**Strojový překlad využívající umělou inteligenci**

1. **Historie a vývoj**

První idea vytvoření strojového překladu byla vyřčena již na **počátku** **50. let** v USA. První vlnu zájmu vyvolal **Warren Weaver**, když navrhl využít počítače také k tomuto praktickém využití a nastínil problémy, se kterými se obor bude potýkat. Právě skutečnost, že v té době již byly dostupné samočinné počítače používané za války ke kryptografickým účelům, inspirovala informatiku ke vzniku tohoto oboru. Američany hlavně zajímal levný a rychlý způsob překladu rusky psaných textů v kontextu studené války. Práce překladatelů totiž nebyla snadno dostupná kdykoliv a kdekoliv.

Prvním pokusem o realizaci byl **Georgetown-IBM experiment**, na němž bylo strojově přeloženo asi 60 ruských vět do angličtiny. Překládal texty slovo od slova a případně upravoval slovosled pomocí jednoduchých naprogramovaných pravidel – tomuto přístupu se říká **přímý překlad** *(direct translation).* Výstup byl velmi kostrbatý, nicméně experiment přilákal zájem široké veřejnosti a především finanční podporu vlády. Kolektiv složený z odborníků IBM a Georgtown University i další týmy tak mohly pokračovat ve vývoji reálně využitelného programu. Práce však neběžela tak, jak předpokládali, a v roce 1966 byla vydána zpráva **ALPAC**, která zhodnotila stav systémů vyvinutých pro strojový překlad a dala na vědomí, že nejsou pro praxi použitelné. Úspěch prvotního experimentu byl totiž dán tím, že se překládaly jen velmi jednoduché věty, použilo se pár pravidel pro překlad a omezený slovník. V USA to pro většinu projektů znamenalo momentální konec rozvoje v oboru až téměř do 80. let, nicméně některé části tohoto projektu získaly jiné využití.

**70. léta**

V Evropě, Rusku a také Japonsku vývoj strojového překladu pokračoval bez přerušení a v Kanadě se podařilo vytvořit program pro překlad předpovědí počasí – **Meteo**. Základem úspěchu byly jednoduché věty, omezená gramatika a slovník – opakovaly se stále ty samé fráze). Obsahoval už ale i náznaky morfologické a syntaktické analýzy. Tento přístup se nazývá **pravidlový** *(rule-based).*

Velmi populárním se pak stal také systém **Systran**, který navázal v USA na předchozí práci a jeho výsledky se využívaly v projektech Apollo, Sojuz a především v Evropském společenství pro překlady anglických textů. Ve svých začátcích ho dokonce používal i Google Translator.

**80. léta**

Pravidlové (či transferové) systémy se dále zlepšovaly a kromě morfologické a syntaktické analýzy se začala využívat i sémantická. Protože už se nedařilo nijak výrazně zlepšovat výsledky strojového překladu založeného na znalosti jazyka, začaly se využívat první překlady založené na **statistických metodách**. Toto bylo podmíněno existencí velkých paralelních vícejazyčných jazykových korpusů.

**90. léta**

S rozmachem levných osobních počítačů v 80. letech se začaly mezi uživateli šířit jednoduché překladové systémy (také prozatím založené na znalosti jazyka). Korpusové a statistické systémy pak bylo možné s příchodem internetu zpřístupnit všem.

1. **Klasifikace způsobů strojového překladu**

* **Pravidlový (znalostní) strojový překlad *(knowledge-based)***

Funguje na základě analýzy textu ve zdrojovém jazyce v několika úrovních. Pokud se text překládá přímo, analýza se téměř neprovádí a pouze se hledají příslušné tvary ve slovníku.

**a) transferové** – obsahuje analýzu zdrojového textu a generátor textu v cílovém jazyce. Analýza může být synt., sém., morf. nebo žádná. Vždy je ale tento systém určen pro dvojici jazyků – každý jazyk má jiné principy analýzy.

**b) interligua** – překlad do reprezentace nezávislé na jazyce, typicky v logickém formalismu. Nemá pevnou podobu, ale musí zachycovat i skutečnosti, které z textu vyplývají (tj. má nějaké znalosti světa). Reprezentace je tudíž jednoznačná.

* **Statistický strojový překlad**

Pracuje s informacemi naučenými z textu. Neprobíhá zde žádná lingvistická analýza, ale běží zde statistický postup – text se rozseká na slova nebo skupiny slov, které k sobě nemusí nutně nijak gramaticky pasovat **(n-gramy)**. Pro překlad nutně potřebuje paralelní korpus, kdy používá techniky strojového učení k vytvoření databáze překladových ekvivalentů spolu s pravděpodobnostmi jejich výskytů (výskyt může být podmíněn kontextem). Překladový model si (např. v případech víceznačnosti) zapamatuje možné interpretace daného slova nebo fráze a pracuje s tím i nadále. Frekventovanější případ poté použije a ber v potaz také kontext. Po překladovém modelu následuje jazykový model, kt. vyhlazuje poměrně hrubý výstup. Je také statistický a pravděpodobnostní a v jeho kompetenci je např. úprava slovosledu. Tím, že má zapamatováno v jakém kontextu se slovo vyskytovalo v jazyce, do kterého se překládá, dokáže celou větu ,,učesat“. Pamatuje si kontext nejen slovní, ale i lemmata morf. značek a můžeme říct, že existuje bigram, ve kt. je předložka a vl. jméno. a na základě porovnávání korpusů zjistí, jak se dá daný řetězec slov přeložit a na základě této informace dokáže překládat.

* **Hybridní strojový překlad**

Jde o kombinaci obou zmíněných přístupů. Program dostane vstup ze statistického způsobu, využije pravidla z pravidlového a kterýkoliv z nich se může použít jako primární překladový systém, jehož výsledky jsou poté vylepšovány tím druhým.

1. **Přístup ke strojovému překladu z hlediska umělé inteligence**

Využitím umělé inteligence ve strojových překladech se výzkumníci začali zabývat již v 70. letech. Hlavním tématem jejich tehdejší práce byl překlad z hlediska sémantiky, kdy se například snažili analyzovat typické situace (nakupování, návštěva lékaře) pomocí ,,scénářů“ či rámců. Nyní je jednou ze zajímavých novinek v této oblasti vznik samostatně se učícího inteligentního agenta na Bristolské univerzitě – systému, který umí sám zlepšovat úspěšnost svého statistického strojového překladu. Lze říci, že zde vědcům nešlo primárně o strojový překlad, nýbrž o vyvinutí hodnotící funkce, která je odolná vůči nepříliš kvalitnímu vstupu a dovede poučit sama sebe na základě vnějšího světa. Zajímavé na jejich experimentu je to, že se odvažují pracovat s texty v surové podobě, kde nikdo nevyznačil, která věta patří ke které ve druhém jazyce, tedy data nejsou nikým tzv. zarovnána. A skutečně se kvalita práce agenta s každým dalším natrénováním zvyšovala.

**Využívané zdroje dat**

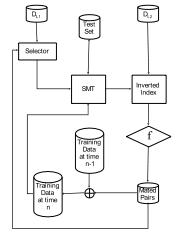
Agent samostatně čerpá data ze dvou pramenů, a to z:

1. **Korpusu**
2. **Webu**

**a) Korpus**

Agent má k dispozici paralelní korpus **Europarl**, kde jsou zápisy z Evropského parlamentu povinně ve všech jazycích. Nevýhodou je, že se aktualizuje jen jednou za čas, ale zato je poměrně obrovský – obsahuje přes milion párů vět. Pro své účely agent využívá španělštinu a angličtinu. Z každého jazyka vezme skupinu vět, každou španělskou přeloží do angličtiny a spočítá si, s jakou přesností se „trefil“ (porovná překlad s originálním anglickým zněním). Věty, které přeložil nejlépe, použije k opětovnému ,,natrénování“ (zaznamená, že dané věty ve španělštině mají tuto podobu v angličtině) a přeloží další soubor jiných. Tímto způsobem se stále zlepšuje, dokud to jde. Dále by si měl sám shánět data (texty v obou výše zmíněných jazycích), ze kterých se bude nadále učit.

**b) Web**

Agent si dokáže sám stahovat z různých elektronických médií vícejazyčný obsah, zejména titulky a perexy (= krátké texty, jejichž účelem je uvést článek a upoutat čtenářovu pozornost). Vstup nemusí být nutně překladem z jednoho jazyka do druhého, tedy není zcela paralelní, zato je **porovnatelný**, tzn. vyjadřuje stejnou informaci. Navíc se zde agent musí stále přizpůsobovat nové slovní zásobě a obratům – zprávy se stále aktualizují, píší je různí redaktoři, tematika je široká, …

Obr. 1: Schematický nákres agenta

1. **Jak inteligentní agent pracuje?**

Schopností tohoto autonomního systému je vylepšovat sám sebe tréninkem na zarovnaném paralelním korpusu – umí sám nacházet další a další data, ta znovu využít k natrénování a takto se stále zlepšovat. Úspěšnost agenta se posuzuje jednak z pohledu **kvality strojového překladu**, jednak z pohledu **spárování přeložených a originálních vět** díky technikám získávání informací z vícejazyčných textů.

Kvalita strojového překladu se měří metrikou Bleu, která porovnává přeloženou a jí odpovídající nalezenou (vzorovou) větu – obě anglické – a určuje jejich podobnost porovnáváním n-gramů. Při spárování vět jde o obor **Cross-Language Information Retrieval** a proces probíhá tak, že se slova z věty přeložené do angličtiny hledají ve všech anglických originálech. Ta věta, která má největší shodu a dostatečný odstup od ostatních, je prohlášena za pojítko mezi původní španělskou a nalezenou anglickou. Samotné svázání vět v obou jazycích se provádí prostřednictvím překladu španělských vět do angličtiny. Tyto překlady se porovnají s anglickými vzory a spočítá se jejich kvalita a blízkost.

Agent se zlepšuje i v případě, že byla trénovací data záměrně zkreslena (zašuměna) tak, že jistá část vět byla ze vzorku odstraněna a nahrazena náhodně vybranými větami z jiných vzorků. Výzkumníci použili velice vyspělý, volně dostupný statistický překladový systém **Moses**, jenž překládá nikoli slovo po slovu, ale celé „fráze“ (n-gramy).

**Podrobný popis činnosti**

Po počátečním natrénování statistického překladového modelu se na základě korpusu anglických vět vytvoří rozsáhlý inverzní index slov, díky němuž je možné rychle vyhledávat a takto určovat podobnost originálních anglických a vět s překlady vytvořenými ze španělštiny. Z korpusu španělských vět se vybere první vzorek vět (rozumné velikosti) a přeloží se už základně natrénovaným systémem Moses.

Každá přeložená věta se nyní vyhledává v korpusu anglických vět také volně dostupným systémem **Lucene**, který vrací nejen výsledné věty, ale i jejich blízkost (podobnost) k překladu, pomocí něhož se hledalo. Do tohoto „skóre“ se zahrnují faktory jako je **závažnost klíčových slov**, která se hledala (pokud se slovo vyskytuje ve velkém množství vět, nebude tak důležité), **počet slov**, ukterá se celkově našla (některé výrazy mohly být přeloženy jinak a nemusely se pak vyskytnout v žádném originálu) a **závažnost nalezeného slova** v dané větě.

Nejlépe hodnocená původní anglická věta, která je takto nalezena, se předloží hodnotící funkci, která je jádrem inteligentního agenta. Rozhoduje, zda se skrze přeloženou „kandidátskou“ větu původní španělská věta podobá nalezené anglické. Toto rozhodování bere v potaz podobnost překladu a anglického vzoru, pamatuje si také to, jestli už náhodou ke vzorové větě nebyl podobný jiný překlad a porovnává se také délka španělského originálu a jeho anglického protějšku, který byl nalezen díky překladu. Když jsou délky podobné, pak teprve se věty spárují.

Obě dvě míry jsou důležité při určování kvality hledání anglického vzoru přeložené věty. „Přesnost“ vyjadřuje, jak velká část z nalezených vět je správně, tedy jaký podíl anglických vzorů z těch najitých skutečně může odpovídat hledané přeložené větě. I kdyby bylo správných překladů mnoho, našlo se jich jen pár, ale z tohoto malého množství byly všechny správně, je přesnost 100 %. Pokrytí, na rozdíl od přesnosti, vyjadřuje poměr počtu správných nalezených vět proti počtu všech vět, které měly být nalezeny.

Při návrhu hodnotící funkce autoři upřednostnili přesnost před pokrytím, protože pro zlepšování kvality trénovacích dat je důležitější jejich správnost, nikoli velikost – zvýšená chybovost by zpomalila vývoj, přílišná chybovost (experimentálně zjištěná asi 60–70 %) by proces prakticky zastavila.

Pro zvyšování přesnosti je určován správný „odstup“ věty, jež je nejlepším kandidátem na spárování s jejím cizojazyčným protějškem, od druhé nejlepší. Byl zvolen takový poměr, který zachovává poměrně slušné pokrytí (pomocí grafu, který přesnost a pokrytí dává do vztahu). Dále uplatňují zásadu dávat do vztahu vždy jen jednu a jednu větu z obou korpusů, nechtějí mít pro jednu větu více protějšků, které mohou být jejím překladem.

Stojí za zmínku, že ačkoli u porovnatelných textů je nesmysl použít absolutní míru Bleu skóre k určení kvality překladu, přesto se s přibývajícími daty spíše zlepšovalo, pokud se použilo srovnání výsledků překladu novinových textů pomocí překladového modelu trénovaného na textech z těchto novin – proti modelu, který se trénoval pouze na náhodných textech z Europarlu.

Cílem tedy bylo navrhnout systém, který se autonomně dokáže přizpůsobovat změnám, takto například agentovi nedělá problém postupně se adaptovat změnám v jazyce způsobeným například u novinových zpráv změnou stylu nebo tématu (když se začne sledovat jiná „kauza“). Zároveň dokáže zlepšovat svou kvalitu tak, že si sám vybere trénovací data, poučit se z nich a opatřit si díky nim ještě lepší data.

1. **Příklad algoritmu inteligentního agenta**

****

**Vysvětlení:**

Zdroje: DL1 (španělský) a DL2 (anglický) korpus, SMT *(Statistical Machine Translation)* učený na paralelním korpusu, míra odstupu *α* , funkce *f* (rozhoduje, jestli k sobě věty patří).

1: Z anglických vět vytvoříme inverzní index (aby se v nich dalo rychle hledat).

2: Dokud máme sady španělských ,,trénovacích vět“, tak…

3: …vezmeme vzorek náhodně vybraných španělských vět.

4: Všechny z nich…

5: …přeložíme.

6: Podíváme se do indexu, vyhledáme v něm anglické věty, které se podobají překladu.

7: Pokud funkce *f* rozhodne, že nejlepší z přeložených vět je dostatečně (pomocí *α*) lepší než druhá nejlepší, tak…

8: …přidá do budoucí trénovací sady dvojici původní španělská věta; a nejlépe pasující anglická věta k té přeložené).

11: K dříve získaným trénovacím datům přidá všechny nalezené dvojice spárovaných vět.

12: Natrénuje překladový systém s těmito daty.

13: Změří úspěšnost tohoto systému (podle skóre Bleu, počtu slov, jež se nepřeložila, a přesnosti a pokrytí při párování vět).

14: Odstraní vzorek španělských vět (který byl právě použit) z trénovacího korpusu, z něhož se bude čerpat v příští iteraci.

1. **Použité zdroje**

*History of machine translation* [online]. Dostupné z: http://en.wikipedia.org/wiki/History\_of\_machine\_translation.

Marco Turchi, Tijl De Bie, Nello Cristianini. *An Intelligent Agent that Autonomously Learns how to Translate* [online]. Bristol: University of Bristol (UK). Dostupné z: https://patterns.enm.bris.ac.uk/files/Turchi\_DeBie\_Cristianini\_wi\_iat\_0.pdf.

Vít Baisa. *Studijní materiál PLIN019 Strojový překlad*. Brno: Masarykovy univerzita.