RAPORTARE ŞTIINŢIFICĂ

FAZA DE EXECUŢIE NR. 1

CU TITLUL Prelucrarea datelor

Avizat,

Coordonator

Denumire: Universitatea din Bucuresti

Reprezentant Legal: prof. dr. Romita Iucu

Semnătură:

Ştampilă:

Director Proiect: prof.dr. Liviu P. Dinu

Semnătură:

Agent Economic

Denumire: Pluriva SRL

Reprezentant Legal: Marius Pascu

Semnătură:

Ştampilă:

Responsabil de proiect: Marius Pascu

Semnătură:

Raport stiintific
privind implementarea proiectului in perioada octombrie – decembrie 2016

Proiect: *Sistem Inteligent de Generare Automată a Răspunsurilor (SIGAR).*

SIGAR: Project 53BG / 2016, funded by Romanian National Authority for Scientific Research and Innovation, CNCS/CCCDI – UEFISCDI, PNCDI III (Programul 2 - Creşterea competitivităţii economiei româneşti prin cercetare, dezvoltare şi inovare. Transfer de cunoaştere la agentul economic "Bridge Grant")

Durata Proiectului: 1 octombrie 2016- 30 septembrie 2018

Director: Prof. univ. dr. Liviu P. Dinu
Contractor: Universitatea din Bucuresti
Etapa: 20 decembrie 2016 (unica)

Obiectivul principal

Principalul obiectiv al proiectului constă în implementarea unei soluţii bazate pe învăţare automată, care va parsa şi analiza baza de date de tichete în vederea construirii unor răspunsuri automate la întrebări sau probleme frecvent întâlnite.

Obiectivul proiectului pentru anul 2016 este:

1. **Prelucrarea datelor**.
   Pentru îndeplinirea acestui obiectiv, au fost planificate următoarele activităţi:
   Activitate 1.1***: accesarea şi pre-procesarea datelor***. În scopul realizarii acestei activităţi următoarele acţiuni au fost planificate şi realizate:
   - Acces la date şi determinarea specificaţiilor
   - Extragerea datelor şi pregătirea acestora pentru procesare
   - Analiza statistică şi cantitativă a datelor

   Activitate 1.2: ***Proiectare software***. Principala acţiune desfăşurată în cadrul acestei activităţi a fost analiza iniţială şi evaluarea unui algoritm de topic modelling.

Rezultatele activitatii in cadrul proiectului in perioada 1 octombrie – 20 decembrie 2016

Rezumatul etapei

Prima etapa a proiectului nostru s-a concentrat pe prelucrarea datelor partenerului proiectului, Pluriva SRL. Aceasta este o companie de produse și servicii informatice înființată acum aproape 20 de ani, al carei principal principal produs este un sistem informatic integrat ce acoperă toate fluxurile de business necesare operării unei afaceri. Pluriva ERP are peste 50 de module, grupate în 11 categorii. Pluriva SRL asigura atat servicii, cat si suport tehnic post implementare. Pentru acest lucru, Pluriva a dezvoltat un sistem de achizitionare prin email a cererilor de la utilizatori impreuna cu solutiile identificate si oferite de specialistii proprii la capatul unui intreg lant de activitati. In cei 5 ani de implementare, au fost achizitionate un numar semnificativ de cereri si raspunsuri. Din practică, s-a observat că utilizatorii formulează de multe ori întrebări în limbaj natural, a căror rezolvare este deja descrisă de multe ori într-un raspuns (tichet) deja existent și stocat în baza de date. Pentru a obține o reacție mult mai rapidă la întrebările utilizatorilor, a fost identificata nevoia unei aplicatii care să sugereze automat o posibilă rezolvare, selectând dintre tichetele deja existente. Acest lucru face obiectivul principal al proiectului nostru.

Pentru a aborda acest lucru, în prima fază a proiectului ne-am concentrat în mod natural pe analiza datelor (întrebări-răspunsuri) existente în baza de date a firmei. In acest scop am primit acces din partea firmei la datele stocate, pe care le-am exportat in scopul analizei intr-un format specific care să poată fi mai ușor de procesat. Avem în vedere faptul că o parte din aceste date sunt reprezentate fie în format HTML extras ca email fie în format text simplu. Așadar, am fost nevoiți să aplicăm un proces de curățare pentru înlăturarea marcatorilor specifici de HTML, lucru care include tag-uri, cod CSS chiar și script JavaScript. Acest proces este esențial pentru a asigura calitatea datelor și pentru a ne asigura că nu includem în analiza cantitativă elemente care nu țin de conținutul în sine.

La o primă vedere, aceste date parcurg un mare spectru stilistic, utilizatorii acoperind o arie largă de probleme iar descrirea problemelor este adesea exprimată folosind termeni tehnici sau neologisme din limba engleză. Răspunsurile primite de la specialiști se regăsesc, însă, într-un registru stilistic restrâns, acest lucru se datorează atât faptului că numarul celor care au răspuns este semnificativ mai mic cât și faptului că există răspunsuri similare pentru întrebări adresate în mod diferit. Din datele puse la dispozitie de firma, am lucrat in prima faza cu un esantion reprezentativ de aproximativ 20000 de perechi întrebare-răspuns, pe care, în această primă etapă, le-am investigat în special prin prisma analizorilor lexicali și cantitativi. Toate tehnicile dezvoltate aici sunt extensibile la intreaga colectie de date precum si la date noi care vor aparea in baza de date a firmei.

Varietatea lexicală se măsoară printr-o metrică rezultat al împărțirii dintre numărul de cuvinte unice și numărul total de cuvinte – type/token ratio logaritmat. În caz de varietate lexicală ridicată, metrica are valori apropiate de 0, caz în care fiecare cuvânt este folosit o singură dată - lucru aproape imposibil în practică pentru texte de dimensiune mare. Valorile mai mici decât 0 indică varietate lexicală scăzută, tabelul următor indică cu claritate diferențele cantiative la nivel comparativ între cereri și răspunsuri.

	Cereri	**Răspunsuri**

Nr. de tipuri	24,220	14,761
Nr. de cuvinte	799,150	282,965
Varietate lexicală	-3.496	-2.953
Lungime medie a textului	735.21	420.25
Lungime medie a cuvintelor	4.913	4.798
Nr. de cuvinte funcționale	180,409	82,850

Observăm că varietatea lexicală este mai scăzută în cazul cererilor decât în cazul răspunsurilor, lucru care indică faptul că agenții care răspund cererilor se folosesc de un limbaj mai diversificat decât utilizatorii sau persoanele care au nevoie de suport tehnic. Deși utilizatorii folosesc cu mult mai multe cuvinte pentru a exprima o cerere, răspunsurile agenților sunt de cele mai multe ori scurte. După cum indică tabelul, lungimea medie în cuvinte a unei cereri este de aproximativ 730 de cuvinte pe când răspunsurile au o lungime medie de 420 de cuvinte.

Varietatea lexicală acoperă cuvintele de conținut precum verbe, substantive, adjective etc. Pentru o comparație la nivel stilistic între cele două tipuri de texte, putem să urmărim cuvintele care nu au sens de sine stătător – conjuncții, preopoziții, adverbe, etc. numite și cuvinte funcționale. Pentru aceasta ne-am folosit de o listă de cuvinte deja existentă pe care am extins-o cu cuvinte specifice datelor de la partener incluzând echivalentul cuvintelor originale cu diacriticele înlăturate, acronime (e.g., www) și prescurtări (e.g., pt., nr., tel., dna). Cuvintele funcționale sunt indicatori de complexitate sintactică și au fost adesea folosiți în literatură ca trăsături de stil. Dacă privim raportul dintre numărul total de cuvinte și numărul de cuvinte funcționale pentru cele două grupuri, observăm, că cererile au raport crescut de cuvinte funcționale (4.429) comparativ cu răspunsurile (3.415). Drept concluzie putem spune că cererile au o complexitate stilistică mai mare decât răspunsurile la nivel sintactic, dar că răspunsurile au o varietate lexicală mai ridicată comparativ cu cererile.

Înlăturarea cuvintelor funcționale este un pas esențial pentru restrângerea spațiului de căutare și creșterea acurateții metodelor de topic modelling. Am studiat posibilitatile de extragere automată a subiectelor din corpus folosind o serie de parametri pentru topic modeling. În urma investigațiilor am constatat că un număr de 50 de topics (subiecte) pot fi extrase optim fără a se suprapune excesiv. Un subiect reprezintă o distribuție peste un vocabular de cuvinte, iar un text poate fi descris ca o mixtură de distribuții de subiecte. Drept algoritm de topic modelling am folosit Latent Dirichlet Allocation – o metodă generativă care este și eficientă față de timpul de rulare și oferă în general rezultate bune. Am antrenat metoda pe o variantă a corpusului care nu conține

punctuație, cuvinte funcționale din lista extinsă, numere sau cuvinte de lungime 1 (o literă). Algoritmul face multiple iterații peste colecția de date, iar textele sunt grupate în bucăți de 1000 de documente pentru optimizarea calculului la fiecare pas.

Câteva astfel de subiecte sunt redate mai jos împreună cu ponderile pentru fiecare cuvânt:

- 1 - 0.114*import + 0.033*facturare + 0.030*dobre + 0.025*mobil + 0.023*omv + 0.023*unele + 0.023*mail + 0.022*delta + 0.021*petrom + 0.019*marketing
- 2 - 0.055*this + 0.046*the + 0.046*and + 0.038*you + 0.030*mail + 0.028*any + 0.023*not + 0.023*to + 0.021*intended + 0.020*of
- 4 - 0.068*unicomp + 0.049*mariana + 0.033*marfa + 0.026*intrare + 0.024*gestiuni + 0.022*contabilitate + 0.022*retur + 0.021*vedea + 0.020*pv + 0.020*transfer'
- 15 - 0.163*aviz + 0.146*client + 0.064*print + 0.053*normal + 0.052*nota + 0.046*predare + 0.038*romania + 0.035*rom + 0.034*seen + 0.028*auchan'
- 23 - 0.043*contractului + 0.034*clienti + 0.029*nevoie + 0.027*martie + 0.020*raport + 0.018*suplimentare + 0.018*randder + 0.017*adaugati + 0.017*persoana + 0.015*introducerea

Dacă unele din aceste subiecte indică operații financiare, contacte sau legături cu terți, putem totuși observa în vocabularul subiectului 2 se găsesc doar cuvinte din limba engleză, cuvinte care au fost menționate în cadrul schimburilor de mailuri dintre clienții care au executat cererea și agenți.

În perioada care vine urmează să lucrăm la optimizări adiționale ale parametrilor algoritmilor, incluzând și un clasificator de întrebări bazat pe utilizarea metricii rank distance in agregarea multicriteriala, metoda introdusa de directorul de proiect. Toate aceste rezultate vor fi folosite in etapele urmatoare ale proiectului.

Activitatea științifică

În anul 2016 activitatea științifică a membrilor proiectului s-a concentrat în special pe analiza și identificarea trăsăturilor determinante cu potențial impact în găsirea unui model de limbă care să ajute la identificarea automată a topicii textului. În paralel au fost proiectați algoritmi care să ajute la identificarea limbii, în special la discriminarea limbilor puternic apropiate. Sunt în curs de analiză tehnici de identificare a stilului și grupare a intrebărilor nu numai pe topici, dar și pe autori.

Într-o etapă relativ scurtă, membrii proiectului au reușit să publice două lucrări la două workshop-uri asociate cu una din cele mai importante conferințe dedicate procesării limbajului natural și lingvisticii computaționale, și anume conferința Coling 2016.

Prima lucrare acceptată (*Vanilla Classifiers for Distinguishing between Similar Languages*, autori A. Ciobanu, S. Nisioi, L.P. Dinu) prezintă rezultatele participării autorilor la a treia ediție a DSL Shared Task (discriminating between similar languages) și articolul descrie rezultatele și metodologia aplicată. Această competiție a avut 2 probe principale (împărțite, la rândul lor, în două subprobe): prima proba cerea să se diferențieze între limbi puternic apropiate (franceza canadiană de cea din hexagon, portugheza braziliana de cea din Portugalia, bosniaca de sârbă și croată,

spaniola din Mexic de cea din Argentina sau de cea din Spania, Indoneza de Malay) și a doua probă se referea la identificarea dialectelor arabe.

Echipa a participat la toate probele și a câștigat un loc 1 (discrimnarea dintre limbi apropiate) si doua locuri 2 (identficarea dialectelor arabe și discrimnarea limbilor apropiate, testul B2).

Al doilea articol publicat A Visual Representation of Wittgenstein's *Tractatus Logico-Philosophicus* (autori A. Bucur și S. Nisioi) are în vedere descrierea unei metode de vizualizarea a datelor pe baza similarităților la nivel textual - http://tractatus.gitlab.io/. Metoda este aplicată în mod particular pe un corpus format din textele filozofice ale lui L. Wittgenstein, dar este aplicabilă pe orice set de date de tip text pentru vizaulizarea și reprezentarea acestora într-un spațiu de căutare bidimensional.

De asemenea, în acest an a fost început în cadrul proiectului un seminar internațional bilunar de lingvistică matematică și computațională, pe parcusrul căruia s-au ținut 7 prezentări. Printre cei care au ținut până acum prezentări, se numără: Andrea Sgarro (Universita di Trieste), Mihaela Balint (Facultatea de Automatica, UPB), Traian Rebedea (Facultatea de Automatica, UPB), Virginio Cantoni (Universita di Pavia), Walther von Hahn (University of Hamburg), Cristina Vertan (University of Hamburg), Grzegorz Kowalski (Institute of Applied Linguistics, University of Warsaw). În cadrul prezentărilor participă cu regularitate atât membrii proiectului, cât și masteranzi, doctoranzi sau postdoctoranzi din Departamentul de Informatică, din alte departamente ale Universității din București, din cadrul Facultății de Automatică a Universității Politehnica din București, membri ai partenerului Pluriva, dar și alți colegi din industrie interesați de tematica abordată. Considerăm că prin această activitate aducem un beneficiu proiectului prin activitatea de diseminare, si, în același timp realizam noi punți de comunicare între mediul academic și industrie.

A fost realizată o pagina web a proiectului (http://nlp.unibuc.ro/projects/sigar.html), astfel încât toate rezultatele așteptate la finalul acestei etape (pagina web, articol trimis spre publicare, raport, analiza datelor) au fost realizate.

Activități administrative și de organizare

Membrii echipei s-au întâlnit periodic pentru a prezenta și discuta aspecte legate de organizare, de repartizarea temelor de cercetare, de colaborare între membrii echipei, de stabilirea conferințelor și jurnalelor în care se trimit spre publicare rezultatele cercetării și de convenirea asupra unui calendar comun cu respectarea termenelor de atingere a obiectivelor proiectului și de livrare a rezultatelor. Intre membrii proiectului si partenerul economic a fost o continua comunicare.

Au fost îndeplinite astfel toate activitățile administrative necesare bunei desfășurări a proiectului.

Activități de diseminare

În perioada scurtă a acestei faze, membrii proiectului au publicat 2 articole științifice în două workshop-uri asociate unei conferințe prestigioase (COLING 2016, categorie A conform

ERA-core) și au organizat un seminar bilunar cu participanți atât din țară cât și din strainatate (2 români, 2 italieni, 2 germani și 1 polonez). De asemenea, a fost întreținută și actualizată pagina oficială a proiectului.

Concluzii

- Obiectivele etapei au fost substanțial realizate.
- Este necesară continuarea cercetării fără modificări în planul curent de realizare.

Publicații și conferințe susținute de către membrii proiectului în anul 2016:

1. Alina Maria Ciobanu, Sergiu Nisioi, Liviu P. Dinu. *Vanilla Classifiers for Distinguishing between Similar Languages.* In Proc. Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3), co-located with COLING 2016, december 11-16, 2016, Osaka, Japan, pages 235-242.

2. Anca Bucur, Sergiu Nisioi. *A Visual Representation of Wittgenstein's Tractatus Logico-Philosophicus*. In Proc. Language Technology Resources and Tools for Digital Humanities (LT4DH), co-located with COLING 2016, december 11-16, 2016, Osaka, Japan, pages 71-75, ISBN978-4-87974-708-2

Director Proiect,
Prof. dr. Liviu P. Dinu