Výsledky dosažené v roce 2007

A. Sekce Pojmenované entity

V oblasti výzkumu pojmenovaných entit v češtině se pracovalo na úpravách anotačního schématu vytvořeného v předešlém roce. Šlo především o změny ve vymezení podkategorií. Nové schéma se použilo k anotaci dalších 2000 vět z Českého národního korpusu a k opravě dosud anotovaných dat. V současné době máme tedy celkem 6000 vět s kvalitně označenými pojmenovanými entitami.

Dále jsme vylepšili metody automatického rozpoznávání a značkování pojmenovaných entit. Rozpoznávač pojmenovaných entit byl rozšířen o nové klasifikační rysy, natrénován a vyhodnocen na nových datech a zbaven některých nedostatků. Byly automatizovány instalace a trénování systému. Probíhá tvorba jeho nové verze.

Vyvinuli jsme pravidla, podle kterých lze z morfologicky anotovaného textu získat následující typy pojmenovaných entit: osobní jména, telefonní čísla, adresy a kalendářní data. Pravidla jsme testovali pomocí dotazů na Českém národním korpusu. Odhad úspěšnosti všech těchto dotazů byl relativně vysoký, v několika případech dokonce stoprocentní.

Pro větší úspěšnost automatického rozpoznávače je výhodné mít seznamy nejrůznějších typů pojmenovaných entit v elektronické formě. Shromáždili jsme tedy místopisná data o České republice, Evropě a světě ve strukturované podobě s informacemi o umístění a vzájemných vztazích. Dále byla shromážděna data o výskytu a četnostech českých jmen a příjmení. Byla navržena a implementována metoda vyhodnocení věrohodnosti, zda formy odpovídající záznamu v těchto datech skutečně odkazují na danou entitu.

Plánovaná vizualizace pojmenovaných entit v TrEdu byla prozatím nahrazena vizualizací pomocí HTML. Zahájili jsme práci na vytvoření WWW rozhraní ke značkování pojmenovaných entit na internetových stránkách. V současné době je hotový algoritmus pro vkládání entit v podobě HTML tagů do prostého textu a rozpoznávání entit v textu v proxy serveru. Zbývá vyladit umísťování značek do správných částí HTML dokumentů.

Výsledky našeho výzkumu v oblasti pojmenovaných entit jsme využili v aplikaci strojového překladu - byly provedeny pilotní experimenty se zpracováním pojmenovaných entit pomocí loglineárního modelu a perceptronu v transferové fázi strojového překladu z angličtiny do češtiny.

B. Sekce Jednotný formát

Práce skupiny zabývající se jednotným formátem lze v roce 2007 rozdělit do dvou částí:

1. Formát PML
Byl navržen PML formát pro pětiúrovňový systém anotačních rovin pro anotaci mluvené řeči ve stylu PDT2.0 umožňující vedle původní mluvené podoby věty zachytit též tzv. rekonstrukci mluvené výpovědi. V této souvislosti byly vytvořeny nástroje pro import výstupního textového formátu rozpoznávače řeči do PML a pro převod mezi PML a XML formátem anotačního nástroje Transcriber.
Vytvořili jsme nový modulární anotační nástroj Med (nahrazující dřívější jednoúčelový Medit a v budoucnu též nástroj Transcriber) založený na PML, sloužící k anotaci vztahů mezi dvěma a více lineárními rovinami (užití ve speech reconstruction, translation alignment, atd.).

Pracovali jsme též na návrhu PML formátu pro český morfologický slovník a valenční slovník PDT. ValLex.

2. Prohledávání a zpracování velkých objemů anotovaných dat
Provedli jsme experimenty s vyhledáváním pomocí dotazovacího jazyka XQuery v anotacích uložených v nativních XML databázích (eXist, Sedna). Z těchto experimetů jsme však spíše vyvodili, že z hlediska výkonu nejsou zatím nativní XML databáze k ukládání a prohledávání stromových anotací v potřebných objemech vhodné.

Daleko nadějněji dopadly experimenty s uložením stromových anotací v relační databázi (testují se současně enginy PostgreSQL a Oracle 10g) založených na indexovém kódování stromů. Experimenty se zatím soustředily na efektivní vyhledávání. Byl vytvořen prototyp režimu pro anotační nástroj TrEd, jenž umožňuje uživatelům formulovat dotazy graficky pomocí dotazového stromu a následně tyto stromové dotazy převádí do jazyka SQL. Kromě grafických stromových dotazů pracujeme též na podpoře dotazovacích jazyků XPath a LPath+. Dále vyvíjíme automatický překladač, který pro danou anotaci v PML na základě PML schématu vygeneruje relační databázové schéma a převede anotaci do SQL databáze (pro účely experimentů jsme relační databázové schéma navrhli částečně ručně).

Vytvořili jsme prototyp nástroje JTred, jenž slouží pro paralelizaci zpracování anotovaných dat v rámci výpočetních clusterů; nástroj na rozdíl od stávajícího NTredu umožňuje efektivní dynamické proudové načítání dat, a tedy jejich zpracování i v objemech překračující operační paměť clusteru.

Řešitelský tým se na pracovišti MFF UK během roku nezměnil. Jedinou změnou bylo jméno jednoho z řešitelů, Jana Votrubce, který se nyní jmenuje Jan Raab. V příštím roce bychom chtěli do týmu opět zařadit Marii Křížkovou. Zmíněná změna a návrh jsou uvedeny ve změnovém listu.
Na spoluřešitelském pracovišti ÚJČ došlo během roku k personálním změnám. Od 1. 11. 2007 se Pavel Štěpán tomuto projektu věnuje již pouze polovinou pracovního úvazku; jeho druhou polovinu převzal Jiří Januška. Dr. Giger působil během posledních dvou měsíců uplynulého roku na Kostnické univerzitě v Německu.
Spolupráce mezi oběma pracovišti, tj. MFF UK a ÚJČ AV ČR, probíhala ve stejném duchu jako v předchozích letech. ÚJČ pracuje na převedení části lexikálního archivu vlastních jmen do elektronické podoby. Dále se soustředí na teoretické otázky, týkající se především homonymie vlastních jmen se jmény obecnými. Řešitelé na MFF se zabývají spíše praktickými aplikacemi.

Mzdové prostředky byly čerpány podle plánu, včetně 10 tis. Kč z FÚUP.
Na MFF se nepodařilo využít celou částku určenou na **OON**. Částečně z nich kryjeme přečerpání 21 tis. Kč z kolonky mezd, zbytek, tj. 70 tis. Kč, převedeme do FÚUP na příští rok.
Díky nedočerpání částky na OON se také nevyčerpala celá plánovaná částka, se kterou se počítalo na pojištění. I tuto částku (25 tis. Kč) převedeme do FÚUP.
Kvůli nečekanému odjezdu pracovníka ÚJČ dr. Gigera se nevyčerpal jeho podíl, takže se jeho plat s příslušnými zákonnými odvody převedl do FÚUP (41 tis. Kč) na rok 2008 a bude vyčerpán po jeho návratu.

Přidělená částka 360 tisíc Kč na **investice** byla použita jako spoluúčast na rozšíření projektem využívaného výpočetního clusteru LRC. Celková cena plně osazeného blade serveru Dell PE 1955, který byl postupně zakoupen v roce 2007, byla 1,594 tisíc Kč (deset výpočetních bladů). Pro potřeby projektu je tak vyhrazen ekvivalent výkonu 2,25 serveru.

Neinvestiční prostředky byly využity na obnovu drobné techniky užívané řešiteli projektu - myši, DVD mechaniky, antivirové řešení pro mailserver (spoluúčast), úpravy pracovních stanic. Část prostředků byla použita na nákup literatury a kancelářských potřeb.

Cestovní prostředky byly čerpány na prezentaci výsledků projektu doma i v zahraničí. Částka 5 tis. Kč převedená loni do fondu účelově určených prostředků (FÚUP) na MFF byla využita na

cestovné. Vzhledem k tomu, že se významná konference ACL 2007 konala v Praze, nevyčerpali jsme letos všechny prostředky určené na cestovné. Z nedočerpaných 19 tis. Kč jsme část použili na nákup neinvestic (11 tis. Kč) a na služby (2 tis. Kč), zbytek převedeme do FÚUP a vyčerpáme je v příštím roce.
Nejvýznamnější konferece, kterých jsme se letos zúčastnili: DAARC 2007 (Portugalsko), MEANING - TEXT THEORY 2007 (Rakousko), TSD 2007 (Plzeň), TLT 2007 (Norsko).

Převod do Fondu účelově určených prostředků
MFF UK - 95 tis. Kč na OON včetně zákonných odvodů a 6 tis. Kč na cestovné. Celkem jde o 101 tis. Kč.
ÚJČ AV ČR - 41 tis. Kč na mzdy včetně zákonných odvodů.
V tabulkách jsou tyto prostředky uvedeny, podle pokynů, jako vyčerpané.

Plán na rok 2008

A. Sekce Pojmenované entity
- Aplikování a testování systému pro automatické značkování pojmenovaných entit na velkých objemech dat (přibližně v řádu milionů vět).
- Dokončení webového rozhraní k detektoru pojmenovaných entit a vizualizačního rozhraní v TrEdu.
- Zahájení experimentu s rozpoznáváním pojmenovaných entit na paralelních česko-anglických datech z korpusu CzEng.

B. Sekce Jednotný formát
- Dokončení a čištění aplikačního rozhraní pro PML.
- Dokončení převodu stávajících zdrojů.
- Dokončení databázového jádra vyhledávacího enginu a dokončení implementace grafického uživatelského rozhraní v TrEdu.
– Propojení systému s dalšími existujícími datovými zdroji a nástroji (včetně zahraničních).

Publikace

Hana Jirka: Lexical Annotation Workbench (LAW), Version 0.7, Univerzita Karlova, 2007.

Lopatková Markéta, Žabokrtský Zdeněk, Kettnerová Václava, Skwarska Karolina, Bejček Eduard, Hrstková Klára, Nová Michaela, Tichý Miroslav Valenční slovník českých sloves. Karolinum, 2007.

Ševčíková Magda, Žabokrtský Zdeněk, Krůza Oldřich. Named Entities in Czech: Annotating Data and Developing NE Tagger. In Lecture Notes In Computer Science: Proceedings of the 10th International Conference on Text, Speech and Dialogue. 4629. Pilsen, Czech Republic: Springer Science+Business Media Deutschland GmbH, 2007, pp. 188-195.

Vidová-Hladká Barbora, Hana Jiří, Hajič Jan, Hlaváčová Jaroslava, Mírovský Jiří, Votrubec Jan: Czech Academic Corpus 1.0, Karolinum - Charles University Press, 2007.

Bojar Ondřej, Cinková Silvie, Ptáček Jan. Towards English-to-Czech MT via Tectogrammatical Layer. In NEALT Proceedings Series: Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007). 1. Bergen, Norway: North European Association for Language Technology, 2007, pp. 7-18.

Bojar Ondřej, Žabokrtský Zdeněk, Češka Pavel, Bena Peter, Janíček Miroslav: CzEng 0.7, ÚFAL

MFF UK, ÚFAL MFF UK, 2007.

Hlaváčová Jaroslava. Korpusové chyby. In Gramatika a korpus / Grammar & Corpora 2005. Prague, Czech Republic: ÚJČ AV ČR, 2007, pp. 77-86.

Hlaváčová Jaroslava. Pravopisné varianty a morfologická anotace korpusů. In Gramatika a korpus / Grammar & Corpora 2007. Prague, Czech Republic: Academia, 2007,.

Linh Nguy, Žabokrtský Zdeněk. Rule-based Approach to Pronominal Anaphora Resolution Applied on the Prague Dependency Treebank 2.0 Data. In Proceedings of the 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2007). Lagos (Algarve), Portugal: CLUP-Center for Linguistics of the University of Oporto, 2007, pp. 77-81.

Lopatková Markéta, Žabokrtský Zdeněk, Kettnerová Václava, Skwarska Karolina, Bejček Eduard, Hrstková Klára, Nová Michaela, Tichý Miroslav: VALLEX 2.5 - Valency Lexicon of Czech Verbs, version 2.5, 2007. V tisku.

Novák Václav, Žabokrtský Zdeněk. Feature Engineering in Maximum Spanning Tree Dependency Parser. In Lecture Notes In Computer Science: Proceedings of the 10th International Conference on Text, Speech and Dialogue. 4629. Pilsen, Czech Republic: Springer Science+Business Media Deutschland GmbH, 2007, pp. 92-98.

Pajas Petr, David Mareček: MEd - an editor of interlinked multi-layered linearly-structured linguistic annotations, UK MFF UFAL, 2007.

Ptáček Jan, Žabokrtský Zdeněk. Dependency-based Sentence Synthesis Component for Czech. In Proceedings of the 3rd International Conference on Meaning-Text Theory (MTT 2007). Munchen - Wien: Verlag Otto Sagner, c/o Kubon & Sagner, 2007, pp. 407-415.

Ptáček Jan. Sentence Synthesis in Machine Translation. In WDS'07 Proceedings of Contributed Papers. MFF UK, Trója, Prague: Matfyzpress, Charles University, 2007,.

Raab Jan. Comparing Prosody Formalisms for Machine Learning. In WDS'07 Proceedings of Contributed Papers. MFF UK, Trója, Prague: Matfyzpress, Charles University, 2007,.

Raab Jan: Morce - Czech morphological tagger, 2007.

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

Ševčíková Magda. Proper Nouns in Czech Corpora. In Proceedings of the Corpus Linguistics Conference Series. Birmingham, UK: 2007,.

Ševčíková Magda, Žabokrtský Zdeněk, Krůza Oldřich. Zpracování pojmenovaných entit v českých textech.Tech. Report TR-2007-36 ÚFAL MFF UK, 2007, 60.

Vidová-Hladká Barbora, Hajič Jan, Hana Jiří, Hlaváčová Jaroslava, Mírovský Jiří, Votrubec Jan Czech Academic Corpus 1.0 Guide. Karolinum - Charles University Press, 2007.

Žabokrtský Zdeněk, Lopatková Markéta. Valency Information in VALLEX 2.0: Logical Structure of the Lexicon., Prague Bulletin of Mathematical Linguistics, 2007, 41-60.

Giger, M., Štěpán, P.: Některá dosud nepopsaná česká deverbální příjmení typu Vybíral, Odložil. Čeština doma a ve světě 15, 2007, s. 90-100.

Giger, M., Štěpán, P.: Pojmenované entity v počítačové lingvistice a vlastní jména. Acta onomastica 48, 2007, s. 44-53.

Giger, M.: In der Falle der schwachen Phoneme: Der phonologische Status der russischen unbetonten Vokale in der Akademiegrammatik von 1980 und anderen Konzeptionen. Sborník prací filozofické fakulty brněnské univerzity A 55, 2007, s. 131-142.

Giger, M.: Jít s sebou a podobná spojení v ČNK a na internetu. Naše řeč 90, 2007, s. 195-202.

Štěpán, P.: Pomístní jména v Čechách utvořená sufixy -ina a -inka z propriálních základů. Acta onomastica 48, 2007, s. 156-164.