Posudek na disertační práci

Název disertační práce:

„JAZYK LÉKAŘSKÝCH ZPRÁV A JEHO INFORMAČNĚ LEXIKÁLNÍ ANALÝZA“

Autorka: **Mgr. Petra Přečková**
Univerzita Karlova v Praze
1. Lékařská fakulta, Praha

Oponent: **Prof. RNDr. Hana Skalská, CSc.**
Katedra informatiky a kvantitativních metod
Fakulta informatiky a managementu
Univerzita Hradec Králové

Cílem disertační práce je informačně lexikální analýza českých lékařských zpráv a využitelnost mezinárodních klasifikačních systémů v českém zdravotnickém prostředí.

Práce vychází z rešerše klasifikačních systémů a výběru nejvhodnějších systémů využitelných pro české zdravotnictví. Vlastní analýza a empirická práce (mapování atributů) jsou založeny na atributech Minimálního datového modelu pro kardiologii (MDMK), vytvořeného v rámci výzkumného centra EuroMISE – Kardio.

Další část práce se zabývá lexikální analýzou lékařských zpráv. Byly analyzovány zprávy psané volným textem různými lékaři a zprávy strukturované, uložené v softwarové aplikaci ADAMEK. Výsledkem této části práce je porovnání shody atributů v obou typech lékařských zpráv. Součástí práce je podrobnější analýza možností mezinárodních klasifikačních systémů SNOMED CT a MKN-10. Práce odkazuje na odborné publikace z poslední doby, které se také zabývají významem i možnostmi mapování atributů mezi klasifikačními systémy jako nezbytným krokem sémantické interoperability. Studuje vlastní možnosti mapování Minimálního datového modelu pro kardiologii na mezinárodní klasifikační systémy, zabývá se praktickými aspekty této problematiky **a předkládá vlastní výsledky v této oblasti**.

Dále práce **prezentuje novou aplikaci metod pro měření diverzity lékařských zpráv**. Nejprve jsou popsány obecné koncepty Diverzity. Na základě analýzy 110 lékařských zpráv s volným textem a 1119 strukturovaných lékařských zpráv byly vytvořeny kategorie atributů a následně vypočítány a porovnány míry Diverzity (Gini-Simpsonova diverzita, Gini-Simpsonova vlastní diverzita, relativní marginální diverzita). Byla stanovena hodnota Diverzity jednotlivých kategorií lékařských zpráv pro různé atributy, testovány hypotézy o shodě Diverzity kategorií v textových a strukturovaných typech lékařských zpráv **a získány vědecky zdůvodněné závěry pro posouzení informační Diverzity v těchto typech zpráv**.

Aktuálnost tématu

Analýza prováděná za účelem využití informace textových dat je aplikační oblastí, která se prudce rozvíjí zejména v posledním desetiletí. K tomuto rozvoji přispěl vývoj infor-

mačních technologií, jejich dostupnost a zejména snaha automatizovat využití textové informace, tedy i ze záznamů ve zdravotnické dokumentaci. Současně s těmito požadavky se zvyšují nároky na kvalitu, srozumitelnost a správnost takto získané informace. Ve zdravotnictví má analýza textových informací význam nejen odborný, ale také etický, celospolečenský, politický (návrh zdravotních opatření, optimalizace služeb veřejného zdravotnictví) i komerční. Automatizovaná přenositelnost informace, zaznamenané formou volného, nebo částečně strukturovaného textu v různých informačních systémech, je však problematická. Interoperabilita mezi různými (zejména zdravotními) systémy, která znamená možnost vzájemného přebírání informací bez zkreslení významu a obsahu předávané informace, je tedy klíčovým problémem. Sdílení klinické informace vyžaduje nalézt způsob, jak automaticky rozpoznat sémanticky ekvivalentní informaci na základě heterogenní informace, zapsané ve zdravotním záznamu.
Zvolené téma, které se touto problematikou zabývá, proto hodnotím jako vysoce aktuální, ale na druhé straně za poměrně náročné, zejména s ohledem na komplexnost a obtížnost obecného řešení, které nelze od individuální práce očekávat.

Použité metody a postupy řešení stanoveného problému

Úvod práce a formulace cílů práce tvoří první dvě samostatné a poměrně stručné kapitoly. Následující třetí kapitola (Materiál a metodika) a kapitola čtvrtá (Výsledky), tvoří stěžejní části práce. Po kapitolách Diskuse a Závěr je uveden přehled literatury, který zahrnuje 64 citovaných zdrojů. Práce se odvolává také na 10 publikací, ve kterých je autorka disertace uvedena jako první autor nebo spoluautor. Následující rozsáhlá příloha (27 stran) prezentuje vlastní výsledky porovnání atributů MDMK, zakódovaných pomocí MKN 10 a SNOMED CT. Ukázky z přílohy jsou též v textu práce.

Třetí kapitola, nazvaná **Materiál a metodika**, je bohatě členěná a je orientována na kódovací a klasifikační systémy, přehled konverzních nástrojů, popis modelu MDMK a softwarovou aplikaci ADAMEK (**A**plikace **D**atového **M**odelu **E**uroMise centra – **K**ardio). V této části autorka prezentuje jednak výsledky své rešerše literárních zdrojů k dané problematice (zde lze vytknout nedůslednost v explicitním uvádění odkazů v textu na použitou literaturu zejména v části 3.2 a obrázků v celé kapitole), jednak vysvětluje zdroje dat pro vlastní analýzy. Přehlednosti koncepce celé práce by prospělo shrnutí rešerše popsané v kapitolách 3.1 a 3.2, vysvětlení významu této části práce na vlastní aplikace a uvedení formálních definic základních, dále používaných pojmů. Jestliže jedním z cílů práce bylo zmapování klasifikačních systémů (rešerše), měly být výsledky této činnosti uvedeny v jiné části práce, případně (kromě charakteristik) mohly být doplněné o přehledné porovnání popsaných klasifikačních systémů s vysvětlením, v čem je tento přehled užitečný pro analýzu vlastních dat.

Přes uvedenou námitku se domnívám, že **touto částí práce doktorandka prokazuje dostatečné teoretické znalosti problematiky analýzy textových medicínských informací, tyto znalosti doplňuje vlastními praktickými zkušenostmi a je takto vybavena k vlastnímu rozboru a hodnocení možností mapování atributů specifické domény na mezinárodní klasifikační systémy, jak je prezentuje v další kapitole. Současně ukazuje vlastní přístup k naplnění cílů práce.**

V kapitole **Výsledky** jsou nejprve popsána východiska **vlastního přístupu, vedoucího k naplnění části stanovených cílů, týkajících se analýzy znaků MDMK**. Pro porovnání strukturovaných záznamů a záznamů textových lékařských zpráv jsou použity absolutní a relativní četnosti (bez intervalů spolehlivosti) výskytů atributů, výsledky jsou prezentovány převážně formou tabulek. Pro doplnění informace z analýzy záznamů

volných textových lékařských zpráv by bylo vhodné podrobněji charakterizovat množinu lékařů, jejichž záznamy byly analyzovány (jejich počet, praxi, apod.).

Výsledky mapování MDMK na SNOMED CT a na MKN 10 jsou uvedeny formou tabulek. Bylo zjištěno, že přibližně 85 % atributů MDMK bylo obsaženo alespoň v některém klasifikačním systému, nejčastěji ve SNOMED CT. Na konci kapitoly 4.2 mohl být uveden přehledný závěr, plynoucí z této empirické části.

Za stěžejní a další nové výsledky které práce přináší, lze považovat návrh nové aplikace měr diverzity pro porovnání různých typů lékařských (textových) zpráv. Tyto výsledky jsou popsány v kapitole 4.3. Jedná se o návrh nové aplikace, která je založená na obecných konceptech diverzit. Autorka navrhuje novou aplikaci a dokazuje použitelnost a vhodnost měr diverzity v této oblasti. Potvrzuje, že atributy v textových lékařských zprávách vykazují vyšší diverzitu, než mají atributy ve strukturovaných lékařských zprávách. Této části práce lze částečně vytknout nekonsistentnost obsahu kapitoly čtvrté (Výsledky), ve které části 4.3.1 – 4.3.4 popisují přehled měr diverzity, jejich definice a vlastnosti, což by spíše patřilo do metodiky. Postrádám také informace o použitých technologiích (softwaru), které mohly být zmíněny v metodice, nebo alespoň v popisu výsledků. Rovněž metody stanovení směrodatných chyb v testech hypotéz o shodě měr diverzity, jejichž výsledky jsou použity v tabulce 22 části 4.3.4, nejsou v práci specifikovány. Kromě vypočítaných měr mohly být též uvedeny intervaly spolehlivosti. Přes tyto námitky lze **ocenit přínos autorky a význam nové aplikace pro kvantifikaci rozdílů diverzit daných atributů** mezi různými typy textových záznamů.

Kapitoly **Diskuse** a **Závěr** shrnují stručně metodiku, vysvětlují logiku použitých metod, zmiňují cenné zkušenosti získané praktickou aplikací mapování atributů a diskutují a shrnují vlastní závěry práce. Některé závěry práce nejsou překvapivé (například větší nesourodost volného textu v porovnání se standardizovaným textem), zde je však cenný návrh nové aplikace míry diverzity, která umožňuje kvantifikaci těchto rozdílů.

Lze konstatovat, že jednotlivé kapitoly práce jsou navzájem logicky provázané, zásadní vytčené cíle byly splněny a práce přináší nové vlastní výsledky.

Formální stránka práce

Práce po formální stránce využívá převážně pečlivě zvolené formulace, místy s překlepy nebo nevhodnou interpunkcí (strany 30, 33, 37, 56, 72), drobnými nepřesnostmi v logické stavbě věty (str. 62), chybějícím zdůvodněním (např. str. 38 poslední odstavec – není jasné, z čeho plyne). Na straně 82 má být zřejmě sdružené pravděpodobnosti (místo společné). Práce téměř neobsahuje formalizmy, obecnější modely ani definice. Grafická úprava práce je převážně velmi pěkná, výhrady lze mít k formátu nadpisu kapitol a podkapitol (např. na str. 58), anglickému nadpisu sloupce tabulky 19, nesprávnému zarovnání jednotek v posledním sloupci tabulek 1 - 9, 11-15, 17 a 18, dále neúměrným vzdálenostem mezi řádky v tabulkách 21, 22 a 23, nedůslednému uvádění odkazů na zdroje obrázků (str. 15, 16, 20, 31, 32, 34, 35, 41, 49 – 51, 54, 55) a častému výskytu osamocených hlásek na koncích řádků. Některé informace nepovažuji za relevantní v dané části (podkapitola Český jazyk 4.2.1 ve Výsledcích) a některé nepřesné („Česky mluví … asi 200 000 osob v dalších zemích"). Podle webu Ministerstva zahraničí, více než dva miliony osob trvale žijících mimo ČR (krajané) uvádí češtinu jako svůj jazyk a dlouhodobě trvale žijících českých občanů (s českým pasem) je dalších 250 000, jedná se tedy o řádové rozdíly. Nicméně záměr uvádění těchto údajů v dané práci není jasný.

K předložené disertační práci mám následující dotazy:

1. Jaké jsou možnosti formalizace popisu procesu mapování atributů mezi systémy? Je vždy nezbytné zapojení experta na nějaké úrovni tohoto procesu?
2. Lze využít míry podobnosti k analýze textových dokumentů lékařských zpráv?
3. Autorka navazuje v práci na výsledky kolektivního výzkumu, speciálně na vytvořený datový model MDMK a aplikaci ADAMEK (Aplikace datového modelu). Jaký je podíl autorky disertace na výsledcích těchto výzkumů a které výsledky na tomto datovém modelu a na aplikaci jsou její vlastní?

Závěr

Práce studuje a řeší vysoce aktuální a obecně náročnou problematiku, cíle práce byly splněny. **Autorka** prezentuje původní, logicky provázané výsledky badatelské práce, kterými **prokazuje** nejen schopnost syntézy získaných poznatků, ale též **způsobilost k odvození vlastních nových vědeckých poznatků a metod,** které prezentuje a jejichž aplikací prokazuje vhodnost navržených postupů.

Za nové vědecké poznatky lze považovat:

1. Návrh možností mapování atributů specifického datového modelu MDMK na mezinárodní klasifikační systémy a jeho praktické ověření.
2. Lexikální analýzu lékařských zpráv psaných volným textem za účelem porovnání s lékařskými zprávami strukturovanými pomocí aplikace ADAMEK.
3. Kategorizaci možných problémů při jazykovém rozboru textových zpráv.
4. Návrh nové aplikace pro měření diverzity lékařských zpráv, psaných v jakémkoliv jazyce.

Doktorandka tak prokazuje schopnost samostatné tvořivé vědecké práce, schopnost aplikace teoretických poznatků a formulování nových výsledků, objektivně zdůvodněných závěrů a na jejich základě odvozených námětů, které dokáže implementovat. Tento závěr mohu kromě výsledků, prezentovaných v předložené práci, podložit také názorem, který jsem získala na základě některých publikací autorky k tomuto tématu, se kterými jsem se seznámila v rámci přípravy tohoto posudku a jejichž citace jsem nalezla v předložené práci.

Dizertační práce prokazuje předpoklady autorky k samostatné tvořivé vědecké práci, proto doporučuji udělení titulu Ph.D. za jménem.

V Hradci Králové dne 14. 8. 2011

Prof. RNDr. Hana Skalská, CSc.