Čeští vědci porazili Google

HOSPODÁŘSKÉ NOVINY

11.10.2013, Rubrika: Téma, Strana: 4, Téma: Univerzita Karlova, Michal Kalina redaktor

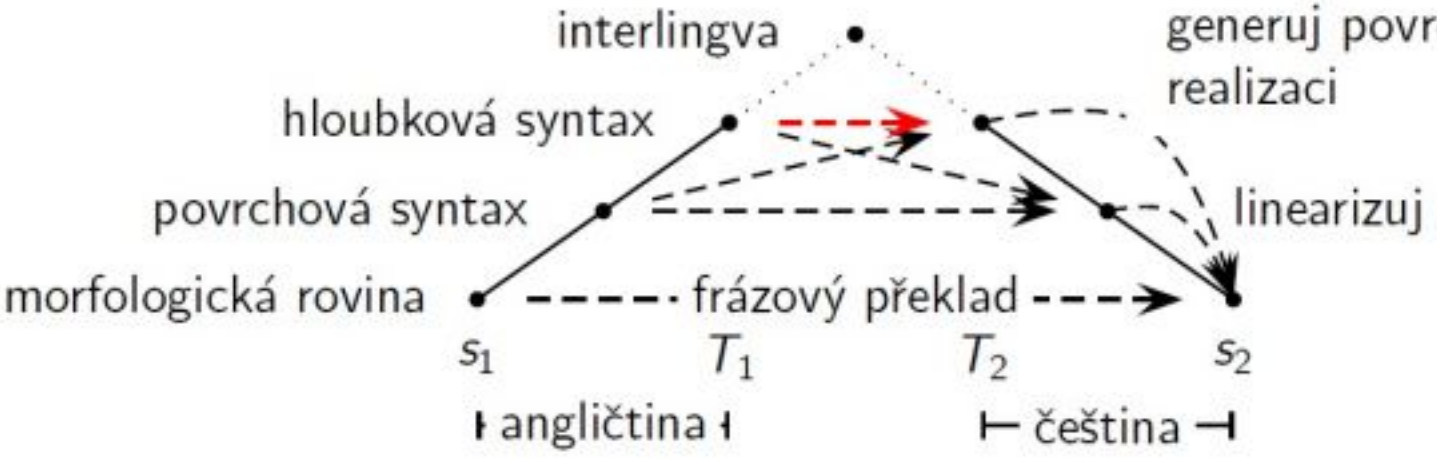
V každoroční překladové soutěži Workshop on Statistical Machine Translation , kterou podporuje Evropská komise, se snaží mezinárodní týmy vylepšit stávající vědecký systém Moses (Mojžíš) či jemu podobné. Pro srovnání jsou zastoupeny i komerční překladače, jako je třeba ten od Googlu.

Letos se ho poprvé při překladu z angličtiny do češtiny povedlo překonat týmu vědců z Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, čímž mimo jiné ukázali, kudy by se komerční produkty mohly ubírat. "Počítače textům samozřejmě nerozumějí, počítače jen simulují porozumění a my se jim snažíme vysvětlit, jak to mají dělat," říká člen týmu Ondřej Bojar z Ústavu formální a aplikované lingvistiky .

Moses vznikl kolem roku 2006 a jde o open-source překladový systém, nebo spíš jeho prototyp. Systém už používají různé evropské firmy, které ho zároveň zdokonalují. Jak tvrdí Bojar, není ale jednoduché se s ním naučit pracovat, a proto se zatím nedočkal většího rozšíření mezi běžné uživatele.

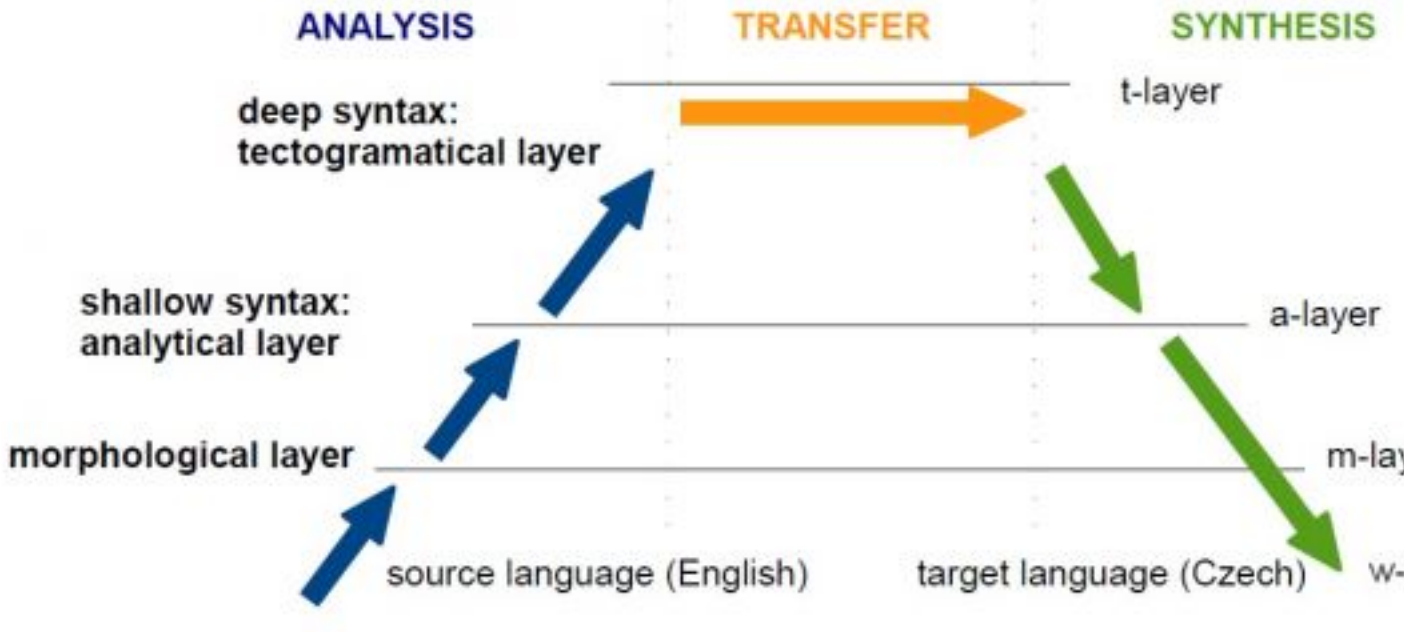
Velcí internetoví hráči, ať už je to právě Google nebo Microsoft, nabízejí zatím nepříliš přesné překladače už několik let. Historie strojového překládání však sahá až do 50. let 20. století, kdy takový systém předvedla poprvé jiná americká společnost – IBM. Tehdy šlo ale o primitivní variantu, prudkého rozvoje se překladače dočkaly až v 90. letech.

Přístupy ke strojovému překladu



Odborníci z Matfyzu se na letošním workshopu rozhodli zkombinovat rovnou dva překladové systémy dohromady a díky tomu uspěli. Nejprve zkušební text vložili do systému TectoMT, který vyvinuli jejich kolegové. TectoMT využívá syntaktický (hloubkový) překlad a převádí anglický větný rozbor na český.

TectoMT: Hloubkový překlad



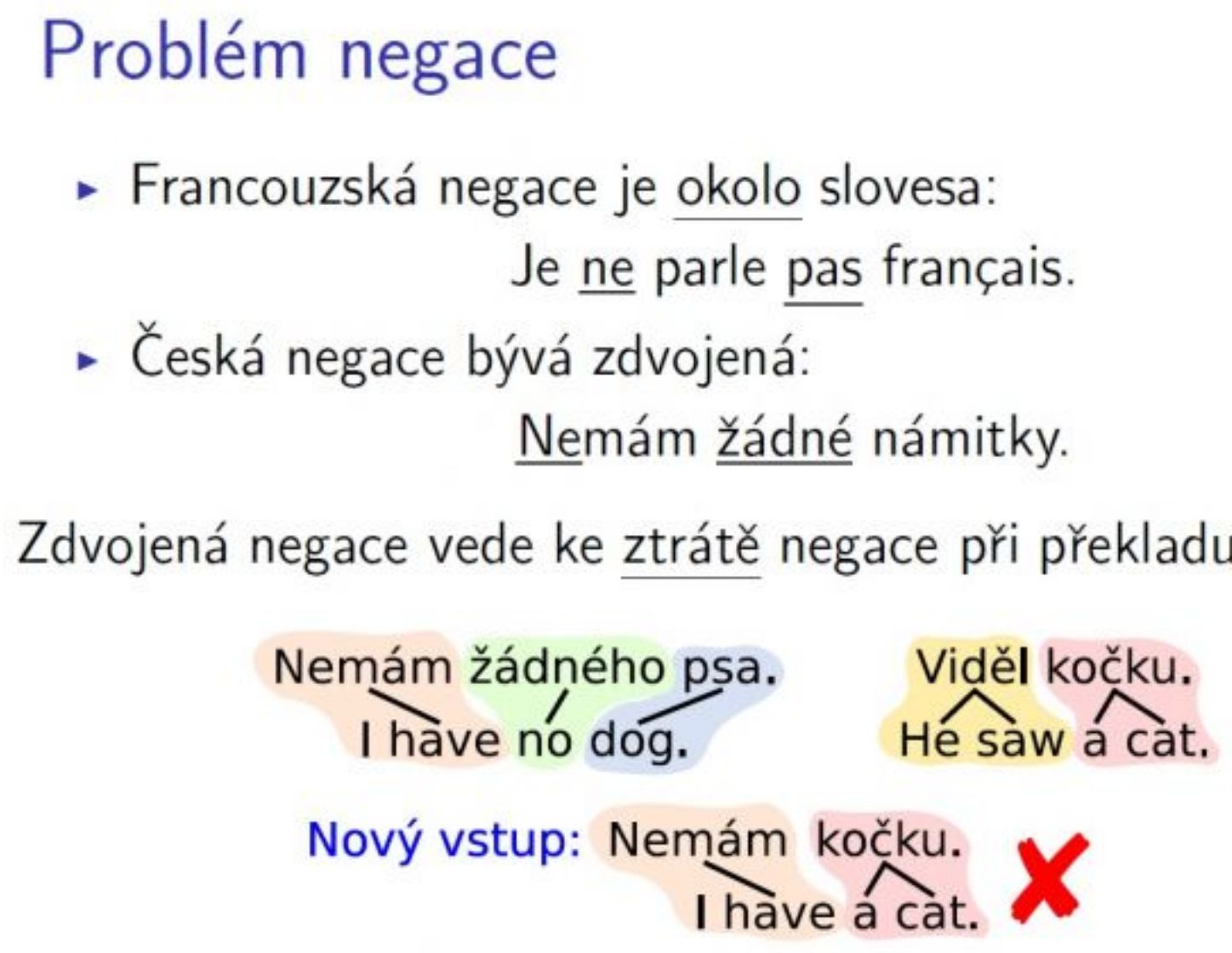
Věty z TectoMT jsou gramaticky správně, shoduje se třeba podmět s přísudkem, ale větám někdy chybí důležitá slova. Proto Češi rozbor prvního systému použili coby vstupní data pro druhý systém – Moses, který překládá na základě naučených frází.

"Je to jako bilingvní kniha, kdy má překládací systém zásobu vět ve dvou jazycích a z nich si vybírá. Zkrátka jde o velkou zásobu dat a počítač se z toho naučí, které posloupnosti anglického jazyka odpovídají posloupnosti českých slov. Spoléhá se přitom na co největší počet i tvarovou pestrost výskytu slov v trénovacích datech," vysvětluje Bojar, jak pracuje.

Pomohl odstraňovač chyb

Nevýhodou frázového systému je, že není schopný samostatně vytvořit jiné tvary slov než ty, které již zná. Navíc nedodržuje gramatiku. Dohromady ale oba systémy poskytují lepší výsledky.

Aby byl výsledek ještě lepší, prošel výsledný text ještě automatickou korekturou pomocí českého systému Depfix, odstraňovače chyb, jenž opravil například špatně přeložené negativní věty a pády.



Podle Bojara Google používá pro překlady několik způsobů, češtinu ale nejspíše zpracovává také frázovým systémem, podobným vědeckému Mosesu, který má dvě složky – překladový model, jenž umí přeložit určité slovo nebo sousloví, a jazykový model, který hlídá slovosled a plynulost výstupu. Veliké množství dat však nestačí na tak kvalitní překlad, k němuž došli čeští vědci.

Na komerční využití zatím nepomýšlí, protože spojení systémů zatím není příliš praktické. Jejich interaktivní použití například na internetu, jak uvádí Bojar, ale možné je a zájemci se již hlásí. Tým z Matfyzu zatím pracuje na dalších vylepšeních, jako je řešení některých chybných výstupů, například zmíněných negativních vět, přímo v rámci systémů.