Úvod do strojového překladu (PV061)

Karel Pala

pala@fi.muni.cz

Centrum ZPJ FI MU

podzim 2019

Výchozí body – vztah SP a NLP

- SP je testovacím prostředím pro NLP
- Techniky vyvinuté v oblasti NLP se ověřují v systémech SP
- SP zahrnuje jednotlivé jazykové roviny:

- Morfologii – slova a jejich tvary
- Syntax – vztahy mezi slovy, struktura věty
- Sémantiku – význam slov a význam vět
- SP mluvené řeči – zahrnuje uvedené výše (ASR)
- V souč. statistické techniky, neuronové sítě

Historie strojového překladu

- C. Shannon, W. Weaver (1948-49): text v čínštině je stejný jako v angličtině, je jen v jiném kódu (naivní)
- Georgetownský experiment – 1956, R-A, P. Toma
- Rusové – O. Kulagina, I. Mel'čuk, 1958, Fr – Ruš.
- Systran, Peter Toma – později oficiální SP systém EU
- Hlasový SP – Verbmobil – 1993-2001, angličtina-japonština-němčina (Tuebingen, 100 mil. Dm)
- V poslední době: pravidlový vs. statistický přístup,
- RBMT v kombinaci se SMT + hybridní řešení
- Google Translator – uplatnění neuronových sítí – lepší výsledky
- Existují rozdíly v kvalitě u jazykových dvojic (A – Č)

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
- Doporučení posílit základní výzkum v oblasti SP
- Zpráva vedla v USA ke snížení finanční podpory v oblasti SP, zpomalení výzkumu
- 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

SP v českém prostředí

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

  (Novák, Konečná, Hajičová, Nebeský1958-60 a dále)
- Pokusy s českým SP z angličtiny – P. Sgall, E. Hajičová, počítače SAPO, LGP, EPOS
- Po r. 1968 rozštěpení pražské skupiny na dvě, FF UK (Novák, Palek, Konečná), MFF UK (Sgall, Hajičová)
- Experimenty se systémem Ruslan, K. Oliva, J. Hajič (VÚMS, Svoboda, sálové počítače)
- V současnosti – ÚFAL MFF UK, J. Hajič, Bojar, systémy EuroMatrix, EUM+
- SP se zčásti věnuje pozornost v CZPJ (V. Baisa)

Příklad

Viz G. Translator

Shrinkage

Úbytek, ztráta, snížení, redukce

woman drive drunk

Systémy strojového překladu I

- pravidlové (RBMT) – vs. statistické (SMT) a s NS
- a) Přímé systémy – 1. generace, doslovný překlad zdroj.text -> MFA -> slovník -> přeuspořádání. -> cílový text, ruská věta *My trebuem mira* se přeloží do ang. jako *We want world* nebo *We want peace*
- Nepřímé systémy – 2. generace
  b) transferové - zdrojov. 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 II

- c) s převodním jazykem – univerzální, multilinguální.
- Zdrojový 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 jako TIL, je jazykově nezávislý,

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

Systémy SP III

d) Statistický SP (SMT)

- Využití velkých dat, paralelních korpusů
- Jazykové modely, n-gramy
- Hlavní představitel: Google Translator a další
- Hybridní – Tecto SP (ÚFAL), Chimera: Mos.
- Rozšíření o neuronové sítě – nová verze Gtranl.

e) Systémy s překladovou pamětí – Trados

- Využití databází již přeložených textů,
- paralelních korpusů
- Používá se pro lokalizaci, práce s terminologií

Některé příklady systémů SP

- TAUM Meteo 1981, ang.-franc., Montreal Univ – práce s podjazykem (počasí) – RBMT
- TAUM Aviation 1985, ang.-franc. – RBMT
- Další pravidlový syst.– Systran (Apollo, US AF, donedávna používán pro EU)
- Hlasový SP, Verbmobil, 1993-2001, ang. něm. jap
- Statistický – Moses, Google Tr., v současnosti
- Hybridní – faktorovaný – EuroMatrix – ÚFAL
- TectoMT – využití tektogramatické roviny
- PRESEMT – EU proj. 2011-2014 – naše účast

Kritéria kvality překladu

- Věrnost – překlad musí přenášet tutéž informaci (význam) jako originál, *A student reads the book.*
- Srozumitelnost – míra jasnosti každé přeložené věty musí odpovídat originálu
- 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 daný systém SP pokrývá

Kritéria kvality překladu II

b) rozsah pokrytí ve slovníku – počet slovních druhů (otevřených, uzavřených) ve slovníku,
c) gramatické pokrytí – procento kompletních vět, jež systém SP 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 pokrytých věcných oblastí (domén) a rozsah jejich pokrytí

Kritéria kvality překladu III

- Stupeň automatizace – míra intervence v překladovém cyklu – čím méně, tím lépe – pre- a posteditace, interaktivní desambiguace.
- Množství času potřebného pro lidský zásah
- Celkový čas potřebný pro úplný překlad
- Míra potřebné preeditace a posteditace
- Práce zkušeného překladatele, preeditor (desambiguátor) 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, důležité pro manuály, předpovědi počasí, zákony a předpisy
- Termíny – jako: *rozdělovač, hlava motoru, státní podpora, daňový poplatník* musí být přeloženy přesně
- Pochopitelnost – míra, v níž je přeložený text srozumitelný pro čtenáře cílového jazyka bez nahlížení do zdrojového textu.

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 – zdvořilostních obratů, bez nichž by se text nedal použít
- je tedy nutná posteditace – podobně v češtině existuje tykání a vykání (není v ang.)
- Podobně – text s odbornými termíny (pro odborníka) je nevhodný pro člověka z ulice, *implicitní presupozice* – nevyslovený předpoklad, *kvantifikace* – číselné vyjádření aj. – stylistika

Kritéria kvality překladu VI

- Uvedené rozdíly je nesnadné kvantifikovat, viz dále.
- Tematická a jazyková portabilita – míra, s níž lze přidat k SP 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 ekvivalentů cílového jazyka.
- Systémy s PJ, u nich pracujeme s jazykově nezávislou reprezentací dané tematické oblasti (domény).

Vlastnosti systémů SP I

- 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.
- Dá se měřit množstvím času potřebného pro:
  - kódování gr. pravidel a slovníkových hesel
  - jejich testování
  - verifikaci a kontrolu, že přidání nezpůsobí nečekané a nežádoucí konflikty.

Vlastnosti systémů SP II

- Zlepšitelnost – míra, v níž systém umožňuje zlepšit úroveň automatizace bez kompromisů v kvalitě překladu,
- je to míra otevřenosti systému: zlepšení bez přebudování designu.
- Ergonomičnost – míra odolnosti systému vůči vzniku chyb, kvalita sw. rozhraní (pokročilost), napojení na strojově čitelné slovníky, odkazy do textu překladu, vazby na databázi překladů (systémy jako TRADOS).
- Integrovatelnost – možnost začlenění do jin.syst.
- Sw. portabilita – přenos na jiné sw. platformy

Statistický SP (evaluace SMT)

Automatické metriky (pokrytí a přesnost)

- 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, meze automatických metrik

- Srozumitelnost a věrnost, už zmíněno

Složky SP – vstupy – výstupy

- Interaktivita na vstupu, řešení víceznačností.
- Psaný vstup – ošetření pravopisu, korigování, interp., oddělovače, převod do výstupního jazyka
  Př.: *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 – odlišný úzus, pomlčky, uvozovky, užití kurzívy, polotučného písma apod.
- Sw. zajištění vstupů a výstupů není jednoduché a je softwarově pracné, samostatná úloha

Morfologie při pravidlovém SP

- Typy jazyků – analytické: angličtina, franc., němč.,
  - syntetické, flektivní: slov.jazyky – ruš., češ., polš.
  - aglutinační: ugrofinské, maď., finština, turečtina,
- Pro každý typ jazyka – vlastní morfologická analýza, tj. pro vstupní větu – zpracování slov, rozpoznání kolokací (MWEs), pak vlastní analýza
- Segmentace slovních tvarů, získání kmenů a gramatické informace (koncovky, alternace),
- Morfologické analyzátory, viz např. MAJKA,
- Struktura morf. analyzátorů v závislosti na jazyce, např. pro češtinu slovník kmenů, koncovkové množiny, vzory

Syntaktická analýza při pravidl. SP

- Rozpoznání větných prvků a vztahů mezi nimi, po identifikaci tvarů slov – mfa a slovník, např. *kopu*
  - k1gMnSc2 (*Nedvěd dal branku z rohového kopu*)
  - k1gFnSc4 (*nedávej to na jednu kopu*)
  - k5eAp1nStPmIaI (*kopu si hrob*)
- Nejprve je potřeba provést desambiguaci:
- 3 významy, pak musíme provést synt. anal. a nějak reprezentovat vztahy mezi prvky ve větě – jak?
- Syntaktický strom vstupní věty – stromové grafy
- vhodný typ formální gramatiky a synt. analyzátor

Synt.analýza při pravidlovém SP I

- Pracuje se s formálními gramatikami: CFG apod.
- Disambiguace: v rámci mfa i synt. analýzy
- Nalezení jednoznačného derivačního stromu
- Dělá to vhodný syntaktický analyzátor (parser)
- Typy synt. analýzy: složková, závislostní aj.
- Modifikace formálních gramatik – zesílení CF formalismu, např. DC gramatiky v Prologu
- Použití valenčních rámců a sémantických rolí dává dobré výsledky – viz faktorový SP u SMT

Sémantická analýza u pravidl. SP

- Potřeba sémantické reprezentace – význam
- Lexikální analýza pokrývá významy slov a kolokací – problém slovníků pro SP, ne u SMT
- Významy ve víceznačných kontextech, např.
- *Kolik to bude stát*? *What will be the price?*
- U RBMT jde o významy celých vět a jejich reprezentace,
- též tu jde o vztah k reprezentaci znalostí
- Analýza promluvy a souvislého textu – vztahy odkazování (koreference, anafora), (zájmena)

Reprezentace znalostí

- Jedna část pravidlového SP využívá znalostí o světě
- Ontologie – hierarchie pojmů a termínů
- Sémantické sítě, WordNet a EuroWordNet
- Encyklopedie, terminologické slovníky
- Znalosti o jazyce, valenční rámce a jejich databáze
- Common sense – neformální znalosti o světě
- KBMT – ne u SMT a NS, jiná metod. orientace

Data pro SP I (slovníky)

- Data pro SP – gramatické kategorie: značky
- Formální gramatika pro analýzu a syntézu (generování cílového textu),
- Lexikální: informace ve strojových slovnících, slova, kolokace (MWEs, víceslovná spojení), např. *škola, vysoká škola, mateřská škola,* WS
- Vztah slovníku a gramatiky – obvykle se tato data v SP systémech drží odděleně
- Lze pro SP použít normální elektronické slovníky – Leda, Lingea, PC Translator? Přímo ne. Jako pomůcky ano.

Data II

- Informace ve slovníku: morfologická, sém. rysy subkategorizace, valence, výběrová omezení
- 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) sl. mívá podobu seznamu všech tvarů (ang.) nebo kmenů (češ.) + mf., synt., SR, inf. potřebná pro výběr alternativ, infce pro syntakt. změny v syntéze – výsledkem značně složitý slovník.

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 transferového dvojjaz. slovníku,
- v praxi: slovníky frekventovaných výrazů, sl.idiomů, sl.nepravid.tv., sl. homografů, mikrosl. – výměnné - zeměd., fyzika, žurnal., IT, sezn. termínů

Práce s literaturou (20 min.)

Systém Moses – 2.10. (prezentace 10 min.)

Chimera (ÚFAL) – 9.10. (10 min.)

Faktorovaný překlad – 16.10. (prez. 15 min.)

TectoMT (Framework Treex) 23.10.

SDL – Trados Studio, překl. paměti – 30.10.

Převodní jazyk – Rosetta – 6.11.

Verbmobil – 13.11.

Systran – test, co umí (chyby) – 20.11.

Google.Translate – (test a eval., chyby) – 27.11.

Osnova

1. Úvod (teoretická východiska překladu, automatického překladu) - kp
2. Historie (od vzniku počítačů) - kp
3. Pravidlové systémy - kp
4. Statistický strojový překlad - pr
   a. Jazykové modely
   b. Paralelní korpusy
   c. Překladové modely
5. Neuronové modely pro strojový překlad - vb
   a. Word embedding
   b. Feed-forward, recurrent NN