

## Počítače se učí lépe překládat

3.6.2006 Lidové noviny str. 38 Věda

Eva Vlčková

Revoluci v automatických překladech z cizích jazyků slibují počítačové programy nové generace.

Zkoušeli jste někdy přeložit cizojazyčný text pomocí dostupných počítačových překladačů? Pokud ano, asi jste zažili rozčarování. Výsledkem je většinou podivná změt' slov, která dohromady nedává smysl.

Schopnosti těchto programů nejlépe otestujete zkušebním překladem z češtiny do jiného jazyka a zpátky. Například rčení „sejde z očí, sejde z mysli“ (anglicky „out of sight, out of mind“) se vám může vrátit jako „slepý idiot“.

Vědci ale slibují, že zanedlouho budou počítačové překlady téměř dokonalé. Podle článku v časopise New Scientist nás v tomto směru čeká přímo revoluce. Dnešní komerčně využívané programy totiž vycházejí jen ze slovníku a základních gramatických pravidel, která jim byla předem zadána. Vícevýznamová slova, nespisovné výrazy a neznalost kontextu jsou pro ně nepřekonatelnou překážkou.

### Revoluční kombinace

Nová generace překladačů má fungovat na zcela jiném principu: odborníci nejprve počítač „nakrmí“ velkým množstvím textu a on si v nich sám najde pravidla. Nikdo je už nemusí ručně vkládat, což byvá pracné a ve výsledku neefektivní.

„Tyto programy pracují na základě statistiky. V lingvistice se začala používat už koncem 80. let, ale teprve vývoj výkonných počítačů umožnil její širší využití v praxi,“ vysvětluje docent Jan Hajič, ředitel Ústavu formální a aplikované lingvistiky MFF UK v Praze.

V posledních letech se statistika začala používat i pro vývoj „samoučících“ překladačů. Do počítače se zadá text jako vstup a jeho překlad v jiném jazyce jako vzorový výstup. Program na základě statistických metod vyhodnotí, jak byla která formulace přeložena, a zapamatuje si nejčastější řešení. Všimá si například i toho, která slova se objevují často poblíž jiných. Dokáže tedy odhadovat výběr správného slova podle kontextu. Když pak zadáme text pouze v jednom jazyce, počítač ho sám podle získaných znalostí přeloží. „Statistika je natolik pružná, že si do jisté míry poradí i s překlepy, nespisovnými výrazy nebo nářečím,“ podotýká docent Hajič.

Použití takového překladače připravuje například americký Google. „Ten může z principu svého fungování „nakrmit“ počítač v podstatě celým internetem, tedy texty, které se objevují na webových stránkách. Proto lze očekávat, že jejich překlady budou velice kvalitní,“ říká český vědec.

Zároveň ale upozorňuje, že sama statistika - tedy „nasypaní“ vstupů a výstupů do počítače při výrobě kvalitních překladačů nestačí. „Pořád je zapotřebí i práce lingvistů, kteří definují, čeho si má počítač při svém učení hlavně všimnout. Budoucnost vidím v kombinaci statistiky s lingvistikou,“ uvádí docent Hajič.

Optimisté by se rádi dočkali i kapesního tlumočnicka ve stylu „babylonské rybky“ ze Stopařova průvodce po galaxii. Přijedete do ciziny, oslovíte kolemjdoucí a můžete s nimi plynule konverzovat, protože vám i jim miniaturní krabička všechno pohotově přeloží.

Na to si ale podle docenta Hajiče ještě počkáme. Americká armáda sice v Iráku a Afghánistánu už podobná zařízení zvaná TransTec testuje, ta ale umí pracovat jen s několika desítkami nahraných frází. „Fungují podobně jako hlasové ovládání mobilního telefonu. Něco řeknete a přístroj hledá, zda se zvuk podobá některé z uložených vět,“ říká docent Hajič. Skutečné překladače by musely mluvenou řeč nejprve převést na text, ten přeložit a pak z něj opět vytvořit řeč.

Programy na převádění řeči do textu a zejména směrem zpátky už existují na docela slušné úrovni. „Například kolegové z Plzně vyvinuli automatické titulkování hokeje pro neslyšící,“ uvádí docent Hajič. Podobné programy se mohou uplatnit pro přepisování dalších pořadů a mluvené řeči obecně.

Hlavní problém miniaturních tlumočnicků je podle Hajiče jinde - spočívá v kapacitě počítačů. „Musejí statisticky zpracovat obrovské množství textu. Na to zatím nestačí ani stolní počítače, které máme doma, natož kapesní PDA,“ říká český odborník. Takže zatímco „chytrých překladačů“ internetových stránek se můžeme dočkat za několik let, k malým pomocníkům do kapsy povede mnohem delší cesta.

V první fázi by se překladače mluveného slova mohly uplatnit na konferencích a podobných akcích, kde řečníci mluví spisovně, zřetelně a často v zaběhlých frázích. Za tímto účelem vyvíjejí němečtí vědci z univerzity v Karlsruhe program zvaný „Lecture Translator“, tedy překladač přednášek.

Učí ho tlumočit pomocí projevů z Evropského parlamentu. Využití pro takové příležitosti by mohlo být snazší než u krabičky do kapsy, protože v zákulisí může stát výkonný počítač, z něhož by se překlad vysílal do sluchátek posluchačů.

Nikdo není dokonalý

Jazykovědci upozorňují, že navzdory vši snaze nebudou automatické překlady nikdy dokonalé. Často je ke správnému porozumění textu nutné znát širší kontext. „Asi se tímto způsobem nebude překládat Shakespeare nebo poezie, ale o to ani nejde. Účelem je především věcně správný překlad,“ říká docent Hajič.

Z podobného důvodu se podle něj neuplatní ani automatizované dabování filmů, jak optimisticky slibuje New Scientist. Občas se mohou objevit kostrbaté formulace, které umožní porozumět, o čem si lidé povídají, ale divák by rušily. Kromě toho umělý počítačový hlas zní zatím přece jen poněkud nepřirozeně. Za přijatelnější považuje český odborník zůstat pro překlady filmů a televizních pořadů u psaných titulků. „Nečekám ani, že se budou plně automaticky, bez lidského zásahu, překládat technické manuály letadel. To by bylo riskantní. Ale třeba pro čtení zahraničních zpráv na internetu budou počítačové programy ideální,“ dodává odborník.

Velké světové firmy mají z obchodních důvodů zájem především o „velké“ jazyky, jako je angličtina, španělština, čínština nebo francouzština. „Například překlady mezi francouzštinou a angličtinou jsou už poměrně kvalitní. Vůbec první pokusy se statistikou v překladu se totiž prováděly se záznamy z kanadského parlamentu, které se povinně archivují v těchto dvou jazycích,“ říká docent Hajič. „S lingvistickým výzkumem češtiny jsme u nás hodně daleko, ne-li dál, než je tomu pro některé z uvedených jazyků, ale byznys je byznys: Česko je malý trh.“

Čeština je navíc v mnoha ohledech složitější. „Například v angličtině lze každé slovo zařadit do zhruba 70 mluvnických kategorií, v češtině do 4500. U statistických metod sice můžeme částečně využít zkušenosti ze zahraničí, což je výhoda proti starším metodám překladu, ale musíme je upravit pro naše potřeby,“ dodává docent Hajič.

Záhadné písmo odolává

Počítačové programy kombinující lingvistiku se statistikou lze využít nejen pro překlady. Dokáží hledat i zákonitosti v neznámých jazycích a písmech. Spočítají například výskyt různých znaků či slov.

Odborníci takto zkoumali i tzv. Voynichův rukopis, knihu z 15. nebo 16. století psanou neznámým písmem. Zabývala se jí už řada kryptologů, ale také studenti katedry počítačů na Johns Hopkins University v Baltimore ve státě Maryland pod vedením docenta Hajiče.

„Písmo zatím nikdo nerozluštil, ale počítačové analýzy ukázaly, že má strukturu, není to jen shluk náhodných znaků,“ uvádí docent Hajič. Studenti porovnávali různé vlastnosti textu s dvanácti jazyky, s žádným z nich se ale neshoduje. Padla i teorie, že by mohlo jít o starou verzi ukrajinštiny bez samohlásek. „Stále není jisté, zda jde o zakódovaný text, nebo o úplně vymyšlený jazyk,“ dodává český odborník.

Podobně zahraniční vědci zkoumají další jevy a hledají v nich pravidla - například náznaky větné skladby ve zvucích delfínů či velryb. „Statistické metody se dají použít na analýzu a modelování nejrůznějších humanitních, technických i přírodních procesů - včetně předpovídání počasí,“ uzavírá docent Hajič.

Hlavní problém tlumočnicků do kapsy spočívá v jejich kapacitě

\*\*\*

Jak zkrotit cizí jazyky

Kvalita dostupných počítačových překladačů se má v blízké budoucnosti výrazně zlepšit. Příčinou je využití statistických metod při tvorbě překládacích programů. Stejně metody vědci využívají i ke zkoumání neznámých jazyků.

Voynichův rukopis

Knihu neznámého autora psanou záhadným písmem si prý kdysi koupil císař Rudolf II. za 600 dukátů. Pravděpodobně šlo o podvrh, jímž kdosi důvěřivého císaře oklamal a vymámil z něj peníze s tvrzením, že má v ruce tajnou alchymistickou příručku. Později se kniha ztratila, až ji roku 1912 objevil americký sběratel a obchodník s knihami Wilfrid Voynich. Dnes je uložena v knihovně Yaleské univerzity, která ji naskenovala a zpřístupnila na internetu.

Foto popis| Automatické překladače budou za pár let pracovat na slušné úrovni, ale tlumočníci o svou práci hned tak nepřijdou. Na malé tlumočnické do kapsy si musíme ještě počkat. Jejich vývoj brzdí malá kapacita kapesních počítačů.

Foto autor| UNIVERSITY, FOTO: YALE ARCHIV // ŠIMON / LN