UNIVERSITATEA POLITEHNICA TIMIŞOARA
FACULTATEA DE AUTOMATICĂ ŞI CALCULATOARE

REZUMAT TEZĂ DE DOCTORAT

cu titlul:

Improving text accessibility and understanding of domain-specific information

(Îmbunătăţirea accessibilităţii textului şi a înţelegerii informaţiei de specialitate)

Autor: Vasile TOPAC

Conducător ştiinţific: prof. dr. ing. Vasile STOICU-TIVADAR

Teza acestei lucrări (ideea de bază):
Adaptarea textelor de specialitate prin identificarea şi explicarea terminologiei (întâlnită atât în formă canonică cât şi în formă derivtă), împreună cu punerea la dispoziţie a unor mijloace de a schimba aspectul textului, pot duce la o accesibilitate şi înţelegere mai bună a mesajului pentru utilizatorii de rând.

Rezumat

Această teză explorează căi prin care se poate îmbunătăţii accessibilitatea şi înţelegerea informaţiei textuale pentru utilizatorii de rând, luându-se în considerare diferite niveluri de limitare ale accesului (limbă, limbajul specializat şi modul de prezentare <aspectul>).

Cea mai mare atenţie este acordată dificultăţii de înţelegere a textului datorate limbajului de specialitate, alegându-se ca şi domeniu de aplicare limbajul medical. Metode şi instrumente pentru recunoaşterea şi explicarea terminologiei medicale, chiar şi atunci când aceasta apare în formă derivată sunt prezentate. Recunoaşterea termenilor în limbajul natural e bazată pe tehnici de tipul *fuzzy matching* combinate cu structuri de date specializate, dezvoltate de către autor. Tehnici pentru îmbunătăţirea preciziei recunoaşterii aproximative (*fuzzy matching*) bazate atat pe metode semi-supervizate sau autonome cât şi pe metode ce au la bază inteligenţa umană (de exemplu platformele de *crowdsourcing*) sunt prezentate. Aceste metode de adaptare a textelor de specialitate sunt implementate în câteva scenarii de utilizare, rezultând o serie de instrumente. Impactul procesului de adaptare a limbajului medical a fost evaluat prin studii cu utilizatori, rezultatele confirmând utilitatea acestor metode şi a serviciilor asociate.

În plus, mijloace de adaptare a modului de prezentare a textului (aspect) sunt explorate. Atât înstrumente pentru adaptarea aspectului textului în general, cât şi instrumente dedicate unor anumite tipuri de utilizatorilor (cum ar fi utilizatori cu disexie sau vedere slabă) sunt prezentate.
Toate instrumentele dezvoltate în această lucrare au fost proiectate acordânduse atenţie deosebită unor factori precum accessibilitate, utilizabilitate şi disponibilitate, ele fiind concepute să ruleze oriunde în mediul online, cu efort minimal.

Obiective

Această teză urmăreşte următoarele obiective:

- Creşterea gradului de **înţelegere a mesajului din textele de specialitate** pentru utilizatorii de rând.
  Obiective derivate din acesta:
  - Recunoaştera cât mai precisă a termenilor de specialitate în limbajul natural
  - Îmbunătăţirea traducerii textelor de specialitate prin adnotarea terminologiei
- **Îmbunătăţirea accessibilităţii şi a lizibilităţii textelor** prin dezvoltarea de metode şi instrumente de adaptare a aspectului textului
  - Dezvoltarea de instrumente specializate pentru uşurarea procesului de citire la persoanele cu dislexie
- Proiectarea serviciilor de adaptare a textului amintite mai sus astfel încât acestea să prezinte un grad ridicat de accesibilitate, utilizabilitate şi disponibilitate. De asemena definirea unui model de accesibilitate universală a textului, şi încadrarea serviciilor dezvoltate în acest model.

Structura tezei (capitole)

1. **Introducerea**
   În această secţiune autorul prezintă motivaţia şi actualitatea temei abordate: *accessibilitatea informaţiei textuale*. Diferitele nivele de limitare a accessului la informaţia textuală sunt prezentate, acordânduse o atenţie specială limbajului specializat şi nivelului de prezentare. Sunte prezentate apoi obiectivele acestei teze, urmate de o trecere în revistă a principalelor metode şi instrumente dezvoltate.

2. **Stadiul curent**
   Stadiul curent al cercetărilor atât pe parte de limbaj specializat cât şi la nivel de prezentare e expus. Ceretearea existentă e încadrată într-un model de accessibilitate universală a textului propus de autor. Sunt prezentate şi servicii funcţionale (din industrie) ce au ca şi ţintă subiecte comune cu această teză. Pe lângă studiile şi proiectele de cercetare prezentate în acest capitol, fiecare din capitolele ce urmează prezintă mai în detaliu aspecte ale stadiului actual pentru respectivul subiect.

3. **Limbaj specializat – studiu de limbaj şi recunoaştere aproximativă (fuzzy matching)**
   Autorul prezintă un studiu al limbajului medical pentru limba română şi engleză cu sopul identificării ratei de apariţie a terminologiei de specialitate în forma canonică comparat cu forma derivată. Studiul relevă o incidenţă mare a terminologiei în forma derivată (mai mare în limba română decât în engleză), ceea ce confirmă nevoia de a folosii tehnici de recunoaştere aproximativă a termenilor. Apoi este prezentată structura de date proiectată de către autor, *FuzzyHashMap*, ca o soluţie pentru identificarea rapidă a termenilor în limbajul natural. Aceasta este o extensie a structurii *HashMap* din limbajul Java. O serie de teste care ilustrează performanţa stucturii sunt apoi prezentate, împreună cu prezentarea altor scenarii în care structura a fost deja folsită.

4. **Limbaj specializat – îmbunătăţirea preciziei recunoaşterii aproximative**
   Autorul vorbeşte în acest capitol despre faptul că, pe lângă beneficiile ce le aduce recunoaşterea aproximativă, există şi efecte secundare datorate naturii tolerante la erorare a acestor tehnici, şi anume recunoaşteri greşite, sau *fals positive*. Autorul prezintă apoi o serie de

tehnici ce au ca şi scop creşterea preciziei recunoaşterii aproximative. Sunt prezentate atât metode ce au la bază tehnici semi-supervizate sau autonome cât şi tehnici ce folosesc inteligenţa umană colectivă. Tehnicile semi-supervizate includ a) antrearea sistemului pentru dezvoltarea unor dicţionare de mapări incorecte ce sunt ulterior folosite pentru filtrarea mapărilor şi b) generarea unui model al mapărilor aproximative şi derivarea unui tipar de hashing din acesta, pentru a fi integrat în structura *FuzzyHashMap*.
Tehnicile ce au la bază inteligenţa umană folosesc feedback agregat de la utilizatorii sistemului propus sau integrează folosirea platformelor de *crowdsourcing*. Teste cu utiliatori ai sistemului (făcute prin chestionare) şi teste cu persoane din spatele platformei de *crowdsourcing* Amazon Mechanical Turk sunt apoi prezentate. Imbunătăţirea preciziei este apoi măsurată prin folosirea metricilor specifice, iar rezulatele arată o îmbunătăţire semnificativă a acesteia.

5. Accessibilitatea limbajului specializat – scenarii de utilizare şi aplicaţii

Acest capitol prezintă o serie de scenarii de utilizare a serviciilor de adnotare a terminologiei din limbaje specializate. Următoarele instrumentele dezvltate de către autor sunt prezentate:

- *text4all terminology interpreter:* instrumentul de bază al acestei secţiuni, folosit pentru a explica terminologia medicală din pagini web existente cu scopul îmbunătăţirii înţelegerii mesajului şi a facilitării procesului de *patient empowerment*,
- *text4all ITS Tagger:* instrument de adnotare a terminologiei bazat pe standartul ITS 2.0 (ce are ca şi scop, printre altele, îmbunătăţirea traducerii textelor de specialitate).
- integrarea serviciului de adaptare a limbajului in proiectul de tele-asistenţă *TELEASIS*
- *text4all term analysis*: instrument de analiză a limbajului specializat din mediul online.

Impactului folosirii instrumentului *text4all terminology interpreter* pentru îmbunătăţirea înţelegerii textelor medicale a fost evaluat prin teste cu utilizatori reali. Testele au confirmat utilitatea acestui instrument. Alte aspecte şi implicaţii corelate cu adaptarea textelor cu specific medical sunt discutate.

6. Accessibilitatea textului la nivel de prezentare

În acest capitol autorul explorează limitările accesării (citirii) textului cauzate de modul de prezentare al acestuia, mai exact de aspectul textului. Elemente precum dimensiunea textului, culori, fonturi, spaţiere şi altele sunt luate în considerare. Instrumente ce permit adaptarea aspectului textului sunt prezentate, şi modul în care acestea pot fii utile diferitelor tipuri de utilizatori sunt explorate. Un instrument dedicat utilizatorilor cu dislexie, dezvoltat de către autor în colaborare cu cercetători specializaţi pe acest tip de probleme, pe baza unor studii ce au folosit tehnici de eye tracking făcute cu cititori cu dislexie, este prezentat. Alte instrumente şi aspecte ale adaptării textului la nivel de aspect sunt discutate şi analizate.

7. Aspecte asupra modului de proiectare a serviciilor propuse

Anumite aspecte ale modului de proiectare şi a modurilor de folosire şi interacţiune ale serviciilor dezvoltate sunt discutate şi analizate. O particularitate cheie a serviciilor e diponibilitatea acestora, ele fiind proiectate ca şi mediatori online (denumite şi proiecte de *transcoding*), funcţionează direct în browser, indepedent de platformă şi nu necesită instalare sau drepturi speciale. Diferite moduri de interacţionare cu instrumentele dezvoltate, precum interacţiunea prin interfaţa grafică sau interacţiunea direct prin adresa URL (aceasta fiind proiectată şi optimizată pentu o utilizabilitate cât mai sporită) sunt prezentate. Instrumentele dezvoltate sunt amplasate într-un model reprezentând o viziune de ansamblu a accesibilităţi informaţiei textuale. Alte aspecte legate de acest model sunt discutate.

8. Concluzii

E expusă o scurta trecere în revistă a celor enunţate în teză. Principalele contribuţii revendicate sunt listate, urmate de o prezentare a serviciilor rezultate, accentul punându-se pe contribuţiile aduse de aceste soluţii. Câteva direcţii de viitor sunt de asemenea prezentate.

Contribuţii

Contribuţii majore:

- Dezvoltarea unor metode şi instrumente asociate de recunoaştere şi explicare a termenilor de specialitate, cu scopul uşurării înţelegerii mesajului. *(Metodele sunt prezentate în capitolele 3 si 4 din teza, iar instrumentele asociate metodelor sunt prezentate în capitolul 5).*
- Proiectarea structurii de date *FuzzyHashMap* (extensie a structurii *HashMap* din limbajul Java) ce permite căutări aproximative de mare viteză. O confirmare în plus a utilităţii acestei structuri este dată de folosirea ei de către alţi cercetători în proiecte de bioinformatică sau information retrieval. *(prezentată în capitolul 3, secţiunea 3.3)*
- Definirea unui model de accessibilitate universală a textului şi ancorarea tuturor metodelor şi serviciilor dezvoltate în acest model. *(în capitolul 7, secţiunea 7.1)*

Alte contribuţii:

- Studii de limbaj pentru aflarea incidenţei terminologiei derivate. *(în capitolul 3, secţiunea 3.2)*
- Studiu asupra eficienţei folosirii platformelor tip *crowdsourcing* pentru validarea recunoaşterilor aproximative. *(în capitolul 4, secţiunea 4.5)*
- Proiectarea şi dezvoltarea serviciilor:
    - *text4all Terminology Interpreter* pentru adaptarea textelor de specialitate. *(în capitolul 5, secţiunea 5.2)*
    - *text4all ITS Tagger* pentru adnotarea terminologiei în formatul ITS 2.0 cu scopul îmbunătăţirii traducerii. *(în capitolul 5, secţiunea 5.4)*
    - *text4all DysWebxia* pentru adaptarea textului atât la nivel de prezentare cât şi la nivel de limbaj (sinonime pentru cuvinte dificile) pentru persoane cu dislexie. *(în capitolul 6, secţiunea 6.3)*
    - *text4all Customizer* pentru particularizarea aspectului textului din pagini web existente. *(în capitolul 6, secţiunea 6.2)*
    - *text4all analyzer* pentru analiza limbajelor de specialitare şi generarea de statistici referitoare la incidenţa terminologiei de specialitate în diferite forme. *(în capitolul 5, secţiunea 5.5)*