Езиковите технологии днес и утре

Галя Младенова Ангелова
Институт по паралелна обработка на информацията
Българска академия на науките

Компютърната лингвистика е сравнително млада дисциплина. Днес езиковите технологии са главната надежда на човека да овладее нарастващото количество текстова информация. Тази статия предлага кратък обзор на областта, с фокус върху разработките създадени в България, и очертава някои насоки за бъдеща работа.

1. УВОД

Компютърната лингвистика е сравнително млада дисциплина, но от гледна точка на информатиката е ветеран с над 50-годишна история. Първите й стъпки са свързани с машинния превод между английски и руски език в началото на студената война. Развитието на структурната лингвистика, от една страна, и по-специално на трансформационните граматики на Чомски, както и напредъкът на дискретната математика, от друга страна, и приложението на крайните и стекови автомати за анализ на изкуствени езици, дават мощен тласък на компютърната лингвистика, тъй като й предоставят някои основни инструменти за действие – а именно, лингвистичните теории за моделиране на строежа на изречението във вид на дърво и съответните автоматни подходи за анализ и обработка на тези структури. Ранните експерименти с компютрите от 70-те и 80-те години на миналия век очертават теоретичните постановки и принципните, неразрешими засега проблеми за обработка на семантиката на естествения език и структурата на кохерентния текст като съвкупност от взаимно-свързани клаузи. През 90-те години на преден план излизат статистическите подходи, които позволяват машинно самообучение от огромни корпуси и решават практически задачи. Днес т. нар. *езикови технологии* – софтуерни решения, ориентирани към крайния потребител - са главната надежда на човека да овладее нарастващото количество текстова и мултимедийна информация. В тази статия се спираме накратко върху принципите на

автоматичната обработка на естествения език, очертаваме възможностите и ограниченията и очакваните в близко бъдеще резултати. Ще използваме примери, достъпни в интернет, за да може читателят сам да проследи описаните идеи и експерименти.

2. ОБРАБОТКА НА ЕЗИКА ЧРЕЗ ПРАВИЛА

Когато въведем една фраза в компютъра, той по принцип 'разбира' единствено факта, че са му подадени низове от символи. Трябва да напишем специална програма, наречена *морфологичен анализатор*, която сравнява въведените низове с предварително подготвен речник и разпознава отделните *словоформи* на думите, споменати в конкретния входен текст. Създаването на морфологичен речник е трудна и трудоемка задача, но е задължителна стъпка при флективни езици с много словоформи за всяка дума [1]. Да разгледаме една демонстрация за българския език, инсталирана на интернет-страниците на проекта Балрик-Линг [2] и да стартираме морфологичния анализатор над демонстрационния текст 1:

> *Зад сините планини на изток, из едно море от светлина и слава, се показваше пламналото лице на майското слънце и събуденото зелено поле широко и весело се къпеше в лъчите му и празнуваше.*

Изведената таблица съдържа резултатите от анализа по словоформи, направен над речник от 10000 думи, които се срещат най-често във вестникарски корпуси от 1999 год. Ето началото на таблицата, която излиза пред нас на интернет-страниците [2]:

Дума (слово-форма)	**Лема (основна форма)**	**Характерис-тики на на лемата**	**Характерис-тики на формата**
зад	зад	предлог	
сините	син	прилаг.; кач.	мн.член.
планини	планина	същ.; ж.р.	мн.нечлен.
на	на	предлог	
изток	изток	същ.; м.р.	ед. нечлен.
из	из	предлог	
едно	-	-	-
море	море	същ.; ср.р.	ед. нечлен.
	море	частица	
....	....	....	....

В речника има 241 маркера за морфологични признаци, дадени в отделен панел на екрана [2]. В червено са оцветени липсващите в речника думи – *едно*, *майското*, *къпеше*. В зелено са показани многозначните словоформи:

- *море* е същ. ср. род, ед. ч-ло, нечлен. или частица,...
- *зелено* е форма на прилаг. *зелен*, в ср. род, ед. число, нечленувано или наречие;
- *широко* е форма на прилаг. *широк*, в ср. род, ед. число, нечленувано или наречие и т.н.

Разпознаването на входните низовете като единици от речника позволява обособяването на *думи*; тоест системата намира думи във входния текст с помощта на предварително зададен речник. Пак чрез речника се разпознават и характеристиките на отделните единици. Морфологичните признаци от речника позволяват по-нататъшен анализ на намерените думи. Тук правим важен извод: лингвистични ресурси (в случая речник) осигуряват набора от граматически категории, в чиито термини се прави анализът на всеки входен текст. След разпознаването на думите системата продължава с анализа. Правила от типа „*ако след словоформа X, която е форма на прилагателно или наречие в речника, следва съществително Y, и X е съгласувано с Y по род и число, то приеми X като прилагателно определящо Y*” позволяват разрешаването на многозначността на думите като части на речта. Чрез прилагане на това правило *широко* и *весело* ще бъдат определени като наречия, защото след тях не следва съществително в ср.р. Виждаме, че самото правило борави с лингвистичните категории от речника (предварително дефинирани в базовите ресурси). Стигаме до основния принцип на автоматичния анализ чрез правила: *целта на анализа е да се строят вътрешни представяния и структури от граматически категории, в чиито термини да се постигне пълно разпознаване на входните единици и да се реши поставената приложна задача.* Граматическите категории и връзките между тях се описват в специални лингвистични ресурси. Най-общо, теориите на компютърната лингвистика обясняват структурата и функционирането на естествения език чрез тези вътрешни представяния и предлагат алгоритми за тяхната обработка. Създаването на лингвистични ресурси също е важна задача, тъй като те кодират граматическо знание или съдържат представителни извадки от текстове.

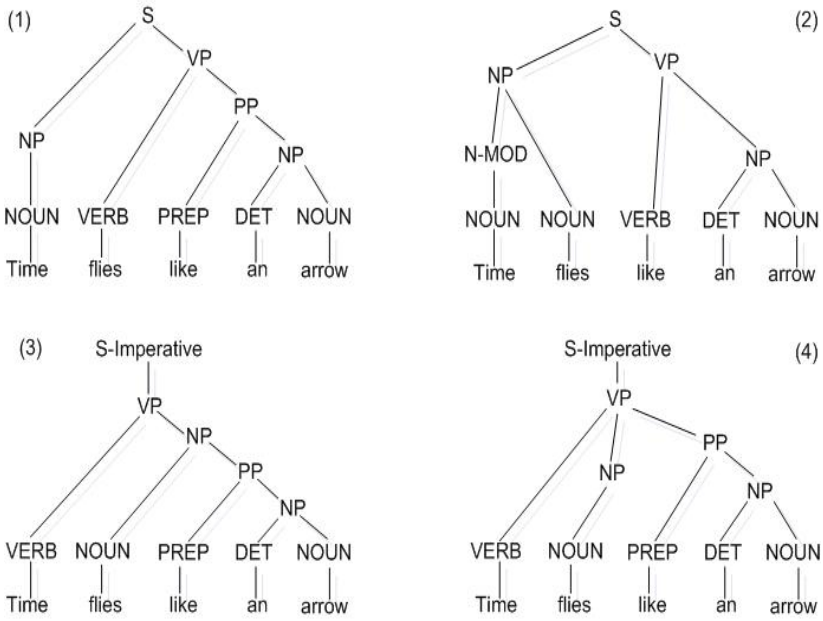
На фиг. 1 са показани класическите нива за обработка на естествения език. Тя очертава и основните компоненти, от които се състои една система за автоматично разбиране или генерация на естествения език. Както вече казахме, като начало се започва с **морфологията** и се разпознават думите в текста (или фонемите в речта). Следва един много важен етап – автоматичният **синтактичен анализ**, който съпоставя на всяко изречение вътрешна структура във вид на дърво от категории. Формалният апарат на синтактичния анализ се основава върху безконтекстните пораждащи граматики на Чомски, които позволяват анализ в реално време поради сравнително простия алгоритъм за проверка дали дадено изречение е граматически правилно (тоест изводимо от правилата на граматиката)[1]. При морфологичния анализ вече видяхме, че резултатът не е единствен. По принцип изходът от синтактичния анализ е още по-многозначен, не само поради самата структурна многозначност в езиковите конструкции, но и поради съчетаването на няколко стотици синтактични признаци и техните стойности при извършване на анализа. Тъй като скоростта на анализа зависи от броя на правилата на безконтекстната граматика, прието е последните да се структурират в разумно и обозримо множество от конструкции. На фиг. 2 показваме един от най-известните примери за синтактична многозначност – анализите (дървета на извода) на изречението *Time flies like an arrow* според Чомски. Чрез този пример ще обясним идеята на Чомски за фразовата структура и композицията на синтактичните конституенти на по-горно ниво от по-долните такива.

Модулът за синтактичен анализ произвежда автоматично синтактични структури като показаните на фиг. 2, за всяко зададено входно изречение, с помощта на безконтекстните правила за извод на формалната граматика. Правилата дефинират коректните синтактични фрази и начините за тяхното композиране в по-общи фрази, които съставят изречението *S*. Някои фрази представят главните конституенти на изречението: *NP* (Noun Phrase), *VP* (Verb Phrase) и *PP* (Prepositional Phrase). Други фрази съдържат по една дума: *VERB*, *NOUN*, *PREP*, и *DETerminer* и на практика са признаци, дошли от морфологичния речник.

---

[1] Поради краткостта на изложението не споменаваме другия често използван синтактичен модел, т. нар. граматика на зависимостите.

Фигура 1. Нива на обработка на естествения език при подходи, използващи правила



Фигура 2. Синтактични структури на изречението '*Time flies like an arrow*'

Различните синтактични структури, показани на фиг. 2, съответстват на различни значения на изречението:

(1) Времето (*същ.*) лети (*глагол*) като (*предлог*) стрела (*същ.*);

(2) Времеви мухи (някакъв специален вид мухи, чието название се състои от две съществителни и първото определя второто) харесват (*глагол*) една стрела (*същ.*);

(3) Заповедна структура: измервай (*глагол*) мухи (*същ.*) които изглеждат като (*предлог*) стрела (*същ.*);

(4) Заповедна структура с две значения, при които предложната фраза *'like an arrow'* модифицира глагола *time*: *Измервай мухи както* (би измервал) *една стрела* и *Измервай мухи както една стрела* (би ги измервала).

Дървото от синтактични конституенти е практически привлекателен модел, тъй като структурата се строи композиционно от отделни 'тухлички'. Например, и в четирите дървета на фиг. 2 се използва правилото:

NounPhrase се състои от *DETerminer* последван от *NOUN* (5)

Така навсякъде е получен анализ, че '*an arrow*' е *NP*. В дървета (1), (3) и (4) виждаме, че

PrepPhrase се състои от *PREPosition* и *NounPhrase* (6)

Тогава '*like an arrow*' е *PP* в три от структурите. При разказните изречения (1) и (2) цялата структура *S* е състои от *NounPhrase* последвана от *VerbPhrase.* И така става възможно да формализираме граматиката на естествения език на фрагменти (т.е. на конституенти), а модулярността на описанието е извънредно полезна при сложното и обемно синтактично знание. Сглобяването на пъзела от тухлички става на етапа на анализа, когато на входа на синтактичния анализатор се подаде конкретно изречение. Програмата преглежда всички възможни правила и се опитва да построи единно дърво за цялото изречение, като се опитва да намести в него подходящите правила. Така например, може да се изпробва правилото (6) при анализа на изречението на фиг. 2 при положение, че първите две думи '*time flies*' вече са определени за *NP* според някое друго правило. Но ако '*like*' е предлог, както се предполага в (6), става невъзможно да се построи цялата конструкция на разказното изречние *S* при положение, че '*time flies*' е *NP*. Тогава анализаторът опитва друга възможност - '*like*' да заеме ролята на глагол, а не на предлог. Наместването на правила успява, поради наличието на правило

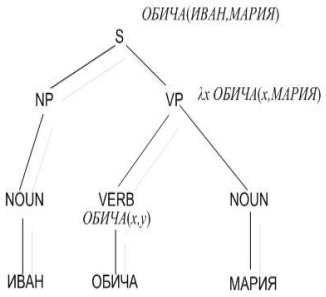
VerbPhrase се състои от *VERB* следван от *NounPhrase* (7)

и така получаваме дървото (2), като алтернативен анализ

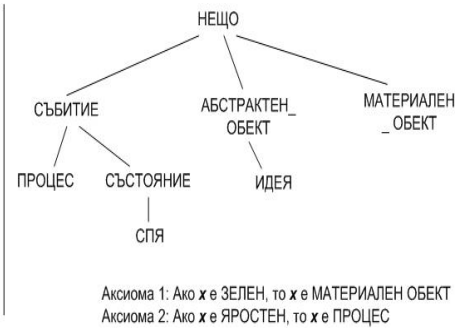
на (1) за разказни изречения. Изпробването на всички възможни правила на граматиката, включително тези описващи структурата на заповедните изречения, ще произведе четирите дървета на фиг. 2. Те са констуирани до голяма степен от едни и същи елементи, които са подредени по всички възможни успешни начини; нека посочим и конституентата (7) в дърво (3), която се среща също и в дърво (2). Обикновено формалните граматики имат стотици правила, които описват в детайли начините за композиране на конституенти в простите изречения и на прости изречения в сложни такива.

Друга привлекателна страна на композиционния подход е, че при съставянето на синтактичните конституенти може да се формира по унифициран начин и една предикатно-аргументна структура, която на етапа на семантичния анализ става *'логическа форма'* на изречението. Тук ще обсъдим композицията на логически изрази в най-простите случаи на съществително, прилагателно и глагол.

Всяко прилагателно или съществително се превръща в едноместен предикат, наименован със самата дума. Например, ако в едно изречение се срещне думата *ЧОВЕК*, програмата за композиране на логическа форма ще произведе $\exists x$ *ЧОВЕК*(x) и ще разглежда получения логически израз като семантика на думата *ЧОВЕК*. При срещане на *УМЕН ЧОВЕК*, с просто правило за конюнкция, се получава $\exists x$ *ЧОВЕК*(x) & *УМЕН*(x). В този случай логическата форма на фразата се получава чрез конюнкцията на едноместни предикати. Глаголите обаче се превръщат в n-местни предикати, като n е броят на задължителните за запълване семантични валенции на глагола. Например, семантиката на глагола *обичам* в изречението *Иван обича Мария* се изразява чрез двуместния предикат *ОБИЧА*(x,y), където x е агентът, а y - обектът. Констуирането на логическа форма на това изречение е илюстрирано на фиг. 3 чрез дървото на синтактичния анализ. Двуместният предикат *ОБИЧА*(x,y) е частично запълнен във върха VP, понеже изречението е анализирано само отчасти, и се е превърнал в λ-израз, който 'чака' да бъде вложен в по-висока синтактична конституента с интегриране на подходящ втори аргумент. Така логическата форма се композира чрез предварително зададени правила за конструиране на логически изрази за отделните конституенти и правила за транформациите им.



Фиг. 3. Композиране на логическа форма



Фиг. 4. Затворен свят с аксиоми, дефиниращи композиции от думи

Преминаването от думи към (вътрешни за системата) логически форми позволява да се контролира логическата коректност на входните изречения. Например, изречението *Зелените идеи яростно спят*[2] е невалидна комбинация от думи в света, показан на фиг. 4. В него само материалните обекти имат цвят – а 'идея' е абстрактен обект, и 'яростни' могат да бъдат само процесите – а 'спя' е състояние. Така, следвайки нивата от фиг. 1, преминаваме автоматично от неструктурирания текст към множествата от логически форми (по една за всяко изречение), които се обработват в контекста на знанието за света с техниките на изкуствения интелект. Веднага трябва да кажем обаче, че преходът към абстрактни вътрешни структури е едно съмнително предимство, тъй като няма големи, публично-достъпни концептуални ресурси от декларативно знание във вида илюстриран на фиг. 4, с чиято помощ да се извършват умозаключения. Самите техники за извод имат своите ограничения при много големи масиви от клаузи и т.н., но всъщност тези две (принципни) технологични затруднения не са главното препятствие пред автоматичния анализ на семантиката на естествения език.

Същинският проблем е, че на 'по-дълбоките' езикови нива се сблъскваме с лингвистични явления, които не са добре изучени и няма психолингвистични и когнитивни теории за тяхното обяснение и моделиране. Тук ще спомена накратко най-важните от тях. Такова явление е прагматиката, която изучава значенията и тяхното

[2] Друг много известен пример на Чомски, за безсмислено изречение с коректна синтактична структура.

функциониране в зависимост от контекста. Нека разгледаме три прости изречения, които лесно могат да се анализират морфологично и синтактично и да се превърнат автоматично в логически форми:

Един студент трябва да учи много.
Студентът трябва да учи много.
Студентите трябва да учат много.

И трите изречения в даден контекст могат да се отнасят към конкретни обекти от света на дискурса. В друг контекст обаче, при т. нар. *обобщено четене* (*generic reading*), и трите могат да се разглеждат като обобщено твърдение, вярно за всички студенти изобщо – бивши, настоящи и бъдещи, независимо от факта, че две от изреченията са в единствено число, а две са членувани. Няма теория за компютърно моделиране на контекста по начин, който да осигури алгоритмичното разпознаване на подобни прагматични отсенки в смисъла на изреченията. С други думи, техниките за умозаключения не разрешават прагматичната многозначност, дори и да разполагахме с формални бази знания от милиони факти за света.

Сериозен проблем е обхватът на значенията на думите, който се преплита по сложен начин със семантиката на явните или неявни квантори, отрицанието, с темпоралните наречия и т.н. По принцип обхватът се изразява чрез скоби в логическата форма. Ето някои примери:

- *Портиерът беше любезен във всеки хотел*. В това изречение се реферира към много портиери, по един във всеки хотел, но самата дума *портиер* е изказана в единствено число. Системата трябва да изчисли от знанието за света, че всеки хотел си има портиер и тогава кванторът за съществуване на променливата, свързана с думата *портиер*, е в обхвата на квантора за всеобщност на променливата, свързана с думата *хотел*;
- *Утре той ще нахрани най-гладното куче*. Тук се въвеждат две най-гладни кучета – едното днес, в момента на говоренето, а другото утре. Само контекстът може укаже за кое от тях става дума, в случай че са различни;
- *Едно гладно куче винаги чака на вратата*. Едно определено гладно куче винаги чака или всеки път чакащото куче е гладно;
- *Всеки мислеше, че България или Румъния ще влязат в ЕС*. Някои са мислели, че България ще влезе, а други – че Румъния ще влезе; или всички са мислели, че една от двете страни ще влезе в Европейския съюз;

- *Ние не видяхме всички деца*. Не сме видели нито едно дете или има поне едно дете, което сме видели.

Горните примери показват, че автоматичното композиране на логически форми може да произведе само предикатите, свързани с конкретните думи от изречението, но остава да се свърши много работа по прецизно наместване на скобите и разрешаване на обхвата на значенията. На практика се генерират всички възможни форми, измежду които се избира една в зависимост от целите и знанията на конкретната система. Този етап е онагледен с трион на фиг. 1, тъй като има избор измежду много възможности.

Алгоритмично-неразрешим проблем е и автоматичната обработка на референията. По принцип естественият език функционира като последователност от линейно-наредени клаузи, тъй като не можем да изкажем всичко наведнъж. Следователно, една система за разбиране на естествения език трябва да може да разпознава различните референции към един и същи обект, които обикновено се изказват по различен начин с цел избягване на повторенията и добавяне на нова информация. Например:

МВР залови хакерите, сринали сайта за детските градини на 4/02/2008. Те са И.П. и Б.Б. на 3 години от град С., които не искат да ходят на детска градина.

Първото изречение въвежда два важни обекта на дискурса: *МВР* и *хакерите, сринали ....* Второто изречение се фокусира върху един от обектите: *хакерите,* които са споменати повторно с *те* (местоименна анафора), и въвежда имената, възрастта и местожителството им. В третата клауза, реализирана като подчинено изречение, се добавя една важна характеристика на обектите: *не искат да ходят на детска градина.* Така във всички клаузи става дума за едни и същи обекти, но поднесената информация е изказана линейно чрез различни думи. Това явление се нарича референция и най-често се реализира за дискурсни обекти, към които се реферира с групи на съществителните (*NPs*). Местоименията са явни указатели за референция. Всички местоимения и съществителни са потенциални кандидати за свързване в поредица от кореферирващи описания.

Обикновено референцията се наблюдава в локалния контекст от няколко съседни изречения. На теория, системата трябва да разпознае всички референции, за да се

постигне разбиране на естествения език от компютъра. На практика обаче това е невъзможно и се работи само за местоименната анафора, чиято обработка е задължителна при машинния превод (при превод на местоименията в единствено число често е необходимо да се смени рода – например, английското *it* за неодушевен обект може да се преведе на български като *той*, *тя*, *то* и др.). Най-добрите алгоритми за английски език работят с успех 75%. Някои местоименни референции са извънредно сложни:

Иван и Мария намериха свещеник и се ожениха. За медения месец те заминаха на Хавай.

Използваните днес алгоритми за разрешаване на анафората биха решили, че на Хавай са заминали *Иван*, *Мария* и *свещеника* (или *свещеника* и *Мария*, понеже са последните два обекта, към които може да се реферира с 'те'). Само със сложни концептуални анализи може да се заключи, че 'те' са *Иван* и *Мария*, но такава обработка е невъзможна практически, още повече че дискурсът е плетеница от неразрешими явления като изброените по-горе. Нека отбележим, че човекът няма никакви проблеми при разбиране и генериране на естествен език, че прави контекстна интерпретация без съзнателни усилия и т.н. Фактически човекът е най-добър на нивата, с които компютърът изобщо не се справя. Но повечето хора не си дават сметка за (несъзнателно) използваните от тях морфологични и синтактични категории, което пък е компютърният подход за анализ на думите и изреченията.

В края на този обзор, който следва изложението в [4], нека скицираме идеята за сегментите като градивна единица на дискурса. Приема се, че кохерентният текст се декомпозира на йерархично вложени сегменти, които представляват фрагменти свързани с една и съща тема [5]. Няма формална дефиниция на понятието сегмент, но има една интуитивна идея, че локално-свързаните изречения се групират по естествен начин. Да разгледаме текста (8):

8.1. Има много начини да разпознаете листо от сребрист клен.
8.2. Първо, то има сребриста лъскавина отдолу.
8.3. Ако го държите в ръка и го движите,
8.4. ще видите как слънцето се отразява от долната му страна.
8.5. Второ, по него има дълбоки резки между върховете.
8.6. Формата му е съвсем подобна на листото от червен клен.
8.7. И трето, ако го счупите в основата, сокът ще бъде млечен.
8.8. Счупете основата му, почакайте 20 секунди и ще видите сока.

Той се състои от един главен и три подчинени сегмента, които са ясно маркирани с вметнатите думи *първо*, *второ*, *трето*. Референция с местоимения се прави само от вложен сегмент към обект в главния и никога към обект от друг вложен сегмент на същото ниво. Например, 'него' и 'му' в сегмент 3 не реферират към 'слънцето' в сегмент 2; 'го' в изречение 8.7 се отнася към 'листото' от изречение 8.1, а не към това от изречение 8.6. Автоматично разпознаване на сегментите е възможно само при ясно видими вметнати изрази както по-горе и при рязка смяна на темата или на глаголното време (които сигнализират смяна на контекста за интерпретация). Според компютърната лингвистика, именно организацията в сегменти осигурява функционирането на референцията (т.е. тя става разбираема за слушателя и читателя).

Разглеждайки нивата на фиг. 1, с оглед изброените по-горе принципни трудности, читателят с основание би се запитал: *Какво изобщо работи в практическите софтуерни системи?* Отговорът е:

- морфологичен и синтактичен анализ над произволни текстове, но почти без семантичен анализ – например при машинния превод, използващ правила; или
- морфологичен, синтактичен и по-дълбок семантичен анализ, но над ограничен брой думи.

Да разгледаме системата за англо-български машинен превод BulTra [6]. Благодарение на много големите си речници, системата поема произволни текстове. Обикновено при машинния превод синтактичната многозначност не се разрешава, а се прехвърля на другия език (тъй като човекът-читател ще се справи с интерпретацията). Ето един пример:

John sees the girl with the telescope се превежда като
Джон гледа девойката с телескопа.

В други изречения, обаче, неразрешената многозначност води до грешен превод:

All rooms have ocean and garden view.
Всички стаи имат океан и градински изглед.

Тази грешка при обработка на английската група на съществителното ((*ocean and garden*) *view*) е типична. Големите системи[3] с прецизни синтактични анализатори разбират, че има два анализа, но приемат за верен по-

[3] Например Systran, www.systransoft.com

честия, понеже няма как да изберат другия без интерпретация в контекста на знанието за света. Референцията също е проблем:

John takes the cup from the table. It was repaired by Jane.
Джон взема чашката от масата. Това бе поправено от Джейн.

По принцип, машинният превод анализира входния текст на морфологично и синтактично ниво (което е достатъчно сложно), като 'прехвърля' получените вътрешни структури на изходния език. Работи се изречение по изречение. Референцията се разрешава само когато във входния текст се срещне местоимение, което трябва да бъде преведено на другия език. За жалост системата BulTra почти не се развива поради липса на достатъчно финансиране. Системи като цитираната вече Systran се изработват за стотици човекогодини и се поддържат непрекъснато, поне с цел обновяване на речниците. Днес машинният превод се прилага предимно за административни и специализирани текстове. Счита се, че той е една от 10-те информационни технологии, които биха променили света.

Като пример за технология, която слиза до по-дълбоко семантично ниво, можем да разгледаме т.нар. *Information Extraction*[4] [7]. Този вид системи са настроени да търсят по едно събитие в големи корпуси, например описание на терористични актове в полицейски доклади. Системите разпознават наименованите единици в текста, тъй като имената на лица, географски обекти, фирми и т.н. са важни указатели за описания да случки от даден вид. Постига се точност над 95% за английския език. След това се вземат само *'интересните думи'*, които сигнализират търсеното събитие, и се работи само за/около тях. Автоматично се разпознават половината кореференции, а точността на разрешените е около 70%. След това се конструира т.нар. сценарий за намереното събитие (template element production). Разпознават се 70-80% от текстовите фрагменти, в които се говори за търсеното събитие. Хората постигат точност 93%. Автоматичното запълване на сценария, еквивалентно на семантичен анализ, се извършва с точност 49-56%. Хората извършват тази задача с точност 81% (вж. повече детайли в [7]).

---

[4] На английски има два термина, които често се смесват при превод: *Information Retrieval*, т.е. изваждане на цели документи от архив или интернет, нещо от рода на *търсене на документи*; и *Information Extraction*, с превод *извличане на информация*.

3. ОБРАБОТКА ЧРЕЗ СТАТИСТИЧЕСКИ МЕТОДИ

Статистическата обработка на естествения език се прилага отдавна, но успехите й се доказват на практика през 90-те години на 20-ти век. Днес методите, използващи правила, все по-често се интегрират със статистическите подходи на различни езикови нива. За краткост ние ги разглеждаме отделно в чист вид и ще споменем само една от най-модерните технологии.

Статистическият машинен превод изобщо не различава езиковите нива, показани на фиг. 1. Той се самообучава директно по огромни корпуси от *паралелни текстове* – т.е. текстове и техните преводи на друг език. За всяка дума или фраза от единия език се натрупват наблюдения как тя се превежда на другия език (с някаква вероятност). На фразите се присвояват съответни условни вероятности и лексикални тежести, пресметнати по всички изречения. Получените данни се прилагат над непознат текст на входния език. Всяка дума или фраза от него се заместват с най-вероятните им преводи с отчитане на контекста (това елементарно обяснение не отразява истинската сложност на процеса). Главното предимство на статистическия машинен превод е това, че се учи от живия текст и от начина, по който превежда преводачът. В момента един от най-добрите прототипи в света е разработен от Преслав Наков за двойката английски-испански език [8]. Тук даваме пример как се държи тази технология за английско-български превод, като обучението е извършено над паралелен корпус от 111 документа с административен текст (45883 двойки изречения с 1017703 думи на английски и 907271 думи на български)[5]. Сериозното обучение при индустриални системи се извършва над 50-100 пъти по-голям корпус.

Английска фраза	*Превод на български език*
in combination	в комбинация, в съчетание
in combining	при съчетание, при съчетание на
in charge of	отговарят за, отговарящ за
both physical and psychological	както физическа, така и психическа
as defined	както те са определени

[5] Авторът благодари на Преслав Наков за предоставените примери.

Превод от човек	*Статистически машинен превод*
Европейска конвенция за взаимопомощ по наказателно-правни въпроси Преамбюл Правителствата, подписали тази конвенция, в качеството си на членове на Съвета на Европа, считайки, че целта на Съвета на Европа е да се постигне по-голямо единство между неговите членове, убедени, че приемането на общи правила в областта на правната помощ по наказателни дела ще допринесе за постигането на тази цел, считайки, че правната помощ е свързана с въпроса за екстрадицията, която вече бе предмет на конвенцията, подписана на 13 декември 1957 година, се споразумяха за следното ..	европейска конвенция за взаимопомощ по наказателно-правни въпроси преамбюл правителствата, подписали този протокол, членове на съвета на европа, считайки, че целта на съвета на европа е постигнето на по-голямо единство между своите членове, убедени, че приемане на общи правила в областта на правна помощ по наказателни дела ще допринесе за постигането на тази цел, считайки, че тази взаимна помощ е свързана с въпроса за екстрадиция, който вече е образувано предмет на конвенция, подписана в 13th декември 1957 година, се споразумяха за следното …

4. СЪВРЕМЕННИ РАЗРАБОТКИ В БЪЛГАРИЯ

Основите на компютърната лингвистика в България са поставени през 1964 год. със създаването на Групата по машинен превод на проф. Александър Людсканов в Института по Математика на БАН. Главната й задача е разработка на машинен превод между руски и български език. По този начин математическата колегия оказва навременна и решителна подкрепа на една нова интердисциплинарна област. През 80-те години на миналия век се създават 3-4 морфологични анализатора, построени с помощта на флективните класове от книгата [9]. Разработват се правописни коректори, базирани върху честотните изследвания на проф. М. Янакиев. Към днешна дата българската компютърна лингвистика е изненадващо продуктивна и десетки научни групи и фирми създават езикови технологии и ресурси за обработка на българския език.

Най-активните научно-изследователски групи в областта се намират в БАН (в ИПОИ, ИМИ, ИБЕ и ИИТ), в Пловдивския университет и в Софийския университет. Напоследък се оформиха научно-изследователски групи в Нов български университет и в Търновския университет. В тази област работят лингвисти, информатици и логици. Активни индустриално-ориентирани частни организации са АПИС, ПроЛангс (разработчиците на БулТра), ОнтоТекст Лаб в Сирма, BMG (от ACT Soft), Сиела, dir.bg, netinfo, както и Българската асоциация по компютърна лингвистика БАКЛ [6]. Създадени са 5-10 много големи морфологични речници на българския език и съответни анализатори към тях, на пазара има няколко програми за корекции на правописни грешки, съществуват поне три прототипни разработки на синтактични анализатори на български изречения, две системи за машинен превод и непрекъснато се правят опити за подобряване на търсенето в архиви от документи на български език. БАКЛ предлага синтезатор на българска реч по зададен входен текст, като продукт ориентиран към граждани с нарушено зрение. Налице е и впечатляващо количество лингвистични ресурси от различен вид, разработени главно в академичните среди.

Секцията за лингвистично моделиране на ИПОИ-БАН разполага със следните ресурси: *(i)* текстов архив от 72 млн. словоформи, *(ii)* няколко вида текстови корпуси с анотация на различно ниво, *(iii)* банка от ръчно създадени синтактични дървета за 15000 български изречения, една от петте най-големи в света, *(iv)* многобройни лексически ресурси за българския и руски език, както и двуезични лексикони, *(v)* 10 млн. словоформи многоезични подравнени паралелни корпуси за междуезикова обработка - български, английски и други славянски езици, *(vi)* частична формална граматика на българския език, създадена във връзка с банката синтактични дървета, а също и *(vii)* ресурс за тестване на системи въпрос-отговор в българо-английски вариант. Като софтуерни средства се използват следните среди за обработка на лингвистични ресурси или неанотирани текстови данни: *(i)* системи за обработка на български език чрез морфологичен и синтактичен анализ и един прототип за въпрос-отговор;

[6] Тук не можем да изброим фирмите разработващи ′търсачки′ за износ в чужбина (предимно на английски език); изброяваме само най-активните разработчици на технологии за българския език.

(ii) среда КЛАРК за създаване и обработка на текстови корпуси; *(iii)* система за намиране на преводни съответствия в двуезични паралелни корпуси и *(iv)* различни софтуерни среди за създаване на базисните речници и граматики на езиковите технологии.

Повечето софтуерни системи у нас, както и голяма част от натрупаните лингвистичните ресурси, се отнасят към нивата на морфологията и синтаксиса на фиг.1. По-долу е описан накратко един прототипен генератор на обяснения на немски и български език, създаден под ръководството на автора в международния проект DB-MAT, финансиран от Фондация "Фолксваген" – Германия. Генерацията не е обсъждана досега, тя организира текста в поредица от свързани клаузи, така че пак се връщаме към фиг. 1.

5. ГЕНЕРАЦИЯ НА МНОГОЕЗИКОВИ ОБЯСНЕНИЯ

Много софтуерни системи ′генерират′ документи чрез съчленяване на фиксирани текстови низове и запълване на шаблони, но тук не става дума за това. От гледна точка на компютърната лингвистика за генерация се говори тогава, когато една система произвежда текст от динамично подаден неезиков вход с определена семантика, чрез обработка на връзки между текстови единици съгласно някаква лингвистична теория за строежа на дискурса и компютърен модел на тази теория. Главното предизвикателство е генерацията на кохерентен дискурс. Тази задача не е решена и до днес по удовлетворителен начин, а и няма психолингвистични теории за това, как човекът генерира естествен език: как подбира релевантната за предаване информация, как планира подредбата на изреченията едно след друго, как изгражда референцията между свързаните изречения и т.н.

Когато системата разполага с формално знание във вид на вътрешни клаузи, тя може да генерира обяснения за него след като реши поне следните проблеми:

- Кои факти следва да се разкажат в конкретния случай. Това зависи и от слушателя, тъй като не всички факти са еднакво интересни за слушателите;
- С кои думи да се изкаже избраното знание. Системата разполага с речник, който съпоставя думи и фрази на концептуалните структури от клаузите. Освен

това генераторите обикновено имат граматически шаблони за изказване на всеки вид клауза. И въпреки това, при строежа на конкретните изречения, са необходими много допълнителни лингвистични елементи: членуване, генериране на референции (най-често местоименни анафори), вметнати думи за подсилване или блокиране на интерпретации на дискурса и т.н.;

- Колко дълга да бъде атомарната клауза на естествен език, тъй като няма директно съответствие между грануларността на фактите в клаузите и изреченията, които ги изказват; и
- Как да се наредят генерираните изречения в параграфи, за да се получи разбираем текст с добър стил, тоест за да се генерира ′хубав′ дискурс.

Концептуалното знание на системата, представено като множество от клаузи, само по себе си не съдържа никакви указания как да бъде разказано в свързан текст. При всяка заявка на потребителя системата решава отново четирите задачи, изброени по-горе. Генерацията на текст предполага наредба на клаузите, която съответства на целите и намеренията на говорителя (т.е. системата) и на модела на слушателя, поддържан от системата. Поради това процесът на генерация стартира с изграждане на план от дискурсни релации, които са нещо като скелет зад генерирания текст, и това е т.нар. *план на текста*. Този план гарантира, че наредбата на клаузите ще предизвика търсения ефект у слушателя, и така ще се изключат нежеланите, погрешни интерпретации. Но тъй като системата не може да знае всичко за слушателя и неговото знание, планът на текста има още една важна функция: той позволява в генерирания дискурс да се включат вметнати изрази и други дискурсни маркери, които да подсилят желания смисъл чрез по-явно вербализиране на дискурсните релации. На практика генерацията работи по обратен начин на процедурите за разбиране на естественя език, както е показано на фи. 1. Преди да минем към конкретни примери, нека цитираме една известна дефиниция [10]: *Един дискурс е кохерентен, ако слушателят разбира комуникативната роля на всеки негов фрагмент, тоест, ако слушателят разбира как говорителят възнамерява да свърже отделните клаузи с всяка друга клауза на дискурса.* Дискурсът не е механична композиция от изречения, тъй като той носи повече информация, отколкото сумата на изграждащите го клаузи.

Наблюдения върху риторичната структура на текстовете ни помагат да построим планове за структурата на някои специфични описания. Съществуват дискурсни релации, които в определени видове текст играят ролята на ′рецепти′ за организация на изреченията. Да разгледаме следните текстове[7]:

Братовчед на коня, зебрата е единственият голям бозайник с козина на ивици от бяло и черно. Има буйна грива и добре окосмена опашка. Тя е най-красива от сродниците си. Дължината на тялото й е 2 метра и 20 см (плюс 75 см опашка). Височината при плешките 1,25 до 1,35 см. Тегло 225 до 420 кг. Продължителността на живот е от 20 до 40 години (в плен).

Щраусът е най-голямата птица в света. Той не може да лети, но има дълги и силни крака. Има малка глава, издължена гола шия и огромни очи. На височина достига 2,75 метра, което му дава възможност да вижда хищниците отдалече.

И двата текста са произведени по една и съща дискурсна схема, която може да се дефинира предварително и да се използва за всяко животно в тази енциклопедия:

- *Въведи името на обекта, неговия клас и най-важните характеристики на обекта и класа;*
- *Опиши дължината на обекта;*
- *Опиши височината на обекта;*
- *Опиши теглото на обекта;*
- *Опиши продължителността на живота на обекта;*
- *Опиши скоростта на движение на обекта,*
- *Опиши други характеристики: зона на разпространение, цвят на козината* и т.н.

Този неформален дискурсен план не ограничава броя на клаузите, а просто показва в какъв порядък да се подредят отделните теми при описанието на всяко животно. Ако няма данни за запълване на някой дискурсен предикат, той се пропуска. При повече налични данни се генерират няколко последователни клаузи. Така при описанието на зебрата като най-важни характеристики са включени три клаузи – за козината, гривата и красотата, а за щрауса са споменати други белези. Горните описания са написани от човек, т.е. и хората следват подобни рецепти.

---

[7] Детска енциклопедия http://www.worldstory.net/bg/, посетено на 11 февруари 2008.

Технологията на схемите като 'замразени дискурсни рецепти' е предложена през 1985 година в [11] и днес е единственият начин за генериране на обяснения, който се прилага в практически системи. Невъзможността за създаване на динамични дискурсни планове, подобни на използваните от човека, е характеризирана със забавното изказване на Йорик Уилкс „*Автоматичното разбиране на естествения език напомня броене от едно към безкрайност, а генерацията – броене от безкрайност към едно*”[8]. По-долу разказваме накратко за едно приложение на този подход при генерация на технически обяснения за пречистване на отпадъчни води.

Генераторът EGEN е изследователска разработка, интегрирана в един прототипен софтуерен модул - работно място за подпомагане на преводача, който превежда технически текстове между български и немски език. При избор на термин от документа за превод, потребителят-преводач може да поиска от системата пояснение на значението му в предметната област (освен справки за лингвистичните характеристики на термина като езикова единица). Това пояснение се планира и вербализира динамично от системата и се реализира чрез генерация. Целта на разработката е да се изследват възможностите за създаване на многоезиков генератор, който е максимално независим от предметната област и лесно може да се прехвърли към друга база знание, стига връзките между лексикалните и концептуални ресурси на системата да са зададени по определен начин. Паралелната генерация на два езика позволява изследване на структурни въпроси като например грануларността на термините в концептуален план, при очевидната разлика между грануларността на многоезичните терминологични единици на български и немски език. Друг въпрос за изследване е свързан със спецификата на класическия подход за подбор на знанието чрез схемите от риторични предикати. При подхода на схемите ‘*знанието, релевантно на дадена тема’* се свежда до ‘*знание, позволяващо на системата за генерация да запълни схемата, която фиксира плана как да се говори по темата*’. С други думи, схемата подбира онова знание, което самата тя може да разкаже и така дискурсният план

[8] Цитирано в Jurafski, D. and J. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

управлява подбора на фактите. За разлика от този подход, EGEN използва техники на изкуствения интелект, които извличат фрагменти релевантно знание независимо от схемата за построяване на текстовото обяснение. Извлеченият екстракт факти се подава към дискурсната схема за запълване на плана на текста и генериране на обяснения. По този начин концептуалният ресурс се обработва независимо от дискурсния план, което осигурява по-слаба обвързаност между фазите на подбор на знанието и построяване на плана на текста. Използват се взаимно-независими концептуални и лексикални ресурси за генерация, а това като цяло облекчава и генерацията в многоезиков план поради възможността по-лесно да се добави нов език.

Генераторът EGEN работи над база знание от концептуални графи. Извличането на факти става чрез операцията *проекция*, която търси концептуални шаблони в базата. Знанието е по-активно използван ресурс в EGEN, в сравнение с класическия подход за генерация чрез схеми; например агрегацията на факти на етапа на микропланирането става чрез операцията *съединение* на концептуални графи. Така механизмът за подбор на знанието е независим както от риторичните цели на системата, така и от предметната област. Повече детайли за прототипа EGEN могат да се намерят в [12] и [13]. Тук даваме само някои примери на генерирани обяснения:

> Ölphasen (Ölpartikel[1]) gehören zu Partikeln[2]. Die[3] Ölphasen sind gekennzeichnet durch Dichte[4]. Die ausschwimmenden[5] und grobdispersen[6] Ölphasen, welche leichter als Wasser sind[7], sind enthalten in Abwasser[8].
>
> Маслените[9] частици са частици. Маслените частици се характеризират с плътност. Маслени частици[10], които се съдържат в отпадъчна вода[10], са изплуващи, грубодиспергирани и по-леки от водата.

В този пример 1 е синоним от лексикона; 2 е надтип от концептуалната йерархия; 3 е определителен член, поставен поради предишното споменаване на обекта Ölphasen; 4 е характеристиката *плътност*; 5 и 6 са характеристиките *изплуващ* и *грубодиспергиран* в съответното съгласуване; 7 е реализирано като подчинено

изречение, тъй като в лексикона не е намерено единично прилагателно за изказването му; `8` е повърхнинна реализация в страдателен залог, тъй като фактът *‘отпадъчната вода съдържа ....”* се вербализира в изречение, където `Ölphasen` са заели ролята на подлог и поради това граматиката предлага само възможност за изказване на конкретния факт в страдателен залог; `9` е членуване на български, което не се среща в немския текст; `10` е пример за неудачно членуване на български (генерирането на членувани форми се оказа нетривиално). По принцип, за да бъде подготвен за вербализация на различни изречения, EGEN разполага с различни шаблони за изказване на прости клаузи. Например генераторът може да произведе

Маслените частици се характеризират с плътност или

Плътността е характеристика на маслените частици

а също и

Сепараторът има отделителна камера

или (при необходимост)

Отделителната камера се съдържа в сепаратора.

По-подробен коментар на проблемите на генерацията и съответните теоретични постановки е даден на български език в [14].

6. ЗАКЛЮЧЕНИЕ

В този текст поднасяме на читателя някои основни факти за подходите на компютърната лингвистика, като опит да обясним сегашното състояние на езиковите технологии. Авторът счита, че принципните затруднения на дисциплината не са широко известни и поради това страничните наблюдатели (както и потребителите) не си дават ясна сметка за ограниченията на методите за алгоритмична обработка на езика. Досега компютърната лингвистика е създала успешни модели само за определен кръг добре изучени лингвистични явления. Всъщност не е сигурно, че днешните компютри – като машини на Тюринг – в обозримо време ще станат по-добри в автоматичната обработка на езика, както не е сигурно и че самите лингвисти, включително когнитивните лингвисти и психолингвистите, ще създадат в близко бъдеще по-всеобхватни теории за функционирането на естествения език при човека. Но от гледна точка на практическите задачи е необходимо да се очертае кръг от софтуерни

приложения, в които скицираните по-горе методи осигуряват полезни информационни услуги за хората.

Първото радикално подобрение със сигурност ще настъпи поради развитието на т.нар. корпусна лингвистика, която изучава езиковите явления в огромни корпуси. Лесно се вижда потенциала на машинния превод, построен върху статистически методи и немаркирани корпуси. Днес активно се създават и маркирани корпуси, например наличната за български език банка от 15000 синтактични дървета, върху която могат да се обучават синтактични анализатори, извличащи автоматично формална граматика на българския език. Същевременно е ясно, че статистическите методи наблюдават това, което се вижда, така че от тях не се очаква да откриват дискурсни релации, например, или да извличат прагматични модели. Несъмнено е обаче, че редица софтуерни системи ще станат по-добри; днес около 2% от софтуера в света интегрира езикови технологии. Тъй като и обработката на реч се основава върху статистически методи, можем скоро да очакваме на нашия пазар и продукти за разпознаване на реч. Българският език е официален в Европейския съюз, поради което трябва да се влагат определени усилия за запазване на неговата идентичност и осигуряване на достъп на всеки гражданин до информация на родния му език. Това обстоятелство също ще доведе до активизиране на разработката на езикови технологии за български език.

ЛИТЕРАТУРА:

[1] Паскалева, Е. *Компютърна морфология – ресурси и инструменти*. ИПОИ-БАН, София 2007, ISBN 978-954-92148-1-9, 150 стр.

[2] Балрик-Линг: http://www.larflast.bas.bg/balric/index/index.htm, от панела вляво: *Морфологични ресурси*, *Анализатор, Демо за българския език*

[3] Осенова, П. и К. Симов. *Формална граматика на българския език*. Институт по паралелна обработка на информацията, БАН, София, България. ISBN: 78-954-92148-2-6, 128 страници. Вж. http://www.bultreebank.org/bgpapers/FormalGrammarBG.pdf

[4] Allen, J. *Natural Language Understanding*. The Benjamins /Cummings Publishing Company, Inc. 1994.

[5] Grosz, B. and C. Sidner. *Attention, Intention and the Structure of Discourse*. Comp. Linguistics, 1986, Vol. 12 , No. 3, pp. 175-204.

[6] Система Bultra http://www.bultra.com

[7] Cunnigham, H. *Information extraction – an user guide*. Research Memo CS-99-07, Computer Science Dept., University of Sheffield, 1999 (http://www.dcs.shef.ac.uk/~hamish/IE).
[8] Nakov, P. and M. Hearst. *UCB System Description for the WMT 2007 Shared Task*, in Proc. of 2nd Workshop on Statistical Machine Translation co-located with ACL-2007, June 23, 2007, pp. 212-215.
[9] Кръстев, Б. *Българския език в таблици и склонения*. София, Наука и Изкуство, 1984.
[10] Mann, W. and S. Thompson. *Rhetorical Structure Theory: Toward a functional theory of text organization*. Text, 1988, Vol. 8 No. 3, pp. 243-281.
[11]. McKeown, K. *Using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press, 1985.
[12]. Angelova, G. and K. Bontcheva. *NL Domain Explanations in Knowledge Based MAT*. Proceedings of COLING'96, Copenhagen, Denmark, Vol. 2, pp. 1016 – 1019.
[13]. Angelova, G. and K. Bontcheva. *DB-MAT: a NL Based Interface to Domain Knowledge*. In: Proc. of the 7th Int. Conference "Artificial Intelligence: Methodology, Systems, Applications (AIMSA-96)", IOS Press, Vol. 35 in the series "Frontiers in AI and Applications", pp. 218-227.
[14]. Ангелова, Г. *Генерация на текстове на естествен език в определена предметна област*. Под печат в сборника 'Компютърна лингвистика' т.1 (Семинар на Асоциацията за национален електронен архив АНАБЕЛА).

Галя Ангелова
Секция за лингвистично моделиране на ИПОИ-БАН
ул. „Акад. Г. Бончев” бл. 25А
1113 София
България
е-майл: galia@lml.bas.bg