Oponentský posudek

na doktorskou disertační práci

Ing. Karla Zváry

Extrakce informací z lékařských textů

pro obor Biomedicínská informatika na 1. lékařské fakultě UK

školitel Doc. Ing. Vojtěch Svátek, PhD.

Oponent: Doc. Ing. Arnošt Veselý, CSc.

Česká zemědělská univerzita, Provozně ekonomická fakulta,

Katedra informačního inženýrství

Předložená práce je věnována problematice extrakce informací z českých lékařských záznamů psaných formou volného, nestrukturovaného textu. Velká část lékařské dokumentace je formou volného textu napsána. Tato dokumentace často obsahuje velké množství cenných údajů. Aby bylo možné tyto údaje ve statistických nebo jiných matematických modelech využít, je třeba požadované údaje z volného textu nějakým způsobem extrahovat. Ručně prováděná extrakce požadovaných údajů z velkých souborů lékařských záznamů je ale velmi pracný, časově náročný a chybami zatížený úkol, který navíc mohou provádět pouze odborníci specialisté. Proto výzkum metod pro automatickou nebo alespoň částečně automatizovanou extrakci údajů a informací z lékařských záznamů je velmi důležitá a aktuální oblast výzkumu, jejíž výsledky snadno naleznou uplatnění v praxi.

Struktura předkládané disertační práce je následující:

V první kapitole autor stanoví cíl své práce, který vidí ve stanovení a výzkumu metod použitelných pro extrakci informací z lékařských narativních zpráv a záznamů. V první kapitole také popisuje strukturu své práce a stav výzkumu v této oblasti u nás i v zahraničí.

V druhé kapitole se autor zabývá legislativním rámcem a standardy, které se pro záznam a předávání lékařských informací v českém prostředí používají. Kapitola je napsána přehledně a z textu je zřejmé, že se autor v dané problematice dobře orientuje.

Třetí kapitola se zabývá automatickou extrakcí informací z lékařských zpráv. Zpráva je nejdříve automaticky rozdělena do tokenů. Tvar tokenů je pak modifikován způsobem definovaným pomocí regulárních výrazů a modifikované tokeny jsou porovnávány se slovníkem nebo s databází lékařských termínů. Při nalezení shody je tokenu přiřazen význam. Tato metoda se standardně používá a autor ji implementoval a testoval na konkrétních datech. Nejlepší výsledky získal při porovnávání rozšířených tokenů s bibliografickým klasifikačním systémem MeSH. Autorovy výsledky potvrzují výsledky ostatních autorů a ukazují na to, že standardní plně automatizovaná metoda extrakce informací z lékařských zpráv není pro svoji velkou chybovost v praxi použitelná.

Čtvrtá kapitola popisuje elektronický zdravotní záznam navržený pro zubní lékařství s interaktivní komponentou zubního kříže s hlasovým vstupem. Zde je informace ukládaná do databáze extrahována z hlasového vstupu. Vytvoření tohoto softwarově značně složitého systému řešil celý tým lékařů a informatiků z Centra biomedicínské informatiky při Ústavu informatiky akademie věd ČR. Autor

disertace byl členem tohoto týmu a navrhl softwarové řešení pro integraci interaktivní komponenty do systému a vytvořil datový model pro uložení dat.

Pátá kapitola navazuje na třetí kapitolu a zabývá se extrakcí informací z narativních klinických zpráv. Zde extrakce není plně automatizována, ale provádí se za přispění lékaře-experta. Extrakce se provádí ve třech fázích. První fází je automatická tokenizace. Následuje tzv. normalizace, která spočívá především v opravě překlepů a v rozšíření zkratek. Normalizaci provádí lékař specialista. Výsledná normalizovaná zpráva může být již přeložena automatickým překladačem do jiného jazyka. Autor na několika příkladech ukazuje, že normalizace zprávy podstatným způsobem ovlivní srozumitelnost přeloženého textu. Vlastní extrakce informace z textu se provádí v poslední třetí fázi, kterou autor nazývá sémantickou anotací. Anotaci provádějí lékaři specialisté pro zvolený klasifikační systém jako je například SNOMED CT nebo MKN 10 za pomoci autorem vyvinutého softwaru, který jejich práci ulehčuje. Třífázová metoda částečně automatizované extrakce informace z lékařských zpráv již dosahuje nízké chybovosti a je využitelná v praxi.

V šesté kapitole autor vyvozuje a diskutuje závěry, které z jeho práce plynou. Závěry práce lze stručně shrnout následovně:

1) Narativní lékařské zprávy mají svá specifika, která nedovolují jejich analýzu standardními metodami, které se používají pro analýzu přirozeného jazyka. Specifika narativních lékařských zpráv také způsobují velkou chybovost při použití plně automatizovaných metod extrakce informací. V tom se autor shoduje s názorem ostatních autorů, kteří problém řešili před ním.

2) Lze navrhnout částečně automatizovanou metodu pro extrakci informací, která má nízkou chybovost a je prakticky použitelná. Autor takovou metodu navrhnul, vytvořil software pro její ověření a metodu na konkrétním případu záznamů z oblasti kardiologie ověřil.

Hodnotící část posudku:

Souhlasím s autorem, že vzhledem k specifikám lékařských narativních zpráv nemůže být metoda extrakce informací plně automatizovaná. Je ale potřeba ji automatizovat alespoň částečně a provádění úkonů, které bude muset udělat lékař specialista podpořit vhodnou počítačovou aplikací.

Autor částečně automatizovanou metodu extrakce informace navrhl a její funkci ověřil na konkrétním příkladu. Ve své práci autor také naznačuje jak automatizaci extrakce informací dále rozšířit. Navrhuje především automatizaci druhé fáze, kterou nazývá normalizace. Zmiňuje také možnost použít pro automatizovanou normalizaci algoritmus schopný se učit. V rámci obhajoby by mohl svou vizi dalšího rozvoje jím navržené metody směrem k její větší automatizaci, blíže vysvětlit.

Jazyková úroveň práce je nadprůměrná. Práce je srozumitelná, i když nutně používá celou řadu oborově specifických pojmů a zkratek. Rovněž formální úroveň práce je velmi dobrá. Překlepů je v práci minimum.

V práci se autor odvolává na celou řadu prací českých i zahraničních autorů, většinou z posledních deseti let. To svědčí o tom, že se autor důkladně seznámil se současným stavem řešené problematiky u nás i v zahraničí a nebudoval svou metodu extrakce informací na zelené louce.

Výsledky své práce autor publikoval v odborném tisku. Výsledky týkající se automatické extrakce informací z lékařských zpráv ve dvou článcích v recenzovaném časopise European Journal of Biomedical Information, výsledky týkající se zubního kříže v impaktovaném časopise Biocybernetics and Biomedical Engineering a výsledky týkající se původní, autorem navržené metody částečně automatizované extrakce informací z lékařských zpráv, v impaktovaném časopise Methods of Information in Medicine.

K disertační práci mám následující konkrétní výhrady a připomínky:

V odstavci 3.2 autor popisuje výsledky, které získal pro automatickou extrakci informací ze souboru lékařských zpráv. Zde mi chybí podrobnější popis souboru, který pro experiment použil.

V odstavci 5.1.4 autor popisuje fázi normalizace. Zde by byla na místě větší míra formalizace použitých pojmů. Autor operuje s pojmy transformace, homogenní transformace, heterogenní transformace, opakovaná transformace, aproximace automatizované transformace, míra chybovosti atd., aniž by byly tyto pojmy přesně vymezeny.

V odstavci 5.1.4 jsou uvedeny vzorce pro výpočet míry chyb *Err* a váhy w_r . Jejich význam není zřejmý a měl by být v textu podrobněji vysvětlen.

Uvedené výhrady však nepovažuji za podstatné pro celkové hodnocení práce.

Závěr posudku:

Práce se zabývá problematikou, která je pro praxi velmi důležitá. Autor prokázal, že se s danou problematikou do hloubky seznámil a že se v ní velmi dobře orientuje. Jím navržená metoda částečně automatizované extrakce informace z narativních lékařských textů je původní a její použitelnost byla autorem demonstrována na konkrétním případě. Disertační práce prokazuje předpoklady autora k samostatné tvořivé vědecké práci.

Disertační práci doporučuji k obhajobě a po jejím úspěšném absolvování doporučuji udělení vědecké hodnosti PhD v oboru biomedicínská informatika.

V Praze dne 10. srpna 2017

Doc. Ing. Arnošt Veselý, CSc.

Legal Office

Official Transcript of Records for

PAVLA CERMAKOVA
(Civic registration number 880208-8149)

The above is an excerpt from the register of student records

Stockholm, September 12, 2017

.....
Gustav Lantto
Archivist