Úvod do strojového překladu – PV061

Vít Baisa, Karel Pala

vit.baisa@gmail.com

pala@fi.muni.cz

Centrum ZPJ FI MU

Historie strojového překladu

- C. Shannon, W. Weaver (1948-49): text v čínštině je stejný jako v angličtině, je jen v jiném kódu
- Georgetownský experiment – 1956, R-A, P. Toma
- Rusové – O. Kulagina, I. Mel'čuk, 1958, Fr – Ruš.
- Systran P. Tomy – oficiální SP systém EU
- Verbmobil – v letech 1993-2001, hlasový překlad angličtina-japonština-němčina (Tuebingen, 100 mil.)
- v posl. době: pravidlový vs. statistický přístup,
- RBMT v kombinaci se SMT + hybridní řešení
- Google Translator – využití paralelních korpusů
- Pokrok, rozdíly u jazykových dvojic, kvalita?

SP v českém prostředí

- Seminář SP na FF UK, B. Palek, P. Sgall,

  Novák, Konečná, 1958-60 a dále
- Pokusy s českým SP z angličtiny – P. Sgall, E. Hajičová, počítače SAPO, LGP, EPOS
- Po r. 1968 rozštěpení skupiny na dvě, FF UK (Novák, Palek), MFF UK (Sgall, Hajičová)
- Experimenty se systémem Ruslan, K. Oliva, J. Hajič (VÚMS, Svoboda, sálové počítače)
- V současnosti – ÚFAL, J. Hajič, EuroMatrix, EUM+
- Zčásti se SP věnuje pozornost i v CZPJ (V. Baisa)

SP – pokračování historie

- Zpráva ALPAC, J. R. Pierce,1964(6), vláda USA
- (Automatic Language Processing Advisory Committee, 7 odborníků)
- Skepticky hodnotila výzkum v oblasti PL (CL) a SP
- Potřeba posílit základní výzkum v oblasti PL a SP
- Zpráva vedla v USA ke snížení finanční podpory v oblasti SP, negativní efekt
- https://en.wikipedia.org/wiki/ALPAC
- UK, Francie, později zpráva JTEC 1992 (J.Tech.C.),
- Velký projekt Eurotra – financován EK 1978-1992
- proj. EuroMatrix a EuroMatrix-plus 2006-09-12)

Příklad

Viz G. Translator

Shrinkage

Úbytek, ztráta, snížení, redukce

woman drive drunk

Systémy strojového překladu

- RBMT- pravidlové
- a) Přímé systémy – 1. generace, doslovný překlad zdroj.text -> MFA -> slovník -> přeuspoř. -> cílový text, např.: ruská věta *My trebuem mira.* se přeloží do ang. jako *We want world*. nebo *We want peace*.
- Nepřímé systémy – 2. generace
  b) **transferové -**
  zdroj.text -> analýza: lex.,mf.,synt. (R_i) -> transfer (R_i -> R_j) -> syntéza: synt., mf. -> cílový text (postred.), novým prvkem je syntaktická (příp. i sém.) reprezentace, mezireprezentace, transferová (převodní) pravidla, jazyková závislost R_i i R_j

Systémy SP …

- c) **s převodním jazykem -** univerzální, multilinguální.
- Zdr.text -> nezávislá analýza -> reprez. v PJ -> nezávislá syntéza -> cílový text,

- poskytuje možnost zpětného překladu a testování - PJ?

- vhodný symbolický systém, logický kalkul, PK1 nebo formule v systému TIL, je jazykově nezávislý,

- přidání nového jazyka vyžaduje přidat jen 2 moduly, u transferových systémů – 4,
- u PJ jsou komplikace s jazykově nezávislými reprezentacemi.
- Systém Rosetta 1986 – http://mt-archive.info/IAI-1986-Appelo.pdf

Systémy SP…

- Statistický SP (SMT)
- Využití velkých dat, paralelních korpusů
- Jazykové modely
- Představitel Google Translator a další
- Hybridní – EuroMatrix
- S překladovou pamětí – Trados
- Využití databází již přeložených textů

Vybrané příklady systémů SP

- TAUM Meteo 1981, ang.-franc. Univ. of Montreal – práce s podjazykem
- TAUM Aviation 1985, ang.-franc., oba RBMT
- Pravidlový – Systran (Apollo, US AF, EU)
- Statistický – Google, Moses, v současnosti
- Hybridní – faktorovaný – EuroMatrix
- PRESEMT – EU projekt 2011-2014
- Hlasový SP, Verbmobil, 1993-2001

Statistický SP (evaluace)

Automatické metriky

- Bleu – kandidátský překlad proti vícenásobným referenčním překladům (viz později)
- NIST – modifikace Bleu, n-gramy
- METEOR – vážený harmonický průměr přesnosti a pokrytí unigramu
- Levenshteinova vzdálenost mezi dvěma slovy je minimální počet editačních kroků (vložení,přesunutí)

Manuální evaluace, viz dále

Kritéria kvality překladu I

- **Věrnost –** překlad musí přenášet tutéž informaci jako originál, např.
- **Srozumitelnost** – míra jasnosti a srozumitelnosti každé přeložené věty,
- **Stylistická vhodnost** – nakolik je cílový text vhodný pro cílového uživatele vzhledem k danému komunikačnímu záměru, japonština
- To jsou základní a zcela obecná kritéria.
- Další parametry pro hodnocení kvality překladu
  a) **jazyková obecnost** – kolik vstupních a výstupních jazyků systém zahrnuje

Kritéria kvality překladu II

b) rozsah pokrytí **ve slovníku** – počet slovních druhů (otevřených, uzavřených) ve slovníku,

c) **gramatické pokrytí** – procento kompletních vět, jež systém beze zbytku analyzuje nebo generuje,

d) procento **negramatických vstupů**, které je systém schopen zpracovat (podle testovacího seznamu),

e) hodnocení **kvality přiřazení** mezi lexikálními jednotkami v obecném slovníku systému,

f) aplikační a **tematická obecnost** – počet věcných oblastí (domén), jež systém pokrývá, a rozsah pokrytí jednotlivých oblastí,

Kritéria kvality překladu III

- **Stupeň automatizace** - míra, v níž musí člověk intervenovat v překladovém cyklu - čím méně, tím lépe - pre- a posteditace, interaktivní desamb.
- Množství času potřebného pro lidský zásah/celkový čas potřebný pro úplný překlad - to je míra automatizace charakteristická pro MT systém,
- lze přihlédnout ke kvalifikaci - posteditor je obvykle zkušený překladatel, preeditor (desamb.) nemusí znát cílový jazyk, nižší kvalifikace - nižší náklady,

Kritéria kvality překladu IV

- **Sémantická přesnost -** míra, v níž přeložený text vyjadřuje stejný význam jako vstupní text - centrální kritérium pro posouzení kvality překladu, je to důležité u manuálů, předpovědí počasí, zákonů a předpisů - výrazy jako - *rozdělovač, hlava motoru, státní podpora, daňový poplatník* musí být přeloženy přesně,
- **Srozumitelnost -** míra srozumitelnosti, v níž je přeložený text srozumitelný pro čtenáře cílového jazyka, aniž se musí dívat do zdrojového textu. Těsně souvisí s sémantickou přesností, počítá s ní.

Kritéria kvality překladu V

- Stylistická adekvátnost (vhodnost) – míra, v níž je cílový text vhodný pro zamýšlené adresáty, např. angl. - japonština – překlad může být srozumitelný i významově přesný, ale nevhodný **sociálně** kvůli **honorifikaci** – použití zdvořilostní frazeologie, obratů, bez nich by text nebyl použit
- Pak je nutná posteditace – podobně v češtině: tykání a vykání
- Podobně – text s výrazy předpokládajícími vyšší vzdělání (pro odborníka) je nevhodný pro člověka z ulice, implicitní presupozice – nevyslovený předpoklad, kvantifikace – číselné vyjádření aj.

Kritéria kvality překladu VI

- Tyto rozdíly je nesnadné kvantifikovat.
- **Tematická a jazyková portabilita –** míra, s níž lze přidat k systému další věcné oblasti a jazyky, dá se měřit **množstvím času** potřebného pro přidání souboru gram. pravidel dalšího jazyka a slovníku termínů pro novou oblast včetně přiřazení ekviv. cílového jazyka.
- Rozdíly: u bin. systémů se závislou analýzou a syntézou mezi dvěma jazyky a u systémů s PJ, kde se přidává jazykově nezávislá reprezentace dané tematické oblasti.

Kritéria kvality systémů

- **Rozšiřitelnost** – míra, v níž MT systém dovoluje hladkou a inkrementální extenzi gramatických pravidel a slovníku a věcné oblasti pro jazyk, s nímž se už v systému pracuje. Závisí to na míře deklarativnosti a transparence použité reprezentace gramatických pravidel a slovníkových hesel a na nástrojích používaných pro údržbu systému.
- Lze ji měřit množstvím času potřebným pro:
  - kódování pravidel a hesel
  - jejich testování
  - verifikaci a kontrolu, že přidání nezpůsobí nečekané a nežádoucí konflikty

Kritéria kvality překladu VII

- **Zlepšitelnost –** míra, v níž systém umožňuje zlepšit úroveň automatizace bez kompromisů v kvalitě překladu, fakticky jde o míru **otevřenosti** systému: zlepšení bez přebudování designu.
- **Ergonomičnost** – míra, v níž systém poskytuje minimum příležitostí pro vznik chyb, pomůcky pro uživatele, kvalita rozhraní (pokročilost), snadnost napojení na strojově čitelné slovníky, hypertextové odkazy do textu překladu, vazby na archiv překladů (viz též systémy jako TRADOS).
- **Integrovatelnost** – možnost začlenění do jin.syst.
- **Softwarová portabilita –** přenos na sw. platformy

Lexikální data I

- Data pro SP (MT) – **gramatická pravidla**, popisují stavbu věty, tj. formální gramatika potřebná pro analýzu a syntézu (generování),
  - lexikální: informace o každé lexikální jednotce (slovníkovém heslu) - slova, kolokace, např. *škola, vysoká škola, mateřská škola,*
- Jde o vztah slovníku a gramatiky – obvykle se tato data v SP systémech drží odděleně – problém: co kam dát?
- Lze pro SP použít normální elektronické slovníky – Leda, Lingea, PC Translator?

Lexikální data II

- Informace ve slovníku: morfologická, subkategorizace, valence, výběrová omezení, SR
- Organizace lex. dat je dána typem SP systému -
  a) systémy s **přímým překladem** - typicky jeden dvojjazyčný sl. - na jedné straně údaje o LJ vstupního jazyka, na druhé straně přiřazení ekvivalentů cílového jazyka,
  b) mívá podobu **seznamu** všech tvarů (ang.) nebo kmenů (češ.) + mf.inf., synt. inf., SR, inf. potřebná pro výběr alternativ, infce pro syntakt. změny v syntéze – výsl. značně složitý slovník.

Lexikální data III

- Nepřímé systémy – moduly analýzy a syntézy jsou od sebe odděleny, oddělené jednojaz. slovníky pro vst. a cílový jazyk, dále dvojjazyčný/é transf.sl., bývají jednodušší než u přímých syst. U každé LJ - mf.inf., POS, SR, výb.omezení, valence
- Časté jsou samost.sl. homografů – *bank (fin.inst., břeh), stát (země, zaujímat polohu, mít cenu).*
- Informace pro výběr cílových ekv. (jeho formy) se často umisťuje do transf.dvojj.slovníku,
- v praxi: slovník velmi četných výrazů, sl.idiomů, sl.nepravid.tv., sl.homografů, mikrosl.- výměnné - zeměd., fyzika, žurnal., IT, terminologické

Vstupy – výstupy

- Obecné sl., interaktivní syst., řeší to víceznačnost.
- **Psaný vstup** – ošetření pravopisu, korigování, interp., oddělovače, převod do výstupního jazyka (*This year, the man, however, and his wife, too, will go on holiday. – Letos ale ten člověk a taky jeho žena pojedou na dovolenou.),*
- Fonty - rozlišný úzus, pomlčky, uvozovky, užití kurzívy, prostrkaného písma, polotučného apod.

Morfologie při SP

- Typy jazyků - analytické: angličtina, franc., němč.,
  - syntetické, flektivní: slov.jazyky – ruš., češ., polš.
  - aglutinační: ugrofinské, maď., finština, turečtina,
- Pro každý typ jazyka – morfologická analýza,

  tj. pro vstupní větu – předzpracování, slova, kolokace, pak vlastní analýza
- segmentace slovních tvarů, získání kmenů a gramatické informace (koncovky, alternace),
- Morfologické analyzátory, viz např. AJKA,
- Struktura morf. analyzátorů, slovník kmenů, koncovkové množiny

Syntaktická analýza při SP

- Při analýze – zjistit prvky věty a vztahy mezi nimi, vstupní text - řetězy znaků, identifikace slov - mfa a slovník = přiřadí slovům nějaké atributy, např. *kopu*
  - k1gMnSc2 (*Nedvěd dal branku z rohového kopu*)
  - k1gFnSc4 (*nedávej to na jednu kopu*)
  - k5eAp1nStPmIaI (*kopu si hrob*) –to nestačí, kromě informace o významu, kterou se budu zabývat až při transferu, je potřeba provést desambiguaci:
- 3 významy, zkusíme provést synt. anal. a nějak reprezentovat vztahy mezi prvky ve větě – jak?
- Syntaktický strom vstupní věty - jak jej lze získat?
  - vhodný typ formální gramatiky a analyzátor

Synt.analýza při SP II

- Jak poznat, že *kopu* je v daném řetězu k1 (jméno)?
- Příklad deriv.stromu (s použitím CFG)

Reprezentace znalostí

- Znalosti o světě – jejich zdroje
- Ontologie, sémantické sítě, WordNet a EWN
- Encyklopedie, terminologie
- Znalosti o jazyce, lex.databáze
- Common sense
- KBMT

Sémantická analýza

- Rozpadá se do dvou částí: lexikální a logické
- Lexikální analýza zahrnuje významy slov a kolokací – problém slovníků pro SP, otázka kvality a zachycení kontextových vztahů
- Logická analýza se týká významu celých vět a jejich reprezentace, též ve vztahu

  k reprezentaci znalostí

Analýza souvislého textu – vztahy odkazování

(koreference, anafora)