■ Povedlo se – z výzkumných záměrů

MEZINÁRODNÍ PROJEKT MALACH ŘEŠENÝ NA KATEDŘE KYBERNETIKY

Na katedře kybernetiky je od roku 2001 řešen velmi prestižní vědecký projekt MALACH (Multilingual Access to Large Spoken Archives), který je finančně zajiš□ován americkou grantovou agenturou National Science Foundation (NSF).

Vznik projektu byl přímou reakcí na film "Schindlerův seznam", natočený v roce 1994. Po uvedení filmu byl režisér Spielberg kontaktován desítkami lidí, kteří prožili hrůzy holocaustu a kteří mu chtěli vyprávět svůj životní příběh. Režisér si uvědomil, že žijící svědci holocaustu jsou již staří lidé, s nimiž v krátké době navždy odejde obrovská historická zkušenost lidstva, která zatím nebyla nijak zaznamenána.

Steven Spielberg se proto rozhodl založit nadaci Survivors of the Shoah Visual History Foundation (VHF) se sídlem v Hollywoodu, jejímž prvotním posláním se stalo vyhledávat po celém světě lidi, kteří se stali svědky holocaustu a kteří jsou ochotni svůj příběh vyprávět ostatním, a uchovat tak svědectví o této hrůzné době pro příští generace. Tato rozsáhlá vyhledávací kampaň byla odstartována v roce 1995. Po celém světě byly ustavovány týmy složené z reportéra a kameramana, které navštěvovaly svědky ochotné předat svoje životní zkušenosti a většinou v jejich bytech s nimi natáčely videorozhovor. Ten byl obyčejně strukturován tak, že přibližně 20 % vyprávění bylo věnováno předválečnému období, 60 % období světové války a zbytek se týkal doby poválečné. Během 4 až 5 let tyto týmy nasbíraly v 57 zemích cca 52 tisíc svědeckých výpovědí. Výpovědi byly namluveny celkem v 32 různých jazycích a průměrná délka jedné výpovědi činila více než 2 hodiny. Celková délka všech pořízených výpovědí se tak přiblížila 120 tisícům hodin! Každý rozhovor byl zaznamenán na VHS kazetu (zvuk ve stereo formátu - jeden mikrofon měl svědek, druhý reportér) a tyto kazety byly zasílány do VHF, kde byly digitalizovány do formátu MPEG-1 (128 kb/s, zvuk 2 x 44kHz). V nadaci VHF vznikl obrovský robotizovaný digitální archiv s celkovou kapacitou 180 terabytů, který by měl být v budoucnu propojen optickou interaktivní sítí s muzei, školami či historickými archivy po celém světě.

Aby mohli historici, studenti, scenáristé a obecně všichni zájemci vyhledávat v archivu potřebné informace, byl zahájen promyšlený a dobře připravený katalogizační proces, při kterém najatí a vyškolení anotátoři výpovědi postupně segmentovali a vzniklé segmenty popisovali vhodnými klíčovými slovy, která byla vybírána z předem připraveného slovníku (tezauru), obsahujícího více než 20 tisíc položek. Tento anotační proces byl ale časově i finančně mimořádně náročný. Pokud by tímto postupem měly být zpracovány všechny výpovědi, stály by veškeré práce přes 150 milionů dolarů. To byl hlavní důvod, proč bylo rozhodnuto přizvat k řešení uvedeného problému odborníky z oblasti zpracování mluvené řeči.

Začátkem roku 2001 jsem byl vyzván kolegy z Johns Hopkins University v Baltimore, abychom se společně s ÚFAL na MFF UK v Praze připojili k týmu výzkumníků z IBM T.J. Watson Research Laboratory, VHF a University of Maryland. Cílem bylo požádat americkou grantovou agenturu NSF o grant na řešení problematiky zpracování rozsáhlého archivu výpovědí svědků holocaustu moderními metodami řečových technologií. Naši kolegové se na nás obrátili pravděpodobně proto, že jsme v předchozích třech letech společně řešili projekt typu Kontakt, v jehož rámci jsme v bezvadné kvalitě připravili rozsáhlé datové zdroje pro trénování systémů automatického rozpoznávání souvislé řeči (tyto zdroje jsou od roku 2000 celosvětově distribuovány nakladatelstvím Linguistic Data Consortium v Pennsylvanii). V tvrdé soutěži dalších amerických projektů z oblasti IT byl náš projekt k radosti všech žadatelů přijat. Úkolem našeho týmu a týmu z MFF UK bylo v projektu zajistit zpracování vybraných jazyků střední a východní Evropy.

Vzhledem k počtu výpovědí namluvených v různých jazycích bylo na první schůzce řešitelů v roce 2001 rozhodnuto, že IBM bude zpracovávat pouze nahrávky namluvené v angličtině (v archivu jich je asi polovina celkového množství), my jsme dostali za úkol zpracovat každý rok jeden jazyk. Bylo rozhodnuto zpracovávat jazyky v pořadí čeština (v archivu je 573 výpovědí), ruština (7052), slovenština (583), polština (1549) a na závěr maďarština (1038).

Měl bych nyní vysvětlit, co bylo předmětem naší práce a v čem spočívalo její řešení. Cílem bylo využít technologie rozpoznávání mluvené řeči ke zpracování výpovědí, tj. k jejich automatickému převodu z mluvené řeči do textové podoby, která by byla základem pro vyhledávání informací ve videoarchivu i k podpoře automatizace katalogizačního procesu.

Širší tým řešitelů projektu MALACH z katedry kybernetiky Fakulty aplikovaných věd.

Abychom vůbec mohli začít pracovat na vývoji jednotlivých systémů rozpoznávání souvislé řeči pro vybrané jazyky, museli jsme získat z VHF potřebná data. Vzhledem k objemu a důvěrnosti dat jsme v průběhu prvních tří let převáželi nahrávky do Plzně na hard discích. Pokaždé, když někdo z řešitelského týmu cestoval do USA nebo do České republiky, dostal 6 až 8 disků a převezl je (v tašce). Za 3 roky jsme takto převezli více než 100 disků s desítkami terabytů dat. U nás se z videonahrávek extrahovaly obě zvukové stopy a zvuková data se zálohovala nejprve na CD-ROMy, později (jak klesaly ceny datových nosičů) na DVD.

Pro návrh a konstrukci systému automatického rozpoznávání spojité řeči je zapotřebí mít k dispozici anotovaná data pro akustické modelování. Anotační práce mohou úspěšně dělat pouze rodilí mluvčí, tj. pro zpracování češtiny rodilí Češi, pro zpracování ruštiny rodilí Rusové atd. Anotátoři museli být vždy v několika sezeních vyškoleni a jejich práce byla stále kontrolována. Činnost anotátorů spočívá v práci na počítači, kdy se sluchátky na uších poslouchají segmenty jednotlivých nahrávek a ve speciálním programu tyto nahrávky "popisují", tj. přesně zapisují, co řečník na záznamu říká, včetně popisu tzv. neřečových událostí, tj. například hlasitých nádechů řečníka, hlasitého "přemýšlení", šustění papírem, šoupání židlí, vzdálené řeči na pozadí, slyšitelného ruchu z ulice ap. Vzhledem k obsahu sdělovaných výpovědí se jednotliví svědci holocaustu dostávali často do velmi vypjatých emočních stavů, takže naši anotátoři museli popisovat i úseky s pláčem, šeptáním, kašláním, smrkáním ap. Obecně lze říci, že řečový projev většiny svědků nebyl příliš kvalitní, což způsoboval do značné míry i jejich věk (průměrný věk řečníka byl cca 75 let). Největší problémy jsme měli se zpracováním ruskojazyčných výpovědí. Tito řečníci žili většinou dlouhou dobu mimo oblast Ruska (nejčastěji v Izraeli, USA a na Ukrajině), kde i poskytli svoji výpověď, přičemž jejich ruština byla velmi často deformována výrazným akcentem, používáním neruských slov i neruských fonémů ap., což značně komplikovalo konstrukci "ruského" systému. Při zpracování čtyř slovanských jazyků jsme překvapivě zjistili, že žádný z těchto jazyků nepoužívá v běžné spontánní mluvě takové množství nespisovných slov jako čeština (v češtině se jinak píše a jinak mluví).

Pokud to bylo schůdné, pak jsme pro anotační práce využívali naše studenty. To bylo možné při přípravě datových zdrojů pro češtinu a slovenštinu. Při zpracování ruštiny s námi pracoval opět náš student (rodilý Rus) a skupina externích spolupracujících Rusů z Plzně a okolí. Při zpracování polštiny jsme v Plzni nalezli jen jednu Polku, ostatní anotátoři byli najímáni v celé ČR, což způsobovalo problémy zejména při jejich řízení. Po domluvě s našimi americkými partnery jsme proto při zpracování zatím posledního jazyka, kterým je maďarština, zvolili odlišný způsob práce. Nalezli jsme partnery na TU v Budapešti, kteří projevili zájem na této úloze pracovat. Následovala návštěva našich pracovníků v Budapešti, kdy jsme školili tamní anotátory i návštěvy maďarských kolegů v Plzni, kdy jsme jim ukazovali celý postup při vývoji systému rozpoznávání řeči. I když je zvolený postup finančně mnohem náročnější než provedení všech prací pouze v naší režii, tato alternativní varianta práce na novém jazyku byla americkou grantovou agenturou finančně podpořena, nebo□ znamená jistou osvětu a vývoz technologie do zemí s menšími zkušenostmi v dané oblasti.

Pro vývoj každého systému rozpoznávání řeči bylo zapotřebí zpracovat vždy 400 patnáctiminutových vzorků řeči od různých řečníků tak, abychom získali co nejpestřejší paletu různých hlasů. Museli jsme se naučit a zpracovat fonetiku jednotlivých jazyků, abychom mohli dobře modelovat jednotlivé fonetické jednotky zpracovávaných jazyků. Pracné a časově velmi náročné byly výpočty spojené s tvorbou akustických modelů jednotlivých systémů, trénování parametrů akustických modelů zde zabralo tisíce hodin strojového času nejvýkonnějších počítačů. Modely jazyka, tj. statistiky řazení jednotlivých slov, nám pro tyto účely dodávali naši partneři z MFF UK v Praze.

Mohu se zadostiučiněním konstatovat, že náš tým zatím vždy splnil včas všechny povinnosti vyplývající pro nás z tohoto výzkumného projektu, a to standardně ve velmi dobré kvalitě. Naši zahraniční partneři velmi oceňují, že funkcionalita vyvinutých systémů je zcela srovnatelná se systémy rozpoznávání mluvené angličtiny, na jejímž vývoji pracovala firma IBM, která je špičkovým světově uznávaným pracovištěm v oblasti řečových technologií.

Až budou výsledky prací na projektu MALACH dokončeny a implementovány v digitálním archivu VHF, bude možné na základě jednoduchých dotazů vyhledávat ve výpovědích konkrétní události (např. vyprávění o slavení židovských svátků, o transportu do koncentračního tábora, o stravě v koncentračním táboře, o setkání s dr. Mengelem ap.) a následně si přehrát úseky tohoto vyprávění. Bude též možné se dotazovat na konkrétní slova, jména, geografické názvy (města, vesnice) ap. Počítá se též s tím, že s využitím vícejazyčného tezauru klíčových slov půjde prohledávat i výpovědi namluvené v jiných jazycích, než v jakém bude formulována otázka. Vzhledem k tomu, že by mělo dojít k propojení digitálního archivu VHF s terminály po celém světě, budou tyto neocenitelné zkušenosti a prožitky z těžkého období holocaustu zpřístupněny v podstatě všem lidem, kteří o ně budou mít zájem.

Práce na projektu MALACH se za dobu jeho řešení dotkly všech pracovníků a doktorandů na oddělení umělé inteligence katedry kybernetiky. Největší tíha výzkumné práce však dlouhodobě ležela na Ing. P. Ircingovi, Ing. J. V. Psutkovi, ml., Ing. J. Matouškovi a též doc. L. Müllerovi a doc. V. Radové.

Projekt MALACH je svým rozsahem, množstvím zpracovávaných jazyků i obecně lidským rozměrem zcela ojedinělým projektem na světě. Jsem rád, že jsme byli k jeho řešení přizváni a že se nám podařilo kvalitní prací zviditelnit naši univerzitu v mezinárodním měřítku.

Prof. Josef Psutka,
FAV, KKY