Elektronická podoba SSJČ

Pavel Smrž, Karel Pala
(Fakulta informatiky Masarykovy univerzity)

1. Úvod

Slovníky jako rozsáhlé zdroje lexikálních dat se dnes převádějí do elektronických verzí:

- slovníkové informace v elektronické podobě potřebujeme při vyhledávání informací,
- automatickém vytváření abstraktů, při strojovém překladu atd.,
- s počítačovými podobami slovníků pracují dnes stále více i lidé. Práce se slovníky v počítačové podobě je rychlejší a pohodlnější, lze je snáze doplňovat a modifikovat,
- el. verze slovníků se dnes budují pomocí vhodných softwarových nástrojů, manuální sestavování slovníků je nepředstavitelně pracné a také velmi drahé,
- výsledkem jsou slovníky v elektronické podobě a pokud možno v jednotném formátu, který vyhovuje stanoveným standardům,
- nové formáty dovolují systematicky kontrolovat konzistenci lexikálních dat a snadněji je modifikovat při přípravě nových verzí.
- většina českých slovníkových dat, která jsou dnes k dispozici (SSJČ, SSČ), nebyla původně určena pro počítačové aplikace. El. verze SSČ (nakl. Leda) vznikla relativně nedávno (1997).

Převádět existující lexikální zdroje do elektronické podoby se vyplatí, i když je to spojeno s obtížemi:

- získávání lexikálních informací je velmi nákladné a představuje i náročný intelektuální výkon,
- převod je často nesnadný, protože relevantní informace nebývají přímočaře přístupné,
- lex. data nebývají dobře strukturovaná a dostatečně konzistentní.
- lex. data obsahují četné chyby díky tomu, že byla pořizována manuálně.
-

Platí obecně, že manuálně vytvořené lexikální zdroje obsahují chyby vždycky.
Lidský uživatel si s těmito nedostatky poradí, ovšem pro počítačové zpracování představují principiální překážku.
Rozumným řešením proto je budovat slovníky s použitím počítačových technologií (tedy nikoli manuálně) a pak je uchovávat v univerzálním, široce dostupném a znovu použitelném formátu. Takové prostředí nyní poskytuje rodina formátů a nástrojů sdružená kolem (deskriptivního) jazyka XML.

2. Jazyk (a formát) XML

XML (eXtensible Markup Language) (Bray et al. 2000) je standardem pro reprezentaci a výměnu (textových) dat. Představuje silný nástroj, který dovoluje:

- obecný způsob značkování všech typů struktur,
- zachycení vzájemných odkazů mezi nimi,

- víceúrovňové zanoření struktur.

XML je tedy velmi vhodným prostředkem pro reprezentaci silně strukturovaných textových (ale i jiných) dat.
Byl vyvinut zejména s ohledem na použití ve webových aplikacích, jde o zjednodušený dialekt SGML (Standard Generalized Markup Language). Je s ním spojena celá škála technologií, které např. dovolují:

- provádět transformace mezi dokumenty,
- definovat omezující podmínky na dokumenty,
- ověřovat struktury a odkazy uvnitř jednoho dokumentu i vzájemné odkazy mezi dokumenty.

Slovníková data typicky obsahují:

- poměrně složité hierarchické struktury,
- ale také relativně nestrukturovaný volný text.

XML formát dovoluje přesně definovat významové vztahy a vhodně měnit způsob, jímž jsou jednotlivé části textu tištěny nebo zobrazovány.
Při práci se slovníkovými daty ve formátu XML můžeme využít existujících mechanismů pro přístup k datům a manipulaci s nimi – obvykle se mluví o rodině standardů XML.
Strukturu textu kódovaného v XML popisuje tzv. definice typu dokumentu (DTD, Document Type Definition). DTD definuje zobecněná pravidla pro strukturu a určuje explicitně, co je v kódování příslušného dokumentu dovoleno.
Pro jazyk XML existují výkonné dotazovací mechanismy, které umožňují efektivně přistupovat k obsahu rozsáhlých dokumentů, např. XQuery (XML Query Language) (Chamberlin et al. 2001).
XML nabízí řadu možností pro standardní výměnu slovníkových dat.

3. Zvyšování informační obsahu

Slovníky obsahují různé typy informací kódovaných různými způsoby. Používá se různých strukturálních a typografických norem pro reprezentaci:

- morfologických informací,
- popisu významu heslového slova,
- homografů,
- lexikalizovaných flektivních variant,
- kolokací, frazeologismů,
- příkladů, kontextů atd.

V lexikální databázi potřebujeme definovat jednoznačný způsob reprezentace všech těchto entit.
Hodnota elektronických slovníků se podstatně zvyšuje, jestliže sdílejí společné značkování, tj. explicitní vyznačení jednotlivých prvků (částí) slovníkových hesel.
Převod dat ze zdrojového do cílového explicitního formátu bývá označován jako proces zvyšování informačního obsahu (up-translation). Z aplikačního pohledu se jedná o cestu od výchozích slovníkových dat k jejich explicitnímu a tedy strojově použitelnějšímu tvaru.

4. Převod Slovníku spisovného jazyka českého z tištěné do elektronické podoby

Projekt převodu SSJČ (osmisvazkového Slovníku spisovného jazyka českého) do formátu XML se realizuje v rámci komplexního grantového projektu GAČR 405/96/K214 (Čeština ve věku počítačů). Na převodu se podílejí dvě pracoviště a celý proces probíhá v několika fázích:

1. V Ústavu pro jazyk český na Akademii věd ČR byla data SSJČ (stránky slovníku) naskenována, pomocí optického rozpoznávání (OCR) převedena do formátu MS Word (*.doc) a dále zkontrolována, aby se nejprve odstranily viditelné chyby vzniklé při optickém rozpoznávání.
2. Laboratoř zpracování přirozeného jazyka (LZPJ) na Fakultě informatiky MU dostala data ve formátu dokumentů MS Word, a to vždy deset stran textu v jednom souboru.
3. Prvním úkolem v LZPJ byl tedy převod z formátu MS Word do základního formátu XML.
4. Data byla převedena s použitím speciálně vyvinutého programu v jazyce Visual Basic, s nímž MS Word pracuje ve formě maker. Díky své jednorázové povaze nebyl tento krok časově příliš náročný.
5. V další fázi se vyhledaly anomálie ve vstupním formátu (např. roztržená slova vzniklá chybným rozpoznáním přechodů mezi jednotlivými typy písma, konkrétně polotučným a normální kurzívou). Nalezené chyby v kódování byly opraveny.
6. Závěrečnou a nejobtížnější fází převodu je transformace mezivýsledku do vlastního formátu XML odpovídajícího již cílovému datovému typu (DTD, Petkevič 2001). Ideálně odpovídá typ elementu ve slovníku přímo některému typu písma, např.

- normální kurzíva vyznačuje definici významu,
- v hranatých závorkách se vždy uvádí [výslovnost],
- určité skupiny údajů jsou tvořeny hodnotami, které musí patřit do předem daného seznamu (např. výčty zkratek, jména autorů).

SSJČ nepracuje se zkratkami konzistentně, např. je typické, že jedna zkratka má několik variant (biol., biolog., styl., stylist.,)
Většina podstatných těžkostí při převodu je dána právě nekonzistencí struktury hesel:

- v rámci jednotlivých hesel se rozdílně uvádějí příklady, kontexty a kolokace,
- hesla v SSJČ se liší ve svých strukturách, tj. na tom místě v hesle, kde má standardně být údaj o významu, můžeme snadno najít jiný údaj.

Tyto inkonzistence pak znemožňují plně automatický převod hesel z původní podoby do podoby plně konzistentní.
V současné fázi proto pracujeme s dvěma variantami XML:
a) nízkoúrovňové formát (viz níže příklad 1) je výhodnější pro opravování nalezených chyb,
b) formát vyšší úrovně -- odpovídá cílovému DTD (příklad 2), je vhodný pro některé dotazy na konkrétní části hesel, i když heslo obsahuje nesprávně rozpoznané prvky.
Nalezené chyby se postupně opravují: nejčastěji jde o nesprávně rozpoznané typy písma, tyto chyby znemožňují automatický převod do výsledného tvaru. Do speciální kategorie patří nekonzistence a chyby vyskytující se už v tištěné verzi slovníku, např. jde o nekonzistence v popisu významů heslových slov nebo chybné struktury hesel. Pokud se na ně přijde, zaznamenávají se odděleně, abychom je kdykoli mohli konfrontovat s původní podobou dat. K odhalování těchto chyb připravujeme speciální nástroj -- analyzátor slovníkových hesel

v SSJČ či SSČ (parciální syntaktický analyzátor pro češtinu DIS, Žáčková, 2001).
Je ovšem otázka, jak daleko lze v těchto opravách jít, a kolik to může stát.

Příklad 1: Nízkoúrovňové kódování dat – zde jsou vyznačeny jen různé typy písma

```
<entry>
<bold>terorismus</bold>
<ital>způsob vlády vymáhající terorem poslušnost;
hrůzovláda, krutovláda,
despotismus:</ital>
<norm>vojenský t.; nesnesitelný t.; demagogie a t.; </norm>
<small>přen. expr.</small>
<norm>to je t., nedejte si to líbit</norm>
</entry>
```

Příklad 2: Formát kódování hesla SSJČ odpovídající cílovému DTD

```
<entry>
  <hw>
       <orth>terorismus</orth>
  </hw>
   <morph>
    <paradig>socialismus</paradig>
  </morph>
  <senses>
    <sense>
      <def>způsob vlády vymáhající terorem poslušnost</def>
      <def>hrůzovláda</def>
      <def>krutovláda</def>
      <def>despotismus</def>
      <eg>vojenský terorismus</eg>
  <eg>nesnesitelný terorismus</eg>
   <eg>demagogie a terorismus</eg>
      <eg>
    <usg
      type=style>přen.expr.</usg>
          to je terorismus, nedejte si to líbit
      </eg>
    </sense>
  </senses>
</entry>
```

5. Manažer lexikálních databází kódovaných v XML

Pro práci se SSJČ a libovolnými dalšími slovníky, jejichž data jsou uložena ve formátu XML byl v LZPJ FI MU vytvořen systém MAXXL. (Karásek, 2000, DP). Jeho hlavní rysy jsou:

- umožňuje efektivní ukládání a vyhledávání slovníkových dat,

- je postaven na architektuře klient/server,
- serverová část data vyhledává a ukládá,
- klientské programy zprostředkovávají komunikaci s uživateli, usnadňují definici dotazů a prezentaci vyhledaných záznamů.

MAXXL je napsán v programovacím jazyce C++ s rozhraními pro další jazyky (Perl, Python, Java). Uživatelé mohou modifikovat výstup tohoto nástroje tak, aby výsledek popisoval zamýšlenou strukturu dokumentu přesně. Lexikální databáze je v systému MAXXL chápána jako množina dokumentů XML. Systém MAXXL reprezentuje data v UNICODE, konkrétně v kódování UTF-8. Lze tak zpracovávat data v libovolné současné abecedě. Systém definuje vlastní dotazovací jazyk. Výsledek dotazu má formu posloupnosti elementů XML nebo přímo sledu slov. K dispozici jsou operátory pro přesnou shodu, prefixové vyhledávání a lokalizace obecných podřetězců. MAXXL umožňuje i tzv. morfologickou expanzi dotazů v různých jazycích – nabízí mechanismus integrace externích morfologických analyzátorů:

- generuje všechny slovní tvary pro daný tvar včetně gramatických kategorií,
- k zadanému slovnímu tvaru nabídne jeho základní tvar (lemma),
- k zadanému slovu přiřadí jeho ohýbací vzor (paradigma).

Systém lze přímo napojit na korpusový manažer Manatee navržený a implementovaný na FI MU (Rychlý 1999). Tak lze splnit požadavek lexikografů – pracovat při budování či úpravách slovníku s reálnými – korpusovými daty, slovníková data lze přímo porovnávat s daty korpusovými, která jsou k dispozici v konkordancích získaných z korpusů

6. Výsledky

Vlastní výsledky jsou k dispozici na CD ROM, který obsahuje data SSČ a SSJČ a prohlížeč MAXXL (s pracovním názvem gslov). Toto CD bylo vytvořeno v Laboratoři zpracování přirozeného jazyka FI MU (pala@fi.muni.cz). V další verzi tohoto CD bude připravena (do konce r. 2001) integrace automatického morfologického analyzátoru ajka (Sedláček, Smrž, 2001) do SSJČ a SSČ (tj. do systému gslov). Oba slovníky tak získají novou kvalitu – dovolí na požádání získávat údaje o:

- paradigmatech heslových slov,
- segmentaci slovních tvarů,
- gramatických kategoriích,
- generovat příslušné slovní tvary,
- a také je lemmatizovat.

Literatura

BRAY, T. et al. 2000. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. http://www.w3.org/TR/1998/REC-xml.
COPESTAKE, A. 1995. ACQUILEX. http://www.cl.cam.ac.uk/Research/NL/acquilex/.
CHAHUNEAU, F. 1994. Current Approaches to SGML Up-translation. http://www.oasis-open.org/cover/fcha.html.
CLARK, J. 1999. XSL Transformations (XSLT). Version 1.0. W3C Recommendation. http://www.w3.org/TR/xslt/.
CLARK, J. 2001. XSL Transformations (XSLT). Version 1.1. W3C Working Draft. http://www.w3.org/TR/xslt11/.

ELLIOTT, L. 2001. How the Oxford English Dictionary Went Online. Ariadne, č. 24. http://www.ariadne.ac.uk/issue24/oed-tech/.
IDE, N. 2000. The XML Framework and Its Implications for the Development of Natural Language Processing Tools. In: Proceedings of the COLING Workshop on Using Toolsets and Architectures to Build NLP Systems.
KARÁSEK, L. 2000. Systém pro tvorbu a presentaci vícejazyčných a výkladových slovníků. Diplomová práce na Fakultě informatiky Masarykovy univerzity v Brně.
Petkevič, V. 2001. Návrh DTD pro SSJČ -- 1.verze, rukopis.
RYCHLÝ, P.,Korpusové manažery a příslušná rozhraní, Disertační práce, FI MU BRNO, 1999
Sedláček, R. -- Smrž, P. 2001. A New Czech Morphological Analyser ajka. In: Proceedings of the 4th International Conference on Text, Speech and Dialogue, September 2001, Železná Ruda, Springer Verlag, Berlin, p.100-107.
Smrž, P., Využití formátů XML pro ukládání lexikálních databází, Sborník konference SLOVKO 2001, říjen 2001, Bratislava.

Obdobné a příbuzné projekty

Jako příklad transformace tohoto typu uveďme převod rozsáhlého výkladového slovníku OED (Oxford English Dictionary) Online (Elliott 2001) do strojově čitelné podoby. V polovině 80. let se nakladatelství OUP (Oxford University Press) rozhodlo vydat druhé vydání svého největšího slovníku, tj. originálního 12 svazkového souboru se všemi dodatky. Elektronická verze byla potřebná pro efektivní práci s pozdějšími revizemi OED. Přibližně 150 písařek přepsalo celý OED a po sérii mnoha oprav byl slovník nakonec v roce 1989 úspěšně vydán. Kódování OED neodpovídalo plně SGML, "vzhledem k unikátnímu obsahu a dlouhému vývoji editorského stylu", nebylo možné svázat celý slovník jednotným formátem. I dnes je OED revidován pomocí vlastního značkování a speciálních softwarových nástrojů vyvinutých na zakázku. Jasně jsou identifikovány definice výrazů, výslovnost, variantní ortografie, etymologie, doklady a jejich datace, včetně jmen autorů a názvů děl, z nichž je citováno. Náklady na vývoj programových produktů OED Online činily 400.000 USD a nakladatelství OUP utratilo přibližně další milion USD marketing, konzultace atd.