ИСТОРИЯ И ВЪВЕДЕНИЕ В КОМПЮТЪРНАТА ЛИНГВИСТИКА

1. Що е Компютърна Лингвистика? (Computational Linguistics / Natural Language Processing)

Компютърната лингвистика е научна дисциплина, която е тясно свързана с последните разработки в информационните технологии. Тя е приложна дисциплина между лингвистиката и компютърната наука. Като наука принадлежи към конюнктивните науки, има връзка с теоретичната лингвистика и с познавателната наука. Като самостоятелно научно направление компютърната лингвистика се оформя през 60-те години. За пръв път се правят опити още през 40-те.

Основната цел на дисциплината е моделирането и симулирането на естествените езици, така че те да бъдат разбираеми за компютрите и в последствие създаване на софтуерни продукти, които имат някакви познания върху тези езици. Въпреки, че съществуващите системи за КЛ са далече от достигането на човешките възможности, те имат многобройни възможни приложения.

Разработките през последния половин век се доминираха главно от постиженията на Чомски в тази област и от големите университети. Главно компютърната лингвистика се разделя на две области – теоретична и приложна.

Приложната част се интересува най-вече от практическия резултат от моделирането на естествения език. Създаването на продукти позволяващи безпрепятствената комуникация човек-компютър е неотложна нужда. Днешните компютри не разбират човешкия език, а хората срещат затруднения да разберат компютърния, който не отговаря на човешкия начин на мислене. Общуването с компютрите на естествен език ще има голямо въздействие върху начина на работа с тях, позволявайки приложението на компютърните технологии във съвсем нови области. Неща, които днес са трудни и отнемат много време могат да станат извънредно прости:

Ще бъде преодолян проблема по общуване на хора използващи различни езици, машинен превод, програми ще управляват потока от електронни съобщения, ще извършват библиотекарски справки, ще правят резюмета, хора в неравностойно положение ще могат да ползват компютрите пълноценно, търсещи машини намиращи винаги точната информация, психотерапия и др..

Въпреки, че днес съществуващите КЛ базирани програми са все още далеч от човешките възможности, те имат многобройни възможни приложения.

Теоретическата част от КЛ се занимава с формалните теории относно това от какви лингвистични способности се нуждае човек, за да говори и разбира езика. КЛ предоставя формални модели симулиращи аспекти на човешката реч и ги прилага в компютърните програми.

Направления в изследванията

1. **Синтактичен разбор**

В това направление са били постигнати резултати още на ранните етапи на развитие на КЛ. Първоначално знанията използвани за направата на разработки в тази сфера се извличат от класическата граматика, но компютърната им интерпретация довежда до възникването на няколко модела имащи различно техническо и теоретическо въздействие. Ключовия проблем, който трябва да се

реши е да се построи логическа структура на изречението и да се удовлетвори изискването тя да се изобрази чрез някакъв вид семантично представяне.

2. Обработка на естествения език

От седемдесетте години насам естествените езици се смятат за най-добрия начин за взаимодействие с компютрите – начин, чрез който могат да бъдат преодолени бариерите между човека и компютъра.

През осемдесетте и деветдесетте години естествената комуникация с компютъра е сведена до графични интерфейси. Изобретяването на интерфейси комуникиращи си с потребителя на една система на неговия естествен език поставя много проблеми за решаване. Практическия резултат засега се състои в представянето на диалогови системи – начин на общуване с компютъра базиран на подразбиране на желанията на потребителя. Ако докато потребителя работи със системата срещне затруднение може да поиска “помощ”. Често системата не отговаря буквално на въпроса, който и е зададен, но тя заключава от какво има нужда потребителя от този въпрос.

В края на деветдесетте вече има системи, които могат да общуват с потребителя гласово – архитектурата им в общия случай включва блок за анализ на речевите съобщения на потребителя, блок за интерпретация на съобщенията, блок за пораждане на смислени отговори и блок за синтез на реч.

3. Морфология и речници

До осемдесетте години обработващите естествени езици програми използвали вградени речници от няколко хиляди думи. Дълго време това се считало за единствения начин за класифициране на думите преди да се извърши граматически анализ. Идеята да се намери начин, чрез който програмата да може да “предсказва” думата, която ще бъде търсена преди тя да е въведена да края мотивира изследванията в областта на морфологичния анализ. Типичните, базирани на този тип анализ, системи се състоят от набор речници на лексикалните елементи – корени, представки, наставки, окончания и правила за формирането на думите.

4. Машинен превод

5. Езикови ресурси

Под това название са обединени всички видове хранилища където се съдържа някакво лингвистично знание. Изучаването на начина за съхранение, стандартите, достъпа и използването на тези ресурси е почти автономна част от КЛ. Най-често срещаните начини на съхранение на данните е чрез речници или сборници (корпуси). Речниците могат да заемат различни форми – обикновени речници базирани на морфологичния анализ, специфични технически речници, речници отчитащи връзката между думите – кои от тях са синоними, омоними и т.н. Корпуса това е колекция от езикови образци съхранявани в специфични формати.

6. Извличане на информация от свързан текст

Функцията на система от такъв тип се заключава в разбирането на текста и отговори на въпроси относно неговото съдържание. Текста “се разчита” само от гледна точка на това какво потребителят иска да знае.

7. Търсещи информация системи

Възникват в края на 50-те и началото на 60-те години като отговор на рязкото увеличение на обема на научно-техническата информация. Според вида на съхранение и обработка на информацията, и според особеностите на системата информационно-търсещите системи се делят на две групи: документални и фактографски.

- документални – съхраняват текста на документите и техните описания (реферати, библиографски карти и т.н.).
- фактографски - занимават се с описание на конкретни факти, при това незадължително в текстова форма – могат да бъдат таблици, формули.

За обезпечаване на търсената информация в информационно-търсещите системи се създават специални информационно-търсещи езици. Информационно-търсещия език е формален език предназначен за описване на отделните аспекти на плана, на съдържанието на документите, съхранявани в информационно-търсещите системи и запитване. Процедурата по описване на документите се нарича индексиране – на всеки документ се приписва формално описание.

8. Хипертекстови системи

Много изследователи разглеждат създаването на хипертекста като начало на нова информационна епоха противопоставена на книгопечатането. Хипертекста се представя във вида на граф във възлите, на който се намират традиционните текстове и техните фрагменти, изображения, таблици и т.н. Възлите са свързани с разнообразни отношения, които се задават от разработчиците на хипертекста или от самия читател. Отношенията дават потенциални възможности за придвижване или навигация по хипертекста. Отношенията могат да бъдат едностранни или двустранни. Съответно двустранните връзки позволяват потребителя да се движи в две посоки, а едностранните само в една. Веригата от възли, чрез които читателя достига до определен текст, образува път или маршрут.

Компютърните реализации на хипертекст биват йерархически или мрежести. Йерархическият (дървовиден) вид на хипертекста съществено ограничава възможностите за преход между неговите компоненти. Мрежестите хипертекстове позволяват използването на различни видове отношения между компонентите.

По начина на съществуване хипертекста се дели на статичен и динамичен. Статичния не се променя в процеса на експлоатация, в него потребителя може да фиксира свой коментар без той да променя същността му. За динамичния хипертекст изменението се явява нормална форма на съществуване. Обикновено динамичния хипертекст се използва там където е необходимо постоянното анализиране на потока от информация.

Отношенията между елементите на хипертекста могат първоначално да се фиксират от създателя или да се пораждат всеки път когато потребителя използва хипертекста. В първия случай става дума за твърда структура, а във втория за мека структура на хипертекста. Твърдата структура технологически е напълно понятна. Технологията на организиране на меката структура се основава на семантичния анализ на близостта на документите един с друг. Това е нетривиалната задача на компютърната лингвистика.

Инструменти и подходи

Съществуват общи принципи на компютърното моделиране на езика, които се реализират във всички модели. В основата им стои теорията на знанията първоначално разработена в областта на Изкуствения интелект. Важни понятия се явяват структурни понятия като “фрейми”, “сценарии”, “план”. Елементите на понятийния апарат на компютърната лингвистика имат антологични и инструментални аспекти**.** Например в антологичен аспект на разделението на декларативни и процедурни знания съответстват различни типове знания, налични в човека – знания “какво” (декларативни) от една страна и знания “как” (процедурни) от друга.

Произход и взаимоотношения с другите науки

Компютърната лингвистика е дисциплина имаща междинна позиция между науките занимаващи се с изкуствен интелект (клон на компютърната наука занимаващ се с човешкото познание), лингвистика, инженерство. Компютърните технологии споделят произхода си с КЛ – синтактичния разбор, който е най-важен за проектирането на компилаторите на програмните езици, е също така основата на всяка обработваща естествените езици програма – и двете са реализации на Теоремата на Чомски. Същата тази теория, заедно с кореспондиращия й компютърен модел, дава приноса си към основната хипотеза на Изкуствения Интелект – че човешкото поведение, обикновено определяно като интелигентно, може да се симулира с компютър. Пътищата на двете дисциплини – ИИ и КЛ – често се пресичат. Поради това, че и двете дисциплини се занимават със симулацията на човешкото поведение, те си поделят усилията по така нареченото “когнитивно моделиране” на различни човешки реакции – което включва и езика. Може би това е причината, че КЛ се разглежда като част от дисциплините, от които произлиза една нова дисциплина наречена Когнитивна наука.

От седемдесетте години насам когато техниките за обработка на езика достигат значително ниво на развитие позволяващо да се реализират някои приложения, инженерите проявяват интерес към тези техники и скоро се оказва, че начините, по които ги прилагат, въпреки че определено са по-малко теоретични и интересни от когнитивна гледна точка, са много по ефективни.

За момента може да се каже, че докато КЛ се интересува повече от коректността и правдоподобността на моделите и начините на представяне на езика, Инжинерингът се интересува повече от тяхната използваемост дори и да е с цената на някои компромиси.

Развитие в различните държави

University of Sussex

- NLCL - the Natural Language and Computational Linguistics group

Една от най-големите групи от изследователи във Великобритания фокусирана върху статистическите и корпус базирани обработки на естествени езици.

Проекти:

- RASP (Robust Accurate Statistical Parsing) – проект занимаващ се с подобряването на акуратността на граматическите анализатори.
- COGENT – контролирано генериране на текст

Columbia University

Columbia Natural Language Processing Group
Проекти:

- AQUAINT – система отговаряща на въпроси, която се фокусира върху проблемите по отговаряне на сложни въпроси – такива, които изискват дефиниции, мнения или биографии като отговор.
- Newsblaster – система помагаща на потребителя да намира новините, от които се интересува най-много. Тя автоматично събира и анализира новините от няколко сайта в мрежата и предоставя интерфейс чрез който могат да се разглеждат резултатите.

Israel Institute of Technology

Laboratory for Computational Linguistics
Проекти:

- Corpus Based Analysis of Hebrew – целта на проекта е да изучава начини за създаване на по-добри инструменти за морфологичен и синтактичен анализ на еврейския език използвайки базирани на корпуси техники.

http://www-a2k.is.tokushima-u.ac.jp/member/kita/NLP/nlp.html

2. Периоди на развитие на компютърната лингвистика:

Исторически, сферата, наречена “реч и езиков процес” е била третирана много различно в компютърната наука, електроинженерството, лингвистиката и познавателната психология. Поради това разнообразие, понятието реч и езиков процес обединява множество от различни, но припокриващи се полета в различни направления: изчислителна лингвистика в лингвистиката, естествен езиков процес в компютърните науки, разпознаване на речта в електроинженерството, изчислителна психо-лингвистика в психологията.

Първоначални стъпки:1940-1950г

Най-ранните прояви в това направление датират от плодотворния в интелектуално отношение период след Втората световна война, който развива и самия компютър.

Периодът от 1940г. до края на 1950г. е свързан с интензивна работа върху две основни парадигми: автоматите и вероятностните модели, извличани от теорията и заключенията на Bayesian (Бейзиан).

През 1950г. от три по-ранни парадигми е дефиниран краен автомат.Първият е модела на Turing (Тюринг)(1930г.) за алгоритмично изчисляване, считан от много за основа на модерната компютърна наука. Машината на Тюринг била абстрактна, с ограничен контрол и входно-изходна лента. С едно движение, машината на Тюринг можела да чете символ от лентата, да напише различен символ върху лентата, да промени положението му и да го премести в ляво или дясно. (Затова машината на Тюринг се различава от крайните автомати главно по възможността си да променя символите върху лентата си). Втората парадигма била работата на Shannon (Шанън) върху информационната теория. Shannon (Шанън)(1948г.) използвал неподвижни машини, почти идентични с крайните автомати, за да моделира свойствата на дискретен информационен канал, подобен на телеграф. Шанън също така приложил към информационния проблем и разработките на Марков, наречени ”дискретни процеси на Марков”.

Дискретният процес на Марков може да се разглежда като машина, с краен брой състояния и възможност за преход от едно състояние в друго.Шанън подобрил тези модели ,така че всеки преход между състоянията генерирал символ, и по този начин създал т.нар. “Марков модел на езика”. Третата основна парадигма била McCulloch-Pitts(МакКълък-Питс) неврон (МакКълък-Питс,1943г.), опростен модел на неврон като вид “изчислителен елемент ”, който може да бъде описан в термините на логиката. Този невронен модел бил двоично устройство, при всяка ситуация или било активно или не. Невронът приемал входяща информация от останалите и се противопоставял ако неговата активност премине някакъв определен праг. Въпреки че той не бил много точен като биологически модел, бил изключително важен като изчислителен модел. Очевидно, МакКълък-Питс неврона е основата на модерните теории за невронните мрежи и свързвания.

По важна е неговата роля в развитието на автоматите или крайните машини. Например, базирано върху теорията на McCulloch-Pitts е откритието на Kleene(Клийн)(1951г.), който дефинира крайния автомат и обикновените изрази и доказва тяхната еквивалентност. Независимо един от друг, базирано върху по-ранните разработки на Тюринг и Шанън, и върху синтеза на електронни вериги, Huffman (Хъфман) (1954), Moore (Мур) (1956), а по-късно и Mealy (Мийли) (1955) дефинират крайния трансформатор.

Като използва идеята на крайния Марков процес от работата на Шанън, Chomsky(Чомски)(1956) първи свързал крайните машини като начин за характеризиране на граматика, и дефинирал краен език като език генериран от крайна граматика G. Тези първи модели довели до сферата на формалните езикови теории, която използва алгебра за дефинирането на формален език като последователност от символи. Това включва контекстно-свободните граматики, първи систематизирани от Чомски(1956) като естествени езици, но независимо от него открити от Backus Баскус(1959), както и от Naur (Наур), Backus (Бакус), Bauer (Бауер), Green (Грийн), Katz (Кац), McCarthy (МакКарти) и Perlis (Пърлис)(1960) при тяхното описание на програмния език ALGOR.

Развитието на вероятностните алгоритми за реч и езиков процес датират от друга разработка на Шанън: метафората на шумния канал и декодирането при предаването на език през медия подобна на комуникационен канал или звукови акустики. Шанън също заел концепцията на антропологията от термодинамиката като начин за измерване на информационната способност на канал или информационното съдържание на език, и създал първата мярка на английския език.

Ранните 50 години също са свързани с първите разработки в сферата на машинното разпознаване на човешка реч. През 1952г., изследователи от Лабораториите ”Бел” създали статистическа система, която можела да разпознава всяка от десетте цифри произнесени от един човек. Системата имала 10 съхранени зависими еталона. Те постигнали 97-99% точност на избиране на еталона, който имал най-голям коефициент на съвпадение с въведения звук.

Двете направления: 1957-1971

В края на 1950г. и началото на 1960г., изследването на речта и езиковият процес се е разделило в две направления: символично и стохастично.

Символичната парадигма се състои от две посоки за изследвания. Първата била въз основа на работата на Чомски за развиване на теорията за формалния език, която продължила от края на 50-те до средата на 60-те. Тази насока на символичната парадигма включва също така и разработки върху формалната лингвистика и дизайна на компилатори,което води до съвременните алгоритми за разбор. Ранните алгоритми за разбор били или botton-up или top-down; в края на 1960г. били открити динамични програмни алгоритми за разбор, а динамичният програмен алгоритъм top-down на Ърли дебютирал през 1970г.

Втората сфера за разработки била в областта на изкуствения интелект.

През лятото на 1956г. Джон МакКарти, Marvin Minsky(Марвин Мински), Клод Шанън и Натаниел Рочестер заедно с група от разработчици организирали симпозиум върху тема, която те нарекли изкуствен интелект. На този етап били разработени ранните системи за разпознаване на естествен език BASEBALL, STUDENT и ELIZA. Те били прости системи, които работили върху единични области главно чрез сравняване на комбинация от еталони и търсели ключови думи чрез прости евристики за разсъждаване и отговаряне на въпроси. В края на 1960г. били развити по-формални модели като крайните логически системи SIR и TLC.

Стохастичната парадигма намерила приложение главно в сферите на статистиката и електроинженерството. Към края на 1950г., метода на Bayesian (Бейзиан) започнал да бъде прилаган за решаване на проблема с оптичното разпознаване. Bledsoe (Бледсоу) и Browning (Браунинг) (1959) създали система за разпознаване на текст, която използвала голям речник и изчислявала степента на подобност на всяка последователност от букви, за съответствие на дума от речника, чрез умножаване на съответната подобност за всяка буква. Mosteler (Мостлър) и Wallace Валъс(1964) приложили метода на Бейзиан за установяване на авторството на статии във “Федералист”.

Те били написани в периода 1787-1788 от Александър Хамилтън, Джон Джей и Джеймс Медисън с цел да убедят Ню Йорк, че Конституцията трябва да се ратифицира. Били анонимно и като резултат, въпреки че някои от 85-те есета със сигурност принадлежали на един или друг от авторите, то за авторството на 12 се водил спор между Хамилтън и Медисън. Мостлър и Валъс приложили метода на Бейзиан: те използвали вероятностни модели за начина на писане на Хамилтън и Медисън, и изчислили максимума на подобие с автора.

През 1960г., ранната работа на Марков и Шанън била разработена в Скрит Модел на Марков и приложена в решаването на проблема с разпознаването на речта. Математиката, необходима за този модел била осигурена от Боум и колегите му в началото на 1970г. Идеята за комбинирането на модела на шумния канал и неговите теоретични техники за декодиране със Скритият Модел Марков била развита и усъвършенствана от изследователи на фирмата IBM, но тя била публикувана доста по-късно, през 1983г.

Няма да бъде преувеличено да се каже, че повечето от фундаменталните алгоритми, стохастичните и символичните системи, които контролират съвременното разпознаване на речта, проверката за правописни грешки и естествения езиков процес били разработени почти напълно до края на 1970г.

През 60-те години се забелязва сериозен напредък в областта на тестоваемите психологически модели на човешкия езиков процес, базирани върху трансформационна граматика, както и първите online сборници от

езиковедски изследвания: сборника на Браун за американски английски, съдържащ един милион примера от писмени текстове от различни жанрове (вестници, романи, академични текстове и др.,), който бил създаден в Университет Браун през 1963-64г.

Естествен езиков процес: 1972-1983

Следващият период е свързан с развитието на същинските алгоритми от сферата, която може просто да бъде запомнена като Естествен Езиков Процес. През 1972 Terry Winograd (Тери Виноград) създал неговата популярна SHRDLU система, която симулирала робот, прикрепен към свят от малки блокчета. Програмата можела да възприема естествени езикови команди (“Премести червеното блокче върху най-малкото зелено”) от невиждана дотогава сложност. Неговата система била и първата, опитала се да систематизира разширена (за времето си) граматика на английския, базирана върху системната граматика на Халидей.

Моделът на Виноград прояснил въпросът със синтактичния/ морфологичния разбор и вече вниманието можело да бъде фокусирано върху семантичните модели за беседване. И наистина 70-те се свързват с голям подем в развитието на семантичните модели. Те включват набор от програми, разработени от Roger Schank (Роджър Шанк), колегите и студентите му от Yale School (Секретното училище). Те създали множество от програми, с цел да моделират разбирането на човешкия език, чрез представянето на тези семантики и концептуални знания като скриптове, планове и цели, и организация на човешката памет (Шанк и Абелсон,1977; Шанк и Рийсбек, 1981; Кълингфорт, 1981; Ленърт, 1977). Другите модели през този период разчитали на по-формални семантични модели, базирани на предикатната логика. Такъв модел е системата LUNAR за въпроси и отговори (Уудс, 1978).

Този период е свързан с много модели на процеса на беседване (discourse processing). Те включват развитието на идеята на Гросз за беседването (Гросз, 1977; Сиднър, 1983), работа върху разрешението на местоименната анафора (Хобс, 1978), и логически базираната работа върху акта на говорене (Перолт и Ален, 1980; Коен и Перолт, 1979).

Накрая, през периода 1979-1982 различни учени от Palo Alto провели операция на унификация, която включвала работата върху определеното просто изречение, функционалната граматика и по-късната работа върху LFG.

Поколението на крайните детерминирани модели: 1983-1993

Началото и средата на 1980 бил добър период за работа върху естественото езиково развитие, с помощта на моделите на Appelt (Апелт), McKeown и др.

Тази сфера не била привлекателна за вниманието на обществото, дори когато се появила оригиналната разработка на Джонсън за крайните детерминирани модели на фонологията и морфологията. Силно отражение оказало откритието на Roland Kaplan и Martin Kay(1981), последвано от Koskenniemi(1983).

Основен проблем на компютърната лингвистика и на проектирането и реализацията на компютърни системи с интерфейс на естествен език е отсъствието на подходящи и адекватни лингвистични бази от знания (ЛБЗ) за съответния език. Задачите, които се формулират и решават в това направление

са свързани с реализацията на софтуерни средства за автоматизирано извличане на лингвистична информация от текстови корпуси и изграждането на ЛБЗ. В това направление се предполага:

- построяване на автоматен модел за морфологичен анализ и синтез на текстове на български език (Коскениеми модел);
- изследване на процесите на словообразуването и създаването на деривационен речник;
- изграждане на статистически управляема граматика на българския език (автоматизирано извличане на знания за лексиката и граматиката по примери);
- реализация на модули за граматическа проверка на български текстове;
- провеждане на изследвания в областта на компютърно подпомаган превод (английски/български) и др.

Основен проблем в компютърен анализ на текстове е автоматичната обработка на непознатите (за съществуващата ЛБЗ) думи и граматически конструкции. В началото на изследванията експериментално се реализира модул за анализ на изрази в произволна алгебрична система с автоматично типизиране на участвуващите елементи. След това, на базата на получените резултати (вкл. и по направление 2) ще се пристъпи към проектирането и реализацията на софтуерна система за анализ на текстове с автоматизирано попълване на съответната ЛБЗ.

В последните години се наблюдава истински бум на приложенията в област на автоматизираните системи за обучение с навлизането на мултимедията и отдалечените компютърни комуникации в ежедневието.

Очертават се две направления:

- А. Компютърно подпомагане на университетското обучение по информатика.

Свързва се със създаването на учебни пособия и/или хипермедийни курсове по следните дисциплини:
компютърна информатика (уводен курс за непрофесионалисти), практикум по информатика, професионални комуникации, обектно-ориентирано програмиране.

- Б. Проектиране и изграждане на интелигентна софтуерна система за разпределено обучение.

Включва следните основни моменти: концептуално описание на предметната област за обучение (чрез понятия, методи за решаване на задачи и тестове, методи и стратегии за обучение и др.), автоматично конструиране на планове за обучение на базата на експертно предсказване на поведението на студента, автоматизиран диалог (с генериране на тестове и задачи и автоматичното им решаване) и др..

Последните години: 2000 – 2004 г.

Българската асоциация за компютърна лингвистика е научна организация с идеална цел, обединяваща хора и организации, работещи в областта на естествения език и езиковите технологии.

Една от основните цели на Асоциацията е да участва и да инициира научно-изследователски проекти в областта на компютърната лингвистика. Тя обединява и организира учени от различни дисциплини и формира екипи от специалисти за изследователска дейност. В рамките на тази дейност вече са

натрупани значителен опит, знания и ресурси в областта на езиковите технологии. Членовете на Асоциацията повече от десет години активно участват в европейски и национални научни проекти.

Основните от тях са:

- 2002-2004 OCoRrect (Cyrillic and Latin OCR correction using electronic dictionaries and sentence context), Volkswagen Stiftung, Ref: I 77 863

Проектът OCoRrect разработва алгоритми и методи за автоматична корекция на текстове на български, немски, английски и руски език, получени след оптическо разпознаване. Методите се основават на използване на големи многоезични електронни речници и на анализ на контекста в рамките на дадено изречение.

- 1999-2002 TELRI II (Trans-European Language Resources Infrastructure) , CA INCO-Copernicus 97/98, PL97-7085

TELRI е инициатива за създаване на действаща инфраструктура между европейски езици и центрове за езикови технологии. Целта е осигуряване на обща платформа и общодостъпни едно- и многоезични езикови ресурси на индустрията, изследователските институти и висши учебни заведения, работещи в областта на компютърната лингвистика.

2001-2004 BalkaNet (A Multilingual Semantic Network for the Balkan Languages), IST-2000-29388

Проектът BalkaNet има за цел изграждането на многоезична лексикална база данни, състояща се от бази на WordNet, за множество централно и източноевропейски езици. Всяка база на WordNet за отделен език ще бъде структурирана в съответствие с базата на WordNet на университета в Принстън и тези на EuroWordNet.

- 2002-2003 Методи за корекция и извличане на структурна информация от текстови документи, Българска асоциация за компютърна лингвистика

Целта на проекта е извличане на структурата и адекватната корекция на текстови фрагменти от електронни документи от различни източници (набран документ, документ след оптическо разпознаване, текст, получен от автоматично разпознаване на реч и др.).

- 1999-2003 Bulgarian INTEX , Българска асоциация за компютърна лингвистика

INTEX е среда за лингвистични изследвания, която включва големи речници и граматики. Тя осъществява обработка на текстове от милиони думи в реално време. Асоциацията разработи INTEX модул за български език, включващ: граматика за разпознаване на граници на изречения, речници на словосъчетанията и словоформите и граматики за различни езикови фрази.

- Последни новини

Българската асоциация за компютърна лингвистика (www.bacl.org) предостави на потребителите безплатен правописен коректор на български език. Софтуерната разработка е дело на колектив, обединен около идеята за създаването на приложни продукти в областта на езиковите технологии. Екипът от утвърдени млади компютърни лингвисти от дълги години участват в редица международни проекти, в рамките на които развиват уникални езикови технологии. Продуктът е платформено независим и би могъл да работи под Linux, BSD, Windows. Настоящата версия е интегрирана в офис пакетите Microsoft Office 2000, XP и 2003 (използващи CSAPI 3). Инсталационният пакет на

програмата е около 5MB и не би забавил работата на компютъра, гарантират създателите.

"Редактор", "Коректор", "Право писец", "Правописник" - това са постъпилите предложения за име на новия софтуерен продукт, но създателите на програмата с благодарност ще приемат предложения за име, което е най-адекватно и със съвременно звучене.

Компютърните лингвисти изтъкват предимствата на техния продукт - освен че е безплатен, в него се използва най-добрата технология, базирана на минимални крайни автомати при съставянето на кандидати за корекция.

Последна кратка бележка

Разбирането на човешкия езиков процес е важна за цялостната сфера на познавателната наука. В допълнение разбирането на човешкия езиков процес може често да бъде полезно в изграждането на по-добри машинни модели на езика. Тъй като ние изграждаме системи за разпознаване на реч за да взаимодействаме с хората , то има смисъл да копираме решение което се държи по начинът с който хора са свикнали.

3. Специалисти, работили за развитието на компютърната лингвистика

Ханс Ускорайт

Ханс Ускорайт е професор по КЛ в университета в Саарланд. По същото време той работи като научен директор в немския център за изследване на изкуствения интелект, където той ръководи лабораторията за обработка на езика. Наред с официалните срещи, той също така е и професор в отдела “Компютърна наука”.

Ханс Ускорайт изучава лингвистика и компютърна наука в техническия университет в Берлин и в тексаския университет – Остен По време на престоя си в Остен, той работи за големия технически проект в лингвистичния изследователски център. През 1987 г. Ускорайт получава своята професорска степен по лингвистика от университета в Тексас. От 1982 г. до 1986 г., той работи като компютърен специалист в центъра за изкуствен интелект в Калифорния. По същото време той е привлечен като главен изследовател, а по-късно и като ръководител на проекта, от центъра за изучаване на езици и информация в университета в Станфорд. В началото на 1986 г. Ускорайт прекарва шест месеца в Щутгарт в IBM-центъра. През декември същата година той става ръководител на проекта “Лингвистични и логически методи за разбиране на немските текстове”. През това време води също така и курсове в университетите в Щутгард.

През 1988 г. Ускорайт е назначен за председател на КЛ в университета в Саарланд и открива отдела “КЛ и Фонетика”. През 1990 г. той става “глава” на новооткритата лаборатория за технически езици. Той е съосновател на принципното изследване “Разделение на специално съвместно изследване”, а също така и съосновател на “Европейската, следдипломна, познавателна и техническа система” , която е съвместна програма с университета в Единбург.

Професорът е президент на европейската асоциация за Логика, Език и Информация. Член е на международния комитет по КЛ; изпълнителен директор е на европейската мрежа за Език и Говор.

Последните му изследователски интереси са за : компютърните модели на естествени езици, напредналите приложения на техниката на езика, приложението и за познавателното и информационно управление, граматическите формалности и техните приложения, семантиката и синтаксиса на ЕЕ и немската граматика.

Джеймс Мартин

Джеймс Мартин е доцент в Департамента за компютърни науки и Института за когнитивни науки в Университета в Колорадо (в Boulder). Той е завършил Техническата гимназия в Бруклин (NY) , където той е израснал. Получава бакалаварска степен по компютърни науки от Колумбийския университет, и докторска степен по компютърни науки от Калифорнийския университет (в Berkeley).

Той е публикувал книга и многобройни статии на различен вид теми, отнасящи се до: изкуствения интелект, когнитивна наука и компютърна лингвистика.

Следните организации поддържат неговите проучвания: National Science Foundation, Boeing, Hewlett-Packard, U S West, the Mars Co., and the Colorado Advanced Software Institute (CASI).

4. Характеристики на изчислителния процес

Изчислителният процес на човешкия език е мултидисциплинарна област и много учебници предлагат въведения в разнообразните прояви на неговите сфери, като: класическия естествен езиков процес, статистическия естествен езиков процес, разпознаването на речта, изчислителната лингвистика и човешкия езиков процес.

Исторически учебниците и курсовете по изчислителен процес на естествения език са били твърде балканизирани. Текстовите данни над ниво дума (разбор, семантична интерпретация, прагматика) на са били анализирани статистически в курсовете по естествен езиков процес в отделите по компютърни науки. Текстовите данни на ниво дума (морфология, лексика) не са били статистически обхванати при изчислителните лингвистични курсове в отделите по лингвистика. Важни проблеми като проверката на правописа са били често напълно игнорирани, подобно на отношението към естествения езиков процес. Но последните 10 години отбелязват голяма промяна в сферата на речта и езиковия процес. Средствата, разработени за разпознаването на речта (Скрит Модел на Марков, N- грами, вероятностни граматики), са били използвани в естествения езиков процес, а тези разработени за теоретичната лингвистика (морфологични трансформатори, крайна детерминирана фонология и морфология) са били приложени към NLP разработките като проверка на правописа. В отговор на този прогрес се създават фундаменталните алгоритми на всяка от тези области, които са статистически и логически, и приложими към речта и писмения език.

Разнообразието на приложенията на езиковия процес е от изключителна важност в последните години. Значимо е да се разберат как езиково свързаните алгоритми (от CMM до унификация, от lambda calculus до преобразувано обучение) могат да се приложат към важни житейски проблеми: проверка на правописа, изследване на документи, разпознаване на речта, web-дизайна,

машинния превод и диалоговите фактори на говоримия език. Целта на традиционните учебници е изграждането на интегриран езиков фактор със софтуерни слоеве, отговарящи на различни нива на лингвистичния анализ. От друга страна се представят важни приложения, съответстващи на всеки отдел, както представеното съответно лингвистично знание. Всяко лингвистично ниво дава възможност да се разберат и моделират определена съвкупност от интересни лингвистични факти.

Преобладаващите статистически алгоритми за езиков процес и развитието на системите за оценяване на речта и езиковия процес, спонсорирани от държавата доведоха до ново тълкуване на самото оценяване. Всяка проблемна област се съпровожда с оценка на системите, която включва изпробване и тестване, утвърждаване и информационно теоретични оценъчни метрики.

5. Цифрови библиотеки

Възникване на цифровите библиотеки

За да могат потребителите в света да използват цифрова информация, библиотеките трябва да проверяват тяхната политика и да следят техния път на работа. Основата на тази нова библиотека е в това, че политиката на електронните документи трябва да се търси в анализите, за това как електронните документи се различават от традиционните, и в анализите, за променящите се роли в институциите в “Веригата от стойности” за документи. В електронната среда библиотеките трябва да се стремят към достигане средствата за информация.

В по-голяма степен електронните публикации трябва да бъдат считани за вариант на традиционна кореспонденция. Това е еволюционен процес, който е адекватен на преходния период, в който сме в момента. Печатните и електронните публикации ще продължават да съществуват едновременно в близкото бъдеще.

Едно от главните свойства на проектите, свързани с Цифрови библиотеки е да помага на потребителя да намери бърза и точна информация. Тези библиотеки могат да бъдат допуснати незабавно чрез многобройни търсения.

Проучването и подобряването на Цифровите библиотеки се увеличава с развитието и навлизането на Интернет, особено на световно разпространената мрежа, която е много удобно средство за поддръжка на проектите за Цифрови библиотеки. Но за съжаление тя няма да работи удобно с по-обемни документи като енциклопедии, речници и репортажи. Затова сега е ясно, че_Цифровата библиотека е сбор от електронни документи в мрежата. Мрежата може да бъде от някакъв тип като локална мрежа или даже Интернет.

Цифрови библиотеки и информационно подобряване

Класическата област на изследване, която се занимава с електронно търсене на документи е Информационното подобрение. През последните 30 г. темата на тази дисциплина произхожда от електронните каталози с текстови и мултимедийни документи. На пръв поглед Цифровите библиотеки и Информационното подобряване не се различават толкова много.

Предимството на търсенето чрез Цифровите библиотеки е начина им на развитие: Информационното подобряване трябва да забрави за общоприетите

условия на доставчиците за професионална информация. То се обърква от голямото разнообразие на сървъри, формати и съмнителни механизми. Една от главните задачи на Цифровите библиотеки е да се справи с тези различия. Мултимедийните информационни техники за развитие на диалога, сложните публикации и информационните системи са няколко от отговорите на този проблем.

В последни дни системите за документалното развитие назначават специализиран персонал, който да подреди документите в списъци. Но много складове за електронни документи не са в състояние да използват този скъп, но ефективен метод. Алтернативата е използването на пособия за автоматично индексиране на целите текстови документи. Те са по-бързи и по-евтини дават предимство на структурната информация (като HTML или SGML), включена в електронните документи. В допълнение те могат да комбинират специфични теми със специфични потребители.

Системите, които използват автоматично индексиране за мултимедийни документи се базират на мрежи, които използват вероятностни оценки за различни доказателства. По този начин доказателства от различен тип могат да бъдат комбинирани в по-общи документи, които засягат един и същ проблем. В допълнение тези системи включват множество от правила, които се активират за да открият специфични пътища за появяване на текущия индекс, или специфична характеристика на цифровия образ и последващо приемане на точната концепция на индекса.

Развитието на технологиите за Цифровите библиотеки, които работят за ефективността на развитието на текущите мултимедийни приложения, изисква постоянен достъп до големи организирани складове на информация, който е все по-важен.

Цифровата библиотека е много важна приложна сфера за информационната инфраструктура В частност, от техническа гледна точка, Цифровите системи ще спомогнат много за развитието на технологията за управление на информация.

В допълнение технологията на Цифровата библиотека трябва да осигурява зависимост, успеваемост, лесен достъп, вътрешна оперативност, сигурност и защитеност.

По същото време Цифровите библиотеки представят частна информация и знание за съхраняването на информационната инфраструктура. Тяхната важност от образователна, социална и икономическа гледна точка е очевидна. В заключение, Цифровите библиотеки ще определят най-важната част от информационната инфраструктура.

За да даде Цифровата библиотека най-достоверна информация за инфраструктурата и за по-ефективно развитие на големи Цифрови библиотеки, научната програма трябва да поощрява и поддържа дългосрочното изследване върху много технически проблеми, които още не са решени. Разбирането на Цифровите библиотеки изисква голям опит, който може да бъде придобит само чрез постоянното развитие на Цифровите системи.

6.NLP – естествен езиков процес

Увод в NLP(Natural Language Processing) - Превод на език

Превеждането на човешки език от една форма в друга винаги е било цел за естествените езикови проучвания (NLP). През 1966 г. ALPAC (Automated Language Processing Advisory Committee)подкрепя доклад за Националната академия на науките, който поддържа тезата, че механичния превод няма да бъде постигнат в близко бъдеще. Но едва 20 години по-късно една практична ефективна система за превод се е появила на пазара, доказвайки че това наистина е възможно. В момента много компании се появяват с относително успешни транслационни системи, които са способни да превеждат различни езици.

С неотдавнашния растеж в Интернет хиляди бизнес и академични статии, написани на различни видове езици от целия свят, се публикуват ежедневно там. Този нарастващ пазар е тласнал индустрията да се обърне към развитието на софтуерни продукти за превод, които са приложими и по Интернет. Това развитие е облагодетелствало обществото по 2 главни насоки. Първо, помага за увеличаване асимилацията на информацията чрез съкращаване на времето, което отнема превода на статия и публикуването и на друг език. С помощта на машинният превод, преведените абзаци биха могли да бъдат сравнително бързо произведени. Макар, че превода може да не бъде 100% правилен, той начертава основната рамка на преведеното, което изисква само някакви малки корекции, за да бъде напълно_коректен. Второ – улеснява използването на информацията, като я прави по-лесна за хората да я четат и осигурява достъп до информацията, написана на други езици. Машините – търсачки могат да бъдат оборудвани с инструменти за превеждане на други езици, които разширяват възможностите на търсачката.

Развитие на естествения език

Развитието на естествения език е двигател на системите, които развиват и анализират писмения или говорим естествен език. То е поле на изкуствен интелект, което се опитва да използва компютрите за превеждане на съдържащата информация на по-широко разпространени езици като английския. По-голямата част от човешкото познание е записана в лингвистична форма, позволяваща на компютъра да разбере човешкия език и това е от голяма полза в улесняване на достъпа до това познание.

Въпреки, че е положен много труд в тази сфера, успехите са малко и ограничени. Главния проблем с развитието на естествения език е познанието за света. Човешкият език е неопределен по природа и всеки народ има много различни интерпретации в езика и неговото използване.

За да разбере компютъра и за да разреши различията между тези езици, той трябва да ги познава отлично. Това обширно познание е просто твърде голямо за да се справи с него днешната технология. Затова развитието на естествения език рядко се използва в по-общи трудове, а е ограничен в домейни.

Има различни нива в естествения език, включващи фонетика – фонеми и звуци, морфология – оформяне на думи, синтаксис – оформяне на думите в изречение, семантика – значението на изреченията и прагматика – как се използват изреченията. Повечето трудове са базирани на синтаксис и семантика.

Едно от най-важните приложения на развитието на естествения език е информационното подобряване, т.е. когато потребителя постави въпрос, той може да бъде разбран и компютъра може да му отговори. Тук познанието за специфични команди не е необходимо повече. Когато му се постави въпрос, компютъра извлича информацията от базата данни. Другите приложения на езика включват обобщаване и превод на текст.

Как работи една система за машинен превод

Различни подходи са били използвани за реализирането на машинен превод, но основните задачи на една система за машинен превод, могат да бъдат квалифицирани както следва:

1. **Анализ на текста**

Тук изходният текст е разбит на отделни сегменти и е анализиран на етапи както следва:

2. **Сегментиране**

Тук изходният текст е разбит последователно на параграфи, изречения и крайни думи.

3. **Морфологичен анализ**

Тук индивидуалните думи се търсят в речници и глаголите се идентифицират.

4. **Функционален анализ**

Определя как всяка дума функционира в изречението и определя тези части на речта за думи, които могат да функционират като повече от една част от речта.

5. **Синтактичен анализ**

Определят това как думите могат да бъдат поставени заедно, за да формират правилни изречения, и определя каква структурна роля има всяка отделна дума в изречението, кои фрази са второстепенни и какви са останалите фрази.

- Източник, цел на трансфера - След анализиране на изходния текст, той е моделиран в независима презентация.
- Цел на преведеното - Крайната цел на езика може да бъде генериране на езикови независими презентации в по-детайлизиран процес, така че той ефективно да връща стъпката за анализиране на изходния текст.

Примери за съществуващи системи за машинен превод

Понастоящем, много продукти базирани на системи за машинен превод бяха произведени. Някои представителни примери са:

SYSTRAN NLP Browser

Това е система, която използва технологията на системата за машинен превод за мултилингвистична обработка на информация. Главно нейната цел е да комбинира ползата от компонентите на машинната транслация, именно чрез огромни речници и граматически разбор, и вече съществуващия търсещ

инструмент (PDIAG) в нов подобрен инструмент - NLP Browser. Той се състои главно от предпроцесор, който синтезира информацията относно документа и я съхранява в база данни, и браузъра, който се състои от 3 компонента: query constructor, който използва разнообразните запитвания, търсачката, за да намери и върне търсените изречения, и viewer to display, които излага изреченията по един от двата езика.

Предпроцесорът работи върху документи отделно от съхраняващия процесор. Той основно преработва документите във файл за търсене, съдържащ подробна информация за морфологията, синтаксиса, семантиката, както и информация за крайния превод. Документите автоматично се разбиват граматически и се превеждат от системата за машинен превод, и граматическите части и изходната информация се съхраняват. Това разделяне на предпроцесорния етап от актуалната информация позволява питането да бъде направено моментално.

Браузър (Query constructor)

Предпроцесорната база данни може да бъде разгледана от два типа въпроси: структурирани и свободни въпроси. Структурираните въпроси, съдържат в себе си официални конструкции с предопределен синтаксис, използвайки структуриран език за да ограничат направените попадения от съществуващите достъпни попадения и още позволяват голяма гъвкавост в изразяването на желаната информация. Свободната форма на въпросите, от друга страна, позволява на потребителя да използва въпроси от най-разпространените езици, като английския и френския. Въпросът може да се състои от проста дума, фраза или дори изречение във форма на изявление. Тези поставени въпроси ще бъдат анализирани от системата за машинен превод по начин много подобен на този на предпроцесорния етап.

Търсачка

В съответствие на двата вида въпроси, се използват два различни механизма за търсене, които представят проучвания базиращи се на точен или относителен подбор. Структурните въпроси прилагат точен подбор, докато всички условия специализирани в питането не бъдат задоволени. Относителният подбор е съответната схема за свободната форма на въпросите. Свободната форма въпроси първоначално преминава през системата за машинен превод и след това се обработват по същия начин както текста база данни е предварително рабработен. След това браузърът представя относителния подбор между вътрешното представяне на въпроса и вътрешната информация на всяко изречение в базата данни. По този начин не само външната прилика, но също така и структурните подобия могат да бъдат комбинирани.Viewer to display - след като търсачката е открила съвпадение между изречение в базата данни и условието на въпроса, изречението е маркирано и извлечено от базата данни, и е трансформирано така че да отговаря на изискванията на езика за да бъде изложено на потребителя.

Други примери за системи за машинен превод са Globalink's GTS Power Translator, LOGOS' Intelligent Translation System and Intergraph's Transcend.

Системи за обобщение на текст

С настъпването на развитието на информацията, електронните документи стават основна медия на бизнеса и академичната информация. Хиляди и хиляди електронни документи се произвеждат и стават достъпни по Интернет всеки ден. За да станат възможно най-ефективни тези документи, е добре да се даде възможност, те да бъдат синтезирани. Затова системите за обобщение на текст имат огромна полза. За да се генерира обобщение, трябва да се идентифицира най-важната информация от документа, изпускайки неуместната информация и незначителните подробности, и да се съберат в ясен компактен доклад. Това обаче е по-лесно да се каже отколкото да се направи, тъй като това включва някои от основните проблеми на NLP. За да се произведе независима система се изискват умения в разбирането на естествения език, семантичното представяне, дискорсивни модели, световни знания. Успехите на независимите системи са много малки и ограничени, за да се идентифицират ключовите пасажи и изречения на документа. Извънредно успешни системи са били произведени за ограничени области, като времето, финансовата област и за медицинска база данни.

Domain Specific подход

1. Извличане на информация

Системата за извличане на информация анализира нестрого оформен текст за да извлече информация, специфична за конкретната област. Това не позволява да се разбере целия текст във всичките вкарани документи, но тя анализира порции документи, които съдържат достъпна информация. Наличието на информация се определя от дефиниран областен наръчник, който трябва да специфицира толкова точно колкото е възможно точно какъв вид информация системата очаква да намери. Има метод, който измисля извличането на информация по термини от условната базата данни. Тук неструктурираните документи се превращат в класифицирани база данни входове, които след това се използват за да запълнят “шаблоните”. Обобщения доклад може след това да бъде внедрен, като използва парчета от готовия текст от “шаблона”.

2. Примери за системи за извличане на информация

Безброй системи за извличане на информация са разработени за специфични сфери. Макар, че повечето от тези системи не се използват постоянно в Интернет, потенциалът им е огромен и внедряването им в Интернет е относително просто.

3. Технически преглед / Научна литература

Разработени са системи, които да контролират техническите статии в сферата на производството на микроелектронни чипове. Тези системи биха могли лесно да бъдат модифицирани към пазара на някои други сфери, чрез просто обработване и разширяване на областните наръчници, за да се използват по Интернет, където много статии могат да бъдат намерени.

4. Извадка от медицинска база данни

Такива системи са били изработени да анлизират и обобщават медицински доклади, чрез извличане на диагнози, симптоми на болеста, физически открития, резултати от тест и терапевтични лечения. Тези системи могат да бъдат използвани в полагането на грижа за здравето на пациентите с помощта на сигурни и качествени знания. Централната база данни, достъпна от различни

медицински центрове, може да се използва за улесняване транспортирането на пациентите, също така и за спешни случай, когато медицинските доклади се необходими веднага. Тези системи биха могли също да бъдат разширени да анализират някои други бази данни, като финансовия отчет например.

Case-Based подход

Тук внедрения документ е комбиниран в корпус от достъпни и недостъпни текстове. Вместо правене на формулиран пакет от областни наръчници от потребителя, системата просто разработва "тренировъчния корпус" на представителни текстове, които потребителя или областния експерт е класифицирал умствено като едно от двете достъпен или недостъпен. Тези предефинирани представяния на текстовете са съчетани с документа, и използват статистически подходи за да определят текстовете, които са достъпни за областта на документа. Главно текстовете, които съдържат само обобщена информация са малко вероятно да бъдат съотносителни с областта, защото подобни случаи ще бъдат намерени в недостъпните, както и в достъпните текстове на тренировъчния корпус. Текстовете, които са твърде специфични са малко вероятни, защото ще има много малко комбинирани случаи. По този начин, използвайки тези статичстически подходи, само представителните текстове, които съдържат информация за специфичната област ще бъдат използвани. Тези комбинирани достъпни текстове, биха могли след това да бъдат използвани за генериране на обобщението на текста. Главно Case-based подходът може да се смята като разширение на основната система за извличане на информация. Проблемът със системата за извличане на информация е, че задържа виртуално цялата информация, която е достъпна за областта, без никаква дискриминация между важната информация, подробностите, и общата информация. Чрез включването на статистически текст в Case-based подхода ние имаме възможност да добием представа за важността и достъпността на информацията изложена в обобщението.

Domain Independent подход

Както е опоменато преди, обобщението на областния специфичен текст е много по-трудно отколкото областната специфична задача. Доброто обобщение трябва да включва най-уместната информация, като се пропуснат подробностите и неуместната информация. Все пак различните парчета информация ще бъдат достъпни за различни хора, в зависимост от техните индивидуални интереси и нужди в областта. По този начин не много пълни сполучливи операционни системи са били разработени. По надолу ще видим някои от най-общите подходи, използвани за производството на някои експериментални системи.

Документ резюме

Тук обобщението е направено чрез отхвърляне на недостъпните текстове на документа, и съдържащо само ключовите пасажи и изречения на документа. Типичната система, главно се състои от две секции – the Reader and the Extractor(четящо и извличащо устройство). The Reader главно чете вкарания текст и го конвертира във вътрешни представяния, имайки предвид появяванията на думата и калкулацията на тежеста й. The Extractor определя конкретните изречения, които трябва да бъдат включени в обобщението, чрез

анализиране на тежеста на думите и тежеста на изреченията, и след това генерира обобщение от вътрешните представяния.

Пример на експериментална система, която използва тези методи е Automatic News Extraction System (ANES) разработена от Лиза Pay. Тази система цели да направи обобщение на новините от много различни източници, постигнала относително добри резултати, въпреки факта че тя е ограничена от смущение, поради което тя е публикувана самостоятелно. Ако е разработена, тази функция би доказала, че сиситемата ще бъде изключително използвана за категоризиране и насочване на информация по Интернет, чрез снабдяване с обобщения на всички достъпни разнообразни документи по мрежата.

Друг пример за областно независимо обобщение на текст, достъпен по мрежата е the NetSumm web страница с инструменти, които са способни да открият ключовите точки в статии и резюмета.

7.Използвана литература

- Speech and Language Processing
  Daniel Jurafsky & James H Martin
- Survey of the State of the Art in Human Language Technology
  Editorial Board:
  Ronald A. Cole, Editor in Chief
  Joseph Mariani
  Hans Uszkoreit
  Annie Zaenen
  Victor Zue
  Managing Editors:
  Giovanni Varile
  Antonio Zampolli
- www.lett.unipmn.it/~ling_gen/fer09.pdf
- http://www.ifi.unizh.ch/CL/CL_FAQ.html