

Gjuha dhe kompjuteri

Musli Musliu

5 shtator 2013

Përmbledhje

Në këtë shkrim shtjellohen disa tema të përpunimit kompjuterik të të dhënave gjuhësore nga fushat e fonetikës, e morfologjisë dhe e sintaksës.

1 Gjuha dhe kompjuteri

Pjesa e dytë e shekullit XX u karakterizua me kompjuterizimin dhe automatizimin e shumë proceseve industriale dhe makinerike. Me zbulimin e kompjuterit dhe makinave automatike shumë procese gjeten edhe ndërrimin e destinacionit të përdorimit. Shtrohet pyetja se si i kuptojnë urdhrat dhe si veprojnë këto makina? Nga gjuha kompjuterike, e cila është një gjuhë që ka një sintakse dhe semantike në vete, gjuhë kompjuteri (të menduara gjuhët programuese).

Në këtë shkrim me shume do të shtjellohen temat e përpunimit kompjuterik të të dhënave gjuhësore. Në këtë lëmi flitet për përpunimin e të dhënave gjuhësore në lëmin e fonetikës, të morfologjisë dhe sintaksës. Analiza sintaksore përqendrohet në programet e ashtuquajtura "Parser", që janë analizatorë sintaksorë të të dhënave gjuhësore. Në lëmin e morfologjisë përqendrimet janë më të mëdha në mësimin dhe analizën e trajtave të fjalëve të cilat përdoren edhe programet për përpunimin e teksteve, po ashtu edhe si programe të veçanta në fushën e gramatikës si, fjala vjen, gjenerimi kompjuterik i foljeve dhe emrave (zgjedhimi dhe lakimi) dhe njohja (analiza) e trajtave të foljeve. Në lëmin e fonetikës përdorimi më i madh gjendet në shndërrimin e tekstit në të folur dhe lehtësimin e gjetjes më të shpejtë të fjalës të cilën e kërkojmë. Këto aplikime mund të bëhen në disa mënyra me anë të bazës së shënimeve, me anë algoritmeve (strukturë e thjeshtësuar e përgjithësuar e përpunimit të të dhënave) në disa raste të ndihmuara edhe nga një pjesë e bazës së shënimeve. Kjo lëmi e shkencës, gjuhësia kompjuterike (linguistika kompjuterike) merret me përgatitjen e të dhënave gjuhësore për përpunim kompjuterik.

Për përdorim të këtyre aplikimeve do të mbështetemi në disa ilustrime që janë dhënë si aplikacione në internet.

2 Kërkimet fonetike dhe kompjuteri

Kërkimi fonetik (në të njëjtin parim është edhe ai morfologjik) është një proces (aplikacion, modul) që është i lidhur (implementuar) për motorët e kërkimit (makinat e kërkimit, si për shembull është google, yahoo në mes shumë të tjerave). Një motor (makina) kërkimi është një modul softuerik (program e algoritëm kompjuterik) që shërben në lehtësime kërkime më të shpejta dhe të zgjeruara të diçkajes që na intereson. Modulet gjuhësore në këtë lëmi shërbejnë si shtesë në disa aspekte, për përdoruesit, kjo do të thotë që ai është në gjendje të shkruaj një apo me shumë shkronja gabimisht dhe motori (makina) e kërkimit do të jetë në gjendje për të dhënë rezultatet ekzistuese alternative.

Për një kërkim fonetik flitet, ndër të tjera, kur programi na propozon një përgjigje që në bazë të algoritmit të implementuar (në rast se kërkimi, vargu i dhënë, nuk është në bazën e makinave të kërkimit) propozimet alternative, synon për të gjetur shprehje të cilat, edhe pse të shkruar ndryshe, fonetikisht janë të njëjta ose të ngjashme. Pra, prej kësaj del se kërkimet fonetike janë fokusuar në dy grupe, të ngjashme dhe përafërsisht të ngjashme, që mund t'i quajmë të "forta" apo strikte dhe të buta, të ngjashme. Çka mund të pritet nga ky lloj i programeve? Nga ky lloj i programeve mund të pritet shumë sa i përket gjetjes së fjalëve të cilat nuk jemi të sigurt se si shkruhen, për fjalët të cilat nuk i dimë si shkruhen dhe ato të një gjuhe të huaj të panjohura. Sa për ilustrim në këtë lëmi kam trajtuar një kërkim fonetik të gjuhës gjermane. Mundësitë e përpunimeve të këtyre algoritmeve janë të ndryshme. Algoritmet e tilla mund të jenë shumë funksionale, të zbatueshme në shumë gjuhë, që do të thotë, i njëjti algoritëm të funksionojë për disa gjuhë, të dedikuara për një gjuhë, të përshtatura për folës të një gjuhe për një gjuhë të caktuar. Mundësia e punës së algoritmeve të fundit, të përshtatura për folës të një gjuhe për një gjuhë të caktuar, mund të ketë rol të anasjelltë, të funksionojnë njëjtë si për gjuhën e dedikuar, ashtu edhe për gjuhën dedikuese.

Në vijim do të jap disa raste për funksionimin e këtyre algoritmeve nga përvoja personale e algoritmit të zhvilluar në vitin 2001. Algoritmi i është përshtatur fonetikës së gjermanishtes (ku një rol të madh ka luajtur edhe shqiptimi i tyre në shqipe). Ky algoritëm i gjuhës gjermane në baza shkencore është në gjendje të njohë fjalët të cilat shkruhen ndryshe, ndërsa shqiptohen njëjtë, si për shembull, "mehr" dhe "Meer", të cilat ortografikisht (në të shkruar dallojnë), mirëpo nga ana fonetike janë të njëjta, kanë një fonemë (pra, në një farë mënyre është përfshirë distanca e levenshtajnit (një algoritëm po ashtu për kërkime gjuhësore). Për dallim nga algoritmi i levenshtajnit, jo edhe të fjalëve siç janë Maus-Haus, Hose-Rose, apo të shqipes gol-pol, mirëpo mund të implementohet shumë lehtë si modul shtesë, gjë që e bën edhe më të përsosur kërkimin). Ndihej dhe i lehtëson përdoruesin që njeh gjuhën gjermane për përdorim dhe gjetje më të shpejtë të një fjale në internet, në një bazë të madhe të shënimeve, siç janë të regjistrave të ndryshëm elektronikë. Ky algoritëm është i përshtatshëm për përdoruesit të cilët nuk e njohin gjuhën gjermane, e sidomos ata të cilët e flasin shqipen, është një ndihmese e madhe, e cila më ka ndihmuar mjaft shumë në përpilimin e këtij algoritmi. Si mund të ndihmojë në kërkime të ndryshe, qoftë ajo në një makinë kërkimi ose në ndonjë bazë të madhe të shënimeve të shqipfolësit? Për shembull, në këtë rast mbiemri Maier, në fonetikë na jep një fonemë e cila në gjuhën gjermane shkruhet në katër mënyra: Meier,

Meyer, Mayer dhe Maier. Kjo ndihmon në lehtësimin e vet gjermanëve, nëse kanë të bëjnë vetëm me të dëgjuar të fjalës si në internet, ashtu edhe në bazat e mëdha të shënimeve (me të shtypur të trajtës), ju mundëson qasje të shpejtë në katër mënyrat, të cilët nuk kanë nevojë të dinë se për cilin "Mayer" bëhet fjalë; në këtë rast eliminohet "Buchstabieren", "spelling" (thënia fjalës shkronjë për shkronjë). Nëse në një regjistër të madh elektronik apo në një makinë kërkimi, kërkohet trajtën "Maier" me ndihmën e algoritmit, si rezultat do të kemi të gjetura të gjitha format e shkruara të përmendura më lart. Sa për ilustrim edhe të njohësve të gjuhës gjermane edhe të atyre që nuk e njohin gjuhën gjermane, po e marrin si shembull edhe trajtën "shuman". Po të bëjmë një kërkim në Google me vargun "shuman", atëherë do të fitojmë një rezultat i cili është i barabartë në formën ortografike e fonetike me "shuman", ndërsa po të implemtojmë në këtë makinë kërkimi edhe modulën e kërkimit fonetik, si rezultat do rezultojnë edhe trajtat pasuese: *Schuman, schuman, schumman, schumann, schumann, shumann, shumann, schumann* apo në mënyre reverzibile, qoftë nga ngatërrimi i të shkruarit apo ngutja të shtypim njërin nga këto trajtat e cekura më lart, si rezultat do të kemi trajtën e duhur të fonemës. Algortimi është shumë i saktë edhe i përshtatshëm për kërkuesit e gjuhës shqipe, të cilët nuk e njohin gjermanishten. Lehtësime tjera janë edhe fonetik-fonetik apo edhe të fjalëve të huaja champion-kampion-shampion. E befasia më e madhe e këtij algoritmi mundëson në rend të parë gjermanofolësve për gjetjen e një fjale shqipe (p.sh një gjermanofolës fjalën "shqip" do të mund të imagjinonte se ka këtë formë "shqip", gjetjen e kësaj fjale e mundëson ky algoritëm i cili në vete është vetëm disa kilobait).

Në të njëjtin parim punojnë edhe me algoritma morfologjike të cilët janë më të paktë. Për shkak të komplikimit të tyre. Janë shumë më të komplikuar dhe kërkohen punë shumë më të madhe sesa algoritmet fonetike (në një varshmëri prej gjuhëve, për gjuhën angleze përpilimi i një algoritmi të tillë është më i lehtë, sesa për gjuhën shqipe, frënge, spanjolle, gjermane etj.), mirëpo në përdorshmëri janë shumë më të dobishëm, sepse shpejtojnë punën, thjeshtojnë mundin, reduktojnë bazën e shënimeve, mund të përdoren për qëllime arsimore dhe analiza gramatikore si dhe në programe drejtshkrimore. Algoritmet e tilla gjejnë trajtat morfologjike të një fjale. Komplikueshmëria e këtyre algoritmeve varet nga natyra e gjuhëve, nga karakteristikat e gjuhëve.

Për një temë të tillë, mund të gjeni artikuj në internet, nëse në makinën kërkuese do të shënoni fjalët, si: Stemming Pattern Matching Fuzzy-Suche Soundex Metaphone.

Për një temë të tillë, mund të gjeni artikuj në internet, nëse në makinën kërkuese do të shënoni fjalët, si: Stemming Pattern Matching Fuzzy-Suche Soundex Metaphone