Evropský den jazyků, 26. září

QTLeap projekt s hybridní technologií pro lepší výsledky strojového překladu

Shrnutí:

Jazyky nejsou pouze řetězce slov

Online systémy strojového překladu se učí jazyky zpracováváním velkého množství vícejazyčných textů metodami statistického učení, tedy propočítáváním pravděpodobností jednotlivých posloupností slov. Jazyky ale nejsou pouze řetězce slov, jsou strukturované gramaticky. Technologická laboratoř jazyků Německého výzkumného centra pro umělou inteligenci (DFKI) proto zahájila vývoj hybridní technologie překladu, která rozšiřuje statistické systémy o jazykové a sémantické znalosti načerpané z internetu, aby dosáhla při překladu lepších výsledků.

Do Evropy v posledních dnech dorazily tisíce uprchlíků. Aby se dokázali začlenit do společnosti, co nejrychleji je potřebujeme naučit jazyky zemí, do kterých přicestovali. Tento nesnadný a rozsáhlý úkol nám zároveň připomíná, že ani evropská integrace ještě není zcela dokončena, zůstaly nám tu právě ty hranice, které je nejtěžší odstranit: jazykové bariéry dusící náš obchod, komunikaci a skutečnou integraci.

Učení se jazykům není jednoduché. Děti se jazyky, ať již mateřské nebo cizí, učí hrou nebo poslechem dospělých, zatímco pro dospělé je učení se gramatických pravidel a slovní zásoby spojeno s velkým úsilím.

Počítačové programy, jako jsou známé systémy online strojového překladu, se učí jazyk spíše dětským způsobem. Používají tzv. metody strojového učení na velkém množství bilingvního (tj. přeloženého) textu ('Big Data') k učení statistických pravděpodobností o tom, jak jsou vybraná slova a řetězce slov přeloženy. Tyto systémy pak produkují přibližný překlad, jehož kvalita se velmi liší v závislosti na vstupních veličinách.

Ale jazyk, to není jen řetězec slov. Spíše se jedná o gramatické struktury, s prvky jako jsou přísudky, podměty a předměty. Existují formy jazykově motivovaného strojového překladu, který používá gramatickou informaci, ale pro všeobecné účely jsou spíše nepružné a tím pádem najdou uplatnění v omezeném rozsahu, hlavně pro blízké jazyky. Např. španělský deník *La Vanguardia* je každou noc překládán do katalánštiny za použití tohoto systému, a to s minimálními lidskými opravami.

Internetové zdroje jako jsou strojově čitelné varianty Wikipedie obsahují tzv. sémantický web a dodávají znalosti o sémantice (významu slov). Tyto zdroje zakódovávají strukturované znalosti o světě, např. "vědí", že Paříž je hlavní město Francie, ale také to, že existuje město Paříž v Texasu.

Technologická laboratoř jazyků Německého výzkumného centra pro umělou inteligenci (DFKI) pracující na projektu EU jménem QTLeap (http://qtleap.eu) vyvinula hybridní

technologii pro strojový překlad, která obohacuje statistické metody o jazykovou znalost a sémantickou informaci z internetu pro vylepšení výsledků překladu.

První prototypy zaznamenaly příslib, ale "jsme teprve na samém začátku v tomto výzkumném směru," říká prof. Hans Uszkoreit, vedoucí Technologické laboratoře jazyků. "Zjistili jsme, že vyhodnotit kvalitu překladu je mnohem těžší než strojový překlad sám. Když se zeptáte několika lidských překladatelů, jejich hodnocení se enormně liší. Bohužel je v dnešních dnech skoro nemožné měřit kvalitu překladu automaticky. To je hlavní výzvou pro celé odvětví, ve kterém intenzivně pracujeme v DFKI."

Jako se dnes naše děti diví, jak jsme před 20 lety zvládli dělat domácí úkoly bez Wikipedie, možná se naše pravnoučata budou divit, jak v roce 2015 zvládli všichni ti, co přišli do Evropy za lepším životem, s námi komunikovat před vynalezením *vše-překladače.*

Pro vice informací a kontakt: http://qtleap.eu nebo hajic@ufal.mff.cuni.cz.