TEZĂ DE ABILITARE

REZUMAT

FLORENTINA HRISTEA

UNIVERSITATEA DIN BUCUREŞTI

FACULTATEA DE MATEMATICĂ ŞI INFORMATICĂ

DEPARTAMENTUL DE INFORMATICĂ

2016

Rezumat

Teza de faţă se referă la întreaga noastră activitate de cercetare postdoctorală. Ea se bazează însă, în mod special, pe acea parte a cercetării noastre care s-a concentrat asupra rolului dezambiguizării automate a sensului cuvintelor în îmbunătăţirea rezultatelor din regăsirea informaţiei. Întrebarea la care ne-am propus să răspundem este următoarea: poate *dezambiguizarea sensului cuvintelor* ("word sense disambiguation" - ***WSD***) să îmbunătăţească rezultatele obţinute în *regăsirea informaţiei* ("information retrieval" - ***IR***) ? Această întrebare a captat, cu predilecţie, interesul nostru ştiinţific în ultimii ani şi a generat o cercetare de amploare şi de natură interdisciplinară ale cărei rezultate originale sunt prezentate în capitolele tezei de faţă.

Capitolul final al tezei de abilitare va arăta modul în care această temă de cercetare îşi găseşte locul în cadrul activităţii noastre de cercetare postdoctorală globale.

Problema menţionată a fost îndelung studiată şi dezbătută în literatura domeniului IR, dar cu rezultate controversate. Şi cu o largă majoritate de răspunsuri negative la întrebarea enunţată anterior. WSD a fost unanim recunoscută ca fiind utilă în multiple tipuri de aplicaţii ale diverselor subdomenii din inteligenţa artificială. WSD în IR a fost pusă sub semnul întrebării şi declarată ca fiind ineficientă. Sperăm că această cercetare va duce la schimbarea perspectivei sceptice care există, în momentul de faţă, asupra acestei probleme.

Performanţa sistemelor de IR este cunoscută ca fiind puternic dependentă de interogări. În mediul motoarelor de căutare, utilizatorii trimit interogări ("queries") în concordanţă cu nevoile lor informaţionale. Ca răspuns la o interogare, sistemul regăseşte şi afişează o listă de documente pe care le consideră relevante şi de interes pentru utilizator, din perspectiva interogării efectuate. Utilizatorul analizează rezultatele regăsite şi decide care dintre ele sunt utile nevoilor sale informaţionale. Interogările pentru care motorul de căutare nu este capabil să furnizeze informaţii relevante sunt considerate dificile ("difficult queries").

Dificultatea interogărilor poate fi legată de un număr de cauze total diferite: ambiguitate, formulare neclară sau prea vagă, lipsa contextului, natura şi structura colecţiei de documente etc. Ambiguitatea termenilor a fost identificată ca reprezentând o cauză majoră a dificultăţii unei interogări. Cercetarea noastră s-a concentrat asupra acelor interogări considerate dificile datorită ambiguităţii cuvintelor care intervin în structura lor.

În pofida faptului că procesul de IR se bazează, în general, pe suprapunerea (coincidenţa) termenilor din interiorul interogărilor şi respectiv al documentelor, fără a fi luat în consideraţie sensul cuvintelor, abordarea noastră s-a bazat pe semantica indusă atât de interogări, cât şi de documentele regăsite. Termenii care intervin nu au mai fost consideraţi ca fiind independenţi, ci ca fiind, mai degrabă, înrudiţi. Principala presupunere pe care s-a bazat abordarea noastră este, prin urmare, aceea conform căreia *contextul* poate îmbunătăţi performanţa sistemelor de IR.

În mod intuitiv, o procesare a textelor mai bogată, care ţine seama de multiple aspecte lingvistice, ar trebui să conducă la o îmbunătăţire a rezultatelor în IR, atât în cazul interogărilor dificile, în general, cât şi în cel al interogărilor ambigue, în particular. De-a lungul timpului, tehnicile specifice domeniului procesării limbajului natural au fost utilizate, în diferite moduri, în IR. Cu toate acestea, în pofida eforturilor depuse, nu au fost constatate îmbunătăţiri semnificative ale eficienţei în IR, mai ales în cazul evaluărilor la scară largă. Suntem de părere că motivul acestui eşec îl constituie faptul că tehnicile procesării limbajului natural nu ar trebui folosite în cazul tuturor interogărilor, ci numai în cazul acelor interogări care necesită, în mod efectiv, o tratare (prelucrare) de acest fel. Propriul nostru studiu şi-a propus să identifice cuvintele ambigue din structura unei interogări şi să decidă ce tehnică de tip WSD trebuie aplicată în cazul punctual al respectivei interogări.

După cum se ştie, deşi procesul de WSD este, în general, uşor pentru subiecţii umani, el reprezintă o adevărată încercare pentru soft. Problema devine şi mai greu de rezolvat în mod automat atunci când un cuvânt ambiguu intervine în cadrul unei "bucăţi" scurte de text, cum este cazul interogărilor dintr-un sistem de IR.

Multe studii argumentează eşecul îmbunătăţirii performanţei în IR prin folosirea tehnicilor de WSD ca fiind datorat ineficienţei algoritmilor de dezambiguizare existenţi, din nou o problemă care se accentuează în cazul fragmentelor scurte de text reprezentând o interogare.

În literatură există numeroase încercări de a se folosi WSD bazată pe cunoştinţe în IR. Acele studii care au raportat rezultate pozitive se caracterizează, însă, prin testări efectuate pe eşantioane mici de date. Atunci când evaluarea a fost adusă la scara unei colecţii de testare ample, îmbunătăţirile erau raportate relativ la un nivel de referinţă ("baseline") slab, aşa cum se comentează, în detaliu, în Cap. 7 al tezei de faţă. În general, ori de câte ori au fost raportate concluzii şi rezultate pozitive, cuantumul de îmbunătăţiri obţinute a fost mic (a se vedea Cap. 7).

Puţinii autori care, mai recent, au exprimat o încredere crescândă în beneficiile aduse de WSD pentru IR, folosesc, cu toţii, o tehnică de WSD supervizată (a se vedea Cap.7 pentru detalii). Este, însă, cunoscut faptul că WSD supervizată nu poate fi folosită pe scară largă în practică, datorită absenţei corpusurilor adnotate / paralele pe care le necesită.

Contrar opiniei tuturor acestor autori, cercetarea noastră propune şi investighează utilizarea unei tehnici WSD nesupervizate în IR. În situaţia în care principala noastră presupunere este aceea conform căreia contextul poate îmbunătăţi performanţa unui sistem de IR, am prezentat în lucrările noastre o abordare care doreşte să identifice cluster-e de contexte similare, cu fiecare cluster desemnând un cuvânt polisemantic folosit printr-un anumit sens al său. Suntem de părere că IR reprezintă o aplicaţie pentru care acest tip de analiză este folositor şi, mai mult, suficient pentru mărirea performanţei. Prin urmare, abordarea noastră nu îşi propune o dezambiguizare directă a sensului, ci o discriminare între sensurile posibile ale unui cuvânt ambiguu care intervine într-o interogare, cu impact în IR.

În vederea pregătirii utilizării tehnicilor de WSD în IR, studiul nostru s-a concentrat asupra a două aspecte fundamentale: tipul de metodă de clustering care ar trebui utilizat pentru WSD nesupervizat şi tipul de cunoştinţe care ar trebui furnizate algoritmului de WSD (lipsit de cunoştinţe) rezultat. Studiile noastre preliminare ne-au încurajat să urmăm ambele linii de investigaţie menţionate, cu scopul testării referitor la interogările ambigue în contextul unei aplicaţii de IR.

Pentru a studia rezultatele dezambiguizării nesupervizate, am luat în consideraţie două tehnici de clustering complet diferite, care au fost testate cu şi fără selecţie de caracteristici. În prezentarea din această teză, parametrii modelului matematic utilizat pentru WSD vor fi întotdeauna formulaţi prin tehnici nesupervizate. Vom compara performanţele algoritmului lipsit de cunoştinţe rezultat în prezenţa şi respectiv în absenţa efectuării unei selecţii de caracteristici bazate pe cunoştinţe. Cercetarea noastră a investigat câteva surse de cunoştinţe de naturi total diferite, considerate adecvate pentru a se realiza selecţia caracteristicilor (a se vedea Cap.3-6), precum şi posibilitatea combinării unora dintre ele (a se vedea Cap. 5).

În fine, ca urmare a acestor demersuri, am propus (2015) o nouă metodă nesupervizată care utilizează WSD (prin discriminare între sensuri) în IR (a se vedea Cap. 7). Metoda pe care am dezvoltat-o se bazează pe spectral clustering şi reordonează o listă de documente regăsite iniţial prin "împingerea" la începutul listei a acelor documente care sunt semantic similare cu interogarea. Testele efectuate de noi corespunzător câtorva colecţii TREC ad-hoc au demonstrat (a se vedea Cap. 7) că metoda noastră este utilă în cazul interogărilor ce contin cuvinte ambigue.

Interesul nostru s-a concentrat asupra îmbunătăţirii nivelului de precizie după 5, 10 şi 30 de documente regăsite (P@5, P@10 şi respectiv P@30), acestea fiind situaţiile cele mai utile aplicaţiilor practice din lumea reală. Am arătat (a se vedea Cap. 7) că, în aceste situaţii, precizia poate fi ridicată cu cel puţin 7,9% deasupra pragurilor de referinţă state of the art.

Drumul care a condus la naşterea acestei metode ce utilizează WSD în IR nu a fost unul uşor[1]. Investigaţia noastră a început (2008) cu utilizarea unui algoritm pentru WSD nesupervizat lipsit de cunoştinţe şi care se baza pe clasicul model Bayesian naiv ca tehnică de clustering. Alegerea a fost motivată de credinţa noastră că potenţialul acestui model statistic, cu privire la WSD nesupervizat, rămăsese insuficient explorat la acea vreme. Cu alte cuvinte, am considerat că dezambiguizarea nesupervizată nu exploatase încă la maximum posibilităţile oferite de acest model statistic simplu, dar puternic. În egală măsură, am fost animaţi de credinţa că este necesar ca modelului Bayes naiv să-i fie furnizate cunoştinţe pentru ca acesta să poată acţiona în mod eficient ca tehnică de clustering pentru WSD nesupervizat. Această din urmă convingere ne-a determinat să plasăm procesul de dezambiguizare semantică la graniţa dintre tehnicile nesupervizate şi cele bazate pe cunoştinţe şi, în cele din urmă, să desfăşurăm întregul nostru demers ştiinţific la această graniţă. Într-adevăr, studiile noastre (2008 – 2013) au arătat că, un

[1] Şi a inclus 7 articole ştiinţifice de revistă, un articol de conferinţă şi o carte, pentru care am fost autor sau coautor în perioada 2008-2015.

algoritm de dezambiguizare simplu, de bază şi lipsit de cunoştinţe, în cazul nostru reprezentat de modelul Bayes naiv, poate acţiona destul de bine atunci când este alimentat cu cunoştinţe în mod adecvat. Aceste cunoştinţe pot fi furnizate în diferite moduri şi pot fi de diverse tipuri. În lucrările noastre am examinat trei surse complet diferite de astfel de cunoştinţe: WordNet (a se vedea Cap. 3), relaţii de dependenţă (a se vedea Cap. 4) şi web N-grame (a se vedea Cap. 5).

Rezultatele pozitive raportate în literatura domeniului IR (2012) în urma folosirii selecţiei de caracteristici bazate pe WordNet, propusă de către noi pentru modelul Bayesian naiv, în cadrul unei aplicaţii de IR (a se vedea Cap. 3), ne-au încurajat să continuăm discuţia cu privire la selecţia caracteristicilor în cazul acestui model statistic. În mod succesiv am studiat şi propus selecţie de caracteristici sintactice bazate pe relaţii de dependenţă şi selecţie de caracteristici bazate pe N-grame culese de pe web, acestea din urmă fiind utilizate pentru prima oară, în WSD nesupervizat, de către noi (2012). Rezultatele noastre cu privire la selecţia caracteristicilor pentru modelul Bayesian naiv, folosit în dezambiguizarea nesupervizată, au fost publicate de Editura Springer (Hristea 2012)[2].

Pasul următor al studiului nostru a fost acela de a lua în consideraţie o tehnică de clustering total diferită, "state of the art", reprezentată prin spectral clustering (a se vedea Cap. 6). Alegerea noastră a fost determinată nu numai de faptul că aceasta este o tehnică de clustering modernă şi cunoscută ca fiind puternică, performantă, ci şi de faptul că ea nu necesită în mod esenţial selecţie de caracteristici. Din contră, testele noastre cu privire la spectral clustering în prezenţa şi respectiv în absenţa selecţiei de caracteristici bazate pe WordNet (a se vedea Cap. 6) arată în mod clar faptul că această tehnică de clustering avansată lucrează mai bine atunci când îşi realizează propria ponderare a caracteristicilor. Spectral clustering este folosit de noi pentru prima oară în WSD şi, relativ la seturile de date pentru care s-a făcut testarea, furnizează cele mai bune rezultate în dezambiguizare. Acest fapt ne-a determinat să implementăm spectral clustering (în prezenţa propriilor caracteristici) în cadrul unei aplicaţii de IR (2013 – 2015; a se vedea Cap. 7).

Astfel, în finalul acestui amplu studiu, ne-am întors la aplicaţia de tip IR descrisă în Cap.3 şi am utilizat WSD nesupervizat realizat cu spectral clustering relativ la aceeasi problemă de IR. Succesul nostru (precizie îmbunătăţită cu cel putin 7,9% deasupra pragurilor de referinţă state of the art) deschide o întreagă linie de investigaţie viitoare şi câteva noi şi promiţătoare perspective în această cercetare (aşa cum se descrie în Postfaţa lucrării).

Încă de la începutul acestei cercetări, scopul nostru nu a fost acela de a câştiga o competiţie de WSD, ci acela de a furniza o metodă (nouă) de dezambiguizare care ar putea ajuta la câştigarea unei competiţii de IR. Din această cauză, testele noastre privitoare la modelul Bayesian naiv implică seturi de date relativ restrânse. În acel stadiu al studiului nostru eram interesaţi doar să

[2] Hristea, F.: The Naïve Bayes Model for Unsupervised Word Sense Disambiguation. Aspects Concerning Feature Selection. SpringerBriefs in Statistics Series, Springer (2012).

"simţim" reacţia acestui model matematic la caracteristici de naturi complet diferite, precum şi la posibile combinaţii ale lor.

Seturile de date implicate în testări au fost alese de noi, în egală măsură, pentru a facilita comparaţia cu *singura* abordare similară (selecţia caracteristicilor pentru modelul Bayesian naiv antrenat cu algoritmul EM) existentă în literatura de specialitate[3] (a se vedea Cap. 1). Acelaşi mediu de testare a fost apoi organizat pentru spectral clustering, tot în scopul facilitării comparaţiilor. Din contră, testele pe care le-am organizat în cazul aplicaţiei de tip IR finale utilizează colecţiile de referinţă pentru evaluare, care există şi au fost create în mod special pentru task-uri de IR[4]. S-a folosit unul dintre cele mai importante seturi de colecţii de testare (a se vedea Cap. 7), care este utilizat pe scară largă de mediile academice şi care provine de la Text REtrieval Conference (TREC) - principalul cadru de evaluare în IR.

La momentul studiilor noastre iniţiale, literatura domeniului IR afirmase deja (Stokoe, Oakes şi Tait 2003)[5] că singura problemă rămasă în dezbatere este dacă o acurateţe în dezambiguizare mai mică de 90% poate conduce la îmbunătăţiri în eficacitatea regăsirii informaţiei, un rezultat despre care suntem de părere că nu poate fi atins de către dezambiguizarea nesupervizată. Prin urmare, încă de la început, ne-am confruntat cu o adevărată provocare ştiinţifică: aceea de a accepta sau a nu accepta părerea predominantă existentă conform căreia WSD nu poate îmbunătăţi performanţa în IR. Am pus sub semnul întrebării, cu tărie, acest punct de vedere al specialiştilor şi ne-am propus să demonstrăm că *tipul de WSD* (nesupervizat) folosit în IR este esenţial, precum şi *tehnica de clustering* pe care acesta o presupune.

Cu excepţia primelor două capitole, introductive şi tratând cadrul general al cercetării, fiecare capitol al acestei teze conduce la o concluzie ştiinţifică distinctă, de sine stătătoare şi, din această cauză, este conceput ca fiind independent. Cu toate acestea, numai parcurgerea tuturor capitolelor, în ordinea lor de aici, poate sugera şirul evenimentelor ştiinţifice din cariera noastră care ne-au condus la succes în cazul aplicaţiei de tip IR finale. Ultimul capitol al tezei arată modul în care această temă de cercetare îşi găseşte locul în cadrul cercetării noastre postdoctorale globale.

---

[3] Ne exprimăm gratitudinea faţă de Profesorul Ted Pedersen de la Universitatea din Minnesota, Duluth, S.U.A., care ne-a furnizat seturile de date necesare pentru efectuarea testelor şi a comparaţiilor prezentate, referitor la adjective şi verbe.

[4] Accesul la datele de testare specifice domeniului IR a fost asigurat de către partenerul nostru francez pentru acest studiu, IRIT (Institut de Recherche en Informatique de Toulouse).

[5] Stokoe, C., Oakes, M.P., Tait, J.: Word sense disambiguation in information retrieval revisited. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03), pp. 159-166. Toronto, Canada (2003).