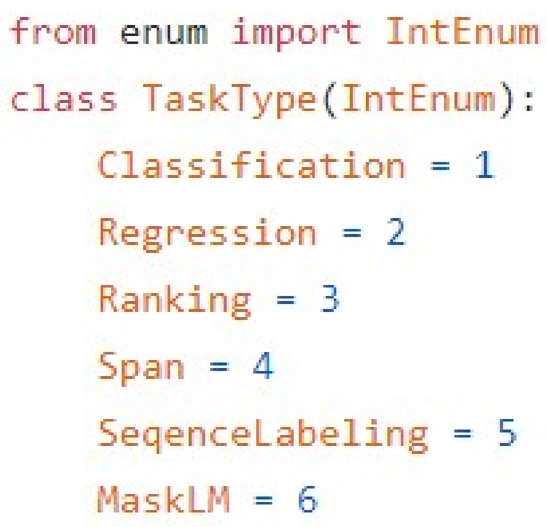


Corpus	Train	Dev	Test	Task	Metrics	Domain
MedSTS	675	75	318	Sentence similarity	Pearson	Clinical
BIOSSES	64	16	20	Sentence similarity	Pearson	Biomedical
BC5CDR-disease	4182	4244	4424	NER	F1	Biomedical
BC5CDR-chemical	5203	5347	5385	NER	F1	Biomedical
ShARe/CLEFE	4628	1075	5195	NER	F1	Clinical
DDI	2937	1004	979	Relation extraction	macro F1	Biomedical
ChemProt	4154	2416	3458	Relation extraction	micro F1	Biomedical
i2b2-2010	3110	11	6293	Relation extraction	F1	Clinical
HoC	1108	157	315	Document classification	F1	Biomedical
MedNLI	11232	1395	1422	Inference	accuracy	Clinical

- 이 중 BERT 형식에 맞게 제공된 data
BC5CDR-disease, BC5CDR-chemical
ChemProt, DDI, HoC

- 각 task별 dataset은 dev, test, train data로 나뉘어져 있음.

MTDNN TaskType

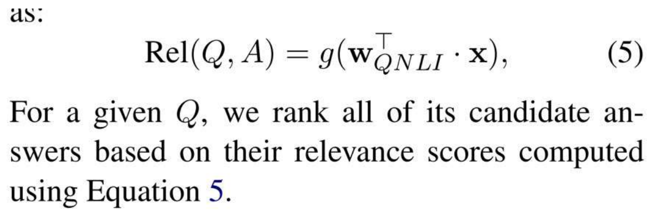


```
from enum import IntEnum
class TaskType(IntEnum):
    Classification = 1
    Regression = 2
    Ranking = 3
    Span = 4
    SeqenceLabeling = 5
    MaskLM = 6
```

1) Classification: 어진 데이터를 정해진 카테고리에 따라 분류하는 task

2) Regression: 연속된 값을 예측하는 task로 주로 어떤 패턴이나 트렌드, 경향을 예측할 때 사용함.
예를 들면, 공부시간에 따른 전공 시험 점수를 예측하는 문제

3) Ranking: 확실하지는 않음.. MT-DNN 논문에 나온 예시



as:

$$\mathrm{Rel}(Q, A) = g(\mathbf{w}_{QNLI}^{\top} \cdot \mathbf{x}), \qquad (5)$$

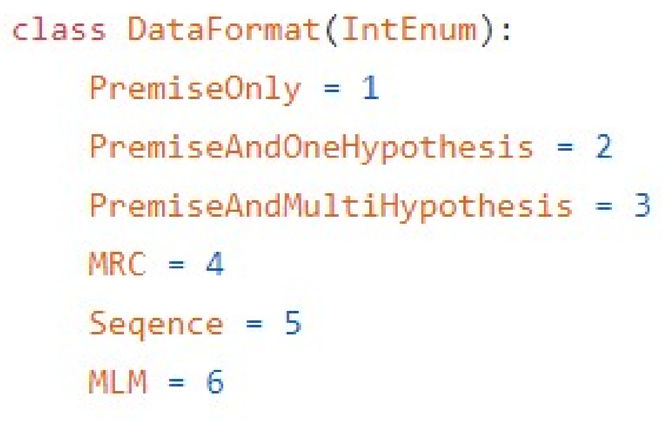
For a given Q, we rank all of its candidate answers based on their relevance scores computed using Equation 5.

4) Span: ex) Squad task

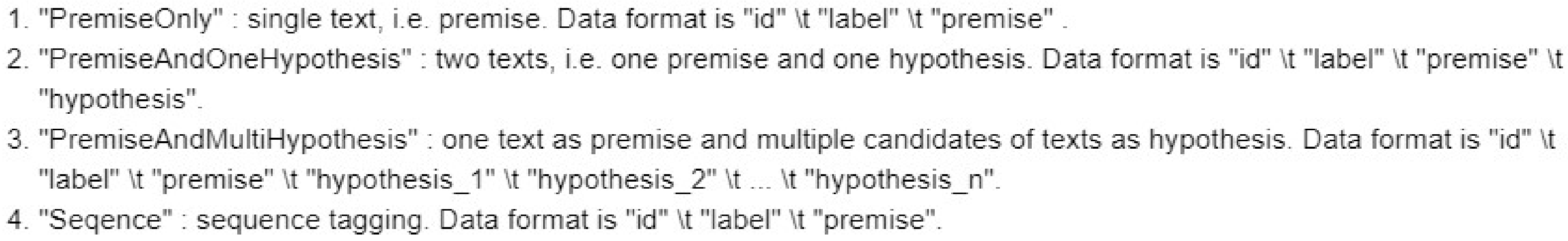
5) SeqenceLabeling: assign a class or label to each token in a given input sequence. 예) NER

6) MaskLM: 주변 text를 보고 mask된 단어 예측
(BERT pre-training에 사용됨)

MTDNN DataFormat



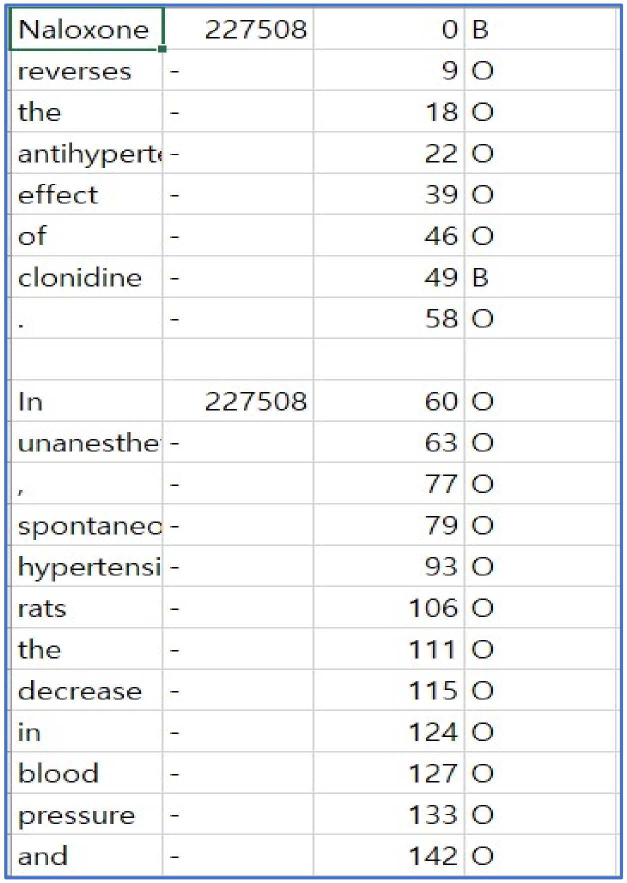
```
class DataFormat(IntEnum):
    PremiseOnly = 1
    PremiseAndOneHypothesis = 2
    PremiseAndMultiHypothesis = 3
    MRC = 4
    Seqence = 5
    MLM = 6
```



1. "PremiseOnly" : single text, i.e. premise. Data format is "id" \t "label" \t "premise" .
2. "PremiseAndOneHypothesis" : two texts, i.e. one premise and one hypothesis. Data format is "id" \t "label" \t "premise" \t "hypothesis".
3. "PremiseAndMultiHypothesis" : one text as premise and multiple candidates of texts as hypothesis. Data format is "id" \t "label" \t "premise" \t "hypothesis_1" \t "hypothesis_2" \t ... \t "hypothesis_n".
4. "Seqence" : sequence tagging. Data format is "id" \t "label" \t "premise".

BC5CDR

BC5CDR is a collection of 1,500 PubMed titles and abstracts selected from the CTD-Pfizer corpus
- 4409 annotated chemicals, 5818 diseases, 3116 chemical-disease interactions

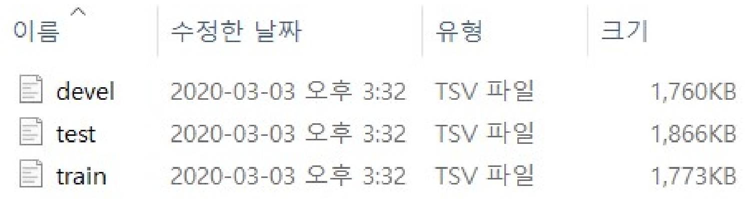


Naloxone	227508	0	B
reverses	-	9	O
the	-	18	O
antihypert	-	22	O
effect	-	39	O
of	-	46	O
clonidine	-	49	B
.	-	58	O
In	227508	60	O
unanesthe	-	63	O
,	-	77	O
spontaneo	-	79	O
hypertensi	-	93	O
rats	-	106	O
the	-	111	O
decrease	-	115	O
in	-	124	O
blood	-	127	O
pressure	-	133	O
and	-	142	O

1) Data 구조
pmid는 문장의 첫 token에만 기입, 나머지는 – 으로

단어	pmid	단어의 시작 index	BIO tagging

2) Dataset 구성 – dev/train/test

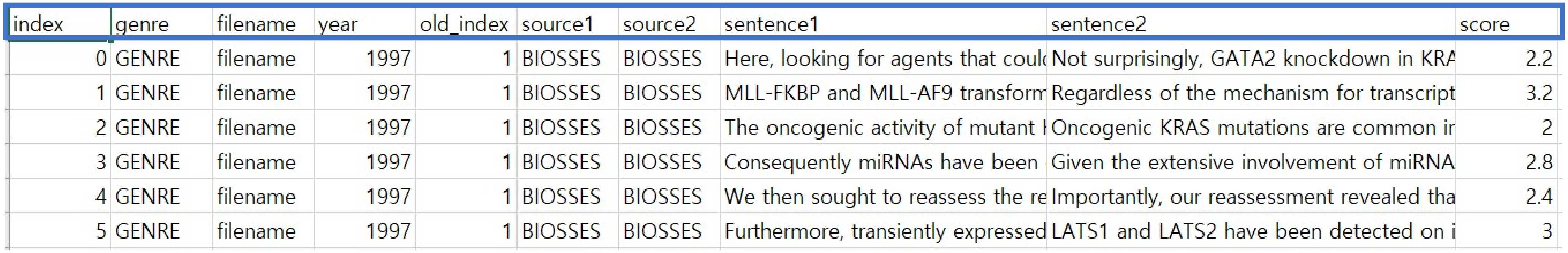


이름	수정한 날짜	유형	크기
devel	2020-03-03 오후 3:32	TSV 파일	1,760KB
test	2020-03-03 오후 3:32	TSV 파일	1,866KB
train	2020-03-03 오후 3:32	TSV 파일	1,773KB

3) 수행하는 task: NER
BC5CDR-disease의 경우 disease에 대해 BIO tagging
BC5CDR-chemica의 경우 chemical에 대해 BIO tagging

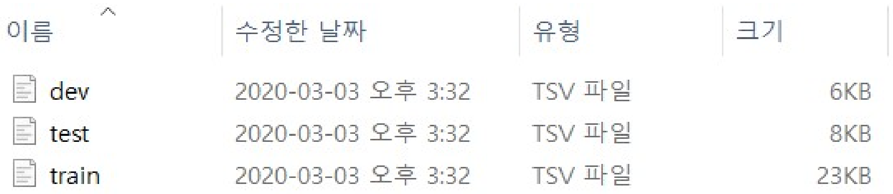
4) MT-DNN에서 해당하는 task 종류와 data format
TaskType: SequenceLabeling
DataFormat: Sequence

BIOSSES



index	genre	filename	year	old_index	source1	source2	sentence1	sentence2	score
0	GENRE	filename	1997	1	BIOSSES	BIOSSES	Here, looking for agents that coul	Not surprisingly, GATA2 knockdown in KRA	2.2
1	GENRE	filename	1997	1	BIOSSES	BIOSSES	MLL-FKBP and MLL-AF9 transform	Regardless of the mechanism for transcript	3.2
2	GENRE	filename	1997	1	BIOSSES	BIOSSES	The oncogenic activity of mutant	Oncogenic KRAS mutations are common i	2
3	GENRE	filename	1997	1	BIOSSES	BIOSSES	Consequently miRNAs have been	Given the extensive involvement of miRNA	2.8
4	GENRE	filename	1997	1	BIOSSES	BIOSSES	We then sought to reassess the re	Importantly, our reassessment revealed tha	2.4
5	GENRE	filename	1997	1	BIOSSES	BIOSSES	Furthermore, transiently expressed	LATS1 and LATS2 have been detected on i	3

2) Dataset 구성 – dev/train/test



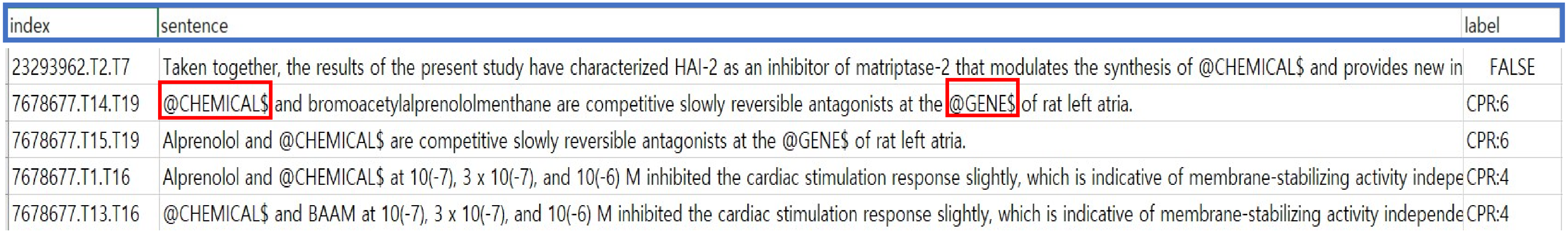
이름	수정한 날짜	유형	크기
dev	2020-03-03 오후 3:32	TSV 파일	6KB
test	2020-03-03 오후 3:32	TSV 파일	8KB
train	2020-03-03 오후 3:32	TSV 파일	23KB

3) 수행하는 task: compute similarity of biomedical sentences
similarity를 이진 labeling 하는 것이 아니라 score를 매김

4) MT-DNN에서 해당하는 task 종류와 data format
; GLUE의 STS-B와 유사
TaskType: PremiseAndOneHypothesis
DataFormat: Regression

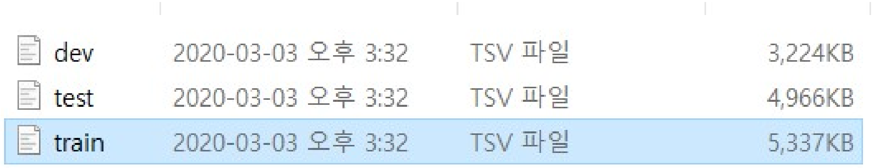
ChemProt

consists of 1,820 PubMed abstracts with chemical-protein interactions

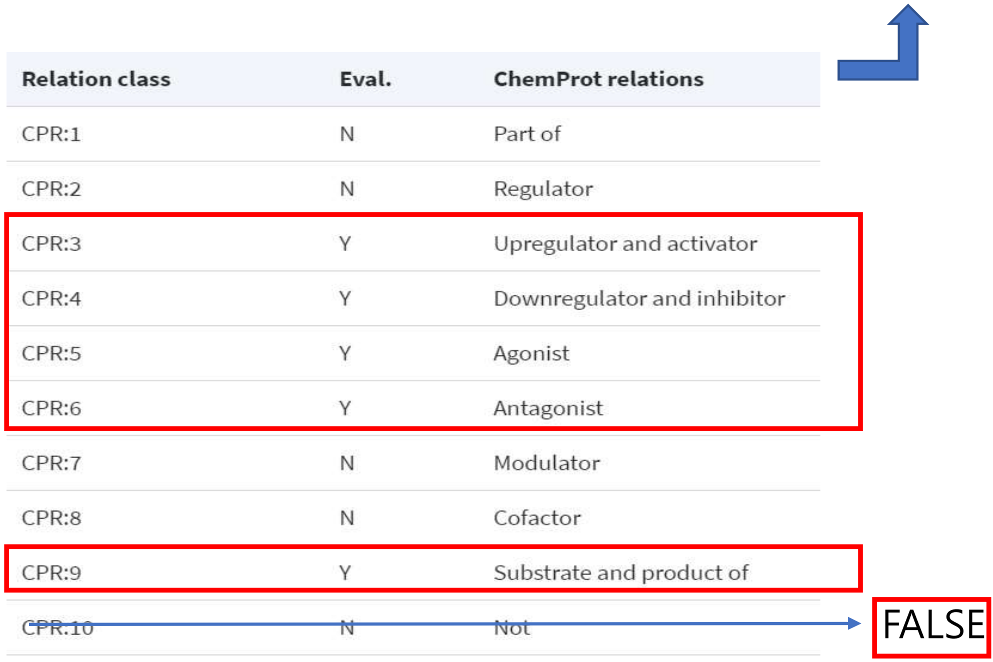


index	sentence	label
23293962.T2.T7	Taken together, the results of the present study have characterized HAI-2 as an inhibitor of matriptase-2 that modulates the synthesis of @CHEMICAL$ and provides new in	FALSE
7678677.T14.T19	@CHEMICAL$ and bromoacetylalprenololmenthane are competitive slowly reversible antagonists at the @GENE$ of rat left atria.	CPR:6
7678677.T15.T19	Alprenolol and @CHEMICAL$ are competitive slowly reversible antagonists at the @GENE$ of rat left atria.	CPR:6
7678677.T1.T16	Alprenolol and @CHEMICAL$ at 10(-7), 3 x 10(-7), and 10(-6) M inhibited the cardiac stimulation response slightly, which is indicative of membrane-stabilizing activity indepe	CPR:4
7678677.T13.T16	@CHEMICAL$ and BAAM at 10(-7), 3 x 10(-7), and 10(-6) M inhibited the cardiac stimulation response slightly, which is indicative of membrane-stabilizing activity independe	CPR:4

2) Dataset 구성 – dev/train/test



dev	2020-03-03 오후 3:32	TSV 파일	3,224KB
test	2020-03-03 오후 3:32	TSV 파일	4,966KB
train	2020-03-03 오후 3:32	TSV 파일	5,337KB



Relation class	Eval.	ChemProt relations
CPR:1	N	Part of
CPR:2	N	Regulator
CPR:3	Y	Upregulator and activator
CPR:4	Y	Downregulator and inhibitor
CPR:5	Y	Agonist
CPR:6	Y	Antagonist
CPR:7	N	Modulator
CPR:8	N	Cofactor
CPR:9	Y	Substrate and product of
~~CPR:10~~	~~N~~	~~Not~~

FALSE

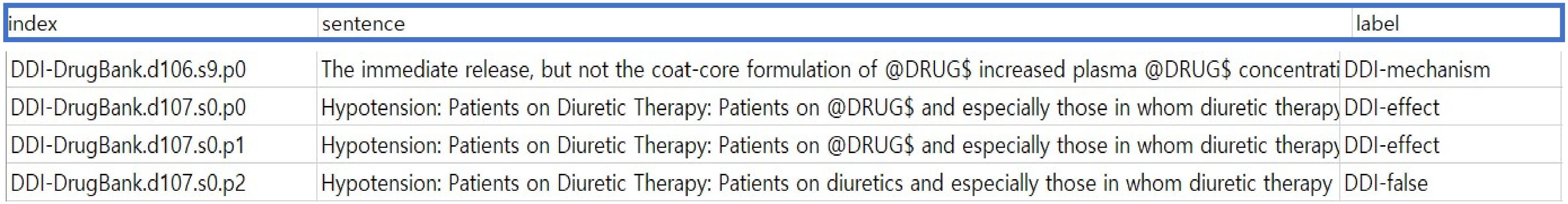
https://academic.oup.com/database/article/doi/10.1093/database/baz054/5498050

3) 수행하는 task: RE
evaluate five relation
-> CPR:3, CPR:4, CPR:5, CPR:6, and CPR:9. + FALSE

4) MT-DNN에서 해당하는 task 종류와 data format
TaskType: PremiseOnly
DataFormat: Classification

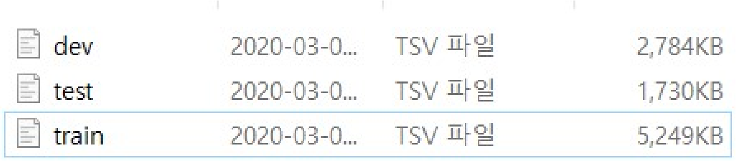
ddi2013-type(DDI)

: collection of 792 texts selected from the DrugBank database and other 233 Medline abstracts



index	sentence	label
DDI-DrugBank.d106.s9.p0	The immediate release, but not the coat-core formulation of @DRUG$ increased plasma @DRUG$ concentrati	DDI-mechanism
DDI-DrugBank.d107.s0.p0	Hypotension: Patients on Diuretic Therapy: Patients on @DRUG$ and especially those in whom diuretic therapy	DDI-effect
DDI-DrugBank.d107.s0.p1	Hypotension: Patients on Diuretic Therapy: Patients on @DRUG$ and especially those in whom diuretic therapy	DDI-effect
DDI-DrugBank.d107.s0.p2	Hypotension: Patients on Diuretic Therapy: Patients on diuretics and especially those in whom diuretic therapy	DDI-false

2) Dataset 구성 – dev/train/test



dev	2020-03-0...	TSV 파일	2,784KB
test	2020-03-0...	TSV 파일	1,730KB
train	2020-03-0...	TSV 파일	5,249KB

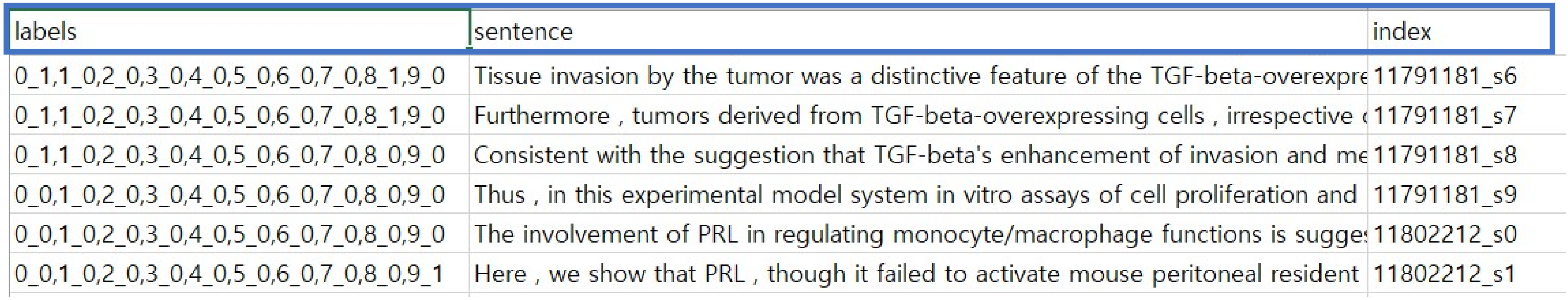
DDI-mechanism	This type is used to annotate DDIs that are described by PK mechanism (e.g. Grepafloxacin may inhibit the metabolism of *theobromine*).
DDI-effect	This type is used to annotate DDIs describing an effect or a PD mechanism
DDI-advise	This type is used when a recommendation or advice regarding a drug interaction is given
DDI-int	This type is used when a DDI appears in the text without providing any additional information

3) 수행하는 task: RE
-> DDI-mechanism, DDI-effect, DDI-advise, DDI-int, DDI-false

4) MT-DNN에서 해당하는 task 종류와 data format
TaskType: PremiseOnly
DataFormat: Classification

Hoc

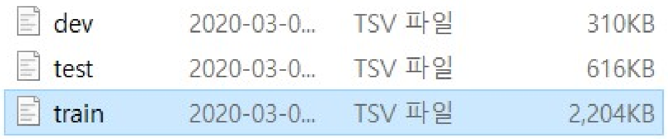
consists of 1,580 PubMed abstracts annotated with ten currently known hallmarks of cancer



labels	sentence	index
0_1,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_1,9_0	Tissue invasion by the tumor was a distinctive feature of the TGF-beta-overexpre	11791181_s6
0_1,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_1,9_0	Furthermore , tumors derived from TGF-beta-overexpressing cells , irrespective (	11791181_s7
0_1,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_0,9_0	Consistent with the suggestion that TGF-beta's enhancement of invasion and me	11791181_s8
0_0,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_0,9_0	Thus , in this experimental model system in vitro assays of cell proliferation and	11791181_s9
0_0,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_0,9_0	The involvement of PRL in regulating monocyte/macrophage functions is sugge:	11802212_s0
0_0,1_0,2_0,3_0,4_0,5_0,6_0,7_0,8_0,9_1	Here , we show that PRL , though it failed to activate mouse peritoneal resident	11802212_s1

Hallmark
0. Sustaining proliferative signalling
1. Evading growth suppressors
2. Resisting cell death
3. Enabling replicative immortality
4. Inducing angiogenesis
5. Activating invasion and metastasis
6. Genomic instability and mutation
7. Tumor promoting inflammation
8. Cellular energetics
9. Avoiding immune destruction

2) Dataset 구성 – dev/train/test



dev	2020-03-0...	TSV 파일	310KB
test	2020-03-0...	TSV 파일	616KB
train	2020-03-0...	TSV 파일	2,204KB

3) 수행하는 task:
10개의 hallmark에 대해 classification 수행

4) MT-DNN에서 해당하는 task 종류와 data format
TaskType: PremiseOnly
DataFormat: Classification