Technologie zpracování přirozeného jazyka pro elektronickou podporu výuky[1]

Pavel Smrž, Tomáš Pitner, Tomáš Gregar

Abstrakt

Tento příspěvek se zabývá aplikací postupů a technologií zpracování přirozeného jazyka (Natural Language Processing - NLP) v oblasti elektronické podpory výuky (e-learningu). Slouží jednak jako přehled dostupných relevantních technologií zpracování jazyka a současně jako příkladová studie ukazující aplikaci základních ideí pro češtinu. Řada diskutovaných technik a postupů je již ve stavu bezprostřední použitelnosti, mnoho (zatím většina) z nich ale na využití teprve čeká.

Úvod

Současné projekty návrhu, realizace a nasazení systémů pro řízení výuky (Learning Management Systems – LMS) se dosud převážně zaměřují na administrativní stránku řízení (evidence studentů, hodnocení) a prezentaci studijních materiálů a jiných pomocných nástrojů. Případně nabízejí nástroje pro on-line či off-line spolupráci v rámci studijních skupin, kontakt s učitelem/tutorem atd.
V této oblasti, zejména pokud jde o tuzemské prostředí, existuje pouze mizivé množství aktivit zaměřených na původní výzkum, který by si dal za úkol integraci pokročilých technologií souvisejících se zpracováním jazyka – asistivních (dialogové systémy, syntéza a rozeznání řeči…), postupů a nástrojů získávání a správy znalostí apod. – do systémů podpory výuky.

Aplikovatelnost technik NLP pro e-learning

Pokusme se nyní stručně identifikovat hlavní kritéria určující relevanci a použitelnost technik NLP pro e-learning. Zcela jistě mezi ně budou patřit:

- použitelnost v synchronním/asynchronním módu;
- složitost;
- integrovatelnost se stávajícími LMS;
- technologická zralost;
- dostupnost pro požadovaný jazyk a portabilita;
- použitelnost pro autory, tutory a/nebo studenty.

Dle posledního kritéria pak můžeme nástroje rozdělit na ty, se kterými bude v rámci elektronické výuky pracovat student (koncový uživatel, subjekt výuky) a se kterými učitel/autor kurzu.

[1]Příspěvek vznikl v rámci řešení projektu Národního programu Informační společnost „E-learning v kontextu sémantického webu“.

Nástroje pro studenty

Vyhledávání

Vzhledem k tomu, že současné LMS zřídkakdy obsahují subsystémy využívající zpracování přirozeného jazyka (a to platí zejména pro jiné jazyky než angličtinu), nemohou podporovat ani základní jazykově orientované vyhledávání – vyhledávání s respektováním morfologie jazyka (většinou fulltextové) – ani získávání výukových materiálů.
Je přitom známo, že čeština spolu s ruštinou jsou jedinými evropskými jazyky, pro něž jsou lokální internetové vyhledávače zohledňující tvarosloví úspěšnější než sice veleúspěšný, ale příliš obecný Google. Na bázi dnes již klasického nástroje vyvinutého v Laboratoři zpracování přirozeného jazyka (NLPlab FI) – morfologického analyzátoru AJKA (Sedláček a Smrž, 2001) – je budován systém umožňující vyhledávat přístupné elektronické učební materiály.

Odhadování jazyka a kódování

Vzhledem k tomu, že zdrojový jazyk dokumentu (a jeho kódování) nejsou vždy explicitně určeny, existuje v rámci zamýšleného vyhledávacího stroje elektronické podpory výuky tzv. *odhadovač jazyka* (a kódování) (language guesser). Ten bude v předloženém dokumentu (na základě porovnávání s texty různých jazyků v různých kódováních – uloženými např. v korpusech) odhadovat správnou kombinaci.

Dotazovací systém

Další jazykový prostředek, který již v našich experimentech ukázal svou užitečnost, je rozšiřující modul umožňující reagovat na dotazy (query expansion module) založený na informacích z českého WordNetu (Pala a Smrž, 2004). Podstatné je v této souvislosti zejména zpříjemnění uživatelského prostředí používaných nástrojů. Uživatelé díky němu mohou vyhledávat nejen všechny tvary slova, ale i synonyma, hyperonyma a jiná sémanticky příbuzná slova, vylučovat kohyponyma apod. Mohou ale také například určovat kontext použití (varianty spisovné, regionální, slohové…), zda se mají vyhledat slova odvozená od vyhledaného pojmu, příbuzná slovesa dle slovesného vidu apod.

Sumarizace obsahu

Velmi užitečným nástrojem bude modul (je právě vyvíjen), který má sumarizovat obsah příbuzných dokumentů. Dílčí aplikací takového modulu by měla být sumarizace zpráv z diskusních fór vyučovaných předmětů. V diskusích provozovaných v současnosti prostřednictvím subsystému pro elektronickou podporu výuky v rámci IS MU jsou u některých předmětů denně vloženy desítky dotazů. Při takové aktivitě studentů může i jen jednoduché prohlížení vláken diskuse týkající se

minulé přednášky představovat nepřiměřenou zátěž. Přitom jakékoli informace získané od studentů mají velmi velkou hodnotu – umožňují například určit, jaké oblasti by se měly vyučovat podrobněji, kde se chybuje, nebo co učitel zapomíná kontrolovat testy. Jednou z nejjednodušších, ale přitom potřebných využití takovéto sumarizace je například kontrola a doplňování seznamu klíčových slov definovaných pro předmět. Využijeme přitom metody založené na statistických testech, známé z počítačové lingvistiky. Automaticky vyextrahovaná klíčová slova jsou pak porovnána a doplněna se seznamem, který zadal autor kurzu.

Automatická klasifikace učebních materiálů

S dostupnými prostředky je již možné provádět automatickou klasifikaci elektronických učebních materiálů do určitých kategorií, resp. zajistit jejich seskupení, pokud pro ně není explicitně vytvořena žádná klasifikační třída. Rovněž lze vyhledávat *podobné* dokumenty nebo doplňující výukové materiály – podobně jako to nabízejí velká vydavatelství. Studujícím je možné nabídnout funkci „nový dokument pro tento den“. Vyhledávání lze aplikovat také na celé výukové kurzy. Tím spíše, že jejich forma („balení“) se stále důsledněji standardizuje a většina aktuálních LMS, jakož i autorských nástrojů podporuje standardy jako jsou IMS, IEEE LOM nebo integrující SCORM.

Personalizace obsahu

Efektivní učení s sebou nese potřebu koncentrace na určité vybrané části kurzů na základě porovnání nabízeného obsahu kurzu se zjištěnými znalostmi studenta. Možná podoba této základní personalizace je založena na následujícím jednoduchém postupu. Jako první úkol studenti dostanou napsat krátké shrnutí o svých aktuálních znalostech z obsahu kurzu. Na základě porovnání výsledků s kapitolami (přednáškami) kurzu je pak každému z nich nabídnuto k dalšímu studiu to, v čem jeho znalosti pokulhávají. Obdobné postupy jsou aplikovány i uvnitř kurzu. Mimo již zmíněného využívání shrnutí je možné získat informaci o úrovni znalostí studentů pomocí statistických metod aplikovaných nad studentovými odpovědmi na otázky.

Asistivní technologie

V budově Fakulty informatiky MU v Brně sídlí celouniverzitní centrum *Teiresiás* pro pomoc studentům se speciálními učebními nároky. Ve spolupráci s tímto centrem je připravován první elektronický výukový kurz, který by byl plně přístupný jak v češtině, tak v Brailově písmu. U kurzů vyučovaných na fakultě informatiky se navíc potýkáme se specifickými potížemi – je nutné například zavést vhodné brailovské ekvivalenty pro matematické výrazy či speciální symboly. Problémem je také převod materiálů, které nejsou v textové podobě – schémata, diagramy, obrázky.

Zodpovídání otázek

Automatické zodpovídání otázek (Automated Question Answering) je také zajímavým úkolem, ve kterém je důležitá morfologická a částečná syntaktická analýza (Smrž a Horák, 2000). Využívá se principů již ověřených v projektu „Encyclopedia Expert“ (Svoboda, 2002). Tento „expert“ je zatím schopen na základě informací získaných z české encyklopedie odpovídat na obecné otázky. K dosažení těchto schopností je definována sada sémantických rámů a odpovídajících syntaktických struktur, které umožňují analýzu nejfrekventovanějších typů otázek. Na zbývající dotazy se odpovídá pomocí fulltextového vyhledávání a identifikace relevantní části dokumentu, která s největší pravděpodobností obsahuje odpověď. Využití stejné strategie je jednoduše možné i v prostředí elektronických učebních materiálů. Vyhodnocení schopností QA modulu ukázalo, že přesnost odpovědí na dotazy, týkajících se určitého kurzu, je vysoká. Dokonce i dotazy na informace, které nejsou plně pokryty předdefinovanými rámci (a odpověď musela být tedy nalezena pomocí fulltextového vyhledávání), jsou zodpovězeny správně s úspěšností okolo 87 %.

Generování testů

Slibné výsledky automatického odpovídače směřují k ideji využít stejné metody zpracování jazyka jiným způsobem – například automaticky generovat otázky, nebo celé testy, založené na obsahu jednotlivých kurzů. Většinou je velmi jednoduché nalézt možné otázky „Co je…“ a ptát se na určité pojmy explicitně v textu definované.

Využití ontologií

Slabým místem popsaného automatického generátoru cvičení a testů je jeho zaměření na faktografii a nemožnost zjištění, zda studenti látce doopravdy *porozuměli*, zjistili, co je v ní nejdůležitější a dokáží získané *vědomosti použít* (funkční gramotnost, budování dovedností).

Pokrok v této oblasti je, zdá se, dosažitelný využitím standardních *ontologií*[2] . Ty jsou – jako jeden ze základních nástrojů pro vytváření sémantického webu – ve světě stále víc vytvářeny a využívány. Ontologie mohou být chápány jako prostředek, jak předat, popsat počítači znalosti obsažené v dokumentech. Formální specifikace pojmů a vztahů mezi nimi většinou využívají skupinu XML standardů: RDF (Beckett, 2003) a OWL (van Harmelen et al., 2003). Právě OWL slouží jako základ našeho posledního výzkumu automatického generování testových otázek na vztahy mezi pojmy dané ontologií. A s narůstajícím počtem ontologií pro různé znalostní domény se bude zvyšovat i potenciál tohoto směru výzkumu.

[2]Stručné představení ontologií a sémantického webu viz (Pitner a Smrž, 2004).

Výuka jazyků

Přirozeně velmi slibné je využití metod NLP v oblasti výuky jazyků. Význam rozsáhlých empirických dat ve formě korpusu – velké sady psaného či mluveného jazyka převedeného do digitální podoby – byl v počítačové lingvistice identifikován již poměrně dávno. Využití korpusů v jazykové výuce je naproti tomu relativně nové. Například výukové kurzy angličtiny na FI využívají *British National Corpus* (BNC) a další přístupné anglické korpusy (např. *Times Corpus*) – viz práce Jamese Thomase. Texty z korpusů mohou být využity například pro automatickou tvorbu prostých testů na anglická slovíčka. Z korpusů jsou vybrány úryvky, z nich vymazána určitá slova, a studenti je musí doplnit.

Ve výuce má velký význam poučení z chyb – vlastních i cizích. Na chyby uživatelů se zaměřil společný projekt *Czenglish* Laboratoře NLP na FI MU a Katedry angličtiny na Filozofické fakultě Masarykovy univerzity. Byl vytvořen elektronický výukový kurz založený na populární knize „English or Czenglish“ věnované studentům angličtiny na vyšší či profesionální úrovni.

Studenti v tomto kurzu dostávají za úkol překládat zadané věty. Pokud jejich překlad neodpovídá těm variantám, které jsou uložen, je zobrazena správná odpověď. Zároveň je ale také odpověď uživatele porovnána s databází špatných odpovědí. Pokud je nalezena nějaká shoda, je zobrazeno vysvětlení chyby. Studenti navíc mohou označit, že i přesto považují svůj překlad za možný. V tom případě je poslána zpráva učiteli, a ten pak rozhodne, zda překlad studenta má být přidán ke správným variantám – což samozřejmě ovlivňuje výsledky testu.

Asistence autorům obsahu – učitelům

Častým postupem při prvotním zpracování podkladů ke kurzu a zejména při jeho inovaci je zakomponování nových materiálů. Analýza dokumentu a ohodnocení jeho „novosti“ vzhledem k existujícímu obsahu kurzu může proces přidávání výrazně zracionalizovat, vede k udržení kompaktnosti a lepší „stravitelnosti“ výukového materiálu.

Pracujeme na expertním systému, který bude sloužit jako asistent autorům. Systém porovná přidávaný dokument s uloženými kurzy a nalezené podobnosti využije k vytvoření odkazů mezi dokumenty – do příbuzných dokumentů, do knihoven, on-line elektronických zdrojů a podobně.

Příprava kurzu s velkou sítí odkazů se může pro autory snadno stát noční můrou. Aby bylo možno sledovat všechny odkazy a vztahy mezi obsahem elektronické výuky, byl navržen a vytvořen systém DEB (Smrž, Povolný 2003). Je to aplikace založená na technologii klient-server, umožňující efektivní ukládání a získávání XML dokumentů. Použitím tohoto systému pak lze zajistit, že veškerý uložený obsah zůstane konzistentní i ve chvíli, kdy se změní odkazovaná informace. Kontroly konzistence jsou také definovány v XML (XSLT styly – Clark, 1999).

Využití zpětné vazby

Prakticky všechny momentálně provozované systémy elektronické podpory výuky (e-learningový modul IS MU používaný na celé Masarykově univerzitě nevyjímaje) poskytují mnoho dat použitelných pro zpětnou vazbu mezi studentem a učitelem. Jsou to především výsledky samotestování (selftestů), hodnocení úloh zadávaných jako domácí úkoly, výsledky písemných prací, (anonymní) studentské ankety, jakož i obsahy diskusí. Málokdy jsou ale dále interpretovány a využívány pro zlepšení práce. Přitom už na bázi poměrně prosté interpretace těchto výsledků s použitím pouze základních metod dolování znalostí lze v autorském systému upozorňovat na potenciálně problematická místa učebních materiálů a dokonce autorovi doporučovat směry vylepšení (zestručnění, rozšíření, oprava věcných a metodických chyb).

Identifikace a rozbor chyb

V některých předmětech se však ukazuje, že detailní analýza výstupů od studentů vytváří extrémně hodnotný materiál, okamžitě použitelný pro optimalizaci obsahu výukových materiálů.

Prakticky byly kupříkladu vytvořeny zvláštní nástroje pro označování a kategorizaci gramatických a slohových chyb ve studentských esejích (Pala et al., 2003). Učitel v elektronických dokumentech označí chyby, které následně studenti musejí opravit, ale přitom i zaznamenat typ chyby a původní (nesprávnou) podobu slova či slovního spojení.

V tomto směru se náš nynější výzkum zaměřuje na využití NLP technik v hodnotícím modulu elektronického výukového systému. Dosavadní zkušenosti říkají, že je relativně jednoduché zaručit zamýšlenou funkcionalitu v případě krátkých odpovědí (ve formě frází, které popisují jednoduché gramatické vzory). Výsledky prvního omezeného experimentu jsou velmi příznivé: jen 31 % odpovědí muselo být kontrolováno manuálně.

(Přibližné) překlady

Ve světě vzniká mnoho relevantních elektronických kurzů pro celou řadu výukových oblastí a velká část z nich je navíc veřejně přístupná na internetu – většinou však v angličtině. Pro snazší „vstřebatelnost“ zejména pregraduálními studenty (nebo v rámci celoživotního vzdělávání) je nutné, alespoň rámcově, část textu přeložit. I zde je možné využít technik NLP, tzv. překladovou paměť (translation memory). Princip této aplikace je stejný jako u standardní lokalizace počítačových programů. Pokud se změní obsah elektronického kurzu, „paměť“ tohoto nástroje pomůže přeložit ty části, které zůstaly z minulé verze stejné, nebo se změnily jen částečně. Manuálně stačí přeložit jen skutečně aktualizované, pozměněné nebo přidané části.

Detekce plagiátů

Posledním, do značné míry již reálně použitelným nástrojem, je experimentální detektor „opisování“. Ukázalo se, že dostatečně efektivní pro odhalení plagiátů jsou již jednoduché metody porovnávání slovních n-gramů (pokud je originál i odvozenina psána v jednom jazyce – zde češtině). Občas se nicméně objeví práce, které jsou otrockým překladem původního cizojazyčného textu. Spolehlivá automatická identifikace takových případů je mnohem složitější. N-gramové metody neposkytují dostatečnou přesnost kvůli rozdílům syntaktických struktur mezi češtinou a angličtinou. Nicméně i toto se může stát jedním cílem v našem budoucím výzkumu

Závěr

Aplikace většiny zmíněných technik využívajících zpracování přirozeného jazyka v oblasti podpory výuky dosud na svoje širší a „profesionálnější“ využití čeká. Pro technologie zpracování jazyka se však jedná o aplikační oblast mimořádně perspektivní. Zejména s ohledem na aktuální vývoj v oblasti sémantického webu, který má obecně k této oblasti velmi blízko.