Využití jazykových technologií v oblasti eLearningu

Vladislav Kuboň, Miroslav Spousta
UFAL MFF UK
Malostranské nám. 25 ¨
118 00 Praha 1

1. Úvod

Neustálý tlak moderní společnosti na zkvalitňování vzdělání a zejména na jeho zpřístupnění co největšímu počtu zájemců vynesl v posledních letech do centra pozornosti i oblast učení na dálku, probíhajícího prostřednictvím moderních počítačových a komunikačních technologií. Pro tuto oblast výuky se ujal i v českém prostředí stručnější název eLearning, který budeme pro jednoduchost používat i ve zbytku tohoto článku. Jedním z největších problémů technologií používaných v této oblasti je kromě zajištění kvalitního obsahu i jeho převod do formátů, používaných standardně pro práci s metadaty. Dalším problémem je vzájemná provázanost jednotlivých učebních objektů, možnost efektivního vyhledávání relevantního obsahu, vícejazyčnost apod.

Mezinárodní výzkumný projekt Jazykové technologie pro eLearning (Language Technologies for eLearning, LT4eL), financovaný Evropskou Unií v rámci 6.rámcového programu, se pokouší prozkoumat možnosti řešení výše uvedených problémů pomocí technologií existujících v oblasti automatického zpracování přirozeného jazyka. Projektu se kromě koordinující univerzity z holandského Utrechtu, účastní i dalších 11 partnerů, z toho 9 univerzit (mezi nimi i Univerzita Karlova), a 2 pracoviště akademie věd. Celkem je v rámci projektu zkoumáno využití jazykových technologií pro 9 jazyků – angličtinu, bulharštinu, češtinu, holandštinu, němčinu, maltštinu, polštinu, portugalštinu a rumunštinu. Projekt si klade následující cíle:

-Zlepšit získávání výukových materiálů,
-podpořit vytváření specifických kursů pro jednotlivé uživatele,
-zlepšit tvorbu personalizovaného obsahu,
-podporovat decentralizaci při spravování obsahu,
-umožnit vícejazyčné vyhledávání.

Ačkoli si projekt klade za úkol vyvinout a vyhodnotit obecně použitelné technologie nezávislé na konkrétním systému pro správu výukového obsahu (Learning Management System – LMS), jsou v jeho rámci vyvíjené funkcionality implementovány do konkrétního systému. Je jím systém ILIAS (www.ilias.de), který se již v minulosti dostatečně osvědčil a je široce používán celou řadou vzdělávacích institucí.

2. Poloautomatické generování metadat

Aby bylo možné efektivně pracovat s výukovými objekty v LMS systémech, je třeba je opatřit popisem na úrovni metadat. Ačkoli na takové případy v praxi můžeme často narazit, dobrým výukovým materiálem pro eLearning prostě není pouhý soubor prezentace v PowerPointu nebo handoutů z přednášky. Efektivní eLearning vyžaduje vzájemné propojení výukových dat na metaúrovni, která zajišťuje rychlou a přehlednou orientaci studenta. V rámci projektu LT4eL se problematika opatřování výukových objektů metadaty řeší ve dvou oblastech – v oblasti automatického generování klíčových slov a v oblasti automatické extrakce definic.

2.1 Získávání a zpracovávání dat

Dříve než je možné jakýmkoli způsobem přidávat metadata, je nutné texty výukových objektů převést do jednotného formátu. Na vstupu je nutné počítat se všemi frekventovanými typy formátů, tedy DOC, HTML, RTF a PDF. Zejména poslední formát je velmi problematický a volně dostupné nástroje pro jeho automatické zpracování a převod do holého textu vykazují řadu chyb zejména ve vztahu k nárokům na zpracování textů ve více jazycích – nezapomeňme, že mezi jazyky zpracovávanými v projektu LT4eL je i bulharština, používající azbuku, a několik jazyků s bohatou diakritikou, která

znemožňuje se spokojit s nástroji na převod PDF do holého textu, které pracují relativně spolehlivě pro angličtinu.
Vlastní tok zpracování vstupního textu se skládá z následujících kroků:

1. Převod z originálního formátu do běžného HTML pomocí volně dostupných nástrojů.
2. Odstranění formátovacích informací, které jsou pro naše účely nepodstatné a převod redukovaného HTML formátu do XML. Jedná se o formát definovaný pomocí DTD nazývaného BaseXML a vyvinutého v rámci projektu.
3. Extrakce textu z BaseXML, jeho segmentace a tokenizace.
4. Lingvistická anotace textu. Pro tuto anotaci používáme pro češtinu morfologickou analýzu Jana Hajiče (viz Hajič 2001) a stochastický tagger (viz Hajič, Hladká 1998). Pro ostatní jazyky se používají obdobné nástroje.
5. Spojením anotovaného textu a formátovacích informací z BaseXML dostaneme konečnou podobu dokumentů. Jedná se o speciální DTD nazvané LT4eLAna. Tento formát slouží jako vstup pro další lingvistické zpracování textů pomocí nástrojů představených v následujících odstavcích.

2.2 Detekce a extrakce klíčových slov

Podrobný popis nástroje na detekci a extrakci klíčových slov je uveden v článku Lemnitzer, Degórski 2006, zde se omezíme pouze na stručný přehled hlavních zásad. Extraktor klíčových slov je založen na předpokladu, že dobré klíčové slovo má určité kvalitativní a kvantitativní charakteristiky. Protože klíčová slova jistým způsobem reprezentují text, měla by se v něm vyskytovat dostatečně často, častěji, než by odpovídalo očekávané pravděpodobnosti jejich výskytu. Protože četnost výskytu nějakého termínu v textu se řídí Poissonovým rozdělením, můžeme jakoukoli pozitivní výchylku proti neutrálnímu stavu považovat za indikátor významnosti daného termínu pro daný text. Druhou důležitou charakteristikou je termín zavedený Katzem v článku Katz 1996. Jedná se o tzv. term burstiness, založený na předpokladu, že se dobrá klíčová slova v textu vyskytují pohromadě. Jakmile se klíčové slovo vyskytne jednou, je následováno několika dalšími výskyty v intervalech, které jsou kratší, než by odpovídalo průměrné četnosti rozdělení výskytů daného termínu v textu. Jakmile tato anomálie skončí, vrací se frekvence k normálu, dokud se termín v textu opět nevyskytne. Toto chováni v podstatě odpovídá tomu, že k důležitému klíčovému slovu se v následujícím úseku textu opětovně odkazuje.
Kvalitativní charakteristiky se projevují zejména v tom, že důležitá klíčová slova se vyskytují často v signifikantních úsecích textu, např. v nadpisech kapitol, v prvních a posledních odstavcích apod. Tohoto faktu využívaly v minulosti nejrůznější nástroje detekující klíčová slova v češtině, z nejstarších uveďme alespoň systém MOZAIKA Zdeňka Kirschnera (viz Kirschner 1983).
Zvláštní péči věnuje extraktor klíčových slov víceslovným termínům. Ruční anotace výukových textů ukázala, že v mnoha případech se klíčové termíny skládaly z více slov, např. v češtině *specifikační jazyk, digitální fotografie, digitální zvukové knihy* či v polštině *wirtualna szkola, system zarządzania nauczaniem* apod. Ve většině případů je víceslovným termínem nějaké rozvité klíčové slovo. Stačí tedy sledovat, zda vybraní jednoslovní kandidáti na klíčová slova nesousedí s nějakým jiným slovem, se kterým by se v textu vedle sebe vyskytovali častěji, než by odpovídalo náhodnému souvýskytu. Toto měření tzv. vzájemné informace pak slouží jako základ pro extrakci víceslovných termínů.

2.3 Hledání definitorických kontextů

Výstup získaný po průchodu automatickým extraktorem klíčových slov (a eventuální ruční úpravě výsledků) je vstupem do dalšího modulu, který se pokouší automaticky hledat definitorické kontexty nalezených klíčových slov. Technicky je na vstupu text ve formátu XML s morfologickými značkami a vyznačenými klíčovými slovy.
Tento vstup je zpracováván jednoduchými regulárními gramatikami, odlišnými pro každý zpracovávaný jazyk. Gramatiky sou implementovány pomocí nástroje lxtransduce (Tobin 2005). Tento nástroj patří do sady LTXML2, vyvinuté na edinburské univerzitě.
Česká gramatika je rozsahem největší mezi gramatikami slovanských jazyků zastoupených v projektu. Jejím jádrem jsou pravidla vyvinutá Nguyenem Thu Trangem na univerzitě v Tuebingenu na základě

pozorování českých hesel ve Wikipedii. Bohužel se ukázalo, že anotované české texty, které byly označkovány a zpracovány pro použití v projektu, nemají jasnou a pravidelnou strukturu slovníkových hesel Wikipedie, proto bylo nutné v několika iteracích gramatiku upravit tak, aby lépe vyhovovala charakteru těchto textů.
Česká gramatika používá 21 pravidel nejvyšší úrovně, rozdělených do pěti skupin. Většina správně rozeznaných konstrukcí patří do kategorie NP je/jsou NP, relativně úspěšná jsou také výrazy s určitými vybranými slovesy – *definuje*, *znamená*, *vymezuje*, *představuje* apod. Přehled počtů jednotlivých typů pravidel je v Tab.1.

Konfigurace	Počet pravidel
NP je/jsou NP	52
NP sloveso NP	45
strukturální	39
NP (NP)	30
NP–/:/= NP	20
Jiné konfigurace	59

Tab.1 Přehled typů českých gramatických pravidel

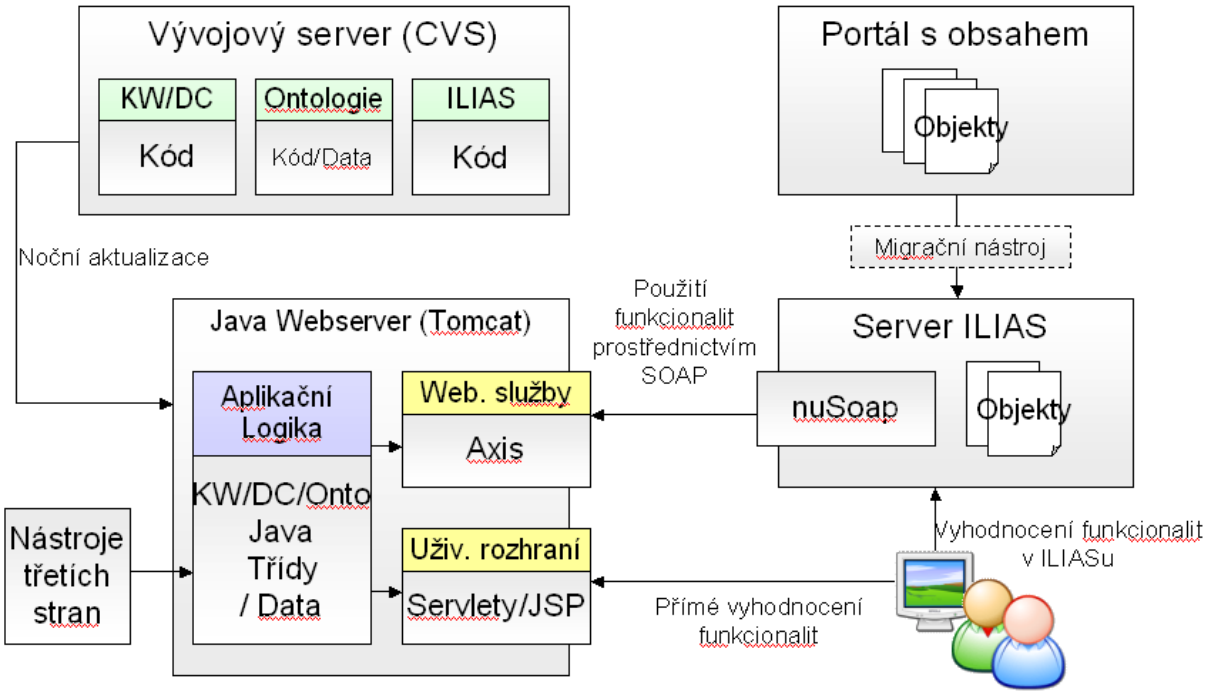
Jedním z testů, které byly provedeny na ručně označkovaných datech, byl test mezianotátorské shody. Výsledek ukázal, že ani mezi lidskými anotátory nepanuje dostatečná shoda o tom, kdy je nějaký úsek textu definicí. Protože srovnatelně nízký výsledek byl zjištěn i pro polštinu, zdá se, že úkol nalezení definic v běžném textu je skutečně objektivně velmi obtížný.
Proto není překvapivé, že přesnost extraktoru definičních kontextů dosáhla pro češtinu pouze 22% při pokrytí 46%.
Existuje však ještě jeden důvod, proč jsou výsledky českého i polského extraktoru tak nízké. Je to bohaté skloňování, zejména u podstatných jmen. Je velmi obtížné vytvořit spolehlivá pravidla pro zachycení definic, pokud mají jednotlivá slovesa různé vazby. Volný slovosled, umožňující, aby v konstrukci NP sloveso NP byla první jmenná fráze předmětem a teprve druhá podmětem, přesnosti extrakce také nepřidá. V obou gramatikách stále chybějí všechny možné konfigurace definic, variant je příliš mnoho. Ukazuje se, že nástroj zvolený pro implementaci gramatik je přece jen více vhodný pro jazyky s pevnějším slovosledem a menší flexí a že by možná stálo za to opustit snahu o používání jednotného nástroje napříč jazyky a podobně jako u morfologického značkování vyvíjet gramatiky pro definitorický kontext pomocí různých prostředků, vhodných pro příslušné typy jazyků.

3. Zapojení sémantiky

Druhým důležitým cílem projektu LT4eL je zapojit do vyhledávání výukových objektů sémantiku. V projektu se tak děje prostřednictvím ontologií, které jsou jak obecné, tak i oborově zaměřené. Při budování ontologií bereme ohled na dva typy uživatelů, na lektory (učitele) i na studenty. Modelovým použitím našeho systému lektorem je příprava nového kurzu pro určitou cílovou skupinu, založeného na existujícím materiálu. Studentský scénář předpokládá použití studentem, který využívá systém k vlastnímu sebevzdělávání neřízenému učitelem.
V obou případech by vyhledávání relevantních informací pomocí sémantického vyhledávání mělo pomoci překonat nedostatky běžného vyhledávání pomocí fulltextových vyhledávačů. Je zcela běžné, že vyhledávače vydají buď příliš mnoho odkazů, ve kterých je velmi těžká orientace, či odkazy ne úplně relevantní, s náhodným souvýskytem vyhledávaných klíčových slov.
V našem systému je každý výukový objekt indexován množinou ontologií, které jsou pro daný objekt relevantní. Vyhledávání zprostředkované ontologiemi potom umožní propojovat jednotlivé dokumenty podle relevance, a to i dokumenty napsané v různých jazycích. Pro otestování tohoto přístupu jsme v projektu zvolili dvě tématické oblasti – využití počítačů humanitně zaměřenými uživateli a eLearning. Vývoj doménových ontologií probíhá prostřednictvím specializace hlavní ontologie. Jako

hlavní ontologii jsme zvolili DOLCE (viz Masolo a další, 2002). Tato ontologie je provázána se slovníčky jednotlivých jazyků a jejich prostřednictvím umožňuje vícejazyčné vyhledávání. Ontologie byla rozšířena o další koncepty odvozené od existujících konceptů, o nadkoncepty existujících konceptů, o chybějící podkoncepty a o koncepty pocházející z anotovaného textu. Po doplnění hlavní ontologie byla vytvořena doménová ontologie, která má v současné době asi 750 doménových konceptů, 50 konceptů z DOLCE a 250 konceptů z OntoWordNetu. Podrobnější informace o použití ontologií je možné nalézt ve článku Simov, Osenova, 2006.
Ačkoli jsou do systému ontologií zapojena i klíčová slova z výukových dokumentů, není jich dost, aby vždy dokázala spojit ontologie s jejich explikacemi v textu. Tento nedostatek je způsoben zejména tím, že zdaleka ne všechna klíčová slova jsou přítomna ve všech jazycích a naopak, některé existující koncepty nejsou v textech vůbec přítomny. Proto je nutné jako mezistupeň zapojit pro každý zpracovávaný jazyk jeho slovník, který se spojí s ontologií, dále mechanismus pro rozeznání lexikálních jednotek v textu a mechanismus pro výběr správných konceptů pro frázi, pokud je tato fráze víceznačná a mohla by být spojena s více koncepty.

4. Integrace do systému ILIAS
Jedním z nejdůležitějších úkolů projektu je zapojit vyvíjené technologie do systému ILIAS (www.ilias.de). Tento systém používá skriptovací jazyk PHP, má otevřené zdrojové kódy, je tedy do něj možné implementovat další funkcionality, vycházející z jazykových nástrojů popsaných v předešlých kapitolách. Stručné schéma postupu integrace nových funkcionalit je znázorněna na Obr.1.



Obr. 1. Integrace nových funkcinalit do systému ILIAS

Celková architektura systému byla navržena s ohledem na maximální flexibilitu a rozšiřitelnost. Bylo potřeba jednak vybrat prostředí pro implementaci nové funkcionality (vyhledávání, extrakce klíčových slov a hledání definitorických kontextů), jednak integrovat danou funkcionalitu do stávajícího LMS ILIAS.

Jelikož výsledek projektu by měl být použitelný také v dalších LMS, byla nová funkcionalita (tzv. Language Technology server, LTserver) implementovaná odděleně od ILIASu, jako Java Web Application běžící na serveru Apache Tomcat. Ke komunikaci s LMS bylo definováno rozhraní webových služeb (SOAP), které slouží pro vzájemnou komunikaci a vzdálené volání jednotlivých

metod LT serveru. Pro ladění a testování LTserveru je k dispozici také textové rozhraní (command-line interface) a grafické rozhraní implementované jako JSP (Java Server Pages). Samotný systém ILIAS bylo potřeba rozšířit o možnost komunikace pomocí web services pomocí knihovny nuSoap.

LT server se skládá z několika oddělených modulů:

- správa dokumentů (součástí je také předzpracování dokumentů popsané v kapitole 2.1)
- extraktor klíčových slov
- extraktor definitorických kontextů
- fulltextové vyhledávání
- vyhledávání podle klíčových slov
- sémantické vyhledávání

Jednotlivé komponenty využívají pro ukládání informací (např. statistiky výskytů slov pro extraktor klíčových slov) knihovnu Berkeley DB Java Edition, vyhledávací moduly využívají fulltextový vyhledávač Apache Lucene (rozšířený pro použití sémantického vyhledávání).

Během procesu integrace jsme narazili na celou řadu problémů, vyvolaných faktem, že jednotlivé nástroje jsou vyvíjeny relativně nezávisle na několika akademických pracovištích. Dochází tak často ke snížení kvaliy a spolehlivosti jednotlivých součástí systému a ztěžuje jejich vzájemné propojení. To se projevuje při testování systému nejrůznějším způsobem – například se ukázalo, že sémantické vyhledávání, ač se jeví jako použitelné samo o sobě, není dostatečně rychlé a bezpečné z hlediska běhu více vláken najednou na web serveru. Důsledkem bylo naprosté zhroucení celého systému v okamžiku, kdy se několik testerů pokusilo spustit sémantické vyhledávání v krátkých časových intervalech (vyhledání jednoho termínu trvalo řádově minuty reálného času).
Tyto počáteční problémy se však podařilo překonat reimplementací problematických modulů, což umožnilo zahájit uživatelskou validaci celého systému..

5. Validace
Ověření, zda nově přidané funkcionality přinesou uživatelům nějaký efekt, je předmětem rozsáhlé uživatelské validace celého systému. Pro tyto účely bylo vyvinuto několik scénářů, testujících buď celý systém komplexně, či pouze jeho jednotlivé části, a to jak z hlediska učitele, tak i potenciálního studenta. Tyto testy budou probíhat iterativně v několika etapách – na základě výsledků první etapy dojde k úpravě modulů a k jejich dalšímu testování.
Při validaci dostanou testeři daný scénář, provedou požadované činnosti a nakonec vyplní dotazník, ve kterém zhodnotí své dojmy. Abychom předešli náhodnému a ne příliš pečlivému vyplňování dotazníku, jsou v něm uvedeny i páry otázek, které se vzájemně negují – v případě, že testovaná osoba vyplní odpovědi náhodně, je značná pravděpodobnost, že bude možné tuto skutečnost odhalit.
Scénáře jsou různorodé, jeden z nich na testování vícejazyčného sémantického vyhledávání, např. vypadá takto:

Ráchel má za úkol připravit prezentaci v MS Office. Během přípravy prezentace zjistí, že
- Existují terminologické rozpory mezi anglickými slajdy použitými v kurzu a jejími německými poznámkami
- Vynechala několik přednášek z Excelu a Powerpointu. Potřebuje tedy najít informace o těchto programech v němčině, použije proto ILIAS, aby

- sjednotila poznámky v němčině a v angličtině, a aby se ujistila, že jsou správně;
- nalezla chybějící německé informace na základě anglických dat;
- mohla zajistit dobrou kvalitu výsledného dokumentu, i když jako nezkušená studentka prvního ročníku není schopna posoudit výsledky vyhledávacích nástrojů typu Google z hlediska správnosti a relevance.

Scénáře jsou lokalizovány do jednotlivých jazyků projektu nejen s ohledem na použitou terminologii a na reálnost daného scénáře, ale zejména s ohledem na testovací data přítomná v systému. Pokud například výukové objekty v určitém jazyce referují k OpenOffice místo MS Office, je tato skutečnost ve scénáři zohledněna.

6. Závěr

Přestože do dokončení projektu LT4eL ještě zbývá několik měsíců, zdá se, že jeho výsledky budou zajímavé. Projekt se zabývá otázkami, které v oblasti eLearningu doposud nikdo neřešil, a i když některé dílčí výsledky jednotlivých modulů nesnesou srovnání s experimentálními systémy, je třeba si uvědomit, že projekt se zabývá zpracováním nijak neupravovaných, v podstatě náhodně vybraných dokumentů (určujícím faktorem při výběru dokumentů byla jejich volná dostupnost nezatížená problémy s autorskými právy). V tomto ohledu se snaží přiblížit co nejvíce realitě a naznačit další směřování potřebného výzkumu. Ve zbývajících měsících do konce projektu v květnu 2008 očekáváme zejména odpověď na klíčovou otázku – zda poloautomatický přístup k anotaci dokumentů spojený se sémantickým vyhledáváním prostřednictvím systému ontologií vůbec má smysl a jak velkou přidanou hodnotu uživatelům eLearningových systémů přináší.

Literatura

Jan Hajič and Barbora Hladká. 1998. *Tagging Inflective Languages. Prediction of Morphological Categories for a Rich, Structured Tagset.* ACL-Coling'98, Montreal, Canada, pp. 483-490.

Jan Hajič. 2001. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Charles University Press, Prague, Czech Republic.

S.M.Katz 1996: *Distribution of content words and phrases in text and language modelling* V: Natural Language Engineering 2 (1996) 1, pp.15-59.

Z Kirschner MOSAIC – A Method of Automatic Extraction of Significant Terms from Texts, V sérii Explizite Beschreibung der Sprache und automatische Textbearbeitung X, MFF UK Praha, 1983.

Lothar Lemnitzer and Łukasz Degórski: *Language Technology for eLearning -- Implementing a Keyword Extractor* přijato k prezentaci na 4. EDEN Research Workshop "Research into online distance education and eLearning. Making the Difference", 25-28 OCTOBER, 2006 v Castelldefels, Spain

Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, Luc Schneider. 2002. WonderWeb Deliverable D17. The WonderWeb Library of Foundational Ontologies and the DOLCE ontology. Preliminary Report (ver. 2.0, 15-08-2002).

Kiril Simov, Petya Osenova: *Applying Ontology-Based Lexicons to the Semantic Annotation of Learning Objects*. RANLP 2007 workshop: Natural Language Processing and Knowledge Representation for eLearning Environments. Borovets, 26. September 2007.

Richard Tobin, 2005. Lxtransduce, a replacement for fsgmatch. University of Edinburgh. http://www.cogsci.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html.