Tian Kang

Ph.D. candidate, Department of Biomedical Informatics, Columbia University
tk2624@cumc.columbia.edu | 917-288-9570 | 622 West 168 Street, New York, NY

SUMMARY

- With 6+ years of experience in Natural Language Processing (NLP) and Machine Learning research in clinical notes, clinical research text and health consumer-generated text.
- Experience in leading development of quality models for both core NLP problems and NLP applications
- Interdisciplinary collaborator, experience in working with clinical practitioners, linguistic researchers and biomedical researchers

EDUCATIONS

Ph.D. in Biomedical Informatics, Columbia University (GPA 3.90/4.00) Expected 2021

- Dissertation title: "Neural-Symbolic Approaches for Delivering Medical Evidence from Free-text Literature to Evidence-Based Practice", advised by Dr. Chunhua Weng.
- The dissertation aims to develop novel language technologies to enable machines to read and understand medical evidence from text-based literature, and help clinicians conduct Evidence-based Medicine at the point of care.

M.A. in Biomedical informatics, Columbia University (GPA3.76/4.00) 2016

- Master thesis: "EliIE: A Machine Learning-based Information Extraction System to Formalize Clinical Research Eligibility Criteria into OMOP Common Data Model".
- Published in Journal of the American Medical Informatics Association (JAMIA) and invited talk in JAMIA Journal Club.

B.S. in Bioinformatics, Huazhong University of Science and Technology, China (GPA 3.50/4.00) 2014

RESEARCH HIGHLIGHTS

Developing a knowledge-infused language model for clinical research text (clinical trial protocols, literature, etc.).

Department of Biomedical Informatics, Columbia University, Oct. 2020 – present

- Aiming to overcome the shortcoming of existing pretrained language models in the highest level of linguistics: knowledge
- Proposed and leading the development of a novel language model learning for clinical research corpus by pretraining knowledge-infused auxiliary task.

***EvidenceMap*: developing semantic technologies and a Natural Language Interface for facilitating clinicians to practice Evidence-based Medicine at the point of care.**

Department of Biomedical Informatics, Columbia University, Nov. 2018 – present

- Proposed a novel representation model for free-text clinical research literature that is computationally efficient and interoperable among different evidence resources (e.g., PubMed, Clinicaltrials.gov)
- Leading the development of multiple semantic technologies and a Natural Language Interface, EvidenceMap, to help clinicians practice Evidence-Based Medicine at the point of care navigating medical evidence in PubMed.
- Conducting usability study to evaluate *EvidenceMap* by collaborating with clinicians.

Developing a Medical Evidence Dependency(*MD*)-informed Self-Attention for Machine Reading Comprehension of Medical Evidence

Department of Biomedical Informatics, Columbia University, March 2020 – August 2020 (GitHub)

- Proposed and led the development of a novel attention model for Machine Reading Comprehension in clinical research literature.
- On multiple public benchmarks for Machine Reading Comprehension in clinical literature, by incorporating MD-informed Self-Attention to existing language understanding models (e.g., BioBERT), the model achieved as large as 30% of performance gain in F1 score and new state-of-the-art performance.

Improving biomedical NLP in low-resource domains by UMLS-based Data Augmentation

Department of Biomedical Informatics, Columbia University, Dec. 2019 – March 2020 (GitHub)

- Led the development of a data augmentation method, called UMLS-EDA, by simple text transformation and incorporation of domain knowledge such as UMLS, to improve biomedical NLP using typical research setting

annotations (small but high quality).

- The proposed augmentation helped improved core biomedical NLP tasks such as NER and sentence classification substantially. The maximum gain on NER task using LSTM-CRF architecture is over 15% in F1 score, outperformed the same architecture but initialized from BioBERT.

Cataloging and mapping treatments for patients from an Online Autism Community
Department of Biomedical Informatics, Columbia University, Sep. 2016 – Jan. 2017

- Applied NLP to understand the treatments used in real world for Autism Spectrum Disorder using patient-generated text from online health communities and compared it to treatment guidelines.
- Presented it at 2017 WWW conference.

***EliIE* (Free-text Eligibility Criteria Information Extraction System): A machine learning-based IE system to formalize clinical research eligibility criteria and facilitate Electronic Health Record screening for trial recruitment**.
Department of Biomedical Informatics, Columbia University, Aug. 2015 – Apr. 2016 (GitHub)

- Led the annotation team of medical professionals and development of an NLP system, *EliIE*, to parse and formalize clinical trial eligibility criteria in order to facilitate trial recruitment.
- Published in the peer-review journal JAMIA, and gave an invited talk at JAMIA Journal Club.

TECHNICAL SKILLS

Languages: Python, R, MySQL, PHP, Java.

Skills: Natural Language Processing, Deep Learning (Tensorflow, PyTorch), Web Application, Data Mining.

PUBLICATIONS

- **Kang, T.**, Turfah, A. Kim, J. Perotte, A. and Weng, C. (2020). Medical Evidence Dependency-informed Self-Attention: Exploiting the Synergy of Symbolic and Neural Approaches. (under review)
- **Kang, T.**, Perotte, A., Tang, Y., Ta, C. and Weng, C. (2020). UMLS-based Data Augmentation for Biomedical NLP with Limited Data. (accepted by *Journal of the American Medical Informatics Association*)
- **Kang, T**., Zou, S., & Weng, C. (2019). Pretraining to recognize PICO elements from randomized controlled trial literature. *Studies in health technology and informatics*, *264*, 188.
- Wei, D. H., **Kang, T.**, Pincus, H. A., & Weng, C. (2019). Construction of disease similarity networks using concept embedding and ontology. *Studies in health technology and informatics*, 264, 442.
- Rogers, J. R., Callahan, T. J., **Kang, T.**, Bauck, A., Khare, R., Brown, J. S., ... & Weng, C. (2019). A Data Element-Function Conceptual Model for Data Quality Checks. *eGEMs*, *7*(1).
- Yuan, C., Ryan, P. B., Ta, C., Guo, Y., Li, Z., Hardin, J., ... **Kang, T.** & Weng, C. (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4), 294-305.
- Butler, A., Wei, W., Yuan, C., **Kang, T**., Si, Y., & Weng, C. (2018). The data gap in the EHR for clinical research eligibility screening. *AMIA Summits on Translational Science Proceedings*, 2018, 320.
- Sen, A., Goldstein, A., Chakrabarti, S., Shang, N., **Kang, T**., Yaman, A., ... & Weng, C. (2018). The representativeness of eligible patients in type 2 diabetes trials: a case study using GIST 2.0. *Journal of the American Medical Informatics Association*, 25(3), 239-247.
- Zhang, S., **Kang, T**., Qiu, L., Zhang, W., Yu, Y., & Elhadad, N. (2017, April). Cataloguing treatments discussed and used in online autism communities. *In Proceedings of the 26th International Conference on World Wide Web* (pp. 123-131).
- **Kang, T**., Zhang, S., Tang, Y., Hruby, G. W., Rusanov, A., Elhadad, N., & Weng, C. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6), 1062-1071.
- **Kang, T**., Zhang, S., Xu, N., Wen, D., Zhang, X., & Lei, J. (2017). Detecting negation and scope in Chinese clinical notes using character and word embedding. *Computer methods and programs in biomedicine*, 140, 53-59.
- Zhang, S., **Kang, T**., Zhang, X., Wen, D., Elhadad, N., & Lei, J. (2016). Speculation detection for Chinese clinical notes: impacts of word segmentation and embedding models. *Journal of biomedical informatics*, 60, 334-341.
- **Kang, T**., Elhadad, N., & Weng, C. (2015). Initial readability assessment of clinical trial eligibility criteria. In *AMIA Annual Symposium Proceedings* (Vol. 2015, p. 687). American Medical Informatics Association.