# Език – обработка на реч и текст.

ВЪВЕДЕНИЕ

Компютърната лингвистика е интердисциплинарна теоретико-приложна наука, която се занимава както с формалното описание на естествения език, така и с разработването и прилагането на компютърните технологии при статистическото и логическото му анализиране и моделиране.

Езиковото компютърно моделиране не е ограничено в конкретен дял или граници в лингвистиката. То обикновено се осъществява от интердисциплинарни екипи от компютърни специалисти, лингвисти, програмисти, логици, математици, специалисти по изкуствен интелект, когнитивна психология и др. Разработват се електронни езикови приложения и системи, които обслужват потребителите при работата им с текстове – такива са програмите за автоматично коригиране на правописа, за автоматичен превод от един език на друг, за категоризиране и резюмиране на документи; за преобразуване на текст в реч и обратно и др., както и програми, обслужващи лингвистичните изследвания и анализи.

***Основните области в компютърната лингвистика са:***

- автоматичен анализ на текстове - Автоматичното анализиране на текстове се използва в области като машинен превод, интелигентно търсене на информация, автоматично отговаряне на въпроси и др.
- генериране на текстове - Компютърната обработка на реч се използва в различни компютърни системи за автоматично преобразуване на текст в реч или на реч в текст, в устройства, управлявани с гласови команди (например управление на автомобил с глас) и др. Автоматичният анализ може да обхваща различни нива от заложената в текста информация. Това определя множество подзадачи и модули като: токънизиране (разделяне текста на определени единици – фонеми, морфеми, графични думи, лексеми, изречения и др.); тагиране (приписване на характеристики на всяка отделна единица – приписване на дадени морфологични, синтактични, морфосинтактични, семантични и др. характеристики); парсиране – морфологичен анализ, синтактичен анализ, разрешаване на различни езиково специфични явления като местоименни и неместоименни анафори, елипси и др. и на различните типове езикова многозначност.

При автоматичния анализ на текстове се използват ресурси, в които е зададено знанието за езика. Такива ресурси са морфологичните речници, граматиките, онтологиите, честотни таблици, информация за синтактичните рамки на глаголите и др.

Друг източник на знания са колекциите от текстове в електронен формат /корпуси/, които се използват за различни типове езикови анализи и заключения, както и за извличане на езикова информация чрез компютърни програми и статистически техники. Този дял от компютърната лингвистика е познат като корпусна лингвистика.

- анализ и синтез на реч - Обработката на реч използва статистически техники, за да се превърнат гласовите команди в текст.

В тази лекция са представени различните компоненти, които изграждат съвременните разговорни агенти, включително алгоритмите, използвани в съвременния етап на развитие на Компютърната лингвистика (КЛ)

Сложните езикови модели трябва да могат да разпознават думи от аудио сигнал и да генерира аудио сигнал от поредица от думи. Тези задачи за разпознаване на реч и задачи за синтез на реч изискват познания по фонетика и фонология; как думите се произнасят от гледна точка на последователности от звуци и как всеки от тези звуци се реализира акустично.


**Разговорни агенти**

Целта на разговорните агенти е да се осигури на компютрите технология за езикова обработка на човешки език / естествен език в компютърно разбираем езиков изход (генериране на естествен език и синтез на реч).

Това е обобщение на просто търсене в мрежата, при което вместо просто да въвежда ключови думи, потребителят може да задава пълни въпроси, вариращи от лесни до трудни. Освен това се предоставят възможности за машинния превод на документ от един език на друг.

Много други задачи за езикова обработка също са свързани с мрежата – автоматичен превод на чужд език, използване на чат бот и други.

Примерни въпроси могат да са:

• Какво означава „космополитен“?
• Коя година е роден Васил Левски?
• Колко окръга има България през 1970г?

• Какво мислят учените за етиката на клонирането на хора?

Това, което отличава приложенията за езикова обработка от другите системи за обработка на данни, е използването на езикови познания. Например програмата Unix wc, която се използва за преброяване на общия брой байтове, думи и редове в текстов файл. Когато се използва за броене на байтове и редове, wc е обикновено приложение за обработка на данни. Въпреки това, когато се използва за преброяване на думите във файл изисква знания за това какво означава дума и по този начин се превръща в система за езикова обработка.

На някои от въпросите за дефиниция или отговор на прости въпроси за факти, като дати и местоположения, вече може да се отговори от търсачките, но отговорът на по-сложни въпроси може да изисква извличане на информация, която е вградена в друг текст на уеб страница, или правене на изводи (извличане на заключения въз основа на известни факти), или синтезиране и обобщаване на информация от множество източници или уеб страници. В този текст се изучават различните компоненти, които изграждат съвременни системи за разбиране от този вид, включително извличане на информация, разграничаване на смисъла на думата и т.н.

Важно уточнение, е че wc е изключително проста система с изключително ограничено и бедно познаване на езика. Сложните разговорни агенти или системите за машинен превод или стабилните системи за отговаряне на въпроси изискват много по-широки и позадълбочени познания за езика.

В останалите лекции се обобщават накратко видовете знания, които са необходими за тези задачи (и други като коригиране на правописа, проверка на граматиката и т.н.), както и математическите модели, които ще бъдат въведени в курса на обучение.

ЗНАНИЯ В ОБРАБОТКАТА НА РЕЧ И ЕЗИК

Обикновено при задаване на въпрос се съобразява и отговора, по отношение на време, род, число и т.н.

Знанията, необходими за подреждане и групиране на думи заедно, попадат под заглавието на синтаксиса, и сложните езикови модели следва да са устойчиви и да отговарят с пълни и смислени отговори, а не само с кратки думи от типа „Да/Не“

Преминавайки отвъд отделните думи, сложните езикови модели трябва да използва структурни познания, за да подреди правилно думите, които представляват неговия отговор.

Следващата последователност от думи няма да има смисъл, въпреки факта, че съдържа точно същия набор от думи както би бил

*Бихте ли отворили вратата на магазина?*

*Съжалявам, че се страхувам,* господине/госпожо*, че не мога.*

*Съжалявам, но се страхувам, че не мога.*

Необходима е информация също и за вида знание, който позволява на сложните езикови модели да определи, че изказването е молба за действие, за разлика от просто изявление за света или въпрос за магазин, държава или вратата.

*Например: Каква част от населението на средновековна/ренесансова България е грамотно/религиозно?*

За да се отговори на този въпрос, трябва да е известно нещо за лексикалната семантика, значението на всички думи (религиозен), както и композиционната семантика (какво точно представлява Средновековна България за разлика от сегашната и кой е периода на средновековие. Трябва също да е известно нещо за връзката на думите със синтактичната структура. Например трябва да е известно, че средновековието обхваща времева крайна точка, а не описание на агента, тъй като съпътстващата фраза е в следното изречение:

*Коя година е тази година?* За да тълкува думи като тази година, системата за отговор на въпроси трябва да проучи по-ранните въпроси, които са били зададени; в този случай предишният въпрос говори за годината, в която се е случило дадено събитие. По този начин тази задача за разрешаване на съотношението използва знания за това как думи като *това* или местоимения като *тя* се отнасят към предишни части на дискурса, като.

• Синтаксис — познаване на структурните връзки между думите от тези нива. Казваме, че някакъв вход е двусмислен, ако има множество алтернативни лингвистични структури, които могат да бъдат изградени за него. Например в турския език не съществува среден и женски род.

• Фонетика и фонология — знания за езиковите звуци всяка от които илюстрира неясната на някакво ниво:

• Прагматика — знание за връзката на смисъла с целите и намеренията тоест вземане на единичен пряк обект или може да бъде с двойно значение.

Разрешаването на неясноти на част от речта и смисъла на думата са два важни вида лексикални проблема

Например, система за синтез на текст към реч, която чете думата от която се нуждае и са морфологично или синтактично двусмислени в своята част на речта.

• Дискурс — знания за езикови единици, по-големи от едно изказване, като начини за разрешаване или разграничаване на тези неясноти. Например да решите дали се отнася до подлога или друго.

• Морфология — познаване на смисловите компоненти на думите Голямо разнообразие от задачи могат да бъдат формулирани като лексикална неоднозначност предизвикани да извършат словесното действие

• Семантика — познаване на значението

В обобщение - ангажирането в сложно езиково поведение изисква различни видове езикови познания.

**Модели и алгоритми**

Едно от ключовите прозрения на последните 50 години изследвания в езиковата обработка е, че различните видове знания могат да бъдат уловени чрез използването на малък брой формални модели или теории. Всички тези модели и теории са извлечени от стандартните набори от инструменти за компютърни науки, математика и лингвистика и трябва да бъдат общо взето познати на тези, обучени модели в тези области. Сред най-важните модели са:

- модели на състоянието;
- системи от правила;
- логика;
- вероятностни модели;
- векторно-пространствени модели.

Тези модели, от своя страна, се поддават на малък брой алгоритми, сред най-важните от които са алгоритмите за търсене в пространството на състоянието като динамично програмиране и алгоритми за машинно обучение като класификатори и други алгоритми за обучение.

В най-простата си формулировка машини модели на състояние са формални такива, които се състоят от състояния, преходи между състояния и входно представяне. Някои от вариантите на този основен модел са детерминистични и недетерминирани автомати с крайни състояния и преобразуватели с крайни състояния.

За мнозина способността на компютрите да обработват език е толкова умело, че хората биха приветствали навлизането на наистина интелигентни машини. Основата на това убеждение е фактът, че ефективното използване на езика е преплетено с нашите общи когнитивни способности.

Алън Тюринг (1950). представя модел, известен като Тест на Тюринг и започва с въпросът какво би означавало една машина да мисли е по същество не отговорим поради присъщата неточност на термините машина и мислене. Вместо това той предлага емпиричен тест, игра в която използването на език от компютъра ще формира основата за определяне дали той може да мисли. Ако машината можеше да спечели играта, ще бъде оценена като интелигентна.

В играта на Тюринг има трима участници: двама души и компютър. Един от хората е състезател и играе ролята на разпитващ. За да спечели, разпитващият трябва да определи кой от другите двама участници е машината, като зададе поредица от въпроси чрез телетайп. Задачата на машината е да заблуди разпитващия да повярва, че е човек, като отговаря като човек на въпросите на разпитващия.

Задачата на втория участник човек е да убеди разпитващия, че другият участник е машината и че те са хора.

При разпознаването на реч се търси правилната дума в пространство от звукови последователности. При синтактичния анализ ние търсим в пространство от дървета за синтактичен анализ на входно изречение. При машинния превод се търси в пространство от хипотези за превод за правилния превод на изречение на друг език. За невероятностни задачи, като например машини на състояние, се използват добре познати графични алгоритми, като например търсене в дълбочина. За вероятностни задачи се използват евристични варианти като първо най-доброто търсене и се разчита на алгоритми за динамично програмиране за изчислителна способност.

По сходен начин статията на Тюринг илюстрира вида взаимодействия, които той е имал предвид при общуването между хора и машини. Очевидно една убедителна симулация на човешки качества не се изисква експертиза във всички области:

За много езикови задачи се разчита на инструменти за машинно обучение като класификатори и модели на последователности. Много често се използват класификатори като дървета на решенията, опорни векторни машини, смесени модели на Гаус и логическа регресия., но могат и се

използват също така скрит модел на Марков модели на Марков с максимална ентропия или условни случайни полета.

**Вероятностни модели**

Вероятностните модели са от решаващо значение за улавянето на всякакъв вид лингвистично знание и биват машинни, формални и логически. Тясно свързани с тези модели са техните декларативни двойници: системи с формални правила.

Сред по-важните, които се разглеждат са регулярните граматики и регулярните отношения, граматиките без контекст, граматиките с разширени функции, както и вероятностните варианти на всички тях. Машини за състояние и системите с формални правила са основните инструменти, използвани при работа със знания по фонология, морфология и синтаксис.

Друг модел, който играе критична роля при улавянето на знанията за езика е **логиката**. Обикновено се разглежда т. нар. логиката от първи ред, известна още като смятане на предикатите, както и такива свързани формализми като ламбда-изчисление, структурни характеристики и семантични примитиви. Тези логически представяния традиционно се използват за моделиране на семантика и прагматика, въпреки че по-новите изследвания се фокусират върху по-стабилни техники извлечени от нелогическа лексикална семантика.

Всеки от моделите може да бъде допълнен с **вероятностни** такива. и да се превърне в претеглен автомат или модел на Марков. Ще бъдат разгледани скрити модели на Марков (HMM), които намират широко приложение в области при маркиране на част от речта, разпознаване на реч, диалог при изправено положение, преобразуване на текст в реч и машинен превод.

Ключовото предимство на вероятностните модели е способността им да решават много видове проблеми с неяснотата за почти всеки проблем с обработката на речта и езика може да бъде преработен като: „при дадени **N избора** за някакъв двусмислен вход, изберете **най-вероятния**“.

Моделите на **векторното пространство**, базирани на линейна алгебра, са в основата на извличането на информация. Обработващият език с помощта на някой от тези модели обикновено включва търсене и много различни на значенията на думите.

## ЕЗИК, МИСЪЛ И РАЗБИРАНЕ

Ясно е, че независимо от това, което хората вярват или знаят за вътрешната работа на компютрите, те говорят за тях и взаимодействат с тях като социални единици.

Най-ранните опити датират от опитите за създаване на крайни автомати и регулярни изрази. Шанън (1948), който пръв прилага вероятностни модели на езика чрез модели като комуникационни канали и речева акустика.

Чомски (1956) първо разглежда крайните машини като начин за характеризиране на граматика и дефинира език с крайни състояния различни, но припокриващи се области в тези различни отдели: компютърна лингвистика в лингвистиката, обработка на естествен език в компютърните науки, разпознаване на реч в електротехниката, компютърна психолингвистика в психологията.

В исторически план обработката на речта и езика е била третирана много различно в компютърните науки, електротехниката, лингвистиката и психологията/когнитивната наука.

Например ELIZA е елементарна програма, която използва съвпадение на шаблони, за да обработи входа и да го преведе в подходящи изходи. Успехът на тази проста техника се дължи на факта, че ELIZA всъщност не трябва да знае нищо, за да имитира лекар или политолог например.
30% е шанса да заблудят човек, който разпитва друг след 5 минуен разпит. Предвид факта, че може да се заблуждават някои от хората през цялото време, не е ясно колко строг е този конкретен стандарт. Независимо от това, критичният проблем за Тюринг е използването на език, както хората го правят, само по себе си е достатъчно като оперативен тест за интелигентност. Дълбоката връзка на ELIZA с идеите на Тюринг е, че много хора, които са взаимодействали с ELIZA, са повярвали, че тя наистина разбира тях и техните проблеми.

Google предоставя междуезични услуги за извличане на информация и превод, където потребителят може да предоставя заявки на родния си език, за да търси в колекции на друг език. Google превежда заявката, намира най-подходящите страници и след това автоматично ги превежда обратно на родния език на потребителя.

Огромното увеличение на изчислителните ресурси, достъпни за средния компютърен потребител, възходът на мрежата като огромен

източник на информация и нарастващата наличност на безжичен мобилен достъп поставиха приложенията за обработка на реч и език в центъра на вниманието на технологиите. Ето някои примери за внедрени в момента системи, които отразяват тази тенденция:

• Големи образователни издателства като Pearson, както и услуги за тестване като използват автоматизирани системи за анализиране на хиляди ученически есета, като ги оценяват и оценяват по начин, който е неразличим от човешките оценяващи.

• Пътуващите, които се обаждат на ботове на доставчици на пътувания, взаимодействат с агенти, които ги водят през процеса на правене на резервации и получаване на информация за пристигане и заминаване.

Алгоритми и системи за автоматична обработка на български език

Това направление включва пионерните разработки върху автоматичната обработка на български език в периода 1980-1985 г. Първата система за автоматично сегментиране на българските словоформи без използване на лексемен речник (речник на българските основи) и без каквито и да е ограничения, наложени върху входния текст, е представена през 1981. Системата реализира алгоритъм за автоматичното сегментиране (определяне на основа и окончание) на словоформите в текста. Алгоритъмът е комбинация от евристични съображения и точно определени множества от езикови елементи (множествата на българските окончания), той е разработен въз основа на анализа на 700 000 български словоформи, които като обем представят съдържанието на няколко тълковни речника и могат да се образуват от 70-те хиляди единици на Обратния речник на българския език. Системата съдържа и съответните обслужващи програми за създаване, поддържане и обновяване на базовата лингвистична информация. Програмната система за автоматичен анализ на български текст реализира за първи път за български език алгоритъм за автоматично снемане омонимията на окончанията (автоматично причисляване на дадена словоформа към определен граматичен клас) през 1985. Системата, работеща в режим на диалог на български език, дава възможност за изследване вида на лингвистичната информация, която може да се извлече от текста в резултат на този вид автоматичен анализ (присъединяване на дадена словоформа към определен граматически клас). Системата реализира Национална конференция по информатика, посветена на 80 г. от рождението на професор Петър Бърнев автоматичен преход „текст →

речник“, включващ сегментиране на български текст; автоматично построяване на различни видове речници и речникови архиви: конкорданси на даден текст или даден автор с добавяне на натрупаната в процеса на обработка статистическа информация, честотни речници, комбинирани речници. По зададена от потребителя заявка, системата реализира търсене и автоматично извличане на словоформи и словосъчетания от даден български текст, с контекст или не, с цел изучаване на поведението на дадената лексическа единица, синтактична конструкция или определен граматичен клас в текста. Системата е разработена в две версии (зависещи от наличните по това време технологични платформи)