Program děkana Fakulty informatiky MU pro podporu studentských výzkumných a vývojových projektů

Návrh projektu

Identifikační kód projektu [1]

M	U	N	I	3	3	/	2	0	2	0	0	8

Základní údaje	
Garant projektu[2] (jméno, příjmení, email)	RNDr. Aleš Horák, Ph.D., hales@fi.muni.cz
Vedoucí projektu[3] (jméno, příjmení, email)	
Název projektu česky	Syntaktická analýza přirozeného jazyka s využitím postupné segmentace věty
Název projektu anglicky	Syntax analysis of natural language using incremental sentence segmentation
WWW stránka projektu	http://nlp.fi.muni.cz/~xkovar3/set
Doba trvání projektu v měsících	12

Složení studentského řešitelského týmu (uveďte jméno, příjmení, e-mail a UČO všech studentů)
Vojtěch Kovář, xkovar3@fi.muni.cz, 139915

Anotace projektu česky (4-5 řádků, tato informace bude zveřejněna na www stránkách FI)
Syntaktická analýza přirozených jazyků je jedním z důležitých kroků ke komplexní analýze výpovědí v přirozeném jazyce. Cílem projektu je poskytnout nástroje pro podporu nového přístupu k syntaktické analýze, který je založen na postupné detekci významných bodů v textu a jeho segmentaci. Projekt zahrnuje návrh a úpravy pravidlového formalismu, implementaci systému, jenž bude tento formalismus využívat, a další rozšíření tohoto systému.

[1] O číslo je třeba požádat pracovníka oddělení VaV FI MU (např. e-mailem).

[2] Garantem projektu musí být zaměstnanec FI. Studenti doktorského studia zaměstnanci zpravidla nejsou. Pokud má být projekt veden doktorandem, je třeba ho uvést jako vedoucího projektu a nalézt formálního garanta z řad zaměstnanců FI (např. školitele dotyčného doktoranda).

[3] Vedoucím projektu je buď garant nebo jím pověřený student doktorského studia. Vyplňujte jen v případě, kdy garant není současně vedoucím projektu.

Anotace projektu anglicky (4-5 řádků, tato informace bude zveřejněna na www stránkách FI)

Syntactic analysis of natural languages is one of the important steps leading to complex analysis of natural language utterances. The aim of the project is to provide tools for a new approach to the syntactic analysis, based on incremental detection of important items in the input sentence and its segmentation. The project includes a proposal of an appropriate formalism, implementation of a system that uses this formalism and further enhancements of the system.

Cíle projektu

Současný stav projektu

Vývoj projektu je v současné době ve fázi funkčního prototypu, dostupného na webové stránce projektu. Dostupná verze systému SET obsahuje základní minimalistický soubor pravidel pro analýzu češtiny a implementaci hlavních algoritmů pracujících s aktuálním formátem a souborem pravidel. Všechny části prototypu jsou původní a jsou vyvíjeny pod licencí GNU GPL.

Vstupem programu je česká věta obsahující lemmatizaci slov a základní jednoznačné morfologické značkování ve formátu, který používá morfologický analyzátor **ajka** [1], vyvíjený v Centru zpracování přirozeného jazyka FI MU. Příklady vhodných vstupů jsou dostupné na webové stránce projektu.

Výstupem programu je syntaktická analýza vstupní věty podle daného souboru pravidel, a to ve formě textového výpisu syntaktických stromů v tzv. hybridním formátu, který kombinuje závislostní a složkové prvky. Je též k dispozici jednoduchý grafický modul pro prohlížení výsledných stromů a chybový výstup, který obsahuje podrobné informace o průběhu vlastní analýzy a je tedy velmi užitečný pro vývoj konkrétních pravidel.

Cíle projektu

Hlavní cíle projektu spočívají v popisu a rozšíření stávajícího prototypu do obecně použitelné aplikace. Konkrétně se jedná o následující:

1. Vývoj souboru pravidel pro češtinu a tedy získání plnohodnotného syntaktického analyzátoru pro češtinu. Srovnání výsledků s existujícími analyzátory a s anotovanými korpusovými daty (zejména s Pražským závislostním korpusem, PDT [2]). Ve třetí fázi projektu (viz časový harmonogram) si klademe za cíl dosáhnout přesnosti analýzy srovnatelné s nejlepšími současnými analyzátory češtiny (nejlepší dosažené výsledky pro češtinu se pohybují okolo 86 % přesnosti [3]); za úspěšné naplnění tohoto cíle budeme považovat dosažení alespoň osmdesátiprocentní přesnosti. Výstup bude ve formě technické zprávy obsahující popis vyvinutého systému pravidel a výsledky srovnání.

2. Dopracování a popis formátu pravidel. Změny ve formátu pravidel oproti prototypu budou již minimální, jejich stávající podoba je dostatečně robustní. Výstupem bude manuál pro použití pravidel uveřejněný na webové stránce projektu.

3. Rozšíření formátů výstupních stromů o čistě závislostní a čistě složkový formát. Hybridní formát stromů lze jednoznačně převádět do závislostního formátu, pro složkový formát je třeba vyvinout způsob převádění neprojektivních větných konstrukcí. Cílem převodu je umožnit kvalitní srovnání vyvinutého pravidlového systému s existujícími analyzátory a anotovanými korpusy pro oba formalismy. Výstupem budou dokumentované programové moduly realizující převod, začleněné do programu.

4. Rozšíření systému o pravděpodobnostní ohodnocení nalezených struktur na základě externích dat, například frekvenčních a kolokačních statistik získaných z jazykových korpusů, dat z valenčních slovníků apod. Budou navrženy formáty pro reprezentaci těchto externích dat a implementovány funkce realizující ohodnocení. Výstupy ve formě popisu datových souborů a programových modulů realizujících ohodnocení.

5. Rozšíření formátů vstupu o jiné značkové sady pro morfologii, zejména o pražský poziční systém značek. Další rozšíření vstupních formátů o vstupy s víceznačným morfologickým značkováním. Výstupy ve formě příslušných komponent začleněných do programu a popisu daných vstupních formátů.

6. Obohacení systému o jiné typy výstupů, než jsou syntaktické stromy: extrakce kolokací, extrakce frází, značkování korpusových textů. Zejména poslední z uvedených implementačních úkolů je velmi důležitý pro další praktické využití v aplikované lingvistice. Výstupy ve formě příslušných dokumentovaných programových modulů.

7. Rozšíření modulu pro zobrazování stromů tak, aby poskytoval uživatelsky příjemné prohlížení výstupních stromů; zejména přizpůsobení pro procházení většího množství stromů při analýze delších textů apod.

Souhrn předpokládaných výstupů projektu:

1. Odladěná aplikace SET použitelná pro návrh pravidel pro syntaktickou analýzu s využitím postupné segmentace věty.

2. Soubor pravidel kompatibilní se systémem SET, který spolu s ním bude tvořit syntaktický analyzátor češtiny.

3. Dokumentace programu, formalismu pravidel a formátů vstupních a výstupních dat; technické zprávy o experimentech provedených se systémem.

Veškeré uvedené výstupy budou zveřejněny pod licencí GNU GPL na webové stránce projektu.

Časový harmonogram

Práce na projektu budou rozděleny do čtyř fází, na každou z nich připadá období 3 měsíce.

První fáze: Dokončení a popis formátu pravidel. Výstup programu ve formě závislostních stromů. První polovina práce na vývoji souboru pravidel pro syntaktickou analýzu češtiny. První srovnání s korpusem PDT.
Výstupy: Dokumentovaná verze programu s implementovaným převodem výstupních stromů, instrukce k použití. Dokumentovaný systém pravidel pro syntaktickou analýzu češtiny (první verze). Krátká technická zpráva o výsledcích srovnání získaného analyzátoru s korpusem PDT.

Druhá fáze: Pravděpodobnostní ohodnocení nalezených struktur na základě externích dat. Kombinace vybraných externích datových souborů s dříve navrženým souborem pravidel, zhodnocení dopadů na kvalitu výstupu. Výstup programu ve formě složkových stromů. Rozšíření modulu pro zobrazování stromů. Rozšíření možnosti vstupů minimálně o pražský poziční systém morfologických značek.
Výstupy: Dokumentovaná verze programu s uvedenými rozšířeními, instrukce k použití. Popis formátu použitého pro začlenění externích dat. Krátká technická zpráva o zlepšení výsledků analýzy s pomocí externích zdrojů.

Třetí fáze: Rozšíření vstupních formátů o vstupy s víceznačným morfologickým značkováním. Druhá, finální polovina práce na vývoji souboru pravidel pro syntaktickou analýzu češtiny. Srovnání s existujícími analyzátory pro češtinu a s anotovanými korpusovými daty.

Výstupy: Dokumentovaná verze programu s možností morfologicky víceznačného vstupu, instrukce k použití. Dokumentovaný systém pravidel pro syntaktickou analýzu češtiny (finální verze). Technická zpráva o výsledcích srovnání získaného analyzátoru s jinými analyzátory pro češtinu a s dostupnými anotovanými korpusovými daty.

Čtvrtá fáze: Další typy výstupů: extrakce kolokací, extrakce frází, značkování korpusových textů. Závěrečné úpravy a ladění programu.
Výstupy: Kompletně dokumentovaný program obsahující všechna výše popsaná rozšíření. Závěrečná zpráva o výsledcích projektu.

Literatura

[1] Sedláček, R.: *Morfologický analyzátor češtiny.* Diplomová práce -- Masarykova universita, Fakulta informatiky. Brno, 1999.

[2] Hajič, J.: *Building a syntactically annotated corpus: The Prague Dependency Treebank.* In : Issues of Valency and Meaning, Prague, Karolinum (1998)

[3] Holan, T., Žabokrtský, Z.: *Combining Czech Dependency Parsers*. In: Proceedings of TSD 2006, Brno, Czech Republic, Springer Verlag

Finanční rozvaha (pouze neinvestiční prostředky, uvádět v tisících Kč na jedno desetinné místo)

	položka	**návrh**	**rozhodnutí komise**
1	Stipendia pro členy řešitelského týmu[4]	38,4	
2	Cestovné[5]	0	
3	Literatura, materiál, drobný majetek[6]	0	
4	**Celkem**	38,4	

Zdůvodnění finanční rozvahy[7]

Účelem projektu je finančně podpořit vývoj svobodného software v oblasti zpracování přirozeného jazyka, konkrétně syntaktické analýzy. Projekt přispěje k interdisciplinární spolupráci lingvistů a informatiků díky (i pro laika) čitelnému formátu pravidel. Také budou posíleny možnosti využití výsledků syntaktické analýzy v jiných aplikacích pracujících s přirozeným jazykem, jako je např. extrakce informací z běžného textu.

Vzhledem k rozsahu plánovaných prací navrhuji čerpání finančních prostředků formou dvanácti měsíčních stipendií ve výši 3200,-. K práci bude využito prostředků Laboratoře zpracování přirozeného jazyka FI MU. Případné náklady na literaturu, možné prezentace výsledků na konferencích či drobný majetek budou hrazeny z jiných zdrojů.

Řešitel projektu v současnosti participuje na grantovém projektu AV ČR T100300414, *Inteligentní metody pro zvýšení spolehlivosti elektrických sítí*. Tento projekt končí v prosinci 2008.

Souhlas garanta

Navrhovaný projekt je v souladu s obecnými cíli *Programu pro podporu studentských výzkumných a vývojových projektů* popsaných v článku 1 Pokynu děkana Fakulty informatiky MU 2/2008. Cíle projektu považuji za reálné a požadované finanční prostředky za přiměřené. Souhlasím s odbornou garancí projektu a přejímám zodpovědnost za kontrolu plnění stanovených cílů projektu. O případných problémech budu neprodleně informovat proděkana pro výzkum a vývoj.

Datum: 14.11.2008 podpis garanta

[4] Maximální výše stipendia je 3.200 Kč měsíčně pro jednoho studenta.
[5] Cestovné lze požadovat pouze v dobře odůvodněných případech, kdy jsou pracovní cesty nevyhnutelnou podmínkou úspěšného naplnění cílů projektu.
[6] Vlastníkem pořízené literatury a drobného hmotného majetku je Fakulta informatiky MU.
[7] Explicitně uveďte, zda (a případně kteří) členové řešitelského týmu participují jako spolupracovníci na nějakém jiném grantovém projektu.