Grantová agentura České republiky – Část C

Zdůvodnění návrhu

Jméno navrhovatele Barbora Vidová Hladká

Název projektu

Automatické určování koreference v textech na základě dat anotovaných netradiční metodou (*PlayCoref*)

Cíle řešení projektu

Cílem projektu je ucelené řešení úlohy automatického určování koreference v textech. Řešení zahrnuje přípravu trénovacích dat, výběr vhodné metody strojového učení, její implementaci a vyhodnocení úspěšnosti. Příprava trénovacích dat bude probíhat netradičním způsobem, a sice prostřednictvím internetové hry, jejíž návrh a implementace jsou také součástí projektu. Cílovými jazyky jsou čeština a angličtina.

Současný stav úrovně poznání

Aplikace typu **komunikace člověk-počítač** mají být navrženy tak, aby práce s nimi byla pro uživatele co nejpohodlnější. Komunikace by měla pokud možno probíhat v přirozeném jazyce, nebo ještě lépe v mateřském jazyce uživatele. Schopnost počítače komunikovat v přirozeném jazyce se tedy musí opírat o co nejpodrobnější **znalosti tohoto jazyka**.

Jedním z nejdůležitějších zdrojů „informací“ o struktuře přirozeného jazyka pro jeho počítačové zpracování jsou bezpochyby **anotované korpusy**, umožňující přenos dosaženého poznání o jazyce do světa aplikací tím, že se na jejich základě budují nástroje na praktické zpracování jazyka. Anotované korpusy a nástroje na nich postavené jsou zastřešeny počítačovou a korpusovou lingvistikou.

Příprava anotovaných korpusů je náročná aktivita ve všech směrech – lidské zdroje, finanční prostředky, časové možnosti. Využijme proto potenciálu internetových uživatelů, které baví si hrát. Nabídněme jim **jazykové on-line hry** se slovy, s větami nebo i s celými články, které, aniž by o tom samotní hráči museli vědět, přinesou cenná data pro úlohy počítačového zpracování přirozeného jazyka.

Založení internetového herního portálu **LGame**[1] s jazykovými on-line hrami (Hladká, Ribarov, 2008) bylo motivováno úspěchy her s obrázky (van Ahn, Dabbish, 2004), které generují data pro úlohu rozpoznávání obrazu (*image processing*) - hry ESP Game[2] (van Ahn, 2006) a Peekaboom[3] (van Ahn,

[1] http://www.lgame.cz
[2] http://www.espgame.org

Ruoran, Blum, 2006). Tento herní portál postupně nabízí jazykové hry, při jejichž hraní vznikají anotované korpusy (data) pro takové úlohy počítačového zpracování přirozeného jazyka, které buď nebyly doposud implementovány (právě pro nedostatek anotovaných dat), anebo již implementovány byly, ale s velmi nízkou úspěšností. První hra, která zahájila provoz herního portálu, je **Shannonova hra**[4]. Zkušenosti z jejího návrhu a implementace jsou přínosné pro návrhy všech dalších her. Hry, které herní portál nabízí, musí splňovat 10 základních charakteristik:

1. Během hry vznikají anotovaná data pro úlohy počítačového zpracování jazyka.	2. Během hry hráči provádějí „nějakou“ akci tak dlouho, dokud nezaznamenají shodu[5].
3. Hraní hry vyžaduje pouze základní znalost gramatiky jazyka, ve kterém se hraje.	4. Pravidla hry jsou formulována nezávisle na jazyce, ve kterém se bude hrát
5. Hru je možné hrát alespoň v češtině a angličtině.	6. Hra je interaktivní v tom smyslu, že v každém okamžiku hráč ví alespoň rámcově, jak si vede jeho protihráč.
7. Hra má dynamický, svěží charakter.	8. Hra je aspoň pro dva hráče.
9. Hra nabízí několik úrovní obtížnosti.	10. Každá hra má svůj vlastní žebříček pořadí hráčů.

Pojem **koreference** v pojetí, které budeme v řešení projektu uplatňovat, a které se uplatnilo i při anotování Pražského závislostního korpusu[6] (PDT 2.0, (Mikulová a kol., 2006)), vychází z pojmu **reference**, tedy ze vztahu výrazů k předmětům nebo situacím reálného věta. Jestliže se v textu vyskytují dva nebo i více výrazů, které poukazují k témuž (osoba, předmět, jev, skutečnost, …), tj. jejich reference je identická, pak je jejich vzájemný vztah označován jako koreference. Posloupnost příslušných výrazů v textu je označována jako **koreferenční řetězec**. V našem projektu, na rozdíl od anotování PDT 2.0, budeme pracovat výhradně s referencí endoforickou, která pracuje s odkazováním v rámci textu, tj. nebudeme pracovat s exoforou, která pracuje s mimotextovými skutečnostmi. Studium koreference je nedílnou součástí analýzy diskurzu. Následující ilustrativní příklad jsme vybrali přímo z PDT 2.0; sledujeme v něm čtyři koreferenční řetězce:

Bělorusko$_1$**: zastavení likvidace **arzenálů$_2$

*Moskva – Běloruský **prezident$_3$** Alexandr **Lukašenko$_3$** nařídil pozastavit likvidaci vojenské **techniky$_2$** na území **republiky$_1$**. Oznámil to v Minsku na čtvrtečním slavnostním večeru k oslavám Dne obránců vlasti. Opatření se týká tanků, letadel, obrněných transportérů a bojových vozidel pěchoty. Podle **Lukašenka$_3$** prý byl tento krok vyvolán ani ne tak nedostatkem finančních prostředků, jako spíše „patrným porušováním vytvořené rovnováhy sil ve světě“. Agentura Interfax soudí, že **prezident$_3$** měl na mysli přání východoevropských zemí vstoupit do Severoatlantické **aliance$_4$**, což by pro **Bělorusko$_1$** znamenalo bezprostřední sousedství s **NATO$_4$**.*

[3] http://www.peakaboom.org
[4] http://www.lgame.cz/shannon
[5] Např. během hry ESP Game hráči vypisují možné „nálepky“ příslušného obrázku tak dlouho, dokud nenapíší stejnou.
[6] http://ufal.mff.cuni.cz/pdt20

Aplikace počítačového zpracování jazyka, které se neomezují pouze na samostatné věty, ale pracují s celými dokumenty (např. strojový překlad, vyhledávání informací, dialogové systémy, systémy pro zodpovídání otázek), se bez určování koreference neobejdou.

Dosavadní celosvětové pokusy o řešení úlohy určování koreference (*coreference resolution*) v anglických textech zatím pro zjednodušení pracují s koreferencí mezi jmennými skupinami. Nejčastěji se data zpracovávají dle metodiky MUC-6[7] (Vilain a kol., 1995; MUC-6, 1995) a MUC-7[8] (Hirschman, Chinchor, 1997). Dle této metodiky je k prostému textu přidána jen nejnutnější informace určující koreferenci. Velká část pokusů využívá při řešení diskriminačního přístupu, neboli řízeného strojového učení s učitelem (Soon a kol., 2001), při kterém je úloha určování koreference formulována jako binární klasifikační úloha. Rozvrstvením celé úlohy do více fází vzrůstá počet voleb algoritmických přístupů, jejichž výčet spolu s porovnáním zmiňuje (Iida a kol., 2005). Porovnání vlivu klasifikačních rysů na kvalitu výsledků zkoumá např. (Uryupina, 2006). V nedávné době se dostavily také úspěchy s neklasifikačními přístupy, např. přeformulováním určování koreference jako úlohy celočíselného lineárního programování (Denis, Baldridge, 2007) nebo využitím neřízeného strojového učení v kombinaci s bayesovským přístupem (Haghighi, Klein, 2007).

Pro češtinu je určování koreference doposud zkoumáno na tektogramatických stromech (významových zápisech) PDT 2.0 (Kučová, Žabokrtský, 2005), (Nguy Giang, 2006), (Němčík, 2006).

Formulace věcného obsahu a cílů

1. Vymezení cílů projektu a strategie jejich řešení

Cílem projektu je ucelené řešení úlohy **automatického určování koreference v textech**. Řešení zahrnuje přípravu trénovacích dat, výběr vhodné metody strojového učení, její implementaci a vyhodnocení úspěšnosti. Příprava trénovacích dat bude probíhat netradičním způsobem, a sice prostřednictvím **internetové hry** CGame, jejíž návrh a implementace jsou také součástí projektu. Cílovými jazyky jsou čeština a angličtina.

Při hře **CGame** budou hráčům předkládány texty, ve kterých budou vyznačovat slova (výrazy), které odkazují k témuž (osobě, předmětu, jevu, …). Hra bude navržena tak, že nebude třeba uživatele zatížit lingvistickými pojmy vztahujícími se ke koreferenci. Hra se stane součástí herního portálu jazykových her LGame. Pro uživatele přinese zábavu s jazykem.

Data získaná během hry budou trénovacími i testovacími daty pro proceduru automatického určování koreference. Tato procedura zajistí aplikace zvolené metody řešení určování koreference a obdržení výsledků měřitelných standardními evaluačními mírami precision, recall a F-measure. Součástí výzkumu je provedení porovnání neřízených (Haghighi, Klein, 2007), příp. optimalizačních (Denis,

[7] http://cs.nyu.edu/faculty/grishman/muc6.html
[8] http://www-nlpir.nist.gov/related_projects/muc/

Baldridge, 2007) metod lineárního programování s různě nastavenými řízenými metodami (Iida a kol., 2005). Důraz bude kladen na optimalizaci nastavení parametrů metody a výběru rysů, s následnou analýzou chyb.

Řešení projektu pokrývá datovou složku a složku nástrojů. Pro každou složku specifikujeme následující cíle a strategie řešení:

DATA

- **Hra CGame**

Formulace pravidel Pravidla budou navržena v souladu s deseti charakteristikami uvedenými výše tak, aby hra byla zábavná a dynamická. Zdůrazňujeme, že hráči nebudou zatíženi žádným lingvistickým pojmem. Formulace pokrývá hesla základní algoritmus hry, bodové ohodnocení, varianta hry s více než dvěma hráči, návrh grafického rozhraní.

Výběr vhodných českých a anglických textů České texty budou převzaty z Pražského závislostního korpusu 2.0. Tato volba je ovlivněna faktem, že PDT 2.0 obsahuje anotaci koreference (na úrovni zájmen), která probíhala na významových zápisech vět, na tzv. tektogramatických stromech (Mikulová a kol., 2006). Volba přinese zajímavé srovnání anotace koreference přímo na textech (ze hry) a na tektogramatických stromech. Podobně i anglické texty budou čerpány z korpusu s ruční anotací koreference, který používá anotační schéma navržené metodikami MUC-6 a MUC-7 v oblasti analýzy koreference. Výběr konkrétního anglického korpusu bude rozhodnut při řešení projektu. Odhad množství textů, které by měly projít aspoň jednou partií hry, při které budou označkovány, se odvíjí od nároků zvolených metod strojového učení na velikost trénovacích dat.

Odhady popularity a přínosu hry Ohlas na hru s textem se dá odhadnout velmi obtížně. Existuje alespoň paralela s nedávno rozběhnutou aktivitou v oblasti zpracování obrazu, která se setkala s pozitivním ohlasem uživatelů internetu a přinesla až překvapivé výsledky[9]. Odhad ohlasu při zpracování textu je ztížen tím, že texty jsou oproti obrázkům v nevýhodě – např. s ohledem na dobu potřebnou ke zpracování informace a udržení pozornosti.

Zájem o hru budeme vyvolávat cílenou propagací. Po prvním zveřejnění hry na internetu počítáme s podporou studentů, kteří odehrají určitý počet partií. Máme přislíbenou mediální podporu internetového vyhledávače Seznam.cz.

- **Data získaná během hry** Při registraci hráče se budou ukládat doplňkové údaje sloužící ke statistickému průzkumu (např. věk, mateřský jazyk**).** Pro každou partii bude zaznamenán její výsledek. Osobní data a výsledky partií budou ukládány ve vybraném databázovém systému (např. PostgreSQL, MySQL). Exportním formátem dat bude formát CSTS – formát založený

[9] „*A total of 13,630 people played the game [ESP Game] during this time [August 9 – December 10, 2003], generating 1,271,451 labels for 293,760 different images. Over 80% of the people played on more than one occasion. Furth more, 33 people played more than 1,000 games (this is over 50 hours playing the game).*“ (von Ahn, Dabbich, 2004). K březnu 2008 bylo hrou sesbíráno cca 10 milionů “nálepek”.

na SGML, který se v počítačové lingvistice používá pro reprezentaci morfologicky a syntakticky anotovaných korpusů[10] .

NÁSTROJE

- **Implementace hry CGame** Samotná implementace bude probíhat podle bodů, které budou součástí specifikace: návrh programovacího jazyka (uvažujeme o PHP5 a Flash); požadavky na°funkčnost (správa uživatelů, řízení průběhu hry, import vstupních dat, export sesbíraných dat); navrhovaná architektura; rozšiřitelnost; bezpečnost.
- **Grafický prohlížeč dat ViewCoref** – Analýzu kvality dat, které vzniknou během partií, usnadní grafický prohlížeč ViewCoref, který názorným způsobem (barevně, indexy, šipkami, …) znázorní koreferenční řetězce. Prohlížeč bude implementován v jazyce Java
- **Automatická procedura pro určování koreference**

Výběr vhodných metod strojového učení Jako základní metodu pro studium, implementaci, aplikaci na data a vylepšování zvolíme metodu shlukování a generování bayesovským způsobem, popsanou v (Haghighi, Klein, 2007). Jak pro řízený, tak i neřízený přístup k problému se zvolí minimálně jedna další, dnes aktivně používaná metoda. V průběhu implementace algoritmů bude zároveň probíhat optimalizace a úprava metod za účelem zlepšení výsledků.

Specifikace další potřebné vstupní jazykovědné informace Předpokladem použití metody kteréhokoli přístupu je získaní potřebných rysů. Zde počítáme nejen s použitím základních statistických rysů, jako jsou např. frekvenční charakteristiky jednotek textu nebo vzdálenosti mezi nimi, ale také s použitím morfologických rysů získaných ze zpracovávaného textu. Pro český i anglický text počítáme s nasazením externích nástrojů, jako je morfologická analýza a tagování (desambiguace morfologické analýzy).

Výběr implementačního prostředí Implementace algoritmů metod závisí na dostupnosti kvalitního a ověřeného software, příp. dílčích knihoven. Pro zpracování textu a tvorbu trénovacích a testovacích dat bude z důvodu efektivity práce s textem nejspíše použit jazyk Perl. Dále bude např. při použití dvoutřídního klasifikátoru stromového typu nebo při generování z rozdělení bayesovským přístupem snaha použít uznávané knihovny, nejspíše přístupné přes balíčky ve statistickému prostředí R či jako C-knihovny. Bude dopsán chybějící kód (pravděpodobně v R, příp. v C) pro dosud neimplementované podúlohy. Důraz bude kladen na co největší modularitu v rámci celého systému a na zpracování uživatelsky přívětivé dokumentace. Výsledný souhrnný "balíček" pro proceduru automatického určování koreference bude poskytnut k volnému použití, včetně sesbíraných dat a ukázky aplikace.

Vyhodnocení Z hlediska strojového učení budou nasbíraná data před vstupem do libovolné implementace nejdříve klasicky rozdělena na trénovací a testovací část. Tímto zajistíme objektivitu při konstrukci i při evaluaci pomocí standardních měřítek *precision*, *recall* a *F-measure*. Naše výsledky porovnáme s výsledky jiných výzkumných pracovišť.

[10] http://ufal.mff.cuni.cz/pdt2.0/doc/data-formats/csts/html/DTD-HOME.html

2. Časový harmonogram

Řešení projektu je naplánováno na **3** roky. Harmonogram uvádíme pro datovou složku a složku nástrojů odděleně s tím, že řazení dílčích úkolů kopíruje jejich provázanost nejen v příslušné složce, ale i napříč oběma složkami:

DATA	NÁSTROJE
1. rok řešení	
Návrh pravidel hry CGame	
Příprava českých a anglických textů pro hru	Specifikace implementace hry CGame Implementace hry CGame Implementace prohlížeče ViewCoref Návrh neřízené metody pro automatické určování koreference
2. rok řešení	
Zveřejnění první verze hry CGame	Implementace neřízené metody
Zkušební partie	Vylepšení hry CGame na základě zkušebních partií
Analýza dat z partií	Návrh řízené metody pro automatické určování koreference
3. rok řešení	
Zveřejnění finální verze hry CGame	Implementace řízené metody a její vylepšení
Zveřejnění dat	Zveřejnění automatické procedury s°nejvyšší dosaženou úspěšností

3. Řešitelský kolektiv

Barbora Vidová Hladká – koordinace projektu, zodpovědnost za řádné čerpání finančních prostředků, zodpovědnost za řešení úkolů v datové složce
Jiří Mírovský – zodpovědnost za řešení úkolů ve složce nástrojů, implementace hry CGame a prohlížeče ViewCoref
Pavel Schlesinger – vývoj, implementace a vyhodnocení procedury pro automatické určování koreference v textech

4. Výstupy a jejich prezentace

- Anotovaná data z projektu budou dána k dispozici k použití v dalších výzkumných projektech jak v ČR, tak i v zahraničí (iniciativa MUC).
- Automatická procedura pro určování koreference bude volně k použití.
- Průběžně budou výsledky projektu předkládány v podobě příspěvků na vybraných (pro tématiku a metodiku projektu relevantních) tuzemských i zahraničních konferencích z oboru počítačové lingvistiky.
- Průběžné výsledky budou zveřejňovány na domovské stránce projektu.
- Budeme pořádat popularizační přednášky o herním portálu mezi školáky a studenty středních škol.

5. Očekávaný přínos projektu v rámci oboru

Projekt se komplexně věnuje jedné z úloh počítačového zpracování přirozeného jazyka, a sice automatickému určování koreference v textech. Určování koreference musí být součástí aplikací, které pracují za hranicí věty, tedy které pracují s celými dokumenty. V rámci počítačového zpracování češtiny nebylo komplexní řešení této úlohy doposud realizováno. Rovněž tak i navrhovaná netradiční metoda sběru dat nebyla doposud uplatněna na žádný jazyk. Nezávislost pravidel hry na ‘hraném’

jazyku otevírá možnosti i pro ostatní jazyky, ne jenom pouze pro češtinu a angličtinu. Data, která se vygenerují při jednotlivých partiích hry, budou poskytnuta i jiným výzkumným projektům. Navážeme kontakt s iniciativou MUC ohledně poskytnutí těchto dat (včetně specifikace úlohy) pro mezinárodní soutěžení v úlohách spojených se systémy automatického zodpovídání otázek (extrakce informací, automatické určování koreference, identifikace pojmenovaných entit).

Podmínky pro řešení projektu

Pracoviště navrhovatele se zaměřuje na oblasti počítačové lingvistiky a automatického zpracování přirozeného jazyka a disponuje dostatečným hardwarovým vybavením a softwarem.

Seznam literatury jako podklad k návrhu projektu

von Ahn, Luis. 2006. Games with A Purpose. *IEEE Computer Magazine* 39 (6), pp. 92-94.

von Ahn, Luis, Ruoran Liu, Manuel Blum. 2006. Peekaboom: A Game for Locating Objects in Images. In *Proceedings of the ACM CHI 2006*.

von Ahn, Luis, Laura Dabbish. 2004. Labeling Images with a Computer Game.In *Proceedings of the ACM CHI 2004*, pp. 319-326.

Denis P., Baldridge. 2007. Global, joint determination of anaphoricity and coreference resolution using integer programming. *Proceedings of HLT-NAACL*, Rochester: pp.236–243.

Haghighi A, D. Klein. 2007. Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague:pp. 848–855.

Hirschman Lynette, N. A. Chinchor. 1997. MUC-7 Coreference Task Definition. *Proceedings of MUC-7*.

Hladká Barbora, Kirila Ribarov. 2008. Play the Language: An Alternative Manner of Collecting Annotated Data, submitted for the ACL 2008, Columbus, Ohio, USA.

Iida R., K. Inui, Y. Matsumoto. 2005. Anaphora resolution by antecedent identification followed by anaphoricity determination. Transactions on Asian Language Information Processing (TALIP), 4(4):417–434.

Kučová Lucie, Zdeněk Žabokrtský. 2005. Anaphora in Czech: Large Data and Experiments with Automatic Anaphora Resolution. In Matoušek, V., Mautner, P., Pavelka, T. (Eds.): *Text, Speech and Dialogue*, 8th International Conference, TSD 2005, Karlovy Vary.

Mikulová Marie, Bémová Alevtina, Hajič Jan, Hajičová Eva, Havelka Jiří, Kolářová Veronika, Kučová Lucie, Lopatková Markéta, Pajas Petr, Panevová Jarmila, Ševčíková Magda, Sgall Petr, Štěpánek Jan, Urešová Zdeňka, Veselá Kateřina, Žabokrtský Zdeněk. Anotace na tektogramatické rovině Pražského závislostního korpusu. Referenční příručka.*Tech. Report 31 ÚFAL MFF UK*, 2006, 183.

MUC-6. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995, SanMateo: Morgan Kaufmann.

Němčík Václav. *Anaphora Resolution*.2006. Master thesis MUNI, Brno.

Nguy Giang Linh. 2006. *Návrh souboru pravidel pro analýzu anafor v českém jazyce*. Diplomová práce MFF UK, Praha.

Soon W., H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Uryupina O. 2006. Coreference Resolution with and without Linguistic Knowledge. *LREC Proceedings*:993–898.

Vilain Marc, et al.1995. A Model-Theoretic Coreference Scoring Scheme, *Proceedings of the SixthMessage Understanding Conference (MUC-6)*, pp. 45-52, November 1995.

Grantová agentura České republiky – Část D

Jméno navrhovatele: Barbora Vidová Hladká

Navrhovatel

Odborný životopis: Mgr. Barbora Vidová Hladká, Ph.D.

Nar. 2. září 1971 v Hlinsku v Č. hladka@ufal.mff.cuni.cz
Bydliště: Semilská 926, Praha 9 – Kbely, 197 00

Zaměstnavatel:
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze, Ke Karlovu 3, 200 02 Praha 2
Adresa pracoviště: Ústav formální a aplikované lingvistiky MFF UK, Malostranské nám. 25, 118 00 Praha 1

Vzdělání:
1989–1994 Matematicko-fyzikální fakulta UK v Praze, obor informatika (Mgr.)
1994–2001 postgraduální studium na MFF UK v Praze, obor matematická lingvistika (Ph.D.)

Zaměstnání:
1996–dosud: Ústav formální a aplikované lingvistiky (Matematicko-fyzikální fakulta UK, Praha) – korpusová lingvistika, strojové učení

Přehled významných publikací za posledních pět let – knižní monografie

Barbora Vidová Hladká, Jan Hajič, Jiří Hana, Jaroslava Hlaváčová, Jiří Mírovský, Jan Votrubec: *Průvodce Českým akademickým korpusem 1.0.* Praha: Karolinum. 2007

Přehled významných publikací za posledních pět let – vědecké stati

Hajičová, Eva – Cuřín, Jan – Hajič, Jan – Kučera, Ondřej – Vidová-Hladká, Barbora. *Jazyk a umělá inteligence: kudy a kam?* Praha: Academia, 2007.

Hajičová, Eva – Hladká, Barbora – Kučová, Lucie. An Annotated Corpus as a Test Bed for Discourse Structure Analysis. In *Proceedings of the Workshop on Constraints in Discourse Structure Analysis.* 2006, pp. 82-89.

Hladká, Barbora – Kučera, Ondřej: A Corpus-based exercise book of Czech, In *Proceedings of the EUROCALL Conference (Mastering Multimedia: Teaching Languages Through Technology),* pp. 111, Coleraine, Northern Ireland, UK, 2007.

Hladká, Barbora – Králík, Jan. Proměny Českého akademického korpusu. *Slovo a slovesnost,* 67:179-194, 2006.

Ribarov, Kiril – Bémová, Alevtina – Hladká, Barbora. When a statistically oriented parser was more efficient than a linguist: A case of treebank conversion., Prague Bulletin of Mathematical Linguistics, 2006, 1, 21-38.

Členství ve výzkumných či jiných pracovních týmech

2004–dosud: hlavní řešitelka grantu GA AV ČR 1ET101120413 „Data a nástroje pro informační systémy“

2007-dosud: členka řešitelského projektu GAUK 207-10/257559 „Elektronická cvičebnice češtiny založená na Pražském závislostním korpusu“

Odborní spolupracovníci

Odborný životopis: Mgr. Jiří Mírovský

Nar. 20. srpna 1973 v Kladně mirovsky@ufal.mff.cuni.cz
Bydliště: Rabasova 1395, Slaný, 274 01

Zaměstnavatel:
Matematicko-fyzikální fakulta Univerzity Karlovy v Praze, Ke Karlovu 3, 200 02 Praha 2
Adresa pracoviště: Ústav formální a aplikované lingvistiky MFF UK, Malostranské nám. 25, 118 00 Praha 1

Vzdělání:
1993–1998: Matematicko-fyzikální fakulta UK v Praze, obor informatika (Mgr.)
2003–dosud: postgraduální studium na MFF UK v Praze, obor matematická lingvistika

Zaměstnání:
2000–2004: Centrum komputační lingvistiky (Matematicko-fyzikální fakulta UK, Praha)
2005–dosud: Ústav formální a aplikované lingvistiky (Matematicko-fyzikální fakulta UK, Praha)

Přehled významných publikací za posledních pět let – knižní monografie
Vidová Hladká, Barbora – Hajič, Jan – Hana, Jiří – Hlaváčová, Jaroslava – Mírovský, Jiří – Votrubec, Jan: *Průvodce Českým akademickým korpusem 1.0.* Praha: Karolinum. 2007

Přehled významných publikací za posledních pět let – vědecké stati
Mírovský, Jiří: Netgraph – Making Searching in Treebanks Easy, In: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), Hyderabad, India, 8th - 10th January 2008, pp. 945-950.
Mírovský, Jiří: Towards a Simple and Full-Featured Treebank Query Language, In: Proceedings of ICGL 2008, Hong Kong, 9th - 11th January 2008, pp. 171-178.
Mírovský, Jiří – Panevová, Jarmila: Learning to Search in Prague Dependency Treebank, In: Proceedings of Grammar and Corpora 2007, Liblice, Czech Republic, 25th - 27th September 2007
Mírovský, Jiří: Netgraph: a Tool for Searching in Prague Dependency Treebank 2.0, In: Proceedings of The Fifth International Treebanks and Linguistic Theories conference, Prague, Czech Republic, 1st and 2nd December 2006, pp. 211-222.
Smrž, Otakar – Pajas, Petr – Žabokrtský, Zdeněk – Hajič, Jan – Mírovský, Jiří – Němec, Petr: Learning to Use the Prague Arabic Dependency Treebank, In: Elabbas Benmamoun. Proceedings of Annual Symposium on Arabic Linguistics (ALS-19). Urbana, IL, USA, Apr. 1-3: John Benjamins, 2005.

Členství ve výzkumných či jiných pracovních týmech
2000–2004: výzkumný pracovník projektu Centrum komputační lingvistiky (MŠMT LN00A063)
2002–2004: výzkumný pracovník mezinárodního projektu MALACH (Multilingual Access to Large Spoken Archives, MŠMT 1P05ME786)
2004–dosud: výzkumný pracovník projektu „Data a nástroje pro informační systémy“ (GA AV ČR 1ET101120413)

Odborný životopis: Mgr. Pavel Schlesinger

Datum a místo narození: 6. února 1979, Třinec schlesinger@ufal.mff.cuni.cz
Adresa: Habrová 300, 739 61 Třinec

Pracoviště:
Matematicko-fyzikální fakulta Univerzity Karlovy (MFF UK), Ke Karlovu 3, 121 16 Prague 2
Ústav formální a aplikované lingvistiky MFF UK, Malostranské nám. 25, 118 00 Prague 1

Vzdělání:
2005–současnost postgraduální studium oboru Matematická lingvistika na MFF UK
1998–2005 Mgr. v oboru Matematická statistika na MFF UK

Zaměstnavatel:
2004– současnost Ústav formální a aplikované lingvistiky – “machine learning”, užití statistických metod v komputační lingvistice

2001-2002 Český statistický úřad – příprava dat a aplikace statistických metod

Publikace - články v odborných časopisech a sbornících z konferencí:

Pavel Pecina a Pavel Schlesinger: *Combining Association Measures for Collocation Extraction.* Proceedings of the 21th International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL 2006), Poster Sessions, Sydney, Austrálie, červenec 2006.

Silvie Cinková, Pavel Pecina, Petr Podveský a Pavel Schlesinger: *Semi-automatic Building of Swedish Collocation Lexicon.* Proceedings of the fifth International conference on Language Resources and Evaluation (LREC 2006), Janov, Itálie, květen 2006.

Jurjen Duintjer Tebbens a Pavel Schlesinger: *Efficient Implementation of Optimal Linear Discriminant Analysis.* Proceedings of the Seminar on Numerical Analysis (SNA'06), Sedlec-Prčice, Česká Republika, leden 2006.

Jurjen Duintjer Tebbens a Pavel Schlesinger: *Improving implementation of linear discriminant analysis for the high dimension/small sample size problem.* Computational Statistics & Data Analysis, 52, 423 – 437, 2007.

Eduard Bejček, Pavel Straňák a Pavel Schlesinger. *Annotation of Multiword Expressions in the Prague Dependency Treebank.* Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), 793–798, Indie, leden 2008.