**Strukturální a lexikální analýza lékařských zpráv**

Zvára Jr., Karel
2011

# Strukturální a lexikální analýza lékařských zpráv

*doktorand:*
ING. KAREL ZVÁRA
Oddělení medicínské informatiky
Ústav informatiky AV ČR, v. v. i.
Pod Vodárenskou věží 2
182 07 Praha 8
zvara@euromise.cz

*školitel:*
DOC. ING. VOJTĚCH SVÁTEK, DR.
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4
130 67 Praha 3
svatek@vse.cz

obor studia:
Biomedicínská informatika

**Abstrakt**

Článek pojednává o výsledcích strukturální a lexikální analýzy lékařských zpráv. V této části zpracování lékařských zpráv jsem prakticky ověřil použitelnost dostupných klasifikačních systémů i obecných nástrojů a databází.

## 1. Vědecká otázka a očekávaný přínos

Hlavním cílem práce je zjištění specifických vlastností českých lékařských zpráv z hlediska možnosti extrahovat z nich konkrétní údaje. Realizace cíle předpokládá splnění dílčích cílů:

1. Zodpovědět otázku „Které vlastnosti českých lékařských zpráv působí největší problémy v jednotlivých nestatistických fázích zpracování přirozeného jazyka?“. Jednotlivými fázemi přitom jsou strukturální analýza, lexikální analýza a slovní rozbor.
2. Navrhnout základní postup pro analýzu česky psaných lékařských zpráv.
3. Pomocí vlastní implementace s využitím externích nástrojů ověřit navržený postup pro analýzu česky psaných lékařských zpráv a základní postup i výsledky publikovat.

Ověřovanou hypotézu jsem formuloval takto: „Z odborných lékařských zpráv psaných v českém jazyce lze pod supervizí odborníka a za použití technologií pro zpracování přirozeného jazyka získávat specifikované odborné informace, například seznam známých alergických reakcí či výsledky biochemických vyšetření.“

Přínosem výzkumu by mělo být přiblížení či přímo implementace nástrojů pro asistovanou extrakci informací z lékařských textů psaných v českém jazyce. Extrahované informace lze následně využít pro potřeby elektronické zdravotnické dokumentace nebo pro využití společně s dalšími technologiemi (např. jako vstupní data do automatů provádějících formalizovaná lékařská doporučení).

Tématu extrakce informací z lékařských zpráv se věnoval Semecký, který v [1] uvedl důvody pro které se zdá, že lingvistická analýza lékařských zpráv nemůže být úspěšná. Semecký v [1] používal především regulárních výrazů pro extrakci číselných hodnot. Na práci [1] navázal Smatana v práci [2], rozšířil přístup Semeckého o lingvistickou analýzu a došel k mírně lepším výsledkům.

Od mé práce očekávám další rozšíření, především vytvoření pracovního číselníku pro kardiologii navázaného na koncepty UMLS [3] a jeho aplikování na dostupné lékařské zprávy.

## 2. České lékařské zprávy

České lékařské zprávy jsou vesměs textové dokumenty. Jejich obsah i forma jsou upraveny zákonem č. 20/1966 Sb ve znění pozdějších předpisů „o péči o zdraví lidu“ [4] (především v § 67b) a vyhláškou č. 385/2006 Sb. ve znění pozdějších předpisů „o zdravotnické dokumentaci“ [5] (vyhláška je závazná, neboť úpravu umožňuje § 67b odstavec 19 zákona).

Styl formátování lékařských zpráv se liší i přesto, že vyhláška o zdravotnické dokumentaci taxativně vyjmenovává obsah zdravotnické dokumentace pro její jednotlivé druhy. Lékaři záznamy ve zdravotnické dokumentaci tvoří obvykle podle šablony, resp. upravením poslední zprávy stejného druhu u stejného pacienta. Takový postup totiž lékařům šetří čas; jednotlivé druhy zpráv obvykle musejí obsahovat velké množství s časem

se jen málo měnících informací jako je identifikace zdravotnického zařízení, administrativní údaje o pacientovi (datum narození, číslo pojištěnce, adresa pobytu), část diagnostické rozvahy (především dlouhodobé diagnózy a známé alergie) a dlouhodobou medikaci (např. léky pro snižování krevního tlaku).

Pro výzkum mám k dispozici sady zpráv ze dvou zdrojů. Při praktickém ověřování postupů proto data z jednoho zdroje využívám pro nastavení ověřovacího pokusu (např. pro vytvoření slovníku) a data z druhého zdroje využívám pro zjištění úspěšnosti metody.

## 3. Strukturální analýza

Strukturální analýza představuje první fázi zpracování textu. Úkolem strukturální analýzy je tokenizace, rozdělení do vět a případně také do vyšších struktur (např. odstavců).

Obvyklým postupem pro strukturální analýzu je rozdělení vstupního textu podle speciálních znaků, tedy symbolů ukončujících slova (mezera, čárka, středník), věty (tečka, otazník, vykřičník). České lékařské zprávy jsou však značně netypickými texty. Obsahují ohromně velké množství zkrácených slov a zkratek.

Při použití běžného přístupu ke strukturální analýze jsem velmi brzy zjistil, že v českých lékařských zprávách je význam speciálních znaků odvoditelný až z jejich okolí. Čeština totiž patří k jazykům s volným pořadím slov. Způsob zápisu lékařských zpráv není striktně standardizován [6].

Ukázka textu v části objektivní nález: *„Akce pravidelná, klidná, 2 ohr. ozvy. Břicho klidné, játra, sleziona nezv., tapot. nebol., jizva po CHE keloidní. Akne po trupu. DK bez otoků a varixů.“*

Výše uvedená věta ukazuje několik typických vlastností českých lékařských zpráv:

- Většina vět neobsahuje sloveso, protože je zřejmé z kontextu. V první větě navíc chybí určení předmětu – jde o akci srdce.
- Druhá věta obsahuje překlep („sleziona“ namísto „slezina“), lékařské zprávy jsou protkány překlepy.
- V uvedených čtyřech větách jsou čtyři zkrácená slova a dvě zkratky.

Problematika zkracování slov není typická jen pro české lékařské zprávy. [7] uvádí, že lékaři jiných odborností jsou schopni správně interpretovat jen asi polovinu užívaných zkratek a zkrácených slov. Podobné potíže uvádí také [8] a z oboru práva též [9].

Některé části lze správně identifikovat až z kontextu. Z toho důvodu jsem se rozhodl ve fázi strukturální analýzy standardizovat konce řádků (CR+LF na CR) a transformovat vstupní text do řetězce objektů (nazývám je kontejnery), přičemž po skončení průběhu v této fázi jsou objekty následujících druhů:

- řetězec alfanumerických znaků (po sobě jdoucích),
- jiný znak (u toho je možné uvést kolikrát za sebou se stejný znak opakuje).

Na získaný řetězec objektů aplikuji metody, které z podřetězce odvozují další druhy objektů. Metody aplikuji i na podřetězce tvořené z takto získaných nových objektů. Tímto způsobem identifikuji:

- numerické řetězce (celé číslo bez znaménka) - číslo,
- separovaná čísla (vždy kombinace: *číslo [separátor číslo]*+),
- datum ve formátu d.m.r (s mezerami či bez mezer za tečkami),
- rodné číslo (kontrola existence data, kontrola součtem u 10-ciferných) – s lomítkem i bez lomítka.

## 4. Lexikální analýza

Úkolem lexikální analýzy je identifikovat jednotlivé základní části textu, tedy slova, hodnoty a podobně. Lékařské zprávy jsou zvláštním druhem volného textu. Hledal jsem proto slovník, který bych mohl využít pro identifikaci slova.

Obecné české korpusy považuji pro tento účel za nevhodné, protože jsou vytvářeny z jiného druhu projevů, obvykle z prózy či novinových článků. Při hledání jsem zjistil, že databáze pro volně šiřitelný slovník pro automatickou kontrolu pravopisu iSpell, je GNU licencí (zajišťující použitelnost pro vědecké účely), a že jeho autor myslel na možné další využití slovníku. Slova tohoto slovníku jsou uspořádána do několika různých souborů, je tak snadno možné identifikovat velké množství jmen a názvů. Pravidla, jejichž využitím iSpell generuje další tvary a odvozená slova, jsou zapsána tak, že odpovídají tvorbě jednotlivých slovních druhů.

V lékařských zprávách je velké množství odborných termínů. U nezkrácených českých slov se mi pomocí rozšířeného slovníku iSpellu podařilo identifikovat slovní druh bez závažnějších problémů i když v mnoha případech nikoliv jednoznačně. Vlastní jména totiž často odpovídají obecnému podstatnému nebo přídavnému jménu (např. Dlouhý či Noha). Pokud jsou taková slova na začátku věty, v části lexikální analýzy není možné řádně klasifikovat slovo.

Pozornost jsem dále upřel na snahu identifikovat odborné termíny, neboť jedním z cílů je zjištění možnosti získat ze zprávy anamnestické informace, především informace o diagnózách, alergiích a výsledcích biochemických vyšetření. Našel jsem celkem tři klasifikační systémy, které by bylo možné využít pro identifikaci jednotlivých odborných termínů.

Prvním testovaným systémem byla anglická verze klasifikačního systému SNOMED CT [10]. Pomocí tohoto klasifikačního systému se podařilo identifikovat termíny, které nebyly zkrácené, a které mají stejné znění v českém i v anglickém jazyce. Vzhledem k odbornosti vstupních lékařských zpráv (kardiologie), tak šlo o tyto konkrétní termíny: „diabetes mellitus“ (SNOMED CT 73211009) a jednotku mmHg (SNOMED CT 259018001). Česká verze SNOMED CT neexistuje, mimo jiné proto, že ani existovat nemůže. Česká republika totiž není členem International Health Terminology Standards Development Organisation (IHTSDO), vlastníka klasifikačního systému SNOMED CT. SNOMED CT není použitelný pro identifikaci lékařských termínů ve volném textu.

Druhým testovaným klasifikačním systémem byla Mezinárodní klasifikace nemocí verze 10 (ICD10, MKN10) v české verzi [11]. Tento číselník byl velkým zklamáním, jeho překlad byl totiž vytvořen jen pro ruční vyhledávání podle kódu diagnózy. Mnoho přeložených textů je totiž složeno ze zkrácených slov, přičemž v některých případech je jedno slovo zkracováno různými způsoby. V tomto záznamu je dvakrát zkráceno slovo „diabetes“, pokaždé jinak: „Diabet.polyneuropat. při diab.“. V některých případech je text kvůli zkracování slov i obtížně čitelný: „J.deg.on.oč.víčka a periok.kr.“. Vzhledem k velmi častému zkracování slov v MNK10 tento klasifikační systém není využitelný pro identifikaci odborných termínů ve volném textu. I kdyby však slova zkrácena nebyla, vzhledem ke skutečnosti, že MKN10 obsahuje jen výčet diagnóz, nebyl by tento číselník využitelný pro využití většiny klinických termínů.

Třetím testovaným klasifikačním systémem byl bibliografický klasifikační systém Medical Subject Headings (MeSH) v české verzi [12]. Pomocí MeSH se podařilo identifikovat průměrně cca 10 termínů na lékařskou zprávu [13]. MeSH není klinicky orientován a tomu odpovídaly také výsledky. Identifikované termíny odpovídaly především označení částí těla, v malé míře měřeným parametrům, v jednom případě diagnóze. Skutečně odborné termíny tak zůstaly neidentifikované.

## 5. Závěr a výhled

Jak uvádím výše, zjistil jsem, že žádný z dostupných klasifikačních systémů není využitelný pro identifikaci odborných termínů. V současné době z části zpráv vytvářím databázi v českých zpráv užívaných odborných termínů mapovaných na koncepty UMLS [3]. Jakmile budu mít zpracovanou základní databázi, otestuji její využitelnost jejím využitím na identifikaci termínů ze všech dostupných zpráv.

## Literatura

[1] J. Semecký a J. Zvárová (školitelka), “Diplomová práce: Multimediální elektronický záznam o nemocném v kardiologii”, *Univerzita Karlova v Praze, Matematicko-fyzikání fakulta*, 2001.

[2] P. Smatana a J. Paralič (školitel), “Diplomová práce: Spracovanie lekárskych správ pre účely analýzy a dolovania v textoch”, *Technická univerzita v Košicích, Košice*, 2005.

[3] “Unified Medical Language System, United States National Library of Medicine”.

[4] “Zákon č. 20/1966 Sb ve znění pozdějších předpisů, o péči o zdraví lidu”.

[5] “Vyhláška č. 385/2006 Sb. ve znění pozdějších předpisů, o zdravotnické dokumentaci”.

[6] P. Přečková, “Language of Czech Medical Reports and Classification Systems in Medicine”, *European Journal for Biomedical Informatics*, Vol. 9, No. 1, pp. 58-65, 2010.

[7] K. E. Walsh and J. H. Gurwitz, “Medical Abbreviations: writing little and communicating less”, *Archives of Diseases in Childhood*, 2008.

[8] H. Yu, G. Hripcsak, and C. Friedman, “Mapping Abbreviations to Full Forms in Biomedical Articles”, *Journal of the American Medical Informatics Association* , Vol. 9, No. 3, pp. 262-272, 2002.

[9] Y. HaCohen-Kerner, A. Kass, and A. Peretz, “Baseline methods for automatic disambiguation of abbreviations in Jewish law documents”, *Lecture Notes in Computer Science, Advances in Natural Language Processing Baseline methods for automatic disambiguation of abbreviations in*

*Jewish law documents* , Springer, Vol. 3230, pp. 58-69, 2004.

[10] "Systematized Nomenclature in Medicine - Clinical Terms, International Health Terminology Standards Development Organisation".

[11] "Mezinárodní klasifikace nemocí, Ústav zdravotnických informací a statistiky / World Health Organization".

[12] "Medical Subject Headings, Národní lékařská knihovna / United States National Library of Medicine".

[13] K. Zvára and V. Kašpar, "Identification of Units and Other Terms in Czech Medical Records", *European Journal for Biomedical Informatics*, Vol. 9, No. 1, 2010.