Počítače mapují holocaust

26.7.2003 Lidové noviny str. 21 Věda

Martin, Uhlíř

Elektronický systém vyhledává ve vzpomínkách pamětníků vyhlazovací apokalypsy * Na projektu spolupracují i čeští vědci

Počítač dokáže převést namluvené vzpomínky pamětníků holocaustu do psané formy a v textu pak vyhledávat. Systém nalezne odpověď i na velmi složité dotazy.

"Nacpali nás do vlaků, které původně měly přepravovat granáty, a těmi nás vezli až do Berlína. Náhle přišel letecký útok. Němci ... se běželi schovat, protože se báli bomb. ... Pro nás však letecký útok znamenal svobodu. Ale co se nestalo - spojenci byli přesvědčeni, že vlak je skutečně plný munice, a chtěli ho zničit. Začali nás bombardovat."

Tak vzpomíná na dramatické události konce války paní Sidonia Laxová, žena, která přestála nacistickou vyhlazovací mašinerii. Její svědectví lze vyslechnout na internetových stránkách nadace Survivors of the Shoah Visual History Foundation (www.vhf.org).

Za hranicemi lidských možností

Lidé, kteří přežili holocaust, nechtějí mlčet. Ukázalo se to v Polsku při natáčení filmu Schindlerův seznam. Za režisérem Spielbergem tehdy chodili například bývalí vězni koncentračních táborů a vyprávěli mu, co je za války potkalo.

Pamětníci však stárnou, umírají a berou si vzpomínky s sebou do hrobu. Režisér si uvědomil, že ze světa mizí cenná svědectví. Založil proto zmíněnou nadaci, aby vzpomínky zachránila.

Pracovníci nadace se rozjeli po světě a začali zaznamenávat výpovědi lidí, kteří prošli soukolím vyhlazovacího stroje. Zachytili přibližně 53 000 výpovědí v různých jazycích, jež představují přes 100 000 hodin videozáznamu.

V takovém oceánu dat se lze jen stěží orientovat. Cílem přitom je, aby se kdokoliv mohl dostat k informacím, které jsou pro něj důležité. Potomci obětí holocaustu by například v oněch desítkách tisíc hodin záznamu měli být schopni najít několik vět, zachycujících vzpomínku na jejich strýce či dědečka.

Řešení se na první pohled zdá být jednoduché: přepišme výpovědi do textových záznamů, v nichž umějí snadno vyhledávat počítačové programy. Jenže kdyby měli všechna svědectví přepisovat lidé, musela by nadace najmout armádu pomocníků a utratit miliony dolarů.

Kdo unikl transportům?

Je nabíledni, že tvrdý oříšek mohou rozlousknout jen počítače. Pomáhají jim nejmodernější přístupy založené na umělé inteligenci a matematické lingvistice. Ani ty však nenabízejí řešení bez náročného, několik let trvajícího výzkumu.

V roce 2001 byl zahájen výzkumný projekt financovaný americkou grantovou agenturou NSF (National Science Foundation). Účastní se jej například firma IBM či John Hopkins University v USA. Stranou nezůstali ani odborníci ze Západočeské univerzity v Plzni a Univerzity Karlovy v Praze.

Úkolem vědců je vytvořit do roku 2006 systém, který by uměl zvukový záznam přepsat do psané podoby, a v té pak vyhledávat. Na příkaz "najdi vše o doktoru Mengelem" by nalezl a přehrál ty pasáže, v nichž jméno "anděla smrti" zaznívá.

Ale nejen to. Učitelé, vědci, studenti či kdokoliv jiný by mohli pokládat i rafinovanější dotazy. Příklad uvádí docent Jan Hajič, ředitel Ústavu formální a aplikované lingvistiky MFF UK v Praze: Zájemci by si mohli vyžádat například svědectví o lidech, kteří se dostali do koncentračního tábora až ke konci války, protože utekli před transporty a několik let se skrývali.

Počítač v roli žáka

Zatímco američtí vědci učí počítače vyhledávat ve svědectvích anglicky hovořících pamětníků, čeští experti dostali na starost některé středoevropské jazyky: kromě češtiny také slovenštinu, polštinu a ruštinu. S rodným jazykem jsou prakticky hotovi, nyní se věnují ruštině.

"U nás na katedře máme studenty z Ruska, kteří nám pomáhají," říká profesor Josef Psutka, vedoucí katedry kybernetiky na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Aby mohl přepsat mluvenou řeč, musí si počítačový systém osvojit schopnost rozeznávat jednotlivá slova a také odlišovat bezvýznamné zvuky, například hluk projíždějícího auta, od

smysluplných sdělení. Účastníci projektu mu proto nejprve převádějí do psané formy stovky výpovědí, na kterých se může učit. Nesmějí vynechat ani rušivé momenty.

"Dohodli jsme se s IBM, že v každém jazyce přepíšeme čtvrthodinové záznamy hlasů čtyř set lidí," upřesňuje profesor Psutka. "Musí se zapsat přesně to, co zaznělo. Včetně hlasitých nádechů a různých zvuků zvenčí. Někde tikají hodiny, jinde mají papouška, který občas pískne - to vše musí být zaznamenáno. Je to mravenčí práce, trvá pro každý jazyk zhruba rok," dodává plzeňský expert.

Počítač pak bude tříbit své schopnosti tím, že srovná nahrávky s jejich přepisy (ty se pořizují ve speciálním programu). Vzorek 400 lidí by měl stačit k tomu, aby se systém vyrovnal s každým dalším hlasem, i když jej předtím nikdy neslyšel.

Neznamená to ovšem, že computer nedělá žádné chyby - úspěšnost přepisu je asi 60 až 65 procent. "Dvě z pěti slov jsou tedy zapsána špatně," vysvětluje docent Hajič. "K tomu, abyste přepsaný text mohl souvisle číst, by to nestačilo. Tři z pěti slov jsou ovšem zachycena správně, a to pro vyhledávání stačí."

Počítač zkrátka dokáže i v nepřesném zápisu najít, co potřebujete. Příslušnou pasáž vám potom přehraje, protože psaný záznam by vám kvůli chybným slovům moc nepomohl.

Problémy s češtinou

Čeští vědci mají proti svým americkým kolegům, kteří pracují s angličtinou, o poznání těžší úlohu. Naučit počítač "rozumět" češtině je obtížnější. "Je to obrovský rozdíl," zdůrazňuje profesor Psutka. "Při spontánním projevu je v češtině přibližně každé desáté slovo nespisovné. Často neřeknete 'mladý muž', ale 'mladej muž' a podobně."

Dalším problémem jsou například pádové koncovky. Američanům ovšem komplikuje život zase odlišný přízvuk lidí z různých oblastí anglofonního světa.

Systém pracující s češtinou dosahuje při přepisu spontánní řeči stejné úspěšnosti jako jeho americký "kolega" - zmíněných 60 až 65 procent.

Odborníci by chtěli nabídnout zájemcům o holocaust ještě jednu vymoženost. Lze si představit situaci, kdy na anglicky položený dotaz počítač nabídne pasáž z nahrávky pořízené třeba v polštině. V takovém okamžiku by měl nastoupit automatický překlad, který by tazateli zprostředkoval odpověď v anglickém jazyce.

Výsledky pětiletého výzkumného projektu mohou sloužit i k jiným účelům, mapování informací o holocaustu nemusí být jediným využitím. "Navštívil jsem plzeňský rozhlas, mají plné sklepy starých pásek a nevědí, co na nich je. Neexistuje nikdo, kdo by měl čas to poslouchat. Přitom tam budou určitě zajímavé pořady," věří profesor Psutka.

Pomoci by opět mohly počítače, které by dokázaly pořídit přepisy zvukových nahrávek, a v těch pak vyhledávat.

Čeští vědci si váží toho, že byli k americkému projektu přizváni. Svědčí to o tom, že počítačová lingvistika je v ČR na špičkové úrovni. Budoucnost oboru přesto nelze vylíčit v růžových barvách. Nenajde-li se ve státním rozpočtu dost peněz na financování současných projektů, může se stát, že se výzkumné týmy rozpadnou a mladí odborníci rozprchnou po světě.

Další informace o možnostech počítačové lingvistiky na straně 22

O autorovi| redaktor LN

Foto popis| Během natáčení filmu Schindlerův seznam za režisérem Spielbergem chodili pamětníci holocaustu a vyprávěli mu své vzpomínky. Režisér založil nadaci, která se stará o to, aby tato cenná svědectví nezmizela ze světa. Nadace pořídila přes 100 000 hodin videozáznamů vzpomínek. Jak se v takovém oceánu slov vyznat? Pomoci musí výpočetní technika.

Foto popis| V natočených svědectvích pamětníků holocaustu bude možné vyhledávat informace pomocí počítače. Projektu se účastní i vědci z Univerzity Karlovy a Západočeské univerzity, kteří spolupracují s v Centru komputační lingvistiky MMF UK

Foto popis| Oskar Schindler (druhý zleva) na jednom ze svých proslulých večírků. Originál seznamu s více než 1200 polských Židů, které zachránil před jistou smrtí byl nalezen koncem devadesátých let, po uvedení Spielbergova filmu

Foto popis| Ti, kteří přežili: to nejsou filmoví hrdinové..Plakát filmu Schindlerův seznam, Steven Spielberg

Foto autor| FOTO: LN-NYNEK GLOS, ARCHIV // KOLÁŽ ŠIMON / LN