UNIVERSITATEA "ALEXANDRU IOAN CUZA" IAȘI
FACULTATEA DE INFORMATICĂ

Abordări Statistice și Bazate pe Inferență în Procesarea Limbajului Natural – Aplicații în Morfologie și Sintaxă

- Rezumat-

Coordonator:

Prof. Dr. Dan Cristea, m.c.A.R.

Doctorand:

Radu Simionescu

Iași

Septembrie 2016

Teza prezinta contribuțiile rezultate din diferite experimente, inovații, analize si dezvoltări efectuate in solitudine sau in colaborare cu cercetători și entuziaști ai ariei NLP (procesarea limbajului natural).

Începutul activității doctorale a fost conturat de trei dezvoltări. Prima este reprezentată de achiziția unui *treebank* Românesc și dezvoltarea unui *parser* syntactic. O a doua este reprezentată de construirea unui segmentator de clauze nominale (NP chunker). Iar cea de a treia constă în îmbunătățirea clasificatorului morfologic (POS *tagger*) Român al UAIC, un instrument folosit pentru adnotarea morfologica automata a textelor Românești care a fost dezvoltat în cadrul anilor de masterat ai doctorandului. Activitățile efectuate pentru a atinge aceste trei scopuri au condus la dezvoltarea Graphical Grammar Studio (GGS), care a acaparat focusul activității doctorale, datorită spectrului său larg de aplicativitate.

GGS este un instrument care astăzi poate fi folosit în multe arii ale NLP, pe diferite nivele de analiză ale limbajului. GGS este un sistem simbolic care poate fi folosit cu ușurință pentru a analiza si adnota date textuale. Întregul său flux de lucru este vizual și necesita din partea utilizatorului un minim de cunoștințe de programare. Deși GGS are multe capabilități care satisfac majoritatea cazurilor de utilizare, oferă totodată posibilitatea de a fi extinse cu ajutorul unui limbaj de *scripting*, pentru a deservi acelor sarcini mai rar întâlnite.

Teza pune accent pe contribuțiile aduse în trei puncte de interes din NLP: dezvoltare de resurse, de instrumente pentru procesare automată și de aplicații.

Din punct de vedere al dezvoltării de resurse, contribuțiile ating diferite subdomenii ale NLP, pe diferite straturi de analiză ale limbajului. Pe stratul morfologic, prezintă noi metode de a extinde în mod automat dicționarul morfologic al *POS tagger*-ului UAIC prin adăugarea de forme vechi, metode validate în baza a două experimente. Într-unul din ele un subset de intrări morfologice diacronice este inferat cu succes. În celălalt, cuvinte vechi sunt corelate in mod automat cu formele noi, cu ajutorul unui model statistic care surprinde regulile ce guvernează evoluția în timp a unor construcții de caractere comune în limba Română.

Deasemenea, tot în segmentul dezvoltării de resurse, teza prezintă contribuția și rezultatele obținute din procesul îndelungat de dezvoltare a *treebank*-ului UAIC, RoDepTreebank (Mărănduc & Perez, 2015). Acest proces a constat în reantrenarea unui *parser* bazat pe platforma MaltParser, pe un corpus *gold*, cu scopul de a îl extinde.

Din punct de vedere al instrumentelor și aplicațiilor NLP, sunt prezentate diverse contribuții.

Dezvoltarea RoDepTreebank a rezultat deasemenea și într-un *parser* sintactic pentru limba Română, realizat pe scheletul oferit de instrumentul MaltParser. Cu această ocazie, s-au făcut experimente și cu un algoritm de parsare propriu, bazat pe un model statistic.

În urma experimentării cu diverse sisteme de învățare automată și bazate pe reguli care sunt populare in NLP, activitatea doctorală s-a concentrat către sisteme simbolice. Modele statistice sunt ușor de obținut, odată ce există suficiente date de antrenare. Și totuși, indiferent de cât de precise ar fi, întotdeauna mai există erori. Observând erorile în adnotare pe care aceste sisteme le produc, lingviștii de obicei vor să aibă posibilitatea de a le controla comportamentul în anumite situații. Acest lucru poate fi realizat în sisteme hibride, care integrează modele statistice cu o manieră de a controla manual procesul. Sistemele bazate pe reguli sunt de obicei adoptate pentru a restricționa, extinde sau suprascrie comportamentul unui sistem statistic. Astfel de sisteme bazate pe reguli pot fi folosite și în solitudine pentru a dezvolta instrumente NLP, fără ajutorul unei abordări statistice.

GGS a fost creat ca un cadru de dezvoltare de sisteme de reguli, aplicabil pentru o varietate mare de sarcini NLP. O descriere potrivită a sa este: Un instrument NLP pentru găsire și adnotare de cuvinte. O expresie GGS poate fi folosită pentru a interoga un corpus și este definită într-o manieră grafica sub forma unei rețele de noduri, care descrie condițiile ce trebuiesc îndeplinite de secvențele căutate. Condițiile sunt de mai multe tipuri. Se pot referi la atributele unui cuvânt, la poziția sa în cadrul secvenței căutate sau la relațiile pe care acesta le are cu alte cuvinte.

GGS a evoluat împreună cu diversele instrumente care au fost dezvoltate sau îmbunătățite cu el. Inițial, un NP chunker pentru limba Română a fost construit. Acest proces a ajutat la identificarea unui set de limitări care a fost depășit prin crearea posibilității de a verifica acordul gramatical în gen număr și caz, între cuvinte.

GGS a fost folosit ca o versiune îmbunătățită a sistemului de reguli al POS-tagger-ului UAIC, care este un sistem hibrid. În acest caz, GGS a fost adoptat ca un sistem de constrângeri.

GGS a fost folosit pentru a dezvolta un segmentator de grup verbal.

Teza prezinta deasemenea experimente cu rezultate încurajatoare, derulate în direcția inferării de gramatici GGS, proces bazat pe exemple de secvențe pe care gramatica inferată trebuie să le găsească într-un corpus.

GGS a fost adaptat pentru a lucra cu volum mare de date prin adopția unui algoritm de căutare care face uz intens de diverse metode te indexare.

Odată cu dezvoltarea de noi capabilități, GGS a dobândit din ce în ce mai multă putere de exprimare. O expresie GGS poate exploata relațiile definite între cuvinte sau între noduri de informație în general. Mai exact, expresiile GGS pot parcurge noduri de informație bazat pe relația de secvențialitate, relația părinte-copil, sau orice altă relație definită printr-o referință, pentru a descrie structuri valide de informație. Pentru a realiza astfel de căutări într-o manieră eficientă, GGS face uz de un mecanism de căutare indexată care este specializat, dar nu limitat, pe căutare de secvențe de cuvinte.

Concluziile tezei subliniază potențialul pe care GGS îl are de a deveni un amplu cadru de dezvoltare de instrumente si de analiza NLP, datorită capacității sale de a căuta structuri complexe de date într-un mod eficient, în special din prisma faptului că utilizatorul poate să definească și să organizeze logica de parsare folosind scheme logice vizuale.

Informația lingvistică poate fi structurată în mai multe straturi. De-a lungul timpului, in domeniul NLP informația lingvistică a fost modelată în multe maniere, însă structura de la bază a fost întotdeauna aceeași. Cărămizile nivelului sintactic sunt de obicei cuvinte încapsulate ca așa numiți tokeni – obiecte ce încapsulează informația sub forma unui dicționar de atribute. Astfel, cuvintele sunt complementate cu proprietăți, care pot fi expuse spre diferite tipuri de analiză. O propoziție este modelată ca o listă de astfel de tokeni. Aceasta reprezentare surprinde relația poziționala dintre tokeni, dar există și alte tipuri de relații care sunt de interes in NLP. Teoriile lingvisticii au avut un influențat în mod direct maniera în care sintaxa este modelată în informatică.

In reprezentarea sintactică se disting doua tipuri: reprezentarea bazată pe constituenți și reprezentarea bazată pe dependențe. La un nivel abstract, noduri de informație pot fi grupate împreună ca noduri mai generice sau pot fi relaționate cu alte noduri, folosind relații etichetate.

Aceste principii generice de modelare a informației sunt folosite pt a reprezenta nu numai sintaxa, dar și semantica. În reprezentările sintactice, din considerente de consistență, numai una din cele doua aspect menționate mai sus, este de obicei adoptată, și foarte rar, o cerință specifică necesită o combinație între cele doua. Deși cele două par a fi opuse, ele sunt de fapt specializări ale unui principiu generic de modelare a informației. Un corpus poate conține structuri sintactice ca o mixtură de constituenți și relații de dependență, iar roluri semantic iar putea fi reprezentate ca relații etichetate între constituenți sau tokeni. Tokenii ar putea referi un bazin comun de informații, precum un dicționar, un tezaur sau o ontologie, care deasemenea respect aceleași principiu de modelare a informației.

În termini mai practici, în NLP, pe nivelul sintactic, nodul atomic de informație tokenul, pe baza căruia se pot stabili alte structuri compuse de informație. Întinderi de tokeni pot fi încapsulați ca noduri unitare de informație și orice nod de informație poate fi relaționat cu un altul, folosind o etichetă.

Procesarea automata a informației lingvistice este o cerință de bază în orice ce analiza NLP. Un instrument care să exploateze faptul ca informația este guvernată de aceleași principii, si care să ofere capacitatea de a defini șabloane într-un mediu atât de complex, multistratificat și arborescent, ar fi de mare interes. Șabloanele ar trebui definite folosind constrângeri peste proprietățile nodurilor de informație și deasemenea peste relațiile care există intre ele. Relațiile ar trebui să acopere următoarele tipuri:

- relații poziționale – utile pentru a impune condiții legate de ordinea în care trebuie să apară nodurile într-o înlănțuire, precum într-o propoziție.
- relații de coexistență în cadrul unui grup – utile pentru a impune condiții pe noduri încapsulate ca un singur nod;
- relații date prin referință – utile pentru a impune condiții pe relații bazate pe referințe, precum relații de dependență.

Odată ce astfel de șabloane ar putea fi definite, ar putea fi folosite pentru a căuta structuri de noduri cu care se potrivesc, într-un corpus dat ca input. Ocurențe extrase astfel ar putea fi adnotate într-o manieră care să expună structura lor internă, in funcție de modul în care s-au potrivit cu șablonul căutat.

Cerințele de imaginate anterior descriu un instrument complex. Există astfel de instrumente, însă aceste sunt dificil de învățat și implică înțelegere avansată de principii de programare. Dorim un instrument simplu și ușor de folosit. Beneficiul adus de reprezentarea vizuală a informației complex relaționată nu poate fi contestat. Vizualizarea unui algoritm poate sa ușureze foarte mult procesul de înțelegere, care de obicei necesită o capacitate nativă de a asimila concepte abstracte. Șabloane sintactice reprezentate într-un mod grafic cu siguranță vor permite crearea de instrumente lingvistice interesante care sunt ușor de înțeles si ajustat.

Multe dintre inovațiile și dezvoltările prezentate în această teză sunt rezultatul colaborării și al coordonării cu alți cercetători și grupuri. În cele ce urmează, subliniem contribuțiile personale ale autorului, în activitățile prezentate.

Implementarea motorului GGS, a Editorului și a componetei AnnotationExplorer a fost realizată în întregime de către autor. Modul în care modelele pot fi descrise de către utilizator, inclusiv adoptarea reprezentării rețelei pentru RTNs și sintaxa expresiilor de potrivire de șabloane, sunt inspirate de NooJ. Cu toate acestea, GGS oferă în plus un set de caracteristici care iese în evidență ca noutate în ceea ce privește utilitatea, funcționalitatea și modul în care a provocările de punere în aplicare au fost depășite. Cele mai importante dintre acestea, sunt:

- Integrarea JavaScript, care a fost proiectată și dezvoltată pentru a permite utilizarea variabilelor, și care, datorită flexibilității sale, poate realiza mult mai mult decât atât. Aceasta permite utilizarea unor structuri complexe de date. Oferă utilizatorului un limbaj de scripting versatil, care conține atât aspecte de programare obiectă orientat și cât și programarea funcțională. Punerea în aplicare a integrării JavaScript a fost o provocare din mai multe aspecte. Atunci cand se face backtracking, contextul JavaScript trebuie inversat. Acest lucru se realizează prin menținerea unei stive de clone ale întregului context JavaScript. Clonarea în sine, reprezintă o altă provocare, datorită caracterului funcțional al limbajului;
- Look-ahead și Look-behind assertions, care sunt inspirate din expresiile regulate. Acestea permit unui șablon să pândească fluxul de jetoane de intrare, ceea ce permite posibilitatea de a explora contextul vecin al unui simbol, fără a îl consuma. Look-behind assertions ascund o importantă provocare de punere în aplicare, deoarece acestea traversează rețeaua în sens invers. Lungimea unei secvențe căutată în acest fel în stânga unui simbol nu este necesar să fie fixă, așa cum este cazul expresiilor regulate. Acest lucru este realizat prin compilarea unei versiuni în oglindă a rețelei și prin alimentarea cu tokeni în ordine inversă;
- Crorss-reference assertions permit traversarea nodurilor de informație, pe baza unor relații între tokeni, relații ce sunt definite utilizând ID-uri de referință. Mai mult decât atât, ele permit invocarea unei căutări suplimentare într-un corpus.
- Capabilitatea de căutare indexată permite utilizarea lui GGS ca instrument de interogare de corpusuri pentru volume mari de date. Acesta adoptă o soluție de noutate în ceea ce privește utilizarea de bitsets pentru a reduce în mod eficient spațiul de căutare pentru un anumit model.

Rezultate interesante au fost obținute în domeniul Deducției Gramaticale. Similaritatea dintre o gramatica simplă de NP chunking, folosită ca o gramatică de inducere, și gramatica inferată, este fascinantă. Algoritmul de contopire de stări prezentat, de asemenea, a fost implementat în întregime de către autor. Algoritmi bazați pe fuzionare de stări nu sunt noi, dar adaptarea acestora la mașinile cu sări reprezentate ca rețele și organizarea structurii rețelei de noduri rezultate, într-un mod ușor de interpretat, este. Acest lucru în sine, reprezintă o provocare separată, care a fost depășită folosind un proces în trei etape, care combina aspecte ale Teoriei Grafurilor cu Teoria Lingvistică Formală:

1. etapa de minimizare a automatului, care se realizează printr-un algoritm specific adaptat la mașinile de stat reprezentate ca rețele;
2. reducerea de arce care se intersectează, care este obținut prin transformarea fiecărui subgraf orientat, maximal, complet și bipartit, care are toate părinții într-o partiție și toți copiii din cealaltă, într-un graf planar. Acest lucru se realizează prin inserarea de noduri goale în rețea, care joacă rolul de hub-uri. Această transformare nu modifică comportamentul rețelei, în același timp făcându-l mai ușor de citit;
3. un pas care se ocupă cu poziționarea nodurilor. Acest lucru se realizează prin așezarea nodurilor pe o grilă și determinarea poziției lor, operațiune bazată pe o traversare în adâncime.

În domeniul instrumentelor și al resurselor lingvistice, contribuții importante au fost făcute, în special în ceea ce privește limba română. Integrarea lui GGS cu POS tagger-ul UAIC, care a permis dezvoltarea unor reguli de dezambiguizare POS fine, reprezintă o altă contribuție personală a autorului.

În domeniul studiilor diacronice, autorul a fost implicat în proiectarea, în testarea metodologiilor experimentale și în executarea a două experimente pe care s-au dovedit metode de succes de recuperare de intrări morfologice diacronice. Experimentul se ocupă cu recuperarea de intrări vechi, care conțin o diferență în rădăcina cuvântului, comparativ cu forma lor contemporană, și a implicat, de asemenea, o validare umană, care a fost realizată de către un grup de 20 elevi. Autorul a condus procesul de validare prin furnizarea de documente și de sprijin, precum și prin asigurarea calității manierei de adnotare.

Multe dintre caracteristicile de noutate ale GGS au fost proiectate pentru a facilita dezvoltarea unui Chunker NP românesc complex, care s-a fructificat ca un alt instrument lingvistic, expus în mod public ca o aplicație și serviciu web.

Prin colaborarea cu lingviști, autorul a contribuit la dezvoltarea UAIC RoDepTreebank și la dezvoltarea unui parser de dependențe antrenat pe acesta, folosind instrumentul MaltParser. Ca și POS Tagger-ul și Chunker NP, parserul a fost, expus în mod public ca o aplicație și serviciu web - un efort atribuit, de asemenea, autorului.

O altă contribuție este reprezentată de suportul oferit de către autor în experimente cu privire la impactul stilului de comunicare textuală non-standardizate, adică textele diacronice și chat-ul modern în datele de antrenare pentru etichetarea POS și parsarea de dependență.

Instrumentele lingvistice românești, care sunt disponibile în mod public ca aplicații și servicii web, tagger UAIC POS, The Chunker UAIC NP și parserul UAIC Dependență, au fost integrate ca și componente UIMA de către autor, ca o sarcină în cadrul proiectului de efort de colaborare - METANET.

Autorul este de asemenea implicat activ în proiectul SSPR, care își propune să dezvolte un treebank românesc în formalismul UD și un parser sintactic, care ia în considerare informații semantice.

Există mai multe îmbunătățiri care sunt avut în vedere ca viitoare dezvoltări pentru GGS. Una dintre cele mai importante este cea de a permite șabloanelor să aibă acces la mai multe resurse în timp ce are loc căutarea de secvențe potrivite. Acest lucru va permite realizarea de Cross-reference assertions în corpusuri multiple, și, de asemenea, va permite utilizarea de resurse externe, cum ar fi baze de date, ontologii, dicționare, tezaur etc. Ca rezultat, GGS va deveni capabil să execute interogări și mai avansate. Un aspect provocator în această dezvoltare este reprezentat de lipsa unei interfețe față de reprezentările interne ale diferitelor resurse.

Procesul de potrivire a GGS poate utiliza informațiile care sunt structurate ca rețele de noduri ierarhice. Dar este în prezent limitat în ceea ce privește modul în care aceste informații pot fi stocate. Pentru NLP, pe stratul sintactic, ne-am concentrat în principal pe un caz de utilizare comună, i.e. nodul atomic de informație este tokenul / cuvântul. Oferim utilizatorului acces prin șabloane GGS la aceste date prin intermediul formatului XML, și oferim mijloacele de a genera o versiune indexată, prin care se poate obține o performanța ridicată în cazul interogării acelorași date, de mai multe ori.

Însă informațiile lingvistice pot fi reprezentate într-o mare varietate de formate. O dezvoltare viitoare a GGS, care reprezintă în prezent o prioritate ridicată, este de a crește cantitatea de formate și tipuri de portale de informație care pot fi accesate de către o interogare. În prezent, numai datele XML pot fi prelucrate dintr-un disc local sau memoria locală. Pentru a permite accesul la diverse resurse externe existente, este necesară crearea unei punți către formatele lor de reprezentare internă. Anumite resurse care sunt de interes în aplicațiile NLP sunt accesibile

ca date externe, cum ar fi baze de date sau servicii web. Este necesară crearea unui adaptor între interogări și modul acestora personalizat de interogare.

Modul actual în care GGS accesează datele de intrare, trebuie să fie decuplate ca un adaptor personalizabil, pentru a permite integrarea și cu alte tipuri de portaluri de informații. Un set implicit de poduri vor fi disponibile în GGS, pe lângă cele actuale, care sunt compuse din formatul XML și a unui sistem de indexare personalizat bazat pe Lucene. Formatul XML oferă deja o mulțime de flexibilitate și poate permite accesarea mai multe resurse externe, care sunt deja disponibile ca XML. Următorul pas este acela de a permite accesul din șabloane GGS la resurse în formatul CONLL, sau accesibile prin utilizarea diferitelor limbaje de interogare, cum ar fi CorpusQueryLanguage (CQL), KoralQL (Joachim & Nils, 2015), XPath sau alte limbaje de interogare de baze de date. Următoarea etapele care urmează să fie atinse în această direcție este aceea de a permite utilizatorilor să acceseze WordNet, FrameNet și eDTLR. Utilizatorii vor avea nevoie, de asemenea, de instrumente pentru a construi propriile lor adaptoare personalizate. Am explorat strategii care ar face un astfel de proces mai ușor (Simionescu & Cristea, 2012), în primul rând deducând automat ce tip de informație este prezentă în date nevăzute anterior. Această ambiție ascunde o mulțime de provocări care nu au fost încă abordate.

O punte către alte tipuri de resurse nu poate fi limitată în mod necesar la portalurile de informații statice. Un utilizator poate decide că o ramură a logicii implicate în luarea unei decizii ar trebui să fie mai bine rezolvată printr-un alt instrument lingvistic, cum ar fi o rețea de neuronală sau de către un alt algoritm personalizat. Adaptoarele GGS ar putea permite integrarea altor instrumente specializate pe anumite sarcini lingvistice.

Domeniul Inducției Gramaticale are un aspect mult mai atrăgător când este privit din punct de vedere al rețelelor de tranziție recursive aplicate pentru limbaj natural. Am reușit doar să observam potențialul de cercetare, cu un prototip promițător. Rămân multe provocări neexplorate, care fac obiectul unor cercetări viitoare în acest domeniu. Cea mai evidentă sarcină este depistarea informațiilor irelevante, care nu este necesară pentru a determina prezența unui anumit fenomen sintactic. Cum s-ar efectua o astfel de abordare într-un context cu o mulțime de informații irelevante sau cum s-ar realiza acest lucru într-o situație în care o mare parte a informațiilor disponibile este prezentă într-o resursă externă, cum ar fi o ontologie, dicționar sau tezaur? Cum s-ar putea obține generalizarea în mod automat?

O altă oportunitate importantă de cercetare în domeniul deducției de gramatici GGS explorează impactul învățării asistate, prin furnizarea de feedback către mecanismul de învățare pentru ipotezele care au un scor scăzut de încredere.

Mai mult decât atât, am explorat doar ideea de a infera șabloane bazate pe relația de adiacență dintre tokeni, folosind o abordare bazată pe fuzionare de stări. În contextul informațiilor structurate ca rețele de noduri ierarhice, inducția de gramatic GGS ar putea oferi mai mult decât o generare de șabloane care se potrivesc cu secvențe de token-uri. Șabloanele nu sunt utile doar pentru identificarea secvențelor de noduri de informații. Următorul caz de utilizare a inducției gramaticale amintește de complexitatea de care mintea umană este capabilă ca să înțeleagă fără efort: Un șablon GGS poate fi proiectat pentru a găsi un singur simbol, care are un rol special într-o rețea relațională complexă de noduri de informații (tokeni și constituenți). Prevăzut cu destule exemple pozitive și negative ale unor astfel de tokeni, un proces de învățare care explorează nu numai relația de adiacenta dintre nodurile de informații ar trebui să stabilească faptul că elementele care separă exemplele pozitive de cele negative sunt conținute nu în atributele tokenilor dați ca exemple pozitive, ci în atributele tokenilor cu care sunt în relații directe sau indirecte.

POS taggerul UAIC a fost integrat cu un sistem bazat pe reguli, care permite intervenţia manuală în comportamentul procesului de dezambiguizare. Regulile au fost proiectate manual, pe baza unor rezultate obţinute în urma analizei erorilor. O oportunitate de cercetare interesantă legată de inducţie gramaticală este aceea de a genera automat şabloane GGS pentru a descoperi contextele în care are loc un anumit tip de eroare. Acestea ar putea fi utilizate pentru a crea automat reguli care corectează erorile generate de procesul de dezambiguizare statistică.

Segmentatorul de grupuri verbale dezvoltat adoptă o strategie de a detecta fraze verbale, fără a impune un set riguros de constrângeri privind structura internă a unor astfel de secvenţe de tokeni. Grupul verbal este detectat, pur şi simplu prin includerea unor cuvinte care sunt aproape de un verb principal, dar ordinea şi combinaţiile în care sunt permise acestea în limba română nu sunt descrise de către gramatica de GGS segmentare. Din acest punct de vedere, gramatica chunking VP consumă cu succes, dar ar putea genera secvenţe incorecte, din punct de vedere lingvistic. O cercetare interesantă cu privire la gramatică, care ar avea ca rezultat o gramatică mai fidela din punct de vedere lingvistic este următoarea: Segmentatorul curent ar putea fi folosit ca o gramatica de inducere pentru a extrage expresii verbale dintr-un corpus şi să le folosească ca exemple pozitive într-un proces de inferenţă. Acest lucru ar genera o reţea GGS care expune constrângerile în ceea ce priveşte ordinea permisă a componentelor care fac parte din grupul verbal românesc. Diferitele căi generate de noduri ar corespunde probabil cu diferite tipuri de grupuri verbale româneşti. Cu unele ajustări reţeaua GGS generată ar putea fi modificată manual pentru a clasifica, de asemenea, secvenţele potrivite.

Metodologiile de recuperare a intrărilor de dicţionar morfologice ale formelor de cuvinte învechite, s-au dovedit eficiente, şi sunt convenabile, deoarece ambele se bazează pe procese semiautomate care utilizează resurse lingvistice româneşti deja existente. Cu toate acestea, pentru a obţine rezultate calitative pe texte vechi, sunt necesare resurse mai cuprinzătoare, care pot facilita dezvoltarea a cel puţin celor mai comune sarcini NLP. Dezvoltarea de corpus necesită o cantitate mare de efort. În cazul corpusurilor diacronice, un grad şi mai specializat de expertiză este necesară, însă acesta este rar. Vom continua colaborarea în activităţi care dezvoltă astfel de resurse, aşa cum este cazul Rodia. Astfel de sarcini întâlnesc în mod constant noi provocări şi necesită sprijin tehnic pentru a ajuta cu progresul dezvoltării.

În acelaşi timp, sunt necesare mai multe resurse lingvistice româneşti, în special în domeniul semantic. Acesta este un segment în care limba română prezintă deficienţe. Astfel de resurse ar putea permite dezvoltarea unui analizator de dependenţă bazat pe semantică, şi crearea de noi oportunităţi de cercetare în ceea ce priveşte limba română, în alte arii ale NLP, cum ar fi traducerea automată. Vom continua colaborarea in activităţi care vizează aceste tipuri de dezvoltări. Vom continua dezvoltarea Ro-PAAS, si ne vom susţine implicarea noastră în proiectul SSPR. Adnotarea semantica nu este o sarcină uşoară;

O altă activitate de cercetare interesantă este fuzionarea segmentatorului de grup nominal cu cel de grup, rezultând iniţial o gramatică ce ar putea identifica clauza. O clauză este, în principiu compusă dintr-un grup verbal şi fraze substantivale. Mergând mai departe cu o astfel de dezvoltare, o gramatică GGS unificată ar putea afişa, de asemenea, structura constitutivă a clauzelor, şi chiar expune aspecte ale structurii discursului narativ. Ne imaginăm că este nevoie de informaţii semantice pentru obţinerea unor rezultate de înaltă calitate, pentru limba română, în special în provocările impuse de ataşamentul prepoziţional.

Această teză se concentrează pe descrierea implicării noastre în dezvoltarea de instrumente pentru prelucrarea limbii române, care au înflorit, pe lângă diferite concluzii, într-o realizare importantă: GGS, un instrument flexibil, care se potriveşte unei game largi de cazuri de utilizare.

Am contribuit la dezvoltarea unor instrumente și resurse lingvistice românești. Am descoperit noi metode de construire a resurselor diacronice prin procedee semiautomate. Am dezvoltat GGS, care poate fi utilizat pentru dezvoltarea altor instrumente, pentru interogarea de corpusuri și pentru a analiza datelor lingvistice. Ne-am aventurat în inducție gramaticală, și am constatat că aceasta promite să permită expunerea structurii care sta la baza informațiilor de suprafață. Am creat acces către cercetători si pasionați ai lingvisticii computaționale la un set de bază de instrumente lingvistice românești: POS Tagger, parser de dependențe și și segmentator de grup nominal.

Informația pură este un flux continuu de date nestructurate. Particulele de lumină, sunet, atomi ai universului, sunt toate unde de informație, care intră în mintea umană prin simțurile noastre. Formele de viață au evoluat pentru a discerne informațiile în unități structurate, un proces care este denumit în mod obișnuit ca generalizare. Acesta este procesul de bază care permite ca informațiile să fie interpretate, prelucrate ulterior, stocate și transmise. Acest proces nu este neapărat o trăsătură a minții, ci mai degrabă o caracteristică a vieții în sine: vedem informații transmise prin diferite mijloace, cum ar fi ADN-ul, hormoni, feromoni, impulsuri electrice, vizuale și fluxuri auditive, etc., în toate tipurile de forme de viață, la fiecare scara a spectrului, i.e. de la viața celulară microscopică la interacțiunea macroscopică în speciile sociale, dintre care, ființele umane sunt, de departe, cele mai evoluate.

Comunicarea este fundamentală pentru viață. Este fundamentală pentru componentele interne ale unui organism, precum și pentru interacțiunea dintre organisme. În afară de comunicarea biologică internă obișnuită, necesară pentru a susține viața, un sistem emoțional a evoluat ca o soluție pentru o calitate ridicată de viață, în special la mamifere. În plus, la om, un limbaj verbal complex, a evoluat. Limbajul verbal nu este folosit doar pentru comunicare între indivizi, ci și pentru comunicare internă, ca un instrument pentru dezvoltarea și rafinarea gândurilor foarte bine organizate, care se pot exprima apoi în exterior sau se pot menține pe plan intern sub formă de credințe de bază ale unui individ. Noi credem că limbajul verbal este motivul fundamental pentru capacitatea crescută a minții umane de a prelucra și structura informații în rețele complexe de unități de generalizare. Acest lucru a permis dezvoltarea gândirii analitice și de colaborare a persoanelor în sarcini complexe, pentru a dezvolta instrumente pentru îmbunătățirea calității vieții și pentru a explora natura lumii noastre.

Nu toate formele de comunicare cer ca informațiile să fie generalizate mai întâi și interpretate înainte de transmitere. Interpretarea poate fi realizată în întregime de către receptor, așa cum este cazul unei ființe umane care vizionează o fotografie – un instantaneu al datelor vizuale brute. Dar, în cazul comunicării lingvistice, informația este mai întâi generalizată folosind cuvinte și structurată în propoziții, înainte de a fi livrată. Există multe dezbateri cu privire la ce este un cuvânt. Nouă ne place să credem că cuvintele sunt semnale care au fost atribuite în mod convențional și în mod natural de către grupuri mari de indivizi, unor structuri comune de informații interne ale creierului. Aceste semnale nu sunt neapărat sub formă de sunet. În comunicarea orală, limba este ajutată de tonalitate, ritm, tempo și variație de volum. În comunicarea față-în-față este, de asemenea, ajutată de repere vizuale, care compun un segment consistent din ceea ce se numește comunicare non-verbală – un limbaj separat în sine. Natură umană a inventat forma de transmitere a limbii prin scris. Comunicarea textuală conține doar informație generalizată, digitizată. Pt unele aspecte legate de tonalitate, a apărut un limbaj al punctuațiilor. Conceptele identice sunt mapate de limbi diferite pentru diferite semnale acustice și reguli de combinare. Cuvintele pot fi, de asemenea, cartografiate de către persoanele surde folosind așa-numitul limbajul al semnelor, o formă de comunicare, care face uz de mâini, degete, gesturi și expresii faciale. Persoanele oarbe și surde folosesc semne tactile ca o formă de comunicare. Indivizii orbi învață să citească scrieri tactile.

Limba poate fi transmisă în diverse moduri, prin diferite canale. Există, cercetări efectuate chiar și în direcția de comunicare directă între două creiere. Indiferent de canal, structura sa este legată de aceleași principii de generalizare, care sunt fundamentale pentru viață. Informațiile lingvistice sunt segmentate în unități care sunt etichete ale unor concepte complexe, comune. Unitățile pot fi legate prin relații și unități multiple de informații pot fi încapsulate împreună ca unul. Rețele de noduri ierarhice de informații sunt astfel formate.

În afară de limbă, care poate fi văzută ca o invenție evolutivă, computerul este, probabil, următoarea realizare cea mai importantă de acest fel uman. Este un instrument care schimbă viața socială într-un mod fundamental. Este o extensie a minții, într-un sens. Calculatoarele nu pot procesa încă date nestructurate complexe, dar pot procesa, stoca, prin transfer rapid și avea acces la un volum mare de informații structurate, capacități pe care mintea umană nu le are. Combinate cu limba, tehnologia informației a permis comunicarea instantanee a mesajelor lingvistice pe distanțe mari și de a face cunoștințe accesibile de oriunde de pe pământ. Acest lucru a transformat modelul anterior al lumii noastre, adică societățile izolate comunicând prin conducte foarte înguste, într-o legătură globală de comunicare.

NLP este domeniul inteligenței artificiale, care are drept scop angrenarea dintre invenția evolutivă a limbajului verbal și recentele progrese în domeniul IT. NLP promite să permită traducerea automată între limbi în timp real, contribuind în continuare la legătura globală de comunicare. Promite să permită o interacțiune bazată pe limbaj natural cu software-ul calculatoarelor. Promite să permită computerelor să preia unele dintre sarcinile noastre cele mai plictisitoare, permițând, și forțând totodată, natură umană să se concentreze asupra gândirii creative ca să cucerească noi perspective asupra universului. Dar ceea ce mă fascinează cel mai mult este că promite să împuternicească calculatoare cu capacități mai apropiate de cele ale minții umane, ceea ce le-ar permite să discearnă cantități uriașe de informații pentru a descoperi cauze profunde și soluții la cele mai mari probleme ale lumii.

GGS este doar un pas mic în această mare viziune. El adresează in mod direct scheletul limbajului permițând traversarea informației care este structurată pe principiile generalizării. Acesta este motivul pentru care poate fi folosit in atât de multe sarcini NLP. Șabloanele GGS sunt momentan promovate ca o manieră de a controla comportamentul unui parser. Inferarea de gramatici GGS reprezintă drumul catre învățarea automată de șabloane.