

March 23

Kenneth Church

Conference papers

<https://aclanthology.org/events/acl-2022/#2022acl-long>

↑ up


pdf (full) **bib (full)** **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**

pdf **bib** **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**
Smaranda Muresan | Preslav Nakov | Aline Villavicencio

pdf **bib** **abs** **AdapLeR: Speeding up Inference by Adaptive Length Reduction**
   Ali Modarressi | Hosein Mohebbi | Mohammad Taher Pilehvar

pdf **bib** **abs** **Quantified Reproducibility Assessment of NLP Results**
 Anya Belz | Maja Popovic | Simon Mille

pdf **bib** **abs** **Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings**
Sangwon Yu | Jongyoon Song | Heeseung Kim | Seongmin Lee | Woo-Jong Ryu | Sungroh Yoon

pdf **bib** **abs** **AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level**
 Amit Seker | Elron Bandel | Dan Bareket | Idan Brusilovsky | Refael Greenfeld | Reut Tsarfaty

pdf **bib** **abs** **Learning to Imagine: Integrating Counterfactual Thinking in Neural Discrete Reasoning**
Moxin Li | Fuli Feng | Hanwang Zhang | Xiangnan He | Fengbin Zhu | Tat-Seng Chua

pdf **bib** **abs** **Domain Adaptation in Multilingual and Multi-Domain Monolingual Settings for Complex Word Identification**
 George-Eduard Zaharia | Răzvan-Alexandru Smădu | Dumitru Cercel | Mihai Dascalu

Examples of Invited Talks

https://github.com/kwchurch/Benchmarking_past_present_future/blob/master/README.md

Video is different from an oral talk

- People can stop a video
- But it is hard to click on links in a video
- Video can be used as a teaser to encourage people to read the paper, or visit a github

Posters

<https://www.youtube.com/watch?v=W97N8wGShVg>



Virtual: GatherTown



Training on Lexical Resources

Kenneth Church, Xingyu Cai, Yuchen Bian
Baidu, USA



gft (general fine-tuning): A Little Language for Deep Nets (Unix Philosophy: Less is More)

Standard 3-Step Recipe

- 1. `gft_fit`: `model = classify(Labels) - text1`
- 2. `gft_fit`: `model = classify(Labels) - text1`
- 3. `gft_fit`: `model = classify(Labels) - text1`

Examples of 1-line GFT Programs

- `gft_fit`: `model = classify(Labels) - text1`
- `gft_fit`: `model = classify(Labels) - text1`
- `gft_fit`: `model = classify(Labels) - text1`

Agenda

Syn/Ant Binary Classification

From Words to Texts

- ANEC: Multitask Expressions
- OOVs: Out of Vocabulary words
- Multi-Lingual
- Negation

Leakage with Standard Benchmarks

VAD Regression

VAD = Valence, Arousal, Dominance

Training on Follows Thesaurus

Results (R1)

word1	word2	gold	synonym	antonym
ancient	oldfashioned	0	good	bad
blame	disapprove	0	good	evil
clearly	confoundly	1	good	benevolent
debt	liability	0	good	terrorist
demure	modest	0	good	terrorist
profitable	fruitless	1	bad	terrorist
revolve	origins	0		
rotation	order	0		
vanity	selfishness	1		

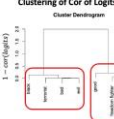
Table 1: Inference: synonymy iff $y_1 > y_2$

From Words to Text (Undesirable Biases)

Logits (top); Cor of Logits (bottom)

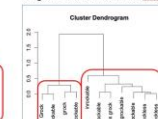


Clustering of Cor of Logits



From Words to Text (OOVs: Out of Vocabulary)

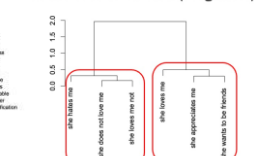
Google Translations of Variants of Grock



Delta (Proposed - MoE)

	adj	noun	verb	follows
adj	-0.013	-0.202	-0.139	-0.016
noun	-0.068	-0.040	-0.065	-0.031
verb	-0.046	-0.107	0.003	0.016
follows	0.089	0.006	0.078	0.281

From Words to Text (Negation)



Conclusions

- Proposed fine-tuning deep nets on lexical resources
- Thesaurus (syn/ant classification)
- VAD Regression
- Proposed method is competitive with MoE baseline, and
- Generalizes better to Follows (1898)
- Words → Texts
- Proposed method can be applied at inference time to MWEL, OOVs and longer texts in multiple languages
- On a cautionary note,
 - found evidence of leakage
 - in standard benchmarks as well as Follows (1898)
 - Work based on bad benchmarks
 - may need to be restricted
- To address concerns with leakage,
 - we introduced a new task: VAD regression
 - Since VAD is fully-connected, we could study sampling methods
- Transfer is more effective
 - when splits are large
 - and representative of one another
 - In such cases,
 - reduces training loss (in fine-tuning)
 - also reduces loss on other splits
- Proposed method:
 - effective for pairs of words in training set
 - but less so for pairs of unseen words

Morpheme Diagnostic

Group words by affixes

Plot y for pairs in each group

$y(w_1, w_2) = |VAD(w_1) - VAD(w_2)|$

Red baselines:

• 0: distance for maximally similar pair

• $\sqrt{2}$: distance for random pair

Observations:

• VAD varies systematically:

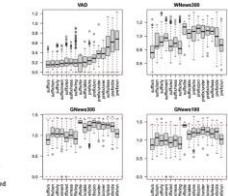
• Small (similar to MWEL) - ~ 0.1 , ~ 0.2

• Large (dissimilar to MWEL) - ~ 0.3 , ~ 0.4

• Word2Vec is large (almost everywhere)

• Almost all pairs of words are far apart

• Even words that are morphologically related



Reading Group

- Good opportunity to learn about some new topic
 - Volunteer picks a small set of related papers (often just one paper)
 - Everyone should read the papers before the reading group
 - Volunteer presents paper
 - Group discusses (perhaps interactively)
- Presentation is different from a conference paper
 - Often more directed to interests of group
 - Often more balanced (critical)
 - How are the results in the paper relevant to needs of group
 - Opportunities to improve over results in paper
 - Often more focused on a small piece of paper
 - Such as a dataset, benchmark, etc.

Pubtator

<https://www.ncbi.nlm.nih.gov/research/pubtator/?view=publication&pmid=36950551&query=p53&page=1>

PubTator

p53

NIH NLM

MENTIONS

group ▼ sort ▼

type freq

Search...

GENE

CDK1 398

CD4 20

TP53 18

SPP1 18

PD-L1 18

more

DISEASE

TUMOR 95

NSCLC 6

LUNG ADENOCARCINOMA 3

LUNG CANCER 3

BREAST CANCER 2

more

CHEMICAL

S5C 7

PYRIMETHAMINE 6

TMB 5

DDI 2

PMID36950551 • PMC10025485

2023

Prognostic and immunological characteristics of CDK1 in lung adenocarcinoma: A systematic analysis

Du Q, Liu W ... Huang D • Front Oncol

BiocXML

Background

Cyclin-dependent kinases (CDKs) play a key role in cell proliferation in lung adenocarcinoma (LUAD). Comprehensive analysis of CDKs to elucidate their clinical significance and interactions with the tumor immune microenvironment is needed.

Methods

RNA expression, somatic mutation, copy number variation, and single-cell RNA sequencing data were downloaded from public datasets. First, we comprehensively evaluated the expression profile and prognostic characteristics of 26 CDKs in LUAD, and CDK1 was selected as a candidate for further analysis. Then, a systematic analysis was performed to explore the relationships of CDK1 with clinical characteristics and tumor immune microenvironment factors in LUAD.

Results

☒ BIOCONCEPTS

☒ GENE

☒ DISEASE

☒ CHEMICAL

☒ MUTATION

☒ SPECIES

☒ CELLLINE

NAVIGATION

TITLE

INTRODUCTION

MATERIALS AND METHODS

RESULTS

DISCUSSION

DATA AVAILABILITY STATEMENT

AUTHOR CONTRIBUTIONS

CONFLICT OF INTEREST

PUBLISHER'S NOTE

SUPPLEMENTARY MATERIAL



Chih Hsuan Wei

Staff Scientist at NCBI / NLM (National Library of Medicine)

Verified email at ncbi.nlm.nih.gov

Natural language processing Machine learning Deep learning BioNLP



Cited by

[VIEW ALL](#)

	All	Since 2018
Citations	4178	2934
h-index	29	25
i10-index	44	39



TITLE	CITED BY	YEAR
-------	----------	------

[BioCreative V CDR task corpus: a resource for chemical disease relation extraction](#)

541 2016

J Li, Y Sun, RJ Johnson, D Sciaky, CH Wei, R Leaman, AP Davis, ...
Database 2016

[PubTator: a web-based text mining tool for assisting biocuration](#)

537 2013

CH Wei, HY Kao, Z Lu
Nucleic acids research 41 (W1), W518-W522

[Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation \(CDR\) task](#)

353 2016

CH Wei, Y Peng, R Leaman, AP Davis, CJ Mattingly, J Li, TC Wieggers, ...
Database 2016, baw032

[tmChem: a high performance approach for chemical named entity recognition and normalization](#)

272 2015

R Leaman, CH Wei, Z Lu
Journal of cheminformatics 7 (1), S3

Index of /pub/lu/PubTatorCentral

Name	Last modified	Size
Parent Directory		-
PubTatorCentral_BioCXML/	2022-01-09 14:06	-
tmp/	2021-08-13 03:39	-
AvailablePMIDsinPubTator.txt	2023-03-22 21:44	294M
README.txt	2022-05-30 18:37	6.1K
bioconcepts2pubtatorcentral.gz	2022-12-17 23:54	5.1G
bioconcepts2pubtatorcentral.offset.gz	2023-02-17 06:05	30G
bioconcepts2pubtatorcentral.offset.sample	2022-04-11 13:56	30K
bioconcepts2pubtatorcentral.sample	2019-08-19 10:32	50K
cellline2pubtatorcentral.gz	2022-12-17 23:56	40M
cellline2pubtatorcentral.sample	2019-07-29 11:00	44K
chemical2pubtatorcentral.gz	2022-12-17 23:55	1.4G
chemical2pubtatorcentral.sample	2019-07-29 11:00	52K
disease2pubtatorcentral.gz	2022-12-17 23:55	1.8G
disease2pubtatorcentral.sample	2019-07-29 11:00	57K
gene2pubtatorcentral.gz	2022-12-17 23:56	730M
gene2pubtatorcentral.sample	2019-07-29 11:00	42K
mutation2pubtatorcentral.gz	2022-12-17 23:56	61M
mutation2pubtatorcentral.sample	2019-07-29 11:00	42K
species2pubtatorcentral.gz	2022-12-17 23:56	527M
species2pubtatorcentral.sample	2019-07-29 11:00	44K


[HHS Vulnerability Disclosure](#)

<https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/>








Point of Pubtator Example



- Many people are interested in NER
 - Named entity recognition
- Practical applications?
 - Not so clear, but pubtator looks like it should be useful
 - Still, not so obvious that this is a useful application
 - BERT has many more citations (and more awareness) than pubtator
- Accessibility
 - Pubtator is super-open
 - Supports both ad hoc queries as well as bulk download
- Community building: encourage others to join in on the fun



What is NER (Named Entity Recognition)?


 **Hugging Face**




Search models, datasets, users...



 Models  Datasets  Spaces  Docs  Solutions Pricing  Log In 

 Datasets: **con112003**  like 39

Tasks:  Token Classification Sub-tasks: **named-entity-recognition** **part-of-speech** Languages:  English Multilinguality: **monolingual** Size Categories: **10K<n<100K** Language Creators: **found**

Annotations Creators: **crowdsourced** Source Datasets: **extended|other-reuters-corpus** License:  other

 Dataset card  Files and versions  Community 5

 Dataset Preview 

Split

train

id (string)	tokens (sequence)	pos_tags (sequence)	chunk_tags (sequence)	ner_tags (sequence)
"0"	["EU", "rejects", "German", "call", "to", "boycott", "British", "lamb", "."]	[22, 42, 16, 21, 35, 37, 16, 21, 7]	[11, 21, 11, 12, 21, 22, 11, 12, 0]	[3, 0, 7, 0, 0, 0, 7, 0, 0]
"1"	["Peter", "Blackburn"]	[22, 22]	[11, 12]	[1, 2]
"2"	["BRUSSELS", "1996-08-22"]	[22, 11]	[11, 12]	[5, 0]
"3"	["The", "European", "Commission", "said", "on", "Thursday", "it", "disagreed", "with", ...]	[12, 22, 22, 38, 15, 22, 28, 38, 15, 16, 21, 35, 24, 35, 37, 16, 21, 15, 24, 41, 15, 16, ...]	[11, 12, 12, 21, 13, 11, 11, 21, 13, 11, 12, 13, 11, 21, 22, 11, 12, 17, 11, 21, 17, 11, ...]	[0, 3, 4, 0, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 7, 0, 0, 0, 0, 0, 0, 0, 0]
"4"	["Germany", "'s", "representative", "to", "the", "European", "Union", "'s", "veterinary" ...]	[22, 27, 21, 35, 12, 22, 22, 27, 16, 21, 22, 22, 38, 15, 22, 24, 20, 37, 21, 15, 24, 16, ...]	[11, 11, 12, 12, 13, 11, 12, 12, 11, 12, 12, 12, 12, 21, 13, 11, 12, 21, 22, 11, 13, 11, 1, 13, ...]	[5, 0, 0, 0, 0, 0, 3, 4, 0, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 5, 0, 0, 0, 0, 0, 0, ...]
"5"	["\"", "We", "do", "n't", "support", "any", "such", "recommendation", "because", "we", ...]	[0, 28, 41, 30, 37, 12, 16, 21, 15, 28, 41, 30, 37, 12, 24, 15, 28, 6, 0, 12, 22, 27, 16, ...]	[0, 11, 21, 22, 22, 11, 12, 12, 17, 11, 21, 22, 22, 11, 12, 13, 11, 0, 0, 11, 12, 11, 12, ...]	[0, 3, 0, 0, 0, 1, 2, 2, 2, 0, 0, ...]
"6"	["He", "said", "further", "scientific", "study", "was", "required", "and", "if", "it", ...]	[28, 38, 16, 16, 21, 38, 40, 10, 15, 28, 38, 40, 15, 21, 38, 40, 28, 20, 37, 40, 15, 12, ...]	[11, 21, 11, 12, 12, 21, 22, 0, 17, 11, 21, 22, 17, 11, 21, 22, 11, 21, 22, 22, 13, 11, ...]	[0, 3, 4, 0]
"7"	["He", "said", "a", "proposal", "last", ...]	[28, 38, 12, 21, 16, 21, 15, 22, 22, 22, 22, ...]	[11, 21, 11, 12, 11, 12, 13, 11, 12, 12, 12, ...]	[0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 1, 2, 0, 0, 0, ...]