Projekt Malach dokončen. Archiv nahrávek zpřístupní knihovna MFF UK

Projekt Malach dokončen. Archiv nahrávek zpřístupní knihovna MFF UK

Název mezinárodního projektu Malach vznikl zkrácením spojení Multilingual Access to Large Spoken Archives. Slovo malach má ale v hebrejštině i poetičtější význam, a to kraloval či ustanovil za krále. Cílem projektu bylo otestovat možnosti automatické indexace multimediálního archivu videonahrávek pamětníků holokaustu, jednoho z největších digitálních archivů na světě. České nahrávky měl na starosti tým z Ústavu formální a aplikované lingvistiky MFF UK. Projekt Malach nám představili jeho tvůrci prof. RNDr. Jan Hajič, Dr., a Mgr. Pavel Pecina, Ph.D., za fakultní knihovnu hovořila PhDr. Petra Hoffmannová.

Malach je mezinárodní projekt. Jak se do něj podařilo zapojit vašemu týmu?

JH: Já jsem měl dobré kontakty na univerzitu Johns Hopkins v Baltimoru, kde jsem tři semestry učil. Univerzitu oslovila někdy kolem roku 1999 nadace Survivors of the Shoah – Visual History Foundation. Tato nadace byla založena v roce 1993 Stevenem Spielbergem, který chtěl natočit vzpomínky lidí, kteří přežili holokaust. Což se také zrealizovalo. Zjistilo se ale, že ruční indexace nahrávek by trvala desítky let, a zkoušelo se, jestli by to nebylo možné automatizovat. Tak sepsali grant a obrátili se na další instituce – jednou z nich byl i jmenovaný Johns Hopkins – aby jim s tím pomohly. IBM, které bylo rovněž přizváno, řeklo, že by si vzalo na starosti jen anglické nahrávky, a tak hledali někoho, kdo by jim pomohl s ostatními jazyky. A na Johns Hopkins Univerzity věděli, že my se této oblasti věnujeme, a oslovili nás. My se specializujeme spíše na textovou podobu, a proto byla ještě navázána spolupráce se Západočeskou univerzitou v Plzni, skupinou profesora Josefa Psutky, která se věnuje audiu.

Ředitel Ústavu formální a aplikované lingvistiky prof. RNDr. Jan Hajič, Dr.

Nahrávání výpovědí 52 000 svědků z 56 zemí ve 32 jazycích trvalo pět let. V roce 1999 byl proces indexace interview téměř dokončen. Jak probíhala práce s materiálem v rámci projektu Malach, a jak se o ni jednotlivé instituce podělily?

JH: Firma IBM pracovala nezávisle na nás na anglických nahrávkách, východoevropské jazyky měli na starosti v Johns Hopkins (a tedy i my). Univerzita v Marylandu se podílela na samotném systému vyhledávání, připravovali seznam témat, která sloužila k testování a vyhodnocování. My jsme měli za úkol zajistit rozpoznávání českých nahrávek, později i dalších jazyků.

Kolem roku 2005 začala práce na druhé části projektu. Již jsme měli převedenu řeč do textů a v těch jsme potřebovali vyhledávat. Požadavky na vyhledávání nebyly ale standardní, protože v dotazech nefigurovala jen klíčová slova, ale i popis situace v rozsahu zhruba jednoho odstavce. Přijde třeba historik a chce všechny nahrávky, v nichž se mluví o tom, jak lidé přišli do ghetta a obstarávali si tam jídlo, než šli do transportu. A zajímá ho to jen pro lidi z jižní Moravy. V původním návrhu bylo, že slova z tezauru se budou k jednotlivým pasážím přiřazovat ručně, a nejenom klíčová slova, ale i hodnotící shrnutí. Zkusili to u 10 % anglických nahrávek a zjistili se, že by to byla práce na desítky let.

PP: Indexace těch pouhých 10 % trvala 4 roky, stálo to 8 milionů dolarů a zpracování jedné hodiny nahrávky trvalo 35 hodin.

„Mohli jsme konstatovat – umíme i v takto obtížných nahrávkách vyhledávat relevantní úseky," říká Mgr. Pavel Pecina, Ph.D.

JH: Proto se objevil požadavek na automatizaci celého procesu. Systém byl zjednodušen tak, že jeden člověk poslouchal bez zastavení hodinu nahrávky a při tom ji rovnou označoval slovy z tezauru. Celé to nesmělo trvat déle než hodinu a pět minut. A takto jednoduše byl nakonec označen celý archiv.

Co konkrétně bylo úkolem českého týmu?

JH: Prvním úkolem bylo převést audio do textové podoby. To dělali tři lidé tady v Praze a celý tým v Plzni. Dále se musel zajistit překlad tezauru z angličtiny do dalších jazyků. A pak proces vlastního vyhledávání.

Co šlo využít ze systému připraveného pro angličtinu?

JH: Teorii, ta je všude stejná. Systémy, co jsme měli, ale nešly příliš na tuto specifickou oblast použít.

PP: Téma holokaustu a 2. světové války má totiž velice speciální slovník, který se liší od toho, co posloucháme v televizi nebo je na internetu. Jde o geografické názvy, jména lidí...

JH: Software a zejména slovníky a jazyková data musely vzniknout nově, aby se systém vylepšil a zmenšila se chybovost.

Jak konkrétně systém převodu audiostopy do textové podoby funguje?

JH: My víme, že jazykový systém jsou poskládané frekvence. Existují základní a přídavné frekvence a my z toho dokážeme přibližně poznat, o jaké písmeno jde. Každý signál se převede do seznamu čísel, která znamenají, jak v tom signálu byly určité frekvence přítomny. To jde dnes docela rychle. Síly frekvencí se pak začnou porovnávat s fonémy, které předem někdo nahrál, a hledá se, čemu se nejvíce blíží. Protože výsledkem jsou jen pravděpodobnosti, čeká se, až těchto fonémů bude celá řada. Pak se použije slovník, který výskyt fonémů omezuje. Například máme slovo, které se přepíše jako l-e-f, s menší pravděpodobností to může být r-e-f nebo c-e-f. Ve slovníku pak systém vyhledává, jaké podobné české slovo existuje, je to ale stále jen s určitou pravděpodobností. Tato slova jsou výsledkem akustického modelování. Tím se zabývali v Plzni. Pak musí přijít na řadu jazykové modelování, které řekne, jaké jsou možnosti řazení slov v češtině a v dané doméně. Na tom už jsme dělali i my. My jsme tedy hledali nejlepší posloupnost těchto slov ze všech možností a snažili se spočítat pravděpodobnost pro celou posloupnost. Protože i ta nejlepší slova mohou tvořit

naprostý nesmysl. Mám-li to shrnout, tak jazykový model nám říká, jakou pravděpodobnost mají posloupnosti slov v češtině.

S jakými problémy jste se potýkali?

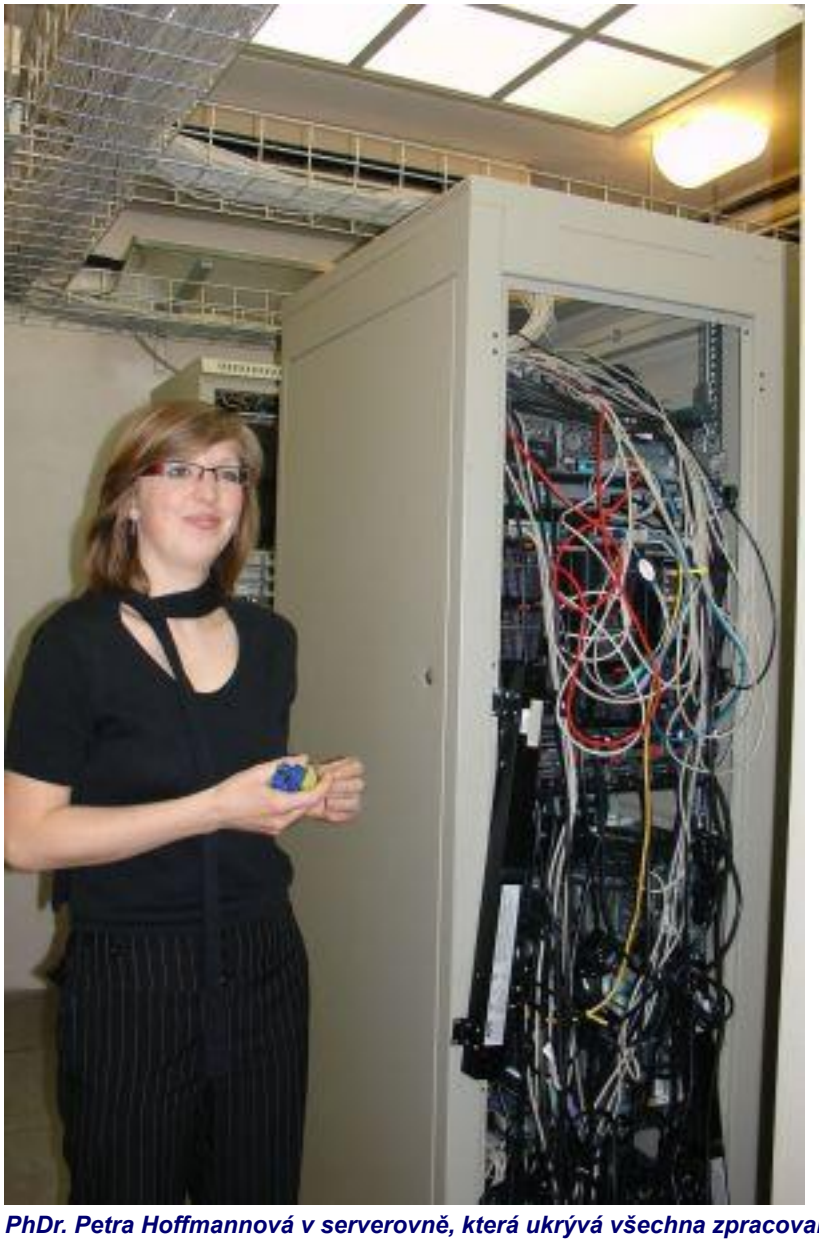
JH: My jsme už nějaké jazykové a Plzeňáci akustické modely měli. Ale když jsme je pustili na konkrétní texty, tak jsme zjistili, že to nefunguje dobře. Bylo to tím, že nahrané texty byly naprosto specifické. Jednalo se o spontánní řeč, byla tam nová slova nebo jejich kombinace. Takže v rámci tohoto projektu jsme museli najít správná data pro předělání jazykového modelu, aby lépe vyhovoval našim textům.

PP: Já jsem tedy musel s kolegou na internetu vyhledat texty, které by byly těm původním podobné. Tím jsme zdesetinásobili počet původních dat v systému.

JH: Chybovost ale byla okolo 35 %, což byl nejlepší výsledek, jakého jsme dosáhli.

PP: To je sice každé třetí slovo, ale ještě před tím byl udělán průzkum, že pokud bude počet chyb do 40 %, tak je v takto převedených rozhovorech možno úspěšně vyhledávat. V angličtině dosáhli chybovosti jen o několik procent nižších. Například náš systém rozpoznal větu jako „doktor Jařab napsal skupinku“ a správně tam mělo být „neschopenku“. Nebo „upytlačila jsem“ bylo vysloveno tak, že my jsme to zaznamenali jako „upekla jsem“.

PH: Nejzajímavější chybou bylo asi slovo „hypermangan“, které systém rozpoznal jako „Hitlerova maminka“...



PhDr. Petra Hoffmannová v serverovně, která ukrývá všechna zpracovaná data

JH: Problémem je, že rozpoznávání normálně probíhá v reálném čase. Pokud by bylo na rozpoznávání více času, byl by výsledek o něco lepší. Ale nikdy ne perfektní.

Jak jste „naučili“ systém vyhledávat podle zadaných kritérií?

PP: Nejprve v Marylandu vytvořili přes sto témat. Ta vznikla tak, že pozvali na univerzitu studenty, historiky, dokumentaristy, prostě lidi, kterým je tato problematika blízká, a poskytli jim na několik hodin přístup k archivu. Oni

pak měli specifikovat témata, která je zajímala. Tato témata byla přeložena do všech používaných jazyků a pak se vyhledávala v nahrávkách. My jsme zkoumali, jak dobře s tím umí systém pracovat a s jakou úspěšností dokáže úseky dokumentu požadované uživateli najít. Proto jsme najali několik studentů, kteří nám ručně v nahrávkách vyhledávali příslušná témata. To trvalo od roku 2006 do září 2007. Studenti pomocí vyhledávacích metod našli úseky, které odpovídají vyhledávaným tématům. My jsme na ta data aplikovali náš vyhledávací systém a porovnávali jsme, jak dobře umí napodobit práci lidí. To jsme vyhodnotili a mohli jsme konstatovat – umíme i v takto obtížných nahrávkách vyhledávat relevantní úseky.

V čem se vyhledávání liší od toho běžného třeba na internetu?

PP: Systém nehledal jen dokument, ale i příslušnou pasáž. Pokud někoho zajímalo téma pochod smrti v Plzni, tak my mu nevyhledáme jen sedm nahrávek, kde je o tom zmínka, ale i konkrétní místo, kde se o tom hovoří. Dalším rozdílem je, že k vyhledávání se nepoužívají jen přepisy textů, ale i automaticky přiřazená klíčová slova tezauru. Existovaly tedy dva druhy informací. My jsme také zjistili, že jednomu tématu mluvčí většinou věnuje dvě až tři minuty. Proto jsme celou výpověď rozdělili do 2,5minutových úseků, které pak systém procházel a hledal, zda obsahují či neobsahují požadované téma. Tím jsme lépe definovali začátek a konec úseku, kde se o tématu hovoří.

Systém je tedy hotový. Co bude s projektem dál?

PP: Doposud byl projekt ve výzkumné fázi. Nyní přijde na řadu jeho zpřístupnění uživatelům.

PH: V září by se měl v naší knihovně otevřít přístupový bod k digitálnímu archivu USC Shoah Foundation (archiv interview shromážděných nadací Survivors of the Shoah se mezitím totiž přestěhoval na Univerzitu Jižní Kalifornie, USC), umožňující prohledávání všech 52 000 interview za použití klíčových slov. A to ve všech 32 jazycích, v nichž jsou nahrávky zaznamenány. Podobná centra jsou zatím ještě v Německu a Maďarsku.

JH: Část nahrávek bude uložena v kopii přímo u nás, to budou ty, o něž budou mít lidé největší zájem, a část v Kalifornii, protože nahrávky jsou opravdu rozsáhlé a my bychom tu neměli odpovídající kapacity. I tak bude potřeba vybudovat poměrně rozsáhlý počítačový systém, aby se tam alespoň malá část záznamů vešla. S ukládáním dat máme vůbec veselé historky. V roce 2000 jsme řešili problém, jak do Česka data z Ameriky vlastně dostat. Zjistili jsme, že po internetu to nejde, to by trvalo léta. Nejjednodušší a nejlevnější nakonec bylo nakoupit disky, doletět do Ameriky, nahrát je tam a dovézt zpátky. Nelítalo se tam samozřejmě jen kvůli tomu, ale kopírovalo se to během nějaké porady, které jsme mívali. Jednou jsme měli problém i při kontrole na letišti, protože tašku plnou disků jsme převáželi krátce po 11. září 2001...

Komu bude nově zprovozněné centrum zejména sloužit?

JH: V Americe se ukázalo, že o přístup k archivu je velký zájem. Zajímá dokumentaristy a filmaře obecně, je ale cenným zdrojem informací a materiálů také pro pedagogy, historiky, psychology, lékaře, právníky a další profese. Pro nás je zajímavé, že budeme mít přístup ke všem datům, což je pro další vývoj automatizovaných systémů důležité. Tolik desítek tisíc nahrávek bychom sami nikdy nesebrali. My je budeme dále používat pro náš jazykový výzkum.

Přístupový bod k digitálnímu archivu USC Shoah Foundation bude v nové studovně na ochoze (**Lucie Kettnerová**)