Výsledky dosažené v roce 2008

A. Sekce Pojmenované entity

V roce 2008 jsme se zaměřili mj. na přesnější a prakticky použitelné rozpoznávání geografických pojmenovaných entit v českých textech. Byl vytvořen software, který analyzuje vstupní text v HTML (z technických a autorskoprávních důvodů byla pro tento účel částečně reimplementována česká morfologie), s vysokou úspěšností v něm na základě různých rejstříků a heuristických pravidel rozpozná geografická jména, barevně je v HTML textu zvýrazní a doprovodí je odkazem na vygenerovaný obrázek s vyznačením daného místa na mapě České republiky. Systém se v současnosti používá na KÚ Libereckého kraje.

Začali jsme také revidovat seznam zeměpisných názvů v morfologickém slovníku. K tomu účelu se využívá speciální editor pro práci s morfologickým slovníkem, vyvinutý na ÚFALu. Při doplňování skloňovacích vzorů zeměpisných názvů se současně ověřuje funkčnost editoru a na základě zkušeností se stále zlepšuje.

Dalším rozšířením byl nástroj na automatické vytváření dvojic zeměpisných názvů a příslušných přídavných jmen, typu Liberec - liberecký. Automatické přiřazení výrazně zefektivní veškeré aplikace, včetně populárních automatických překladů.

Věty s ruční anotací pojmenovaných entit, které byly v rámci tohotoprojektu anotovány v předcházejících letech, se začaly využívat ve výucepředmětu Úvod do strojového učení (v počítačové lingvistice) (NPFL054 na MFF UK). V rámci tohoto předmětu mají studenti za úkol implementovat vlastní rozpoznávač pojmenovaných entit založený na metodách strojového učení. Jedna z úspěšných studentských implementací, založená na klasifikační metodě Support Vector Machines, byla již integrována do prostředí pro vývoj strojového překladu TectoMT, který je vyvíjen na pracovišti MFF UK.

Do prostředí TectoMT byl také letos integrován rozpoznávač pojmenovaných entit pro angličtinu Stanford Named Entity Recognizer, díky kterému bylo možné zvýšit úspěšnost anglicko-českého překladu implementovaného v TectoMT. S tímto překladovým systémem jsme se zúčastnili soutěže v překladu pořádané v rámci Workshop on Statistical Machine Translation při významné mezinárodní konferenci ACL.

Jeden z klíčových předpokladů pro vývoj kvalitního překladového systému je možnost automatického alignmentu velkého objemu paralelních textů (přiřazování sobě odpovídajících slov/částí textu v originálním a přeloženém textu). Za účelem zvýšení úspěšnosti alignmentu byla provedena ruční anotace alignmentu na cca 2000 párech českých a anglických vět, vybraných mj. s důrazem na vysoký výskyt pojmenovaných entit. Tyto anotace sloužily jako trénovací a testovací data pro nově vyvinutý aligner na tektogramatické rovině, který je založený na perceptronu. Tento aligner i výše uvedené rozpoznávače pojmenovaných entit jsou kromě strojového překladu v současnosti používány také při vývoji vznikajícího česko-anglického paralelního treebanku CzEng.

B. Sekce Jednotný formát

V tomto roce jsme se soustředili na vytváření nástrojů pro práci s jazykovými daty ve formátu PML. Jde zejména o zbrusu nový dotazovací jazyk PML Tree Query (PML-TQ) pro dotazování nad stromovými strukturami, uloženými ve formátu PML. Uskutečnili jsme dvě implementace nového vyhledávače, pomocí databáze (SQL) i bez ní (BTrEd). Dotaz je nyní možno graficky vykreslit v již existujícím nástroji TrEd, pro nějž jsme vytvořili kompletní grafické rozhraní k vyhledávači. Dále existuje jednoduché textové rozhraní pro příkazovou řádku a programové rozhraní pro Perl. Nový

dotazovací jazyk je podobný již existujícímu nástroji NetGraph, je však podstatně silnější. Pro lepší kompatibilitu jsme sestrojili i nástroj, který umí převést dotaz vytvořený v NetGraphu do PML-TQ.

Anotační nástroj TrEd, který se již používá k práci s daty ve formátu PML, jsme rozšířili o nové možnosti vykreslování. Je nyní také rychlejší, což je pro lingvistickou práci s velkými objemy dat podstatné. Další podstatné vylepšení TrEdu spočívá v rozšíření o uživatelské toolbary, side-panel a další funkce, které zpříjemňují, zrychlují, a tím zefektivňují práci s nástrojem.

Formát PML i nástroje se masově využívají na našem pracovišti při práci s treebanky. Využíváme je i ve výuce, zejména v předmětu Pražský závislostní korpus (NPFL075). TrEd je používán i na řadě podobných pracovišť v zahraničí.

Řešitelský tým se na pracovišti MFF UK rozrostl o doktoranda Davida Kolovratníka, který nahradil Mgr. Ševčíkovou ve druhé polovině roku po jejím odchodu na mateřskou dovolenou. Stará se o nový morfologický analyzátor a editor, který jsme začali používat k opravám zeměpisných pojmenovaných entit ve slovníku a k doplňování nových. Změna je uvedena ve změnovém listu.
Na spoluřešitelském pracovišti ÚJČ nedošlo během roku k personálním změnám, avšak ke konci r. 2008 ukončil svou činnost na projektu doc. M. Giger v souvislosti se svým odchodem na universitu v Basileji.
Spolupráce mezi oběma pracovišti, tj. MFF UK a ÚJČ AV ČR, probíhala ve stejném duchu jako v předchozích letech. ÚJČ se zabývá teoretickými problémy českých pojmenovaných entit, MFF využitím dat pro praktické aplikace, v poslední době především pro systémy automatického překladu. ÚJČ také pokračuje ve skenování a digitalizaci archivu vlastních jmen.

Mzdové prostředky byly na pracovišti MFF čerpány podle plánu, v závěru roku došlo k nedočerpání částky 3669 Kč. Tu převedeme do FÚUP a vyčerpáme příští rok.
Vzhledem k poměrně vysoké částce převedené z loňského roku se nám ani letos nepodařilo využít celou částku určenou na **OON**. Soustředili jsme se proto lépe na vytvoření plánu čerpání OON v letošním roce. Částku 62 665 Kč, kterou převádíme do FÚUP, tak vyčerpáme spolu s plánem letos. OON budou využity na pomocné programovací práce studentů v oblasti automatických procedur na rozpoznávání pojmenovaných entit.
Díky nedočerpání mzdových prostředků se také nevyčerpala celá plánovaná částka, se kterou se počítalo na pojištění. I tuto částku (30 tis. Kč) převedeme do FÚUP.
Na pracovišti ÚJČ byly mzdové prostředky čerpány dle plánu (včetně dočerpání prostředků převedených do FÚUP za r. 2007, které byly v souladu se svým původním určením použity v r. 2008 na plat doc. Gigera).

Investiční prostředky byly využity na rozšíření paměťové kapacity dvou uzlů clusteru, které jsou primárně využívány pro řešení projektu. Tyto servery nyní disponují paměťovou kapacitou 32GB každý. Pro tento účel byla vyhrazena třetina investičních prostředků (50 tis. Kč).
Ostatní prostředky umožnily přestavbu dvou diskových polí, která slouží pro potřeby více projektů (další projekty přispěly na výměnu disků v těchto polích). Uvedené změny zajistily bezpečnost dat, jakožto hlavních produktů výzkumu. Původní pole již dosáhla konce své životnosti.

Neinvestiční prostředky byly využity na nákup paměťových kitů pro výpočetní servery a částečně také na obnovu pracovních stanic pracovníků. Část prostředků byla použita na nákup kancelářských potřeb. Nevyužité prostředky z kolonky Služby a Drobný majetek (dohromady 31 tis. Kč) jsme použili pro dofinancování prostředků na cestovné, které jsme v loňském roce podcenili.
Ústav pro jazyk český použil neinvestiční prostředky na nákup kancelářských potřeb a dalšího drobného materiálu a na služby spjaté s digitalizací archivu vlastních jmen.

Cestovní prostředky byly čerpány na prezentaci výsledků projektu doma i v zahraničí. Vyčerpali

jsme i nedočerpanou částku 6 tis. Kč, která byla převedena na cestovné do FÚUP v roce 2007. Na cestovné jsme tedy utratili celkem plánovaných 50 tis., 6 tis. z FÚUP a 32 tis. převedených z Neinvestičních prostředků, dohromady tedy 88 tis. Kč.
Nejvýznamnější konference, kterých jsme se letos zúčastnili: Coling 2008 (Manchester, Anglie), EAMT (Hamburg, Německo), HCC 2008 (Bellagio, Itálie), LREC 2008 (Marrakeš, Maroko), TSD 2008 (Brno), Workshop on Statistical Machine Translation (Columbus, Ohio, USA).
Na financování cest se podílely částečně i jiné projekty pracoviště MFF.

Převod do Fondu účelově určených prostředků
MFF UK - 66 tis. Kč na mzdové prostředky (3 669 Kč mzdy + 62 665 Kč OON), 30 tis. Kč. na zákonné odvody. Celkem jde o 96 tis. Kč.
ÚJČ AV ČR - 42 tis. Kč na služby spojené s digitalizací archivu českých vlastních jmen (zejm. skenování).
V tabulkách jsou tyto prostředky uvedeny, podle pokynů, jako vyčerpané.

Plán na rok 2009

A. Sekce Pojmenované entity
* Z ručních anotací vzniklých v průběhu projektu (zejména anotace vět z Českého národního korpusu a anotace paralelních vět v češtině a angličtině) vytvoříme veřejně dostupné balíčky.
* Z rozpoznávače pojmenovaných entit v českých textech bude vytvořena veřejně dostupná knihovna v jazyku Perl.
* Novou verzí rozpoznávače pojmenovaných entit budou označkována data v korpusech PDT a CzEng.
* Budeme pokračovat ve využití nástrojů pro pojmenované entity v experimentech s automatickým překladem.
* Pokračování ve zpracování méně rozšířených českých vlastních jmen (zejména pomístních jmen) a další onomastické tématiky
* Pokračování digitalizace archivů vlastních jmen

B. Sekce Jednotný formát

* dokončení již začaté práce na specifikaci formátu PML ve verzi 1.3, která přinese nové možnosti v popisu datových typů (šablony)
* vytvoření kompletní dokumentace k PML-TQ a implementovaným rozhraním
* dokončení implementace PML-TQ v BTrEdu (výstupní filtry)
* další rozšíření možnosti PML-TQ o uživatelem definované relace
* návrh a implementace modulárního systému rozšíření pro PML-TQ (např. o funkce a relace specifické pro nějaký treebank)
* pokračování v optimalizacích stávajících nástrojů

V řešitelském týmu dojde v roce 2009 k drobným změnám, jak personálním, tak i v rozložení úvazků. Změny jsou zaznamenány ve změnových listech. Neplánujeme žádné přesuny finančních prostředků.
V týmu působícím v ÚJČ nahradí doc. Gigera dvě doktorandky (Mgr. Jana Steinerová a Mgr. Žaneta Procházková) a jedna pracovnice výzkumu a vývoje, Mgr. Martina Zirhutová.

Publikace

Ševčíková Magda. Proper Nouns in Czech Corpora. In Proceedings of the Corpus Linguistics Conference Series. Birmingham, UK: 2008, pp. 1-10.

Raab Jan, Hajič Jan, Spoustová Drahomíra: Compost Czech, UK MFF UFAL, 2008. http://ufal.mff.cuni.cz/compost

Duběda Tomáš, Raab Jan. Pitch Accents, Boundary Tones and Contours: Automatic Learning of Czech Intonation. In Lecture Notes in Computer Science: Proceedings of the 11th International Conference, TSD 2008. 5246. Berlin Heidelberg: Springer-Verlag, 2008, pp. 293-301.

Spoustová Drahomíra, Hajič Jan, Raab Jan, Spousta Miroslav: Compost English, UK MFF UFAL, 2008. http://ufal.mff.cuni.cz/compost

Mareček David, Žabokrtský Zdeněk, Novák Václav. Automatic Alignment of Czech and English Deep Syntactic Dependency Trees. In Proceedings of the Twelfth EAMT Conference. Hamburg: HITEC e.V., 2008, pp. 102-111.

Lopatková Markéta, Žabokrtský Zdeněk, Kettnerová Václava. Valenční slovník českých sloves. Praha: Nakladatelství Karolinum, 2008.

Hlaváčová Jaroslava, Hrušecký Michal. "Affisix" Tool for Prefix Recognition. In Lecture Notes in Computer Science: Proceedings of the 11th International Conference, TSD 2008. 5246. Berlin Heidelberg: Springer-Verlag, 2008, pp. 85-92.

Pajas Petr, Štěpánek Jan. Recent Advances in a Feature-Rich Framework for Treebank Annotation. In The 22nd International Conference on Computational Linguistics - Proceedings of the Conference. 2. Manchester: The Coling 2008 Organizing Committee, 2008, pp. 673-680.

Vidová Hladká Barbora, Hajič Jan, Hana Jiří, Hlaváčová Jaroslava, Mírovský Jiří, Raab Jan: Czech Academic Corpus 2.0, LDC - Linguistic Data Consortium, ÚFAL MFF UK, 2008.

Pajas Petr, Štěpánek Jan: PML Tree Query 0.5alpha, ÚFAL MFF UK, 2008. http://ufal.mff.cuni.cz/~pajas/pmltq/

Hlaváčová Jaroslava, Lopatková Markéta. Variants and Homographs: Eternal Problem of Dictionary Makers. In Lecture Notes in Computer Science: Proceedings of the 11th International Conference, TSD 2008. 5246. Berlin Heidelberg: Springer-Verlag, 2008, pp. 93-100.

Novák Václav. Semantic Network Manual Annotation and its Evaluation.Ph.D. thesis., 2008, 129 s.

Bojar Ondřej, Janíček Miroslav, Žabokrtský Zdeněk, Češka Pavel, Beňa Peter. CzEng 0.7: Parallel Corpus with Community-Supplied Translations. In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). Marrakech, Morocco: European Language Resources Association (ELRA), 2008,.

Žabokrtský Zdeněk, Ptáček Jan, Pajas Petr. TectoMT: Highly Modular MT System with Tectogrammatics Used as Transfer Layer. In ACL 2008 WMT: Proceedings of the Third Workshop on Statistical Machine Translation. Columbus, Ohio:

Association for Computational Linguistics, 2008, pp. 167-170.

Novák Václav, Hall Keith. Inter-sentential Coreferences in Semantic Networks: An Evaluation of Manual Annotation. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008). Marrakech, Morocco: European Language Resources Association, 2008, pp. 695-704.

Žabokrtský Zdeněk, Bojar Ondřej. TectoMT, Developer's Guide.Tech. Report TR-2008-39 Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague, 2008, 50 s.

Ševčíková Magda. Pronouns Introducing Content Clauses. In Grammar & Corpora / Gramatika a korpus 2007. Prague, Czech Republic: Academia, 2008, pp. 277-284.

Bojar Ondřej, Cinková Silvie, Ptáček Jan. Towards English-to-Czech MT via Tectogrammatical Layer. Prague Bulletin of Mathematical Linguistics, 2008, 90 s.

Hlaváčová Jaroslava, Kolovratník David. Morfologie češtiny znovu a lépe. In Informačné Technológie - Aplikácie a Teória.Zborník príspevkov, ITAT 2008. Seňa, Slovakia: PONT s.r.o., 2008, pp. 43-47.

Vidová Hladká Barbora, Hajič Jan, Hana Jiří, Hlaváčová Jaroslava, Mírovský Jiří, Raab Jan. The Czech Academic Corpus 2.0 Guide., Prague Bulletin of Mathematical Linguistics, 2008, 41-96.

Křen Michal, Hlaváčová Jaroslava. Corpus as a Means for Study of Lexical Usage Changes. In Proceedings of the 13th EURALEX International Congress. 2008, pp. 437-447.

Štěpánek Jan. Pražský závislostní korpus. In Varia XV. 2008, pp. 581-587.

Ptáček Jan. Two Tectogrammatical Realizers Side by Side: Case of English and Czech. In Fourth International Workshop on Human-Computer Conversation. Bellagio, Italy: [http://www.companions-project.org/events/200810_bellagio.cfm], 2008, pp. 1-4.

Štěpán, P.: Sufix *-dlo* v pomístních jménech v Čechách, Acta onomastica 49, 2008, pp. 333–343.

Giger, M.: Der "gehen"-Prospektiv im Slovakischen: Semantik und Grammatikalisierung. Vyjde v: Weiss, D. (ed.): Slavistische Linguistik 2006. Referate des 32. Konstanzer Slavistischen Arbeitstreffens in Zürich-Boldern. München. (Slavistische Beiträge). 25 s. [odevzdáno do tisku 03/08]

Giger, M: Partizipien als Exportschlager. Zum Einfluss des Russischen auf andere slavische Sprachen im 19. Jhdt. Vyjde v: Kosta, P. (ed.): Slavistische Linguistik 2007. Referate des 33. Konstanzer Slavistischen Arbeitstreffens in Potsdam. München. (Slavistische Beiträge). 32 s. [odevzdáno do tisku 04/08]