

OPONENTSKÝ POSUDEK DISERTAČNÍ PRÁCE

UNIVERZITA KARLOVA V PRAZE

1. LÉKAŘSKÁ FAKULTA

Název práce: Extrakce informací z lékařských textů

Autor: Ing. Karel Zvára

Obor: Biomedicínská informatika

Cílem práce je „zjištění specifických vlastností českých lékařských zpráv z hlediska možnosti extrahovat z nich konkrétní informace“. Dílčí cíle se věnují specifickému problému nestatistické fáze zpracování lékařských zpráv. Vlastní práce směřuje k návrhu postupů, které předcházejí překladu a sémantické anotaci českých lékařských zpráv.

Práce řeší problém návrhu a vývoje metod pro přípravu vstupního textu a popisuje vlastní softwarovou implementaci pro interaktivní korekci a sémantickou anotaci narativních lékařských zpráv. Popisuje návrh a ověření třífázové metody předběžného zpracování narativních lékařských zpráv a odkazuje na vlastní publikované výsledky navržených postupů.

Práce předkládá řešení aktuální problematiky, která se vztahuje k návrhům systémů a datových modelů pro aplikace ve zdravotnictví, dále k automatizaci analýzy textových dokumentů českých lékařských zpráv a extrakci informací.

Pro problém extrakce informací z českých textů dosud není vyhovující univerzální, ani dílčí řešení s aplikací. Většina odborných prací a aplikací se zabývá anglicky psanými texty. Specifickému tématu českých lékařských textů se dosud věnovalo pouze několik autorů. Předložená **práce předkládá, realizuje a vyhodnocuje vlastní návrhy a řešení několika problémů, které souvisejí se systémy datových modelů a automatizací analýzy české zdravotnické dokumentace.**

Metody řešení vycházejí z literární rešerše, přehledu specifik lékařských textů a ze souhrnu poznatků k automatizované analýze textu. Práce se také zaměřuje na **praktické aspekty realizovatelnosti** cílového řešení. Popisuje právní rámec pro vytváření a užití lékařské dokumentace, přináší přehled různých typů standardů pro uchování elektronické zdravotní dokumentace a nakonec předkládá různé koncepty klasifikačních systémů a číselníků.

Stěžejní části práce se věnují dvěma typům problémů, souvisejících s elektronickým zdravotním záznamem a jeho využitím. Jedním z řešených problémů je aplikace pro orální medicínu s interaktivní komponentou zubního kříže (Lifetime Dent Cross). Druhý typ se týká automatické lingvistické analýzy textů lékařských zpráv a vlastního návrhu třífázové metody předběžného zpracování. Automatická lingvistická analýza vychází z popisu základních principů analýzy textových dokumentů. Vlastní softwarové rozpoznávací nástroje umožní identifikovat jednotlivé tokeny v rámci použitých dostupných klasifikačních systémů. Postupy a výsledky analýz narativních lékařských zpráv byly publikované.

Třífázovou metodu předběžného zpracování (tokenizace, normalizace a ověřování) implementoval autor disertace v softwarovém nástroji TOCESA, jehož vývoji se věnuje dlouhodobě. Výzkum se softwarem vyžadoval spolupráci lékařů, kteří provedli normalizaci a sémantickou anotaci na vzorku 49 lékařských zpráv. Výsledky jsou v práci podrobně popsány

a ukazují na poměrně malé inter-individuální rozdíly v anotaci mezi oběma lékaři (neshoda byla zjištěná u 3,07 % z 1500 hodnocených případů).

Vlastní výsledky. Disertace popisuje a shrnuje teoretické i praktické aspekty automatické analýzy textů českých lékařských zpráv. Touto problematikou se autor disertace dlouhodobě zabývá teoreticky, vytvářel a ověřoval softwarová řešení, **řadu poznatků časopisecky publikoval, nebo se na publikacích v této oblasti podílel.**

Dalším specifickým výsledkem je dlouhodobá spolupráce na návrhu a vývoji technologie, návrh datového modelu a podíl na realizaci softwarového řešení aplikace pro orální medicínu s interaktivní komponentou zubního kříže. Publikace, která studuje vlastnosti aplikace, je součástí příloh disertace.

Nejnovější poznatky a výsledky automatické analýzy narativních klinických zpráv jsou implementované v softwarovém nástroji TOCESA a popsané v impaktované publikaci, která tvoří přílohu disertace.

Formální náležitosti práce. V práci nelze přehlédnout některé prohřešky formálního rázu, například odlišné formátování textu kapitoly 2 a nepřívětivou úpravu formátů u většiny tabulek, časté osamocené hlásky na konci řádku a občasně překlepy.

Kapitola 4 (Elektronický zdravotní záznam) narušuje logiku obsahové návaznosti umístěním mezi dvě kapitoly, které se týkají analýzy textů. Podobný problém je v autoreferátu na str. 15 (poslední odstavec).

V přílohách disertace jsou dvakrát vloženy strany 207, 209 a 210 z publikace.

Výkladový slovník (str. 78) je spíše slovník zkratk, avšak obsahuje jen některé zkratky, použité v práci (pokud tento výběr vysvětlených zkratk měl nějaký důvod, mohl být zmíněn). Formáty citované literatury často neodpovídají normě (například nejsou uvedené strany časopiseckých článků, z odkazů na sborníky není jasné, zda odkazují na celý sborník, nebo jen na určitý článek ve sborníku apod.).

Autoreferát disertace obsahuje překlepy a mohl obsahovat jenom literaturu základní, citovanou v autoreferátu, nikoliv všechny publikace, citované v disertaci.

Obrázky autoreferátu disertace na straně 14 (porovnání překladu pomocí původní a nové verze překladače Google) mohly být v textu vysvětleny a pro srovnání možná doplněné o původní i normalizovaný text zprávy.

Za slabší stránku práce bych označila neujasněnou (nevysvětlenou) strukturu koncepce práce ve vztahu k vymezeným cílům. Upřesnění by napomohlo porozumět návaznostem jednotlivých částí práce navzájem i jejich vazbě na jednotlivé cíle.

Význam práce spočívá v návrhu nových metod a vytvoření nástrojů, které podporují práci specialisty v etapě konverze českých lékařských zpráv do tvaru, který umožní automatizované zpracování bez rizika zkreslení původního významu lékařské zprávy.

Upravený tvar zprávy lze využít pro automatizovanou extrakci strukturované informace a pro následné vložení extrahované informace do elektronického zdravotního záznamu.

Předložené řešení se zabývá teoretickou i praktickou stránkou problému. Výsledky práce lze považovat za ojedinělé v oblasti analýzy české narativní lékařské zprávy.

Práce uvádí také dřívější výsledky autora z oblasti analýzy lékařských zpráv nástroji pro zpracování přirozeného jazyka a realizaci návrhu metod pro ukládání strukturovaného zdravotního záznamu, který je již v praxi ověřený.

Otázky k obhajobě:

Lze navržené řešení předzpracování textu považovat za univerzálně využitelné v každém lékařském oboru? Pokuste se o odhad reálné proveditelnosti a ekonomické náročnosti vytvoření databáze normalizovaných lékařských zpráv pro některý obor (odbornou ambulanci).

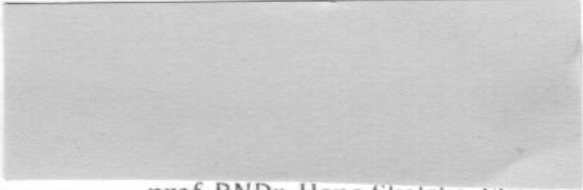
Lze provést srovnání náročnosti a univerzálnosti Vašeho řešení s konceptem ep SOS?

Závěr:

Stanovené cíle byly splněny, současně jsou specifikovány podmínky, navrženy metody a vytvořeny softwarové nástroje, které umožní extrahovat odborné informace z lékařských zpráv.

Stěžejní výsledky práce prokazují vysokou odbornou úroveň znalostí a aktuálního stavu řešené problematiky, prokazují schopnosti analýzy, návrhu teoretického i realizaci nového vlastního řešení.

Autor disertační práce prokázal předpoklady k samostatné tvořivé vědecké práci a udělení titulu Ph.D.



prof. RNDr. Hana Skalska, CSc.
Katedra informatiky a kvantitativních metod
Univerzita Hradec Králové

V Hradci Králové, 11. 8. 2017

Legal Office

Official Transcript of Records for

PAVLA CERMAKOVA
(Civic registration number 880208-8149)



Karolinska
Institutet

The above is an excerpt from the register of student records

Stockholm, September 12, 2017

.....
Gustav Lantto
Archivist