Laboratoř zpracování přirozeného jazyka

Projekt Laboratoře zpracování přirozeného jazyka na Fakultě informatiky Masarykovy university je zaměřen na získání teoretických i aplikovaných výsledků v oblasti syntézy a rozpoznávání mluvené řeči (češtiny), lexikálních databází, representace znalostí a representace významu výrazů přirozeného jazyka a také využití metod strojového učení pro desambiguaci korpusových dat. Podstatným cílem pedagogickým je na základě zmíněného výzkumu vyškolit řadu postgraduálních i pregraduálních studentů v nově vznikajícím pomezním oboru, pro který se ujal název „jazykové inženýrství" (language engineering).

Základní informace

Projekt Laboratoře zpracování přirozeného jazyka na Fakultě informatiky Masarykovy university je zaměřen na získání teoretických i aplikovaných výsledků v oblasti syntézy a rozpoznávání mluvené řeči (češtiny), lexikálních databází, rerpesentace znalostí a representace významu výrazu přirozeného jazyka a také využití metod strojového učení pro desambiguaci korpusových dat. Podstatným cílem pedagogickým je na základě zmíněného výzkumu vyškolit řadu postgraduálních i pregraduálních studentů v nově vznikajícím pomezním oboru, pro který se ujal název „jazykové inženýrství" (language engineering).

Rozbor problému a shrnutí současného stavu poznání

Počítačové zpracování přirozeného jazyka představuje oblast, která je dlouhodobě zpracovávána jednak s cílem získat přirozený komunikační prostředek mezi člověkem a počítačem a jednak použít sílu výpočetního systému na řešení problémů lingvistické povahy. Relativně pomalý postup v této oblasti jde na vrub složitosti samotného objektu — přirozeného jazyka — a je determinován potřebou vytvořit kvalitní a rozsáhlou platformu lingvistických znalostí, dat a nástrojů, které dnes existují jen zčásti. V neposlední řadě brání pokroku i skutečnost, že dnes zatím neexistují standardy umožňující a regulující opakované a sdílené použití již existujících zdrojů jazykových dat a jejich začlenění do aplikačních systémů
V rámci projektu je tedy jednou z našich hlavních snah podstatně rozšířit dosavadní poměrně omezené zdroje strukturovaných jazykových dat o českou lexikální databázi použitelnou v rámci systému WordNet a terminologický thesaurus pro oblast IT a CS, který by měl být vzorem i pro další obory. Tento postup zároveň klade základy pro vytváření tolik potřebných standardů. Výzkum se bude srovnávat s podobnými projekty v rámci EU, takže bude respektovat směry tam rozvíjené.

Cíle projektu

Hlavním cílem projektu je zřízení laboratoře zpracování přirozeného jazyka, v jejímž rámci se bude souběžně řešit 5 vzájemně na sebe navazujících úkolů, tvořících určitý celek (viz níže). Projekt je orientován na získání nových teoretických poznatků v oblasti zpracování přirozeného jazyka a také na vytvoření nových souborů specializovaných a bohatě strukturovaných jazykových dat pro češtinu, dále lexikografických databází a terminologických thesaurů, které mohou opakovaně sloužit v řadě softwarových produktů pro inteligentní zpracování textu v přirozeném jazyce. Druhá skupina cílů zahrnuje výchovu mladých vědeckých pracovníků a její posílení na FI MU. K nejvýznamnějším cílům projektu patří také vyvinutí softwarových prostředků, které budou použitelné pro podporu přístupu nevidomých k výpočetní technice a k usnadnění jejich výuky na FI MU i jinde ve formě studijního pracoviště, a to v přímé spolupráci se Sjednocenou organizací nevidomých a slabozrakých.

Vědecký význam projektu

Vědecký význam projektu spočívá v propojení jinak samostatných směrů výzkumu v oblasti zpracování přirozeného jazyka do komplexnějšího celku: to povede jak k novým konkrétním výsledkům, tak i metodologickému obohacení dosud užívaných přístupů. Projekt je koncipován tak, aby přinesl nové teoretické poznatky v oblasti reprezentace znalostí a významu, dále také v oblasti desambiguace korpusových textů technikami strojového učení (teoreticky přínosná je již sama aplikace metod induktivní inference na desambiguaci textu). U rozpoznávání mluvené řeči dojde k potřebné aplikaci nových poznatků o syntéze a rozpoznávání češtiny do konkrétních softwarových realizací. V oblasti terminologie a lexikálních databází počítáme se získáním velmi potřebných konkrétních souborů nových jazykových dat spolu s nástroji pro jejich vytváření a zpracování.

Očekávané výsledky

- syntéza české řeči nové generace
- zvukový, hlasově ovládaný hypertextový systém
- vzorové studijní pracoviště pro zrakově postižené studenty
- prostředky pro konverzi skript a ostatních textových učebních pomůcek do hypertextového systému pro zrakově postižené
- terminologický thesaurus a nástroje pro semiautomatické zpracování terminologie ve volných textech (rozpoznávání a vyhledávání termínů na základě sémantických vztahů)
- nástroje pro tvorbu českého terminologického thesauru a vybudování jádra českého WordNetu (v rozsahu do 30 000 heslových slov)
- zpracování systému Gentzenovy přirozené dedukce pro transparentní a intenzionální logiku, dále pak nové originální výsledky velmi potřebné a dobře použitelné v oblasti umělé inteligence a zpracování přirozeného jazyka jako celku
- nástroje pro desambiguaci textu technikami strojového učení, komprese textu na základě jeho morfologických vlastností, generování konkordancí s využitím gramatického značkování, určování příbuznosti hesel

Předpoklady pro úspěšné zvládnutí úkolu

Stupeň připravenosti k řešení úkolů předložených v rámci projektu laboratoře zpracování přirozeného jazyka je podle našeho názoru dobrý. Následující body jasně dokumentují tuto naši připravenost a kvalifikovanost řešitelského týmu:

- Díky těsné spolupráci s pracovišti na FF UK, Ústavem národního korpusu a MFF UK máme přístup ke korpusovým datům ve vznikajícím Českém národním korpusu.
- Na FI MU je založen samostatný korpus textů z oblastí IT a CS v rozsahu téměř 10 mil. slovních tvarů; ten je předpokladem pro tvorbu terminologického thesauru.
- Máme již zkušenosti s automatickým gramatickým značkováním českých korpusových textů (cca v rozsahu 250000 slovních tvarů a jejich desambiguací. Máme k dispozici kvalitní lemmatizační program pro češtinu LEMMA, který spolehlivě pokrývá více než 95% slovní zásoby současné češtiny.
- Disponujeme výchozími daty pro tvorbu českého WordNetu (elektronická verze slovníků českých synonym [Pala, Ševeček, Všianský, 1995]).
- Dva z řešitelů mají kvalitní znalosti v oblasti strojového učení, jemuž se věnují již řadu let, a byli také organizátory mezinárodních sympózií o strojovém učení.
- V oblasti syntézy a rozpoznávání mluvené řeči se rovněž můžeme opřít o dlouhodobě kvalitní výsledky, které získal a také publikoval doc. RNDr. Ivan Kopeček.

Časový plán řešení

Harmonogram projektu se dá rozdělit do následujících etap:

- Tvorba softwarových nástrojů a technik potřebných pro jednotlivé úkoly (pomocné programy pro vyhodnocování výchozích jazykových dat - psaných i akustických.
- Sběr dat s využitím existujících materiálových zdrojů (korpusy, elektronické slovníky). Pokládáme za přirozené, že začátek této fáze se bude překrývat s předchozí etapou.
- Testování získaných řešení na vhodně zvolených ověřovacích datech a shrnutí poznatků získaných v jednotlivých úkolech formou publikační činnosti.

Úvod do korpusové lingvistiky

Úvod

Korpusová lingvistika je nové odvětví lingvistiky, které se objevilo relativně nedávno až díky počítačům a informačním technologiím. Existují softwarové nástroje, které umožňují třídit a klasifiovat, analyzovat a vyhodnocovat jazyková data v rozsahu, který by nebyl manuálně nikdy možný. To má ovšem značné metodologické důsledky: bez počítačů a informačních technologií bychom sotva mohli dospět k takovému typu poznání jazyka, jaké je dnes možné. Nyní lze podrobně zkoumat v podstatě libovolné jazykové jevy a pokoušet se o jejich opravdu přesné a adekvátní generalizace, proti nimž byly dřívější popisy jazyka jen intuitivními (to ale nemusí znamenat, že vždy chybnými) aproximacemi. Hromadnost a velikost zpracovávaných dat vede ke kvalitativním změnám v metodologii takové empirické vědy, jímž je současná lingvistika.

Kdy vznikla korpusová lingvistika

Na teoretické rovině to bylo nejspíše v 50. letech, kdy někteří američtí lingvisté (Harris, Hill) dospěli k názoru, že *korpus – dostatečně velký soubor přirozeně se vyskytujících jazykových dat* – je nutným a dostačujícím empirickým základem pro vytvoření popisu daného přirozeného jazyka (jeho gramatiky); přitom intuitivní evidence a introspekce byla odsunuta až na druhé místo, ne-li vůbec na poslední.
Korpusová lingvistika v novém pojetí začala vznikat nenápadně počátkem 60. let (Quirk, 1960, Kučera a Francis, 1967). Quirk začal pracovat na *http://www.ucl.ac.uk/english-usage/, Survey of English Usage, SEU*. V rámci SEU se počítalo i se zpracováním mluvené angličtiny, nebyl však orientován počítačově. O něco později začal pod vedením Čecha H. Kučery a Američana N. Francise na Brown University v USA vznikat počítačový korpus současné americké angličtiny – *Computation Analisis of Present-Day American English*, obsahující jen psané texty.

Dnes je již korpusů v jednotlivých jazycích celá řada a jejich rozsah i počet roste – jen u angličtiny to začíná klasickým miliónovým *Brown Corpusem* až po nedávný *http://info.ox.ac.uk/bnc/, British National Corpus – BNC* obsahující *100 miliónů* slov a v rámci *http://titania.cobuild.collins.co.uk/, COBUILDu* v Birminghamu vytvořený korpus *http://titania.cobuild.collins.co.uk/, Bank of English* (J. Sinclair) čítající nyní *220 miliónů* slovních forem a připravený k rozšíření na 500 miliónů.
Plný rozkvět korpusové lingvistiky však nastává teprve v poslední době a to díky prudkému vývoji v oblasti informatiky, informačních technologií a hardwaru. Lze očekávat, že s rozvojem textových procesorů, strojově čitelných textů, slovníků, multimediálních a počítačových sítí budou do konce století k dispozici korpusy čítající *miliardy* slovních forem.

Co je korpus

V současnosti se *korpusem* rozumí *rozsáhlý vnitřně strukturovaný a ucelený soubor textů daného jazyka elektronicky uložený a zpracovávaný.* Texty jsou v korpusu strukturovány a organizovány se zřetelem k využití pro určitý cíl, vůči němuž pak je korpus považován za reprezentativní.
Podle účelu existují různé typy korpusů. Podle zdroje textů mohou být korpusy psaného nebo mluveného jazyka, všeobecné nebo specializované na určitý styl, publicistický nebo odborný. Většina korpusů s ohledem na svou reprezentativnost obsahuje v různém poměru zástupce všech možných kategorií textů. Podle uložených dat mohou korpusy obsahovat pouze holé texty nebo texty různě označkované (anotované). Značkované korpusy samozřejmě poskytují více informací o jazyku, a proto je snaha korpsy značkovat. To lze provádět buď ručně, což je ale velice nákladné, nebo automaticky (strojově), což může někdy znamenat zanesení jisté míry nepřesností do značkování. Proto se také mnoho výzkumů v korpusové lingvistice zabývá právě automatickým značkováním textů.

Využití korpusů

Korpusová data jsou použitelná pro odborníky v řadě oborů:

- psychology
- sociology
- sociolingvisty
- odborníky v oblasti masové komunikace a médií
- lexikografy a lingvisty
- překladatele (strojový překlad)
- tvůrce učebnic a referenčích příruček (gramatiky, slovníky)
- v oblasti umělé inteligence (porozumění v přirozeném jazyce, reprezentace znalostí aj.)

Mistři Evropy v jazyku i řeči

Důvody vzniku programu Euromasters in Speech and Linguistics

Pokusy vyvinout systém, který by používal lidskou řeč jako prostředek pro komunikaci s počítaci, ukázaly, že k dosažení úspěchu je potřeba pokročit jak v oblasti rozpoznávání řeči, tak zpracování přirozeného jazyka (NLP) celkově. Při pohledu na růst v oblasti jazykového inženýrství se výrazně zvyšuje poptávka po lidech, kteří jsou schopni pracovat v týmech složených ze specialistů na mluvenou řeč, NLP i informatiku. ¡p¿ Členové těchto týmů by měli být schopni pracovat společně a měli by tedy rozumět hlavním aspektům dalších příbuzných oborů. Těžko bychom hledali nějaký existující studijní obor zaměřený na mluvenou řeč nebo na NLP, jenž by propojil tyto dvě oblasti, a to i v případech, kdy jsou oba vyučovány v rámci jedné instituce.

Participující organizace

V rámci platformy programu Evropské unie Sokrates vytvořilo a implementovalo sdružení universit v Aalborgu, Athénách, Barceloně, Bonnu, Brně, Edinburghu, Erlangenu, Essexu, Lausanne, Saarbruckenu, Sheffieldu, Stuttgartu, Patrasu a Utrechtu program studia, který umožní studentům získat kvalifikaci potřebnou pro týmovou práci v jazykového inženýrství.

Organizace studia

Kvůli právním bariérám znemožňujícím realizovat magisterský stupeň na celoevropské úrovni jsme se rozhodli pro schéma, které kombinuje národní systémy s požadovaným obsahem studia. Student, který studuje v ramci

programu „European Masters in Speech and Linguistics", nehledě na zemi nebo vzdělávací systém, opustí universitu s diplomem v rámci národních zvyklostí a s osvědčením dokládajícím, že student splnil všechny požadavky pro Masters.

Zaštítění

Tento projekt posvětily Evropská asociace pro mluvenou řeč (The European Speech Communication Association) a Evropské shromáždění asociací pro počítačovou lingvistiku (European Chapter of the Association for Computational Linguistic). Jejich prezidenti budou výše uvedené osvědčení podepisovat.

Obsah studia

Obsah Euromasters je vymezen následujícími tématickými oblastmi:

- Teoretická lingvistika
- Fonetika a fonologie
- Kognitivní modely pro zpracování řeči a jazyka
- Zpracování přirozeného jazyka
- Zpracování řečových signálů
- Rozpoznávání řeči na pomocí statistických vzorů
- Aplikace jazykového inženýrství

Ke každé této oblasti byl vytvořen rozsáhlý kontrolní seznam, který slouží k posouzení, jak jednotlivé výukové kursy pokrývají program European Masters.
Navíc se očekává, že každý student tohoto programu stráví alespoň tři měsíce v zahraničí. Nedílnou součástí studia je taktéž ověřování dovedností v praxi, nejlépe v průmyslu. Chtěli bychom také vždy jednou ročně uspořádat Velikonoční (Letní) školu, na které by se sešli všichni studenti tohoto programu. V jejím rámci jsou naplánovány intenzivní studijní kursy, presentace na průmyslových aplikacích a - v dlohodobém horizontu - i pokusy se zkoušením.

Účast university

Katedry, které si přejí účastnit se programu EuroMasters, musí poskytnout následující informace:

- Popis kursů vypisovaných katedrou s ohledem na předepsaný obsah EuroMasters - *http://www.cstr.ed.ac.uk/Euro*
- Následovat má průchod studenta studiem včetně odborných stáží a získaných národních diplomů.
- Pokud se to požaduje, je potreba navrhnout, jak si může student doplnit chybějící obsah jinde, tj. kursy jiných universit.
- Popis, jak bude výměna studentů realizována (nejčastěji to bude v rámci programu SocratesÉrasmus, aby bylo možno se vyhnout problémům s poplatky).
- Procedury, které provede zkouškova komise katedry (oboru), než student obdrží osvědčení European Master.

Přihlášky kateder budou projednány Radou tohoto projektu (Master Board).

Účast studentů

Studenti, kteří chtějí získat osvědčení European Master, mohou tento program absolvovat na libovolné zúčastněné universitě. Optimální volba závisí téměř výlučně na kvalifikaci a předchozích studiích oboru. Možnosti se v současné době pohybují mezi plným čtyřletým studiem (bez předchozích požadavků) až po dvou nebo třísemestrální studium, které však již vyžaduje předchozí kvalifikaci.
Další informace najdete na *http://www.cstr.ed.ac.uk/EuroMasters/*

Obsah

Rejstřík