CharacterBERT

Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters

Hicham El Boukkouri [1], Olivier Ferret [2], Thomas Lavergne [1], Hiroshi Noji [3], Pierre Zweigenbaum [1], Junichi Tsujii [3]

[1] Université Paris-Saclay, CNRS, LIMSI, Orsay, France | *{elboukkouri,lavergne,pz}@limsi.fr*
[2] Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France | *olivier.ferret@cea.fr*
[3] Artificial Intelligence Research Center (AIRC), AIST, Japan | *{hiroshi.noji,j-tsujii}@aist.go.jp*

Table of contents

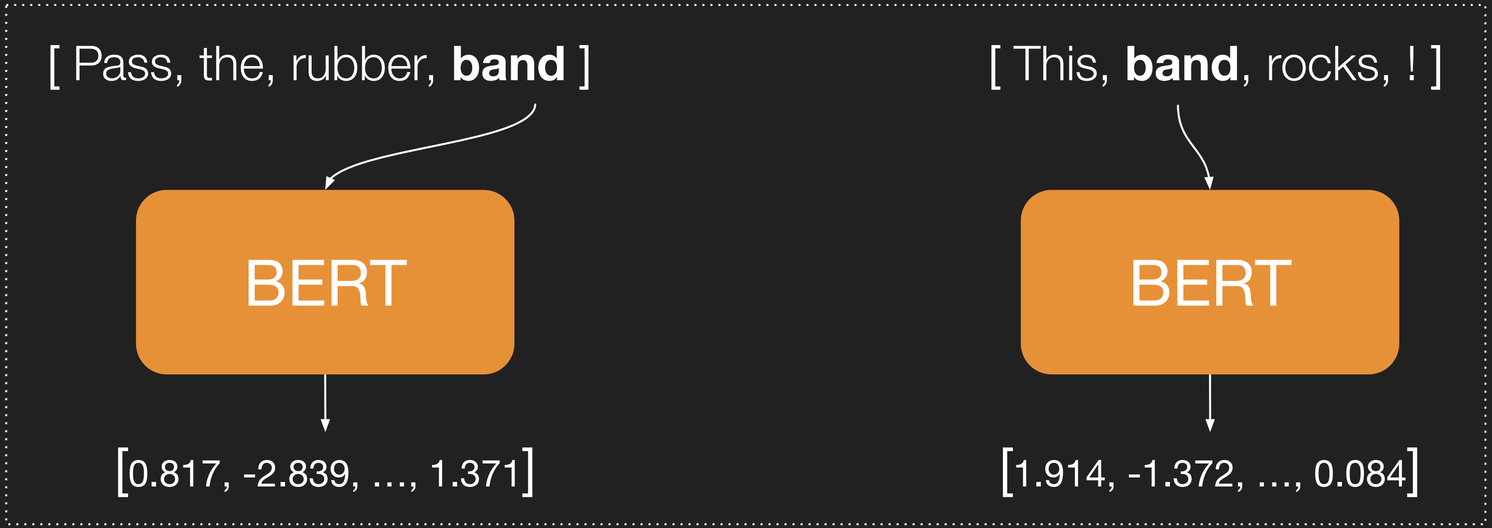
- Reminder on BERT & WordPieces
- Raised Issues
- Proposed Solution: CharacterBERT
- BERT vs. CharacterBERT
- Conclusion

Reminder on BERT & WordPieces

BERT: General Idea

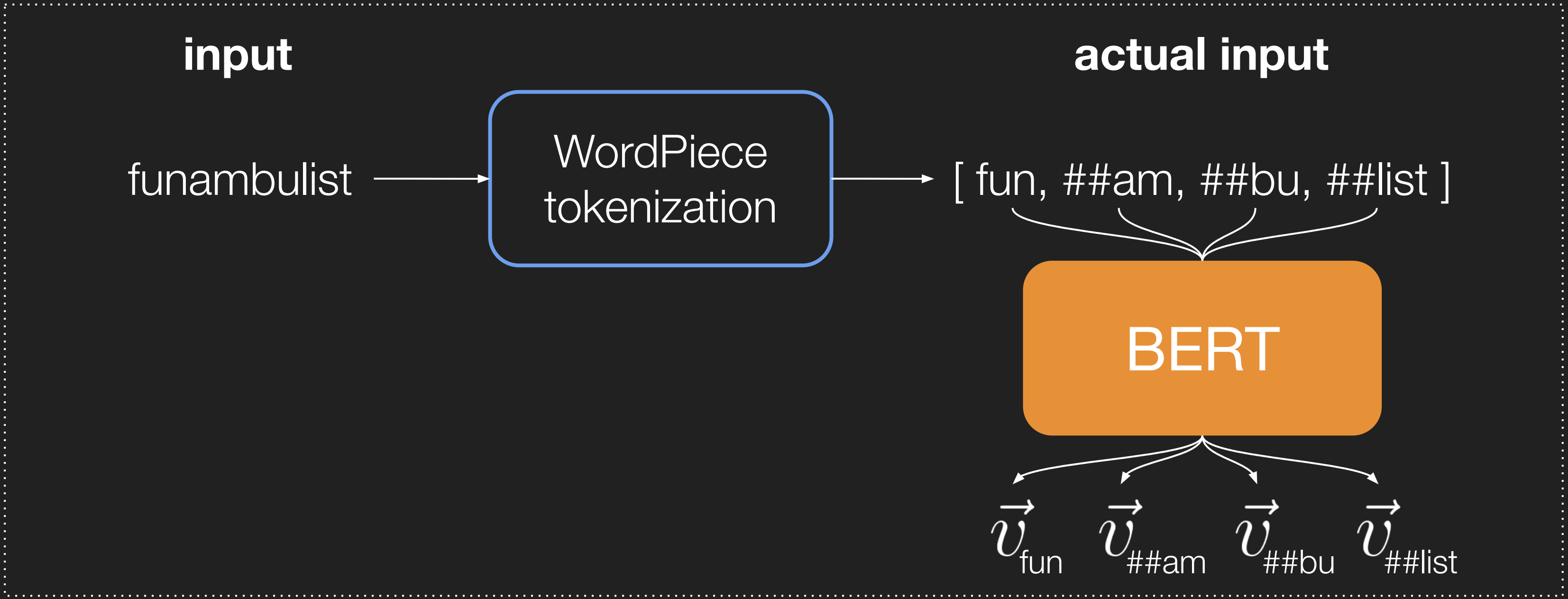
Neural language model based on Transformers[1]

→ Contextualized embeddings



BERT: Handling OOVs with WordPieces

Tokenizes unknown tokens into known WordPieces



Raised Issues

Unnecessary Complexity

- More convenient to handle actual “words”
- Added complexity to
  - Word Similarity Tasks: how to aggregate ?
  - NER Tasks: which wordpieces to tag ?

Specialized Domains

- Re-training original BERT
  → keep a general-domain vocabulary

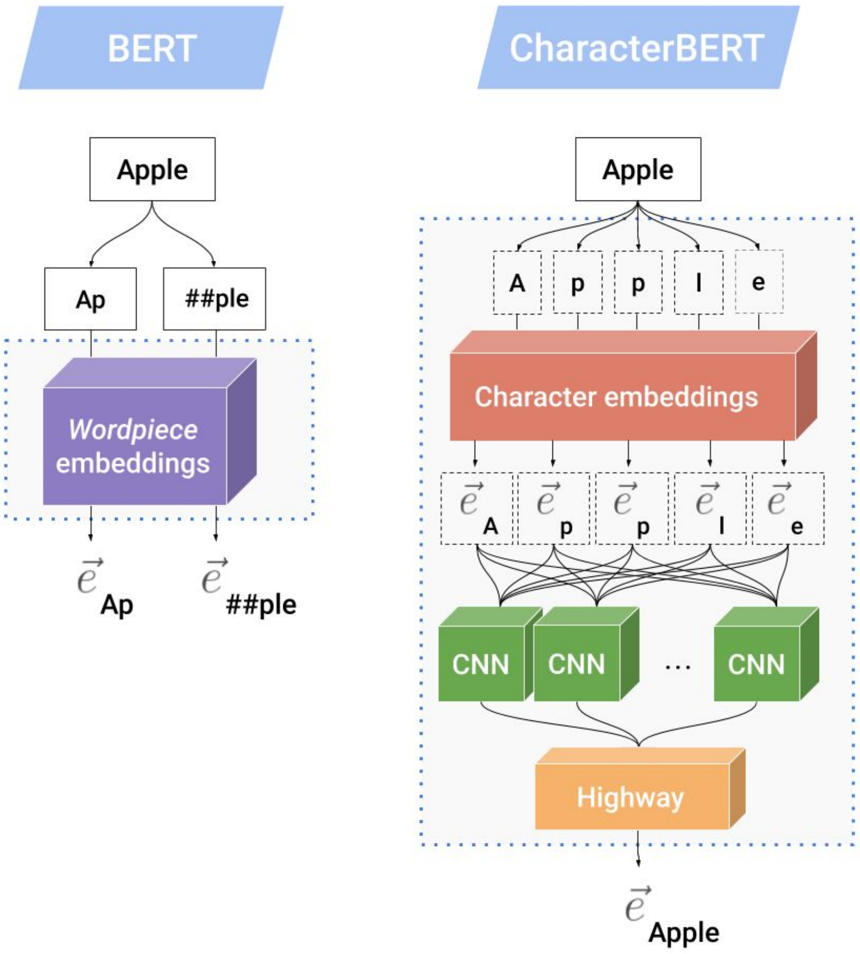
Reference	Medical Vocabulary	General Vocabulary
paracetamol	[paracetamol]	[para, ce, tam, ol]
choledocholithiasis	[choledoch, olithiasis]	[cho, led, och, oli, thi, asi, s]
borborygmi	[bor, bor, yg, mi]	[bo, rb, ory, gm, i]

Tokenization of medical terms produced by vocabularies from different domains
General Vocabulary: BERT's vocabulary | Medical Vocabulary: PubMed & MIMIC-III

Proposed Solution

CharacterBERT[2]

- Inspired by ELMo[3]
- Drops the WordPieces
- Uses a CharacterCNN
  → Open-Vocabulary
  → Character-level
  → 1 word = 1 vector



Context-independent representation in BERT vs. CharacterBERT

BERT vs. CharacterBERT

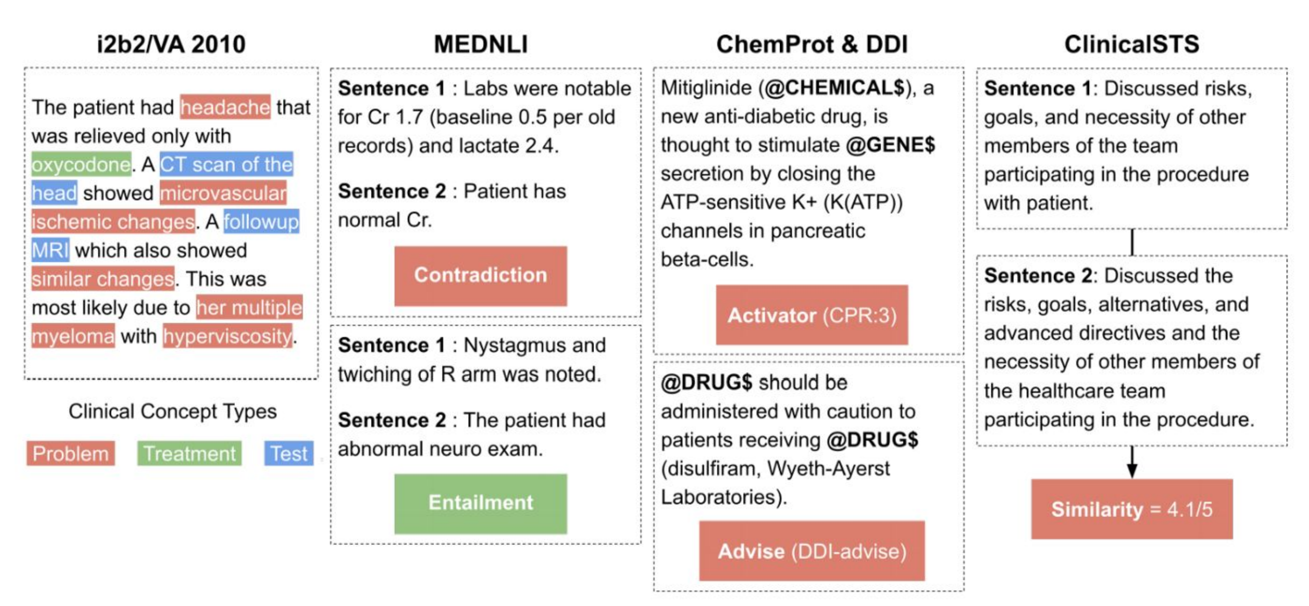
Pre-training

- Same pre-training conditions*
- Medical models = general models + re-training

Corpus	Composition	# documents	# tokens
General	Wikipedia (EN)	5.99×10^6	2.14×10^9
	OpenWebText	1.56×10^6	1.28×10^9
Medical	MIMIC-III	2.09×10^6	0.51×10^9
	PMC OA abstracts	2.33×10^6	0.52×10^9

Evaluation

- Multiple evaluation tasks (clinical & bio-medical)



Evaluation

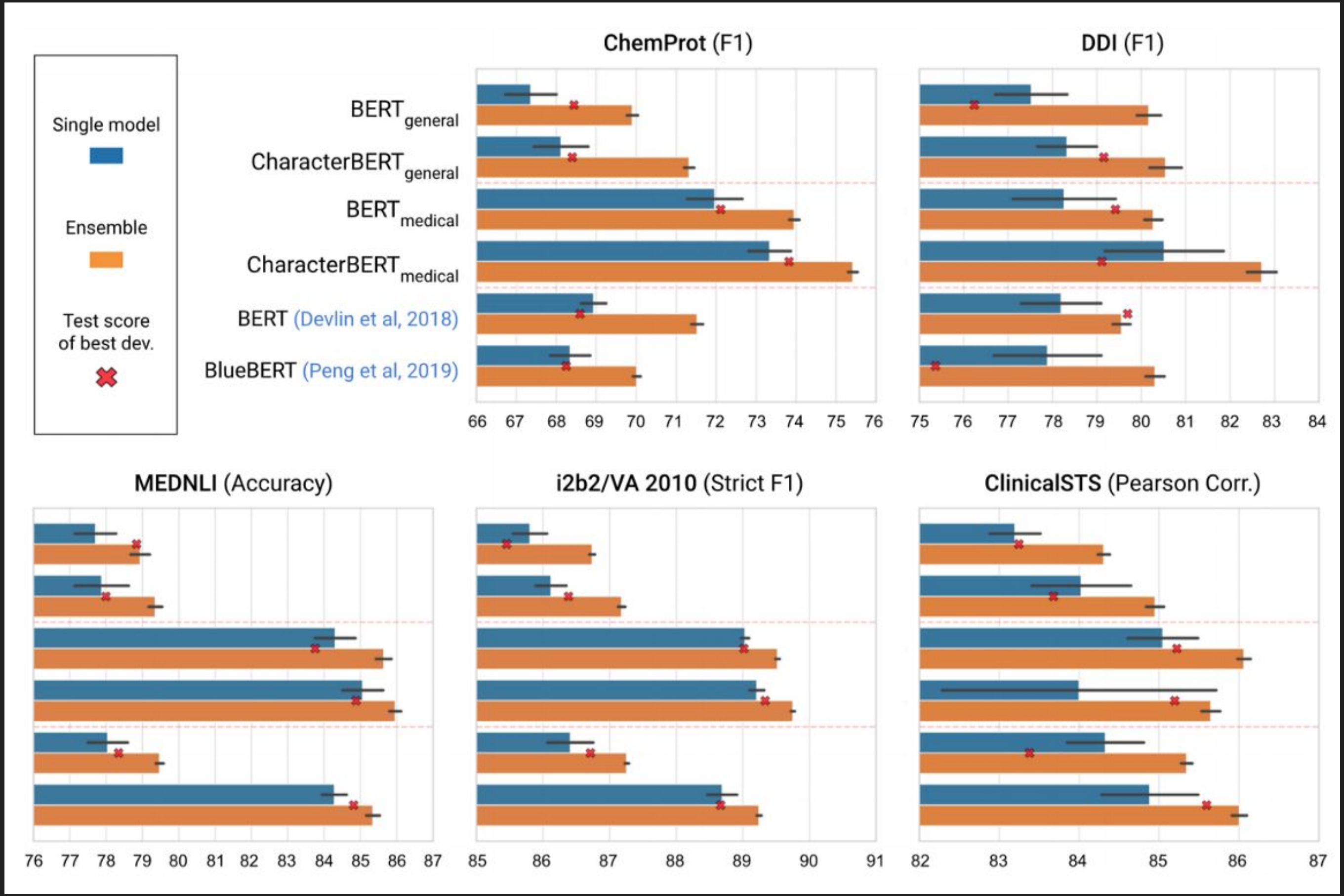
- Account for variance: 10 random seeds

→ Single model score: mean ± std

→ Ensemble model score

  - STS: average similarity
  - Other tasks: majority vote

Results



Performance on the test set: single model in blue, ensemble in orange
and best model on validation set in red.

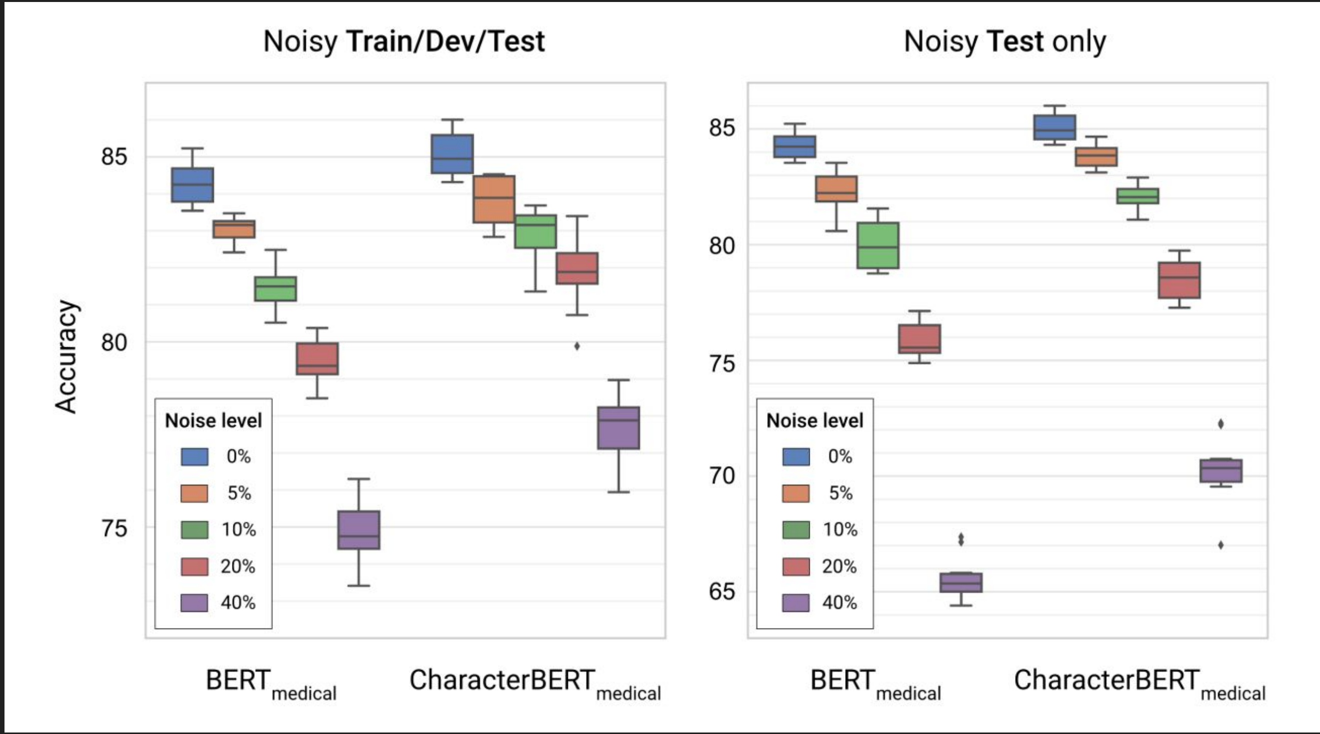
		ChemProt (F1 score)			DDI (F1 score)			MEDNLI (Accuracy)			i2b2/VA 2010 (Strict F1 score)			ClinicalSTS (Pearson Correlation)		
		Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
$\text{BERT}_{\text{general}}$	S	66.94	67.04	67.84	76.93	77.40	77.99	77.43	77.67	77.94	85.72	85.86	85.97	83.07	83.22	83.35
	E	69.81	69.88	69.98	79.99	80.14	80.35	78.71	78.90	79.15	86.71	86.73	86.77	84.27	84.33	84.36
$\text{CharacterBERT}_{\text{general}}$	S	67.89	68.24	68.38	77.89	78.17	79.03	77.18	78.09	78.50	85.99	86.18	86.28	83.63	83.91	84.09
	E	71.27	71.30	71.40	80.39	80.54	80.88	79.20	79.29	79.47	87.15	87.17	87.23	84.94	84.97	85.03
$\text{BERT}_{\text{medical}}$	S	71.76	71.93	72.23	77.54	77.93	79.15	83.79	84.25	84.69	88.99	89.01	89.08	84.80	84.98	85.20
	E	73.85	73.94	74.01	80.14	80.20	80.38	85.51	85.65	85.78	89.49	89.51	89.55	86.01	86.08	86.12
$\text{CharacterBERT}_{\text{medical}}$	S	72.84	73.44	73.78	79.18	80.38	81.70	84.56	84.95	85.58	89.14	89.24	89.30	82.92	84.80	85.15
	E	75.31	75.40	75.50	82.44	82.74	83.01	85.83	85.97	86.11	89.73	89.75	89.77	85.54	85.62	85.76
BERT (Devlin et al, 2018)	S	68.67	68.82	69.18	77.67	78.08	78.83	77.67	78.02	78.29	86.23	86.54	86.61	83.97	84.44	84.65
	E	71.46	71.54	71.64	79.46	79.49	79.61	79.41	79.47	79.54	87.23	87.26	87.28	85.32	85.37	85.40
BlueBERT (Peng et al, 2019)	S	68.25	68.31	68.69	77.55	77.89	78.74	84.07	84.25	84.55	88.47	88.73	88.87	84.39	84.98	85.39
	E	69.93	69.98	70.10	80.26	80.33	80.43	85.25	85.41	85.44	89.22	89.24	89.28	85.95	85.99	86.06

Same results in 1st, 2nd and 3rd quartiles of the score distribution over the 10 seeds.

Robustness to Noise

- Noisy versions of the MedNLI task
  - Noise = delete, swap, introduce random chars.
  - Noise level = probability of changing a char.

Robustness to Noise



Downsides of CharacterBERT

- Single downside: longer pre-training (108% slower)

Fine-tuning (w/ Tesla V100-PCIE-32GB)

Avg. +19%

	i2b2	MEDNLI	STS	DDI	ChemProt
BERT	3:36:20	1:09:29	0:02:58	1:32:42	2:42:36
CharacterBERT	4:29:01	1:22:40	0:04:12	1:19:43	3:25:31
Relative difference	**+24.35%**	**+18.97%**	**+41.57%**	**-14.01%**	**+26.39%**

Inference (w/ Tesla V100-PCIE-32GB)

Avg. -14%

	i2b2	MEDNLI	STS	DDI	ChemProt
BERT	0:11:16	0:00:11	0:00:01	0:00:31	0:02:28
CharacterBERT	0:10:57	0:00:10	0:00:01	0:00:22	0:02:44
Relative difference	**-2.81%**	**-9.09%**	**0.00%**	**-29.03%**	**+10.81%**

→ pre-trained models are available [link]

Conclusion

Conclusion

- CharacterBERT: CharacterCNN instead of WordPieces
    - More convenient
    - Improved performance (⅘ tasks)
    - Improved robustness

→ Price: slower pre-training

→ Possible solution: contrastive pre-training ?

Thank you for listening 😊

Code & Pre-trained models:

https://github.com/helboukkouri/character-bert

References

[1] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[2] Boukkouri, Hicham El, et al. "CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters." arXiv preprint arXiv:2010.10392 (2020).

[3] Peters, Matthew E., et al. “Deep contextualized word representations.” arXiv preprint arXiv:1802.05365 (2018).

Corpora

Johnson, Alistair, et al. “MIMIC-III Clinical Database” (version 1.4). PhysioNet (2016), https://doi.org/10.13026/C2XW26.

PMC OA corpus: https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Evaluation tasks

Chemprot: Krallinger, Martin et al. “Overview of the BioCreative VI chemical-protein interaction Track.” (2017).

DDI: Herrero-Zazo, María, et al. "The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions." Journal of biomedical informatics 46.5 (2013): 914-920.

ClinicalSTS: Wang, Yanshan, et al. "MedSTS: a resource for clinical semantic textual similarity." Language Resources and Evaluation 54.1 (2020): 57-72.

I2b2 2010: Uzuner, Özlem, et al. “2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text.” Journal of the American Medical Informatics Association 18.5 (2011): 552–556.

MedNLI: Shivade, Chaitanya. “MedNLI-A Natural Language Inference Dataset For The Clinical Domain” (version 1.0.0). PhysioNet (2019), https://doi.org/10.13026/C2RS98.

Other relevant references

El Boukkouri, Hicham. "Ré-entraîner ou entraîner soi-même? Stratégies de pré-entraînement de BERT en domaine médical." Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL. ATALA, 2020.