CENTRUM ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA NA FI MU

Datum konání: 5. 11. 2021

Místo konání: FF MUNI, místnost D51

Název přednášky: Centrum zpracování přirozeného jazyka na FI MU

Přednášející: doc. RNDr. Aleš Horák, Ph.D., a doc. Mgr. Pavel Rychlý, Ph.D.

Počet účastníků: ~13

Zpracovali: Vlček Tomáš

V pátek 5. listopadu nás v rámci cyklu odborných přednášek navštívili doc. RNDr. Aleš Horák, Ph.D., a doc. Mgr. Pavel Rychlý, Ph.D., z Fakulty informatiky MU. Docent Horák je vedoucím Katedry strojového učení a zpracování dat. Docent Rychlý je vedoucím Centra zpracování přirozeného jazyka. Oba se dlouhodobě zabývají počítačovým zpracováním přirozeného jazyka a souvisejícími oblastmi, jako je strojové učení, a oba, ještě s profesorem Karlem Palou, stáli u vzniku Centra zpracování přirozeného jazyka. V současnosti v Centru zpracování přirozeného jazyka vedou dohromady přes 50 různých projektů.

Jako první se slova jal docent Rychlý. Začal krátkým představením Centra zpracování přirozeného jazyka, které na Fakultě informatiky Masarykovy univerzity funguje již 24 let. Centrum zpracování přirozeného jazyka se, jak z názvu vypovídá, zabývá jak samotným výzkumem v oblasti NLP, tak i spoluprací s průmyslem při vývoji aplikací pro reálné využití. Několik takových projektů nám poté představil.

Jedním z prvních takovýchto projektů byla spolupráce s firmou Seznam na zdokonalení zpracování dotazů v jejich vyhledávači. Dotazy jsou při vyhledávání morfologicky analyzovány, aby byla doplněna případná chybějící diakritická znaménka a aby každému slovu bylo přiděleno lemma. Oproti klasické analýze jsou slovům přidělována specifičtější lemmata, aby nedocházelo ke ztrátě významu a nepřesným výsledkům vyhledávání. Jelikož musely být analyzovány milióny slov za sekundu, starší analyzátor Ajka, který například doplňoval diakritiku poměrně pomalu, musel být nahrazen novější Majkou.

Následně mluvil o spolupráci se společností Lexical Computing, která se zabývá tvorbou jazykových korpusů a vývojem nástrojů pro práci s nimi. Jejím nejznámějším produktem je korpusový manažer Sketch Engine, který se v oblasti lexikografie stal de facto standardem. Centrum zpracování přirozeného jazyka s Lexical Computing blízce spolupracuje od roku 2004. Lexical Computing je strategickým partnerem FI MU: Zaměstnanci Lexical Computing běžně vyučují na FI MU a podílejí se na vedení diplomových a bakalářských prací.

Centrum zpracování přirozeného jazyka se společně s Ústavem pro jazyk český AV ČR podílelo na vývoji Internetové jazykové příručky. CZPJ pracovalo jak na vývoji technické stránky, tak i na přípravě některých dat. Aktuálně se stará o její provoz.

Dále docent Rychlý mluvil o projektu DEB, což je skupina více jak 10 mezinárodních projektů, které se zabývají počítačovou lexikografií a mezi které patří například prohlížeč slovníků DEBDict, výkladový a překladový slovník CZJ znakový slovník či The Oxford Dictionary of Family Names in Britain and Ireland. Informoval nás o spolupráci se společností Deloitte, pro kterou Centrum zpracování přirozeného jazyka vytvářelo systém pro získávání informací ze záznamů schůzek s klienty, a o projektu CityDoc, ve kterém šlo o dolování informací z dokumentů správních řízení vydávaných městskými a obecními úřady a následném vyznačování zasažených oblastí na mapě.

Posledním tématem, o kterém docent Rychlý mluvil, byl strojový překlad, jímž se v současné době zabývá. Centrum zpracování přirozeného jazyka se aktuálně věnuje problematice statistického strojového překladu a spolu se společností Smith pracuje na tvorbě hybridního překladače.

Poté se slova chopil docent Horák, který poměrně podrobně rozebral dvě témata: detekci anonymního autorství a rozpoznávání manipulativních technik v textu.

Podle něj můžeme detekci anonymního autorství chápat jakožto tři příbuzné problémy (seřazeny podle složitosti):

A. Ověření autorství: Porovnáváme texty a zkoumáme, jestli dva dokumenty napsal stejný autor (1 : 1). Srovnáním vícero textů můžeme ověřit, zda dokument doopravdy napsal ten, kdo je za autora pokládaný (1 : n).
B. Nalezení autorství: Z množiny autorů se snažíme určit, kdo napsal daný text.
C. Shlukování dokumentů podle autorství: Soubor textů se snažíme rozdělit do skupin podle možného autorství.

V současnosti se všechny tyto problémy snaží řešit výpočetní stylometrie. Zkoumaný dokument je nejprve nutné morfologicky a syntakticky označkovat. Potom se z charakteristik jako je délka slov, vět či celého textu, hloubka syntaktických stromů, bohatost slovní zásoby, pravopisné chyby, opakování slov či slovních druhů apod. vytvoří vektor označovaný jako stylom či writeprint, který se poté porovnává se stylomy ostatních textů. Centrum zpracování přirozeného jazyka pracovalo na projektu pro Ministerstvo vnitra ČR, ve kterém šlo o vytvoření statistického modelu určování autorství česky psaných textů a za který CZPJ dostalo Cenu ministra vnitra za mimořádné výsledky v oblasti bezpečnostního výzkumu.

Nakonec krátce popsal společný projekt Fakulty informatiky, Fakulty sociálních studií a Právnické fakulty MU Manipulativní techniky propagandy v době internetu. CZPJ se na tomto projektu podílelo tvorbou softwaru, který v textech stažených z internetu (v tomto případě z proruských zpravodajských serverů) pomocí některých stylometrických metod detekuje osm různých manipulativních technik (*demonizing*, *relativizing*, *fear mongering*, *labelling*, *confabulation*, *emotions*, *argumentation*, *blaming*).

Docent Horák a docent Rychlý nám přiblížili práci Centra zpracování přirozeného jazyka – nejenže nám sdělili, jaké v CZPJ řešili projekty a čím se aktuálně zabývají, také se s námi podělili o své zkušenosti a postřehy z oboru. Naše počítačovělingvistické rozhledy byly rozhodně rozšířeny.