„Езикът в ерата на дигиталните технологии и изкуствения интелект“-
академично слово на проф. д-р Петя Осенова
по случай патронния празник на Софийския университет
„Св. Климент Охридски“
25 ноември 2019 г.

Уважаеми господин Ректор,
Уважаема госпожо заместник министър-председател,
Уважаеми господин заместник-председател на Народното събрание,
Уважаеми колеги и гости,

Напоследък във фокуса на политиките на ЕК и респективно на Европейския съюз се появи знаковото понятие ‘изкуствен интелект’. Много от вас знаят, че това понятие всъщност не е никак ново (то съществува поне от 60-те години на миналия век), но някак изведнъж - с увеличаването на компютърната изчислителна мощ - то се изпълни с ново съдържание (или просто се преосмисли в новите условия) и показа реалните си възможности успешно да подпомага редица важни дейности с приложения в електронното образование, неинвазивната медицина, единния дигитален пазар, борбата с фалшивите новини, екологията и т.н. Изкуственият интелект (както и естественият впрочем) обаче има и своята тъмна страна. Много от методите, които се използват, макар да дават изключително добри резултати, представляват своего рода ‘черни кутии’. По тази причина процесите в тях трудно могат да бъдат разбрани, проследявани и управлявани, а това смущава, защото е важно да се знае как се взима дадено решение. Затова в момента в световен мащаб се работи интензивно по посока на преодоляване на тези слабости.

Дигиталните технологии, от друга страна, позволиха масовия ни достъп до различна по тип информация, тъй като и самата информация вече е основно дигитална и следователно - лесно преносима. Приближихме се не само до различни тематични области, но и до различно време и пространство (напр. приложенията и игрите с добавена (историческа) реалност. Освен това в помощ на забързаното ни ежедневие се появиха виртуалните асистенти (като АЛЕКСА на компанията Амазон), с които общуваме, за да получим съвет или помощ относно нашето банкиране, пътуване или здраве. Разбира се, трябва да се отчете и негативният страничен ефект, при който огромните, разнообразни и бързо разпространяващи се потоци дигитална информация (особено чрез социалните медии) доведоха и до по-големите възможности за манипулация и пропаганда сред хората поради незнание или с цел заблуда.

По образование и призвание съм филолог-езиковед, но от 20 години насам живея професионално в интересния интердисциплинарен свят на езиковите технологии - и по-специално на езиковите технологии за българския език (в тясно сътрудничество с колегите от ИИКТ-БАН). Разглеждам езиковите технологии като съществена част от дигиталните технологии и като **хуманитарната сърцевина на изкуствения интелект.** Езиковите технологии се отнасят най-общо до събирането, подготовката, обработката и оценката на езикови данни (текст, реч, образи), както и до разработването на софтуерни модули за анализ и генерация на текст, извличане на ново знание и други. Езиковите технологии обаче успешно подпомагат и неезикови приложения (напр. мултимодалните приложения, роботиката, интернет на нещата, електронното правителство и др.). Те позволяват търсенето и извличането на релевантно съдържание (напр. за бизнес аналитика, за сравняване на европейското и националното право, за подпомагане на лекарите при изготвяне на портфолиа на пациентите и много други); проследяването на важни тенденции или мнения по определен въпрос (напр. в историята, философията, политологията, социологията). Самите езикови ресурси включват основно два типа: корпуси и речници. Корпусите са структурирани текстови архиви, събрани в определен формат, за да улеснят изпълнението на някаква задача или множество от задачи. Благодарение на тях можем да обозрем много данни, да тестваме своя хипотеза или да намерим нови явления, за които не сме и подозирали, че съществуват. Тези структурирани архиви могат да бъдат обогатени с лингвистична информация: част на речта, подлог, сказуемо, значение на думата и т.н. По този начин - чрез внедрени в текста езикови модели - става възможно извличането на по-точна информация от данните. Напр. голяма част от термините са именни словосъчетания (*кръвна захар),* а за да извлечем информация за дадено събитие, използваме думите и техните граматически характеристики (напр. *Лекарите от болница „Ив. Рилски" извършиха животоспасяваща операция на мъж с рядък мозъчен тумор).* Речниците, от своя страна, кодират лексикалния запас на един език и дават идеята за потенциалната свързаност с други думи в текста - затова те подпомагат добавянето на лингвистична информация към корпусите - лексикални значения, парадигматична информация (синонимия, антонимия, част и цяло, вид и подвид), граматическа информация, информация за участниците в дадена ситуация и др.

Обработващите средства се разделят най-общо на методи с правила и на статистически методи. Методите с правила бяха популярни през 70-80-те години на миналия век. Те разчитат на езиковото моделиране и затова са с

висока степен на точност, но невинаги обхващат всичко. Езиковите модели се строят на базата на лингвистични теории. Тези теории са много важни за постигането на системност при кодиране на езиковите данни. Статистическите методи, които разчитат на обучение чрез определен алгоритъм и се славят с добро покритие, придобиха особено голямо значение с увеличаването на мощността на изчислителната техника. Сред тях са популярни методите с учител и методите без учител. При методите с учител (supervised) се използват езикови корпуси-еталони (т.е. с правилно въведена информация), върху които се обучават алгоритмите, докато при методите без учител (unsupervised) се използват данни без предварително въвеждане на знание за езика. По принцип целта е да се развиват методите без учител, защото по този начин по-бързо и по-директно се постигат добри резултати над големи неструктурирани данни, но в действителност всички приложения в някакъв свой етап не могат да минат без знанието за естествения език.

Напоследък модерни станаха методите с невронни мрежи, които обработват езикови данни, представени като вектори (word embeddings). Невронните мрежи получават като вход определени данни, обучават се върху тях, използвайки различни слоеве с памет и механизми за контекст, и после автоматично анализират други данни. Превръщането на думите и изреченията, а и цели документи, във вектори, позволи на алгоритмите да вземат предвид контекстите им на употреба, да изчислят близост или отдалеченост в този контекст и да постигнат нечувани досега резултати. Така в компютърните науки само преди 3-4 години беше направен значителен пробив в обработката на големи масиви от данни.

Бих искала обаче да обърна внимание на важния въпрос за мястото на езика в политиките, свързани с изкуствения интелект и дигиталните технологии. По мое мнение ролята на езиковите ресурси и средства е **централна** в ерата на дигиталните технологии и изкуствения интелект. Бих искала да илюстрирам това с няколко мита от областта, в която работя:

Мит 1: Има един езиковонезависим модел, който е директно и автоматично приложим за всички езици:

- всъщност голямото количество приложения са доста езиково зависими и без необходимите ресурси за даден език те са неуспешни.

Мит 2: От големите данни всичко нужно може да се научи чрез машинно обучение и следователно няма нужда от данни-еталони и от лингвистична информация в тези данни:

- всъщност съвсем не се научава всичко; например не се научават детайли, особености, редки случаи, т.е. явленията с

ниска честота или ранк. Това е поради фрагментарността и имлицитността на знанието дори в големите бази данни. Голяма част от човешкото знание, необходимо за разбирането на големите данни (текст например), не се съдържа в тези данни.

Мит 3: Едни и същи данни, алгоритми, модели дават едни и същи добри резултати без значение каква е областта, кой е езикът и каква е целта:

- всъщност често се налага адаптиране на ресурсите, моделите, подходите и дори алгоритмите според поставената задача за разрешаване. Необходима е и също така намеса на експерти в областта.

Мит 4.1: Проблемът с машинния превод от всички езици към всички езици е решен и ние свободно комуникираме помежду си.

- всъщност в момента най-добри са моделите от почти всеки език към английски, но не толкова в обратната посока - от английския към даден език. Също така не е добър машинният превод между двойки други езици - дори сродни - например по-добър е преводът между български и словашки, когато мине през английски. Също така, по-добър е преводът в определена тематична област, отколкото преводът изобщо.

Мит 4.2: Машинният превод от всички езици към всички езици е достатъчно добър и няма нужда от преводачи (или преводачите са редактори на машинния превод)

- всъщност машинният превод подпомага преводача да се справи с огромните количества данни за кратко време, но е по-беден лексикално и граматически от човешкия превод.

Полумит 5: Създаването на езикови ресурси е много скъпо и нелеко начинание. Затова е ненужно или нежелателно.

- Факт е, че създаването на ресурси и средства за даден език е трудоемко и скъпо начинание, но всъщност инвестирането в подобни ресурси се отплаща многократно. За справка - приложенията за английския език, които са доста развити благодарение на постоянните инвестиции в езикови технологии.

За щастие, в съвременността голямо значение имат и методите, основани на знание (включително и лингвистично). Тези методи също разчитат на невронните мрежи и на векторизацията на данните, но те отчитат и експлицитно кодираното експертно знание. Такова приложение е например снемането на многозначност. Тази задача е важна за много реални приложения. Например при определянето на терминологията, присъща за определена предметна област (химия, физика, право, медицина, политология, педагогика и др.), при системите за превод, и др.

Неслучайно живеем и в ерата на свързаните отворени данни. Ние вече преживяваме бума на огромните, разнородни (дори подвеждащи и фалшиви) количества данни, които са непосилни за критическо наблюдение и осъзнаване от човека. Затова сега идва ред на свързването им в мрежа от знания с цел постигането на съвместимост и по-добър информационен поток от данните към човека. Ако има нещо, което все още да липсва при машинната обработка, това е достатъчното и свързано знание във всички негови форми. Мрежата от свързани знания (Knowledge graph, използващ Linked Open Data) използва семантични технологии, за да свърже лингвистичното, експертното и енциклопедичното знание. Например, ако искаме да извлечем автоматично информация от примера: *Петър Стоянов спечели трофея в Москва,* е необходимо знание за света. То идва от информацията, че Петър Стоянов е името не само на един от президентите на Република България, но и на български сумист. Само граматиката не е достатъчна за разбирането на даден контекст. Най-проблемни остават знанията за света, които включват ситуативното знание, човешкия опит, културната памет, пресупозитивните връзки.

Похвално е, че съществуват инициативи, които разглеждат езика като данни, т.е. като обект на науката за данните (data science). Трябва, разбира се, винаги да се отчита фактът, че езикът е преди всичко социално явление. Той изразява голяма част от човешкото знание, но същевременно голяма част от релационното и ситуативното знание остава скрита. Затова езикът е толкова различен от обектите на изследване в химията, физиката, математиката, биологията. Затова той толкова трудно се поддава на формализация, но пък моделирането му (макар и с редицата съпътстващи несъвършенства) остава една доста перспективна област.

Вярвам, че бъдещето е в комбинираните методи на работа с данни. Както алгоритмите, така и езиковите модели имат своето място в зоната на изкуствения интелект.

Според мен силата на езиковите технологии е именно в интердисциплинарните тематики, цели и задачи - и оттук - в смесените екипи от специалисти: математици, компютърни инженери, езиковеди, филолози, юристи, политолози, медици, журналисти, педагози, социолози, икономисти и много, много други.

За българския език вече има създадени много лингвистични ресурси и средства за автоматичен анализ. Те се вграждат в различни приложения и са тихите помагачи в социално значимите задачи за обществото. Разбира се, предстои да се създадат още, както и да се интегрират наличните в

мрежа от знания. Щастлива съм, че съм била и че продължавам да бъда част от подобна ключова мисия.

Честит празник!