AKADEMIE VĚD ČR

ZPRÁVA O PRŮBĚHU PRACÍ V ROCE 2005

na programovém projektu „Informační společnost“

(Doporučený rozsah zprávy je 1-2 strany textu. Při její formulaci se soustřeďte na ***zdůvodnění použitých*** *účelových finančních* ***prostředků*** *a* ***vyjádření k dodržení časového postupu řešení*** *uvedeného v návrhu projektu.)*

Projekt sestává ze dvou základních částí:
Pojmenované entity (named entities – NE)
Jednotný formát jazykových dat

Výzkum pojmenovaných entit byl zahájen studiem už existujících, zejména zahraničních prací o tomto tématu. Následoval návrh klasifikace NE, která by byla použitelná pro anotaci českých textů. První verze návrhu byla zpracována s využitím 1000 vět z Českého národního korpusu. Návrh byl dále upřesňován, výsledkem je dvouúrovňová klasifikace NE (první obecnější úroveň klasifikace je upřesněna druhou úrovní, např. geografická jména na první úrovni a jména řek, jména měst, jména oblastí atd. na druhé). Tento přístup umožní přidat anotací nějakou informaci i tam, kde si anotátor nebude typem NE zcela jist (podspecifikované značkování). Souběžně byla navržena technická realizace anotačního prostředí. Finální řešení spočívá ve využití jednoduchého, řádkově orientovaného textového formátu. Jednotlivé výskyty pojmenovaných entit mohou být anotovány vkládáním jednoduchých textových značek v libovolném textovém editoru.

V části formátové byly stanoveny základní principy a požadavky na nový datový formát založený na XML. Na jejich základě byla vytvořena specifikace prvotní verze formátu nazvaného PML, obsahující v základní podobě všechny stanovené rysy (princip stand-off anotace, propojení anotačních vrstev, typování, odkazy). Byl navržen obecný formát pro formální popis strukturní reprezentace jednotlivých rovin anotace v PML, tzv. PML schémata, která byla implementována pro zachycení spodních rovin reprezentace, zejména pro potřeby českého a anglického jazyka (tokenizace, segmentace a morfologická rovina). Pozornost byla věnována hlavně potřebám Pražského závislostního korpusu, ovšem i některých dalších aplikací, jako jsou projekty směřující k automatickému přepisu mluvené řeči. Přihlíželo se i ke specifickým potřebám dalších jazyků (např. arabštiny, kde se zdá výhodné k morfologické analýze přistupovat odlišně, pomocí tzv. morpho-trees).

Na základě konkrétního požadavku vydání Pražského závislostního korpusu (PDT) verze 2.0 v XML byla na základě nového formátu implementována rovněž schémata reprezentace závislostních rovin PDT 2.0 (analytické a tektogramatické), což umožnilo nový datový formát otestovat v praxi. Navržený PML formát pro PDT 2.0 sklidil kladné ohlasy od některých zahraničních kolegů, kteří měli možnost se s PDT 2.0 setkat jakožto beta-testeři a porovnat ho se starým CSTS formátem užívaným v PDT 1.0.

Dále byly učiněny první kroky k přizpůsobení základních anotačních nástrojů (zejména editor TrEd) na datový formát PML a byly vytvořeny základní podpůrné nástroje pro vývoj PML schémat (např. XSLT šablony pro převod PML schémat do XML validačního jazyka Relax NG).

Postup prací

Řešení obou částí probíhalo paralelně. V první polovině roku jsme provedli rešerše dostupných zdrojů. V části pojmenovaných entit tuto práci provedli převážně pracovníci spoluřešitele ÚJČ AV ČR. Ti také řešili většinu teoretických problémů v oblasti NE. Ve druhé polovině roku mohly být zahájeny anotační práce s reálnými texty, které prováděli brigádníci. OON byly tedy čerpány převážně ve druhé polovině roku, kdy již byly stanoveny podmínky pro ruční anotace a pomocné programátorské práce. V polovině roku 2005 se změnil platový řád UK, z čehož plynou rozdíly v poměru mezi základní a pohyblivou částkou mzdy jednotlivých pracovníků. Další rozdíly spočívají v rozdílném chápání tabulek 1a a 1b v původním plánu a ve výkazu za rok 2005. Celková suma za mzdy však byla vyčerpána podle plánu. Věcné náklady byly využity na pojištění (510 tis. Kč), režii (295 tis.) a ostatní provozní náklady (knihy, softwarové licence, spotřební materiál apod.) v celkové výši 112 tis. Kč.

Dílčí výstupy
Závěry rešerše existujících datových formátů, klíčové výsledky dosažené při tvorbě datového formátu, formální specifikace aktuálního návrhu formátu PML a program jeho dalšího vývoje byly shrnuty v technické zprávě vydané v rámci řešení této části projektu: P. Pajas, J. Štěpánek, A Generic XML-Based Format for Structured Linguistic Annotation and Its Application to the Prague Dependency Treebank 2.0, 2005, Technická zpráva č. TR-2005-29.

Výstupem první fáze zpracování pojmenovaných entit je zpráva o třech částech: (1) Definice NE (způsoby vymezení), (2) Využití NE v NLP, (3) Dostupné české elektronické zdroje vlastních jmen využitelné pro zpracování NE. Navržená klasifikace NE i technické řešení anotace byly popsány v dokumentu "Pokyny pro anotování pojmenovaných entit". Podle těchto pokynů bylo do konce roku 2005 anotováno 2000 vět (cca 6000 instancí NE) z Českého národního korpusu (ČNK). Rovněž byla zpracována podrobná studie "Návrhy pro řešení homonymie českých příjmení typu Vybíral, Odložil při automatickém morfologickém značkování". Předběžné experimenty s implementací disambiguačních pravidel na ČNK ukazují, že asi desetiprocentní chybu ve značkování těchto jmen v ČNK lze pomocí navržených disambiguačních pravidel snížit na cca 1%. Této části projektu se týká i další technická zpráva vydaná z prostředků projektu: D. Zeman, Manual for Morphological Annotation, ver. 2 (TR-2005-27).

Cestovní prostředky byly využity na prezentaci výsledků na domácích i zahraničních konferencích a seminářích. Někteří pracovníci se zúčastnili odborných konferencí s cílem získat zkušenosti od odborníků z jiných institucí. Byly navázány kontakty s pracovišti, kde se řeší podobné problémy v oboru NE, např. v Bratislavě, Saarbrueckenu a Varšavě.

Spolupráce
Spoluřešitelské pracoviště ÚJČ AV ČR se zúčastnilo podle plánu pouze části řešící NE. Provedlo již zmíněnou rešerši zdrojů, vypracovalo zásady pro anotování NE v českých textech a vyvinulo první dílčí algoritmus na rozpoznávání určitých typů NE homonymních se slovesy.

V roce 2005 bylo v rámci projektu, část pojmenovaných entit, dosaženo všech plánovaných cílů. Ve druhé části se podařilo vyřešit některé problémy dokonce ve větším rozsahu, než jsme očekávali.

AKADEMIE VĚD ČR

PROGRAM PRACÍ NA ROK 2006

na programovém projektu „Informační společnost“

*(Doporučený rozsah zprávy je 1-2 strany textu. Při její formulaci se soustřeďte na **časový postup řešení a cíle, kterých má být v roce 2006 dosaženo**. Současně připojte **komentář k předpokládanému použití** účelových finančních **prostředků**.)*

Dílčí cíle v jednotlivých sekcích:

Pojmenované entity:
- podrobná analýza dat anotovaných v roce 2005,
- ruční anotace dalších vybraných textů,
- vývoj nástrojů na automatickou detekci a anotaci pojmenovaných entit v textu,
- návrh a testování algoritmů na rozpoznávání dalších typů pojmenovaných entit,
- revize stávajícího morfologického slovníku s ohledem na pojmenované entity.

Jednotný formát:
- dopracování některých technických detailů formátu PML (zejména verzování, typová dědičnost, specifikace reprezentace metadat),
- převod dalších datových zdrojů do formátu PML,
- testování aplikovatelnosti PML na širší množinu jazykových dat, než představuje PDT,
- analýza a experimenty s možnostmi zpracování velkých objemů dat v PML formátu (indexace, paralelizace).

Investice

V prvním roce projektu proběhly přípravné práce, které bylo možno realizovat s využitím volných kapacit informační infrastruktury pracoviště řešitele. Z toho důvodu nebyly požadovány žádné investiční prostředky. V dalších fázích počítáme se zpracováváním mnohem většího objemu dat. Zejména bude třeba testovat a posléze i používat PML na řádově větších jazykových korpusech. Z toho důvodu je třeba se podílet na budování výpočetního zázemí pracoviště, aby bylo možno pro potřeby projektu vyhradit potřebný výpočetní čas a prostor na datových úložištích.
Z předpokládaného harmonogramu prací vychází i následující návrh potřebných investičních prostředků pro rok 2006:

Plánované investiční prostředky celkem: 360.000,- Kč
Předpokládané investice:
- 2x upgrade výpočetního serveru (či 2x nový uzel výpočetního clusteru),
- spoluúčast na obnově multifunkčního zařízení pro tištěná data (síťová tiskárna, skener, kopírka, vazačka).

Mzdy

Návrh mezd vychází z nových předpisů pro Karlovu Univerzitu platných od 1.7.2005. Určité nadhodnocení v kolonce pohyblivé složky mzdy vychází z předpokladu, že plánujeme přijmout nového pracovníka – programátora v lingvistické oblasti. Jedním z důvodů je celoroční nepřítomnost kolegy Daniela Zemana, PhD., který stráví rok 2006 na zahraniční stáži v USA. Změny ve složení řešitelského týmu jsou uvedeny ve změnovém listu.

OON

Prostředky OON budou využity na pomocné programátorské práce a ruční anotace textů.

Prostředky na **cestovné** využijeme opět k prezentaci našich výsledků na konferencích doma i v zahraničí. V letošním roce neplánujeme větší cesty, proto jsme snížili celkovou částku na cestovné.

Věcné náklady
Nejpodstatnější částky půjdou opět na režii (607 tis.) a pojištění z mezd (573 tis.). (Na letošní rok bylo ze strany fakulty podstatně zvýšena částka na režii, a to 20% z neinvestic (oproti 10% v roce 2005).) Ze zbylé částky (115 tis.) budeme hradit hlavně spotřební materiál, drobný hmotný majetek a nezbytnou literaturu.

Spolupráce se spoluřešitelským pracovištěm bude probíhat opět pouze v sekci Pojmenované entity, neboť se jedná o téma, se kterým tam mají větší zkušenosti. Předpokládáme, že budou vyvinuty algoritmy na automatické rozpoznání dalších typů pojmenovaných entit. Dále budou průběžně doplňovány dostupné seznamy nejběžnějšího typu pojmenovaných entit - vlastních jmen.

Podle dosavadního průběhu prací očekáváme, že budeme i nadále plnit předběžný harmonogram uvedený v návrhu celého projektu.