Onfram: Nástroj pro lokalizaci biomedicínských ontologií

Kolesa, Petr
2006

Onfram: nástroj pro lokalizaci biomedicínských ontologií

doktorand:
MGR. PETR KOLESA

Ústav informatiky Akademie věd ČR
Oddělení medicínské informatiky
Pod Vodárenskou věží 2
182 07 Praha 8

kolesa@euromise.cz

školitel:
ING. VOJTĚCH SVÁTEK, DR.

Katedra informačního a Iznalostního inženýrství
Vysoká škola ekonomická v Praze
Náměstí W. Churchilla 4
130 67 Praha 3

svatek@vse.cz

obor studia:
Biomedicínská informatika

Práce je podporována projektem 1ET200300413 a částečně i výzkumným záměrem AV0Z10300504.

Abstrakt

V článku popisujeme procedury a nástroj Onfram vyvinuté pro usnadnění a zrychlení tvorby lokalizovaných českých biomedicínských ontologií. Náš přístup je založen na vyhledávání konceptů v korpusu medicínských textů a jejich přiřazení protějškům v některé zavedené mezinárodní ontologii (tzv. *základové ontologii*). Takto vzniklá ontologie dvě hlavní výhody: je kompatibilní s některou z mezinárodních ontologií a potenciálně pokrývá všechny fráze používané v českém zdravotnictví.

Nástroj podporuje tvůrce ontologie tím, že se snaží automatizovat některé rutinní úkoly vyskytující se v průběhu tvorby ontologie. Nástroj se především snaží naučit, jak identifikovat koncepty v textu a jak je namapovat na základovou ontologii. Nástroj se učí postupně. Tak jak autor ontologie zpracovává další dokumenty, učící množina se zvětšuje a nástroj poskytuje lepší odhady. Po identifikaci konceptů, předkládá nástroj své návrhy uživateli. Ten buď přijme nebo opraví. Na základě této zpětné vazby nástroj upraví pravidla pro identifikací konceptů. Učení a úprava extrakčních pravidel probíhá pomocí metod zpracování přirozeného jazyka a nástroje na extrakci informací.

1. Úvod

Inteligentní medicínské systémy (IMS) pomáhají lékařům navrhnout a řídit léčbu specifickou pro jednotlivého pacienta a poskytují okamžité návrhy nebo upozornění o změnách v zdravotním stavu pacienta. Navíc napomáhají lékařům organizovat léčbu tak, aby byla v souladu s doporučenými postupy, ať se jedná o obecná doporučení vydávaná odbornými společnostmi nebo o technicko-organizační pravidla konkrétního zdravotnického zařízení. Takový přístup k léčbě vede k zlepšení poskytované péče na straně jedné a ke snížení výdajů za tuto péči na straně druhé. Z jiného úhlu pohledu IMS přebírají rutinní část lékařovi práce a lékař tedy může pracovat efektivněji.

Aby IMS mohly plnit tuto úlohu, musí být schopny pracovat s odpovídající medicínskou znalostí. Tato znalost musí být navíc poskytnuta ve formě, která je počítačově zpracovatelná. Dále musí vhodně popisovat oblast medicíny od elementárních skutečností až po velmi komplexní situace, a zároveň být pokud možno univerzální a znovupoužitelná, protože vytvoření takovéto znalostní reprezentace je nákladné a časově velmi náročné. Pro některé medicínské aplikace, jako například rozhodování v kardiologii, se zdá být nejlepší znalostní formalizací splňující výše vyjmenované *ontologie*.

K předním biomedicínským ontologiím patří *Foundation Model of Anatomy* (FMA) [1], *Systematized Nomenclature of Medicine* (SNOMED) [2] a *Unified Medical Language System* (UMLS) [3]. Všechny z těchto ontologií vznikly na území Severní Ameriky a soustředí se tedy především na podmínky a postupy obvyklé ve zdravotnictví v této části světa. Použití některé z těchto ontologií v aplikacích v českém zdravotnictví naráží kromě jazykové bariéry (všechny uvedené ontologie jsou primárně v angličtině a v různé míře přeloženy do francouzštiny, španělštiny, němčiny a několika dalších jazyků) i na rozdílnost zdravotnictví u nás a v Severní Americe. Příkladem takového rozdílu jsou používané léky: některé zde běžně používané léky nejsou dostupné v Severní Americe, a proto nejsou ani reprezentované v příslušné ontologii. Přímé

použití některé ze zavedených mezinárodních biomedicínských ontologií tedy není možné. Zároveň ale efektivně neexistuje žádná česká medicínská ontologie, a ani není žádná vyvíjena.

Bez vhodné biomedicínské ontologie ale není možné vyvíjet nebo nasazovat IMS do praxe v České republice. Tuto mezeru se snaží překlenout nástroj Onfram a metody, popsané v tomto článku. Onfram je nástroj určený k ulehčení vývoje biomedicínské ontologie vhodné pro vývoj IMS. Tvorba české biomedicínské ontologie je založená na tom, že v anglicky mluvících oblastech již existují IMS, které jako znalostní zdroj přímo využívají některou z výše jmenovaných ontologií.

Nejjednodušší postup, jak vytvořit ontologii, splňující výše uvedené požadavky je identifikovat v korpusu reprezentativních textů povrchové fráze [1] a namapovat je na koncepty v nějaké existující (*základové*) ontologii. Pokud takový koncept v základové ontologii neexistuje, je potřebný koncept vytvořen a autorem vložen do ontologie, včetně odpovídajících relací s ostatními koncepty.

2. Metody

Při vývoji nástroje jsme se zabývali otázkou, jak je možné zrychlit a usnadnit vývoj českých biomedicínských ontologií. Šlo nám hlavně o to, zautomatizovat některé části procesu vytváření ontologie, jako například identifikaci povrchových frází v textu a hledání odpovídajících protějšků v zavedené mezinárodní ontologii. Jsme tak schopni využit úsilí investované do vývoje dané ontologie. Tento způsob vytváření stále vyžaduje odborníka z oblasti tvorby ontologií, ale díky vyvinutému nástroji bude pracovat rychleji.

Česká ontologie vytvořená tímto způsobem není pouhým překladem základové ontologie, ale je odvozenou ontologii, která sice do značné míry přejímá strukturu základové ontologie, ale na rozdíl od ní dokáže popsat procesy v českém zdravotnictví. Odvozená ontologie navíc obsahuje pouze koncepty, které se v českém prostředí skutečně vyskytují. Na druhou stranu díky provázání odvozené a původní ontologie, bude jednoduché propojit systémy, založené na těchto ontologiích.

Ve dvou následujících oddílech stručně popíšeme základní metody, na kterých je Onfram založen. Ve třetí části bude stručně popsán vznik odvozené ontologie a ve čtvrté části rozebereme některé kroky tohoto procesu detailněji.

2.1. Extrakce informací

Obor nazývaný extrakce informací (information extraction, IE) se zabývá metodami automatického získávání informací z volného textu. Existuje mnoho metod a přístupů. V současnosti jeden z nejlepších algoritmů LP^2 [4], je implementován v nástroji AMILCARE [5], [6]. Tento nástroj jsme zvolili, protože podává dobrý výkon, a je vhodný pro interaktivní použití. AMILCARE navíc poskytuje Java API a je tedy snadno integrovatelný. AMILCARE využívá učení učitelem: na základě korpusu dopředu anotovaných textů (v každém dokumentu jsou vyznačeny informace, které chceme extrahovat) generuje systém extrakční pravidla. Na základě těchto pravidel jsou pak z neanotovaných textů extrahovány požadované informace.

AMILCARE není schopen sám o sobě provádět jazykovou analýzu textu. Toto není problém, pokud je text vhodně strukturován. Většina medicínských textů, jak se v průběhu vývoje ukázalo, však není dostatečně strukturovaná. Naštěstí je AMILCARE schopen využít dodatečných informací o textu, včetně jazykových.

2.2. Zpracování přirozeného jazyka

Zpracování přirozeného jazyka (Natural Language Processing, NLP) je obor, jehož zájmem je počítačové zpracování přirozeného jazyka takovým způsobem, jako by počítač do určité míry jazyku rozuměl. Porozumění jazyku počítačem je obecně velmi složitý problém, který neumíme dnes uspokojivě řešit.

[1] Povrchová fráze je část textu, která označuje nějaký konkrétní koncept. Různé povrchové fráze mohou označovat stejný koncept, například *glaukom* a *zelený oční zákal*. Na druhou stranu jedna povrchová fráze může v závislosti na kontextu označovat různé koncepty, *například teplota*: *fyzikální veličina* versus *symptom: zvýšená tělesná teplota*.

Nicméně existují nástroje, které pro praktické využití umožňují s uspokojivou přesností provádět morfologickou a syntaktickou analýzu textu.

Tyto nástroje jsou tak schopny poskytnout dodatečné informace o textu využitelné programem na extrakci informací, a tím zvýšit úspěšnost extrakce. K tomuto účelu využíváme český lemmatizer a morfologický analyzátor, který je součástí Pražského závislostního korpusu [7], [8]. Tento nástroj je schopen poskytnout řadu informací o zpracovávaném textu. V současnosti využíváme především informací o slovním druhu, pádu, čísle a slovním lemma (např. infinitiv u sloves, první pád jednotného čísla u podstatných jmen, zájmen, etc.). Lemmatizace je pro zpracování českých textů velmi důležitá, bez ní je prakticky nemožné ztotožnit různé tvary téhož slova a tedy identifikovat povrchové fráze označující stejný koncept.

2.3. Tvorba ontologie

Proces vytváření nové ontologie se sestává ze tří hlavních fází. V první fázi je zvolen způsob reprezentace znalostí. To zahrnuje rozhodnutí o způsobu reprezentace konceptů a vztahů mezi koncepty. Ve druhé fázi jsou identifikovány koncepty. Ve třetí fázi jsou nalezené koncepty začleňovány do ontologie – jsou definovány různé relace různých typů mezi koncepty. Ve skutečnosti je proces tvorby ontologie daleko komplikovanější – jednotlivé fáze se překrývají.

Pro úsporu času je při tvorbě ontologie racionální využít co největší množství práce, která již byla v oblasti tvorby ontologií vykonána. Proto se snažíme využít existujících mezinárodních ontologií jako základ pro vytvářenou českou ontologii. Nejprve tedy vybereme základovou ontologii tak, aby její struktura a definice typů a vztahů odpovídala zamýšlenému využití české ontologie. Tato část ontologie bude téměř beze změny převzata. Tento přístup je možný díky tomu, že, jak ukazují naše pozorování, existující mezinárodní biomedicínské ontologie ontologie jsou nevhodné k popisu jevů v českém zdravotnictví pouze proto, že neobsahují potřebné koncepty. Struktura typů a vztahů je vyhovující a proto může být převzata.

V druhé fázi je zpracováván korpus textů a jsou v něm vyhledávány povrchové fráze, které budou později mapovány na koncepty ontologie. Formálně je možné vytvořit korpus z libovolných textů, ovšem z hlediska úspěšnosti extrakce informací je vhodné, aby se jednalo o co nejvíce homogenní skupinu textů, ve kterých je informace odvoditelná spíše ze struktury a syntaxe než ze sémantiky. Příkladem takové množiny jsou například příbalové lékové informace. Při vhodně zvolených textech může být vyhledávání povrchových frází částečně zautomatizováno a tedy urychleno. Korpus je nástrojem zpracováván na pozadí tak, jak uživatel prochází další dokumenty v korpusu. Nástroj vyhledává povrchové fráze a uživatel jeho návrhy opravuje a doplňuje. Na základě této zpětné vazby systém zpřesňuje pravidla pro extrakci povrchových frází. Jak nástroj pro extrakci informací, tak nástroj pro zpracování přirozeného jazyka, byli vybrány tak, aby umožňovaly interaktivní reakci na rozhodnutí uživatele.

Ve třetí fázi je každá povrchová fráze identifikovaná v předchozí fázi přeložena do angličtiny a je hledán její protějšek v základové ontologii. Pokud je hledání úspěšné, je povrchová fráze přiřazena konceptu, pokud koncept nalezen není, je potřeba provést namapování ručně nebo v nové ontologii vytvořit koncept nový.

2.4. Extrakce povrchových frází a přiřazování konceptů

Při vytváření ontologie na zelené louce není k dispozici žádný počítačově zpracovatelný zdroj, kromě základové ontologie a jednoduchých slovníků jako například seznam diagnóz. Nejprve dojde k předzpracování korpusu textů: každý dokument je automaticky doplněn lingvistickými informacemi. V současnosti to jsou morfologické značky a slovní lemma. Ale obecně se může jednat o libovolnou informaci o syntaktické nebo sémantické rovině textu, která by mohla pomoct s identifikací povrchových frází. Příkladem jsou informace o struktuře věty.

V následující fázi uživatel prochází a zpracovává jednotlivé dokumenty. Po načtení dokumentu nástroj nejprve označí koncepty na základě nějaké pomocného slovníku, například seznamu léků [9]. Dále jsou aplikována extrakční pravidla generovaná na základě již zpracovaných dokumentů (při zpracování prvního dokumentu žádná taková pravidla neexistují). Povrchové fráze označené v předchozích dvou krocích jsou zobrazena uživateli k posouzení. Uživatel doporučení přijme, opraví nebo doplní. Na základě oprav

učiněných uživatelem, nástroj upraví extrakční pravidla.

V tomto okamžiku jsou všechny povrchové fráze v dokumentu identifikovány, ale mnoho z nich je ve tvaru složených jmenných frází, které musí být před dalším zpracováním rozloženy. Rozklad je v současné době prováděn heuristicky, algoritmem, který bere v úvahu podstatná jména, spojky, interpunkci a závorky ve složené jmenné frázi. Závorky obvykle obsahují příklad obecnějšího konceptu, který závorce předchází. Čárky a spojky oddělují jmenné fráze obsahující stejné podstatné jméno (případně rozvité), ale druhé a případné další výskyty tohoto podstatného jména jsou vynechány. Algoritmus se pokouší rozdělit složenou frázi a znovu zrekonstruovat jednoduché jmenné fráze, z nichž se skládá. Výsledné jednoduché jmenné fráze jsou zobrazeny uživateli jako návrhy, který má možnost je opravit.

Nakonec jsou povrchové fráze přiřazeny konceptům v základové ontologii. Proces přiřazení povrchových frází konceptu může být automatizován, pokud je daný koncept v základové ontologii obsažený. V našem testovacím případě je jako základové ontologie použito UMLS. Povrchové fráze (v češtině) jsou přeloženy do angličtiny pomocí česko-anglického slovníku slovnik.seznam.cz. Překladový modul zkouší překládat frázi jako celek i po jednotlivých slovech. Výsledkem překladu je tedy množina potenciálních výrazů. Tento přístup je možný díky tomu, všechny překlady jsou v dalším kroku testovány proti UMLS databázi a tak jsou vyloučeny všechny nesmyslné překlady.

Překlady povrchových frází jsou testovány pomocí aplikačního serveru UMLSKS [10]. Všechny koncepty, které se shodují s překladem povrchových frází, jsou pak zobrazeny uživateli, aby vybral ten správný. Uživatel má také možnost ručně zadat na který koncept se má povrchová fráze mapovat. Ručního mapování je vždy třeba, pokud se vhodný koncept v základové ontologii nenachází.

Onfram umožňuje vložit nový koncept na základě toho, že se v ontologii nachází *sourozenec* tohoto konceptu. Například antihistaminikum Flonidan, v UMLS neobsažené, bude mít většinu relací stejnou jako lék Zyrtec, tedy koncept v UMLS obsažený. Nástroj umožňuje projít relace, ve kterých se vyskytuje sourozenec (Zyrtec) nového konceptu (Flonidan) a pro nový koncept relevantní vztahy “zkopírovat”.

3. Výsledky

V průběhu vývoje byl nástroj Onfram testován na reálném problému: tvorbě české lékové ontologie na základě příbalových letáků. Nejprve byl vytvořen korpus 200 příbalových letáků, které byly náhodně vybrány z více než tří tisíc příbalových letáků léků dostupných v České republice. Úspěšnost nástroje byla měřena po zpracování poloviny dokumentů. Důvodem bylo, že extrakce povrchových frází je založena na učícím se algoritmu, jehož úspěšnost roste se zvětšující se trénovací množinou, tedy množinou již zpracovaných textů. Ostatní použité metody, jako je zpracování přirozeného jazyka nebo mapování povrchových frází na koncepty, nejsou v aktuální verzi Onframu ovlivněny zpětnou vazbou uživatele a proto jejich úspěšnost nezáleží na množství zpracovaných dokumentů.

Výsledky jsou uvedeny v tom pořadí, v jakém po sobě následují jednotlivé kroky, jsou v procentech a zaokrouhleny na celá čísla.

Nejprve proběhlo obohacení testu o jazykovou informaci. Morfologický analyzátor a lemmatizátor byl úspěšný v 96 procentech případů. To je velmi dobrý výsledek, vzhledem k tomu, že tento nástroj byl vytvořen na základě korpusu obecných textů, jako například novinové články, romány, etc.

Dále byly vyhledány v textu povrchové fráze. Nástroj pro extrakci informací dosáhl úplnosti (recall) 78 procent a přesnosti (precision) 93 procent. Úspěšnost rozkladu složených jmenných frází na jednoduché byla 64 procenta.

Při překladu byl použit on-line slovník slovnik.seznam.cz, který má dobré pokrytí v oblasti medicíny. Sedmdesát osm procent povrchových frází bylo přeloženo správně (v tom smyslu, že hledaný překlad byl v množině možných překladů, kterou vrátil překladový modul). Zhruba 11 procent povrchových frází bylo namapováno přímo prostřednictvím rozhraní UMLS Knowledge Source Server (UMLSKS), které je dos-

tupné na [10]. Toto rozhraní toleruje překlepy stejně tak unifikuje slova, jejichž pravopis kolísá. Tímto způsobem se dařilo najít protějšky slov latinského a řeckého původu, protože pravopis těchto slov se v češtině a angličtině liší jen lehce (například *hypokalemie* versus *hypokalemia*). Celkově bylo automaticky přeloženo a správně namapováno 75 procent povrchových frází. Uživatel pouze prováděl výběr v momentě, kdy bylo na výběr více možných mapování.

Pro zbylých 25 procent povrchových frází nebyl nástroj schopen navrhnout žádný mapování. Tato skupina frází se sestávala z frází, které mají v UMLS protějšek, ale nebyl nalezen (např. *člunkovitá deprese ST* versus *U-shaped ST depresion*) – celkově 17 procent všech frází; a frází, jejichž protějšek v UMLS nebyl schopen nalézt ani zkušený uživatel UMLS ve spolupráci s lékaři – 8 procent všech frází.

4. Diskuze

Nástroj Onfram je navržen především na tvorbu českých biomedicínských ontologií. Nicméně většina jeho komponent je jazykově neutrální a Onfram je navržen tak, že všechny jazykově závislé části, jako nástroje na zpracování přirozeného jazyka a překladový slovník, mohou být snadno nahrazeny. Stejně tak je možné Onfram použít na tvorbu některých ontologií i mimo oblast medicíny. Podmínkou je ale existence vhodné základové ontologie.

Největší slabinou aktuální verze nástroje Onfram je identifikace povrchových frází ve volném textu. V budoucnosti se chceme věnovat vylepšování úspěšnosti extrakce informací. Cestu vidíme jednak v poskytnutí dalších informací programu AMILCARE, jako je například větná struktura, jednak v kombinaci více nástrojů extrakcí informace, jako například AMILCARE a SVM^{light} [11].

Problémem extrakce informací a tvorby ontologií se zabývá několik projektů. Existuje zde tedy několik nástrojů, které poskytují tuto funkcionalitu. Nejznámějšími a nejvíce používanými jsou Protégé [12], nástroj na tvorbu ontologií, a GATE [13], obecný framework pro extrakci informací a zpracování přirozeného jazyka. Pro běžného uživatele zde však existue mezera – nástroj, který by komfortním způsobem spojoval oba nástroje. Proto jsem se rozhodli vyvinout vlastní nástroj, který by splňoval požadavky takového uživatele.

5. Závěr

V článku jsme popsali metody pro vytváření českých biomedicínských ontologiích. Dále jsme popsali nástroj Onfram, který tuto metody implementuje. Ukázali jsme, jakých výsledků nástroj dosahuje na reálném projektu – tvorbě české lékové ontologie. Onfram byl testován na korpusu 200 příbalových letáků. Mapování nalezených povrchových frází na koncepty ontologie je poměrně úspěšné (75 procent).

Náš přístup k vytváření českých biomedicínských ontologií je založen na využití struktury typů a vztahů některé z uznávaných mezinárodních biomedicínských ontologií, které jsou především v anglickém jazyce. Ukázali jsme, že není možné tyto ontologie přímo použít jednak kvůli jazykové bariéře a také kvůli chybějícím konceptům.

Nástroj Onfram zrychluje tvorbu odvozené české ontologie tím, že se pokouší automatizovat některé činnosti při tvorbě ontologie, jako například hledání povrchových frází ve volném textu nebo mapování povrchových frází na koncepty základové ontologie. Díky způsobu vzniku odvozené ontologie je tato do značné míry kompatibilní s původní – základovou – ontologií. Díky tomu, že koncepty jsou hledány v českých medicínských textech, odvozená ontologie obsahuje právě ty koncepty, které jsou v českém zdravotnictví používány.

Onfram se nesnaží plně automatizovat tvorbu ontologie, ani žádné z jeho částí. Cílem projektu je vyvinout nástroj, který pomůže uživateli tím, že redukuje rutinní práci, kterou by bylo jinak potřeba dělat ručně.

Literatura

[1] "The Foundational Model of Anatomy". *http://sig.biostr.washington.edu/projects/fm/index.html*.

[2] "Systematized Nomenclature of Medicine – Clinical Terms". *http://www.snomed.org/snomedct/*.

[3] "Unified Medical Language System." *http://www.nlm.nih.gov/research/umls/*.

[4] Ciravegna F. "(LP)2, an Adaptive Algorithm for Information Extraction from Web-related Texts." *In: Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[5] "AMILCARE Homepage". *http://nlp.shef.ac.uk/amilcare/lp2.html*.

[6] Ciravegna F. "Adaptive Information Extraction from Text by Rule Induction and Generalization". *In Proceedings of IJCAI 2001*, 2001.

[7] "The Prague Dependency Treebank 1.0." *http://ufal.mff.cuni.cz/pdt/*.

[8] Hajič J. "Disambiguation of Rich Inflection - Computational Morphology of Czech". Karolinum, 2004.

[9] Státní ústav pro kontrolu léčiv "Databáze léčivých přípravků". *http://sukl.cz/cs02leciva/index.php*.

[10] "UMLS Knowledge Source Server". *http://umlsks.nlm.nih.gov/*.

[11] Joachims T. "Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning". *In B. Schölkopf and C. Burges and A. Smola (ed.)*, MIT-Press, 1999.

[12] "The Protégé Ontology Editor and Knowledge Acquisition System". *http://protege.stanford.edu/*.

[13] "General Architecture for Text Engineering". *http://gate.ac.uk/*.