Човек или ИИ? Доверие, интерпретация и човешки надзор в ерата на големите езикови модели

Преслав Наков

Големите езикови модели променят начина, по който създаваме, потребяваме и оценяваме информация. Това поражда фундаментален проблем: дали да вярваме на текстове, когато става все по-трудно да определим кой ги е написал и дали съдържанието им е вярно? Ще разгледаме този въпрос от две взаимно допълващи се перспективи: предсказване дали даден текст е генериран от ИИ и използване на големи езикови модели за подпомагане на процеса на проверка на твърдения.

Първо, ще представим ExaGPT - метод за интерпретируемото откриване на машинно генериран текст. ExaGPT обосновава своите предсказания с конкретни доказателства, като идентифицира пасажи в документа, които наподобяват примери на текстове, написани от хора или генерирани от големи езикови модели. Този подход, предоставя на потребителите прозрачни и лесни за разбиране обяснения защо даден текст изглежда като написан от човек или от машина. Ще обсъдим и резултати от мащабно многоезично изследване на човешкото възприятие, които показват, че при подходящи условия хората често могат да различат дали даден текст е генериран от ИИ или от човек, но същевременно „човекоподобен“ не означава непременно „предпочитан от хората“.

Във втората част на лекцията ще разгледаме въпроса дали големите езикови модели могат да автоматизират писането на статии, обясняващи фактологичността на твърдения. Докато повечето изследвания в областта на автоматичната проверка на твърдения приключват с оценка за достоверност, професионалните фактчекъри публикуват внимателно изготвени статии, които обясняват, контекстуализират и аргументират своите заключения. Ще представим Qraft, агентна система, която генерира цялостни фактчекърски статии чрез имитиране на ключовите етапи от процеса, използван от професионалисти, включително подбор на доказателства, планиране на статията, създаване на първоначален текст и последваща редакция. Накрая ще обсъдим подходи за интегриране на експертна обратна връзка в процеса на генериране под формата на диалог между човека и машината.

Лекцията ще се проведе на 29 юни от 14 ч. в зала 228 на ИИКТ-БАН, блок 2.

За лектора

Преслав Наков е професор и ръководител на катедрата по обработка на естествен език (NLP) в Университета за изкуствен интелект „Мохамед бин Зайед“ (MBZUAI). Той ръководител на екипи в Института за фундаментални модели към MBZUAI, разработили Jais — най-добрият арабско-ориентиран LLM с отворен код, Nanda — най-добрият модел с отворени тегла за хинди, и Sherkala — най-добрият модел с отворени тегла за казахски език.

Преди това е бил главен научен сътрудник (Principal Scientist) в Qatar Computing Research Institute към Hamad Bin Khalifa University (HBKU), където е ръководил мащабния проект Tanbih, разработен съвместно с MIT, имащ за цел да ограничи въздействието на фалшивите новини, пропагандата и медийните пристрастия, като помага на потребителите да осъзнават по-добре какво четат и така насърчава медийната грамотност и критичното мислене.

Преслав получава докторска степен по компютърни науки от Калифорнийския университет в Бъркли с подкрепата на стипендия „Фулбрайт“, както и магистърска степен от Софийския университет „Св. Климент Охридски“.

В момента е председател на Европейската секция на Асоциацията по компютърна лингвистика (EACL), секретар на ACL SIGSLAV и секретар на настоятелството на организацията Truth and Trust Online. В миналото е бил програмен председател на ACL 2022 и президент на ACL SIGLEX.

Той е член на редакционните колегии на редица водещи научни списания, сред които Computational Linguistics, Transactions of the Association for Computational Linguistics (TACL), ACM Transactions on Information Systems (TOIS), IEEE Transactions on Affective Computing (TAC), IEEE Transactions on Audio, Speech and Language Processing (TASLP), Computer Speech & Language (CS&L), Natural Language Engineering (NLE), AI Communications и Frontiers in AI. Автор е на монография на Morgan & Claypool, посветена на семантичните отношения между съществителни имена, на две книги по компютърни алгоритми и на над 400 научни публикации.

Носител е на наградата за най-добра статия на ACM WebSci 2022, на наградата за най-добра дълга статия на CIKM 2020, на наградата за най-добър ресурсен труд на EACL 2024, както и на редица други отличия, включително почетни споменавания за най-добра демонстрационна статия на ACL 2020 и за най-добра статия за задача на SemEval 2020, награда за най-добър постер на SocInfo 2019 и наградата за млад учен на RANLP 2011. Той е и първият носител на наградата „Джон Атанасов“ на президента на Република България, на името на изобретателя на първия автоматичен електронен цифров компютър.

Неговите научни изследвания са отразявани от над 100 медии по света, включително Reuters, Forbes, Financial Times, CNN, Boston Globe, Al Jazeera, Defense One, Business Insider, MIT Technology Review, Science Daily, Popular Science, Fast Company, The Register, WIRED и Engadget.