

PLIN009 – Strojový překlad

Pravidlový strojový překlad

Vít Baisa

jaro 2012

14. března 2013

Úvod

- 1 Úvod
- 2 Tokenizace
- 3 Morfologická rovina

Úvod

Rule-based Machine Translation – RBMT

- lingvistické znalosti formou pravidel
- pravidla pro analýzu
- pravidla pro převod struktur mezi jazyky
- pravidla pro syntézu

Knowledge-based Machine Translation – KBMT

- systémy využívající znalosti o jazyce
- obecnější pojem

Knowledge-based MT

- je důležité správně analyzovat kompletní význam zdrojového textu
- ne ovšem *totální* význam (všechny konotace, explicitní a implicitní informace)
- dříve spíše význam systému využívajícího interlinguu
- zde jako ekvivalent pravidlového systému

Rozdělení systémů KBMT

- přímý překlad
 - direct translation
 - nejstarší, 1 krok – transfer
 - Georgetown experiment, METEO
 - zájem o něj rychle opadl
- systémy používající interlinguu
 - interlingua-based
 - dva kroky – analýza, syntéza
 - Rosetta, KBMT-89
- transferové systémy
 - tři kroky (+ transfer)
 - PC Translator

Do 90. let pouze tyto dva typy systémů.

System přímého překladu

- hledají se korespondence mezi zdrojovými a cílovými jazykovými jednotkami (slovy)
- první pokusy s překladem EN-RU
- všechny složky jsou striktně omezeny na konkrétní jazykový pár
- typicky se skládá z velkého překladového slovníku a monolitického programu řešícího analýzu a syntézu
- nutně dvojjazyčné a jednosměrné
- pro překlad mezi N jazyky potřebujeme $N \times (N - 1)$ přímých dvojjazyčných systémů / modulů

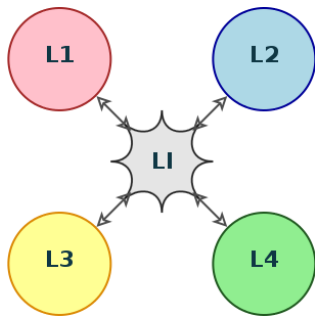
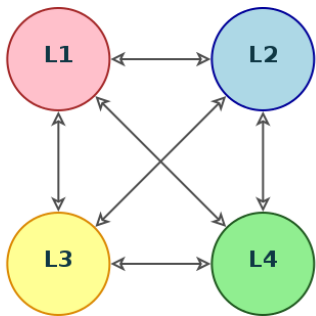
Přístup pomocí interlinguy

- předpokládá, že je možné SL konvertovat do sémanticko-syntaktické reprezentace, která je (částečně) nezávislá na jazyku
- interlingua musí být jednoznačná (unambiguous)
- z této podoby (interlingua) je generován TL
- analýza SL je jazykově závislá, ale nezávislá na TL
- analogicky syntéza TL
- SL a TL nepřijdou do styku
- pro překlad mezi N jazyky potřebujeme $2 \times N$ modulů

Transferové systémy strojového překladu

- provede se analýza po jistou úroveň
- transferová pravidla převedou zdrojové jednotky na cílové
- ne nutně na stejné úrovni
- převod na (nejčastěji) syntaktické úrovni dovoluje zavádět kontextová omezení u přímých překladů nedostupná
- na cílové straně se pak generuje cílový řetězec
- systém linearizace
- při hlubší analýze dochází ke stírání rozdílů mezi interlingua-based a transfer-based systémy
- značná část obou systémů se může překrývat

Interlingua vs. transferové KBMT



Tokenizace

- 1 Úvod
- 2 Tokenizace**
- 3 Morfologická rovina

Tokenizace

Co to je?

- rozdělení vstupního řetězce do tokenů
- token = řetězec znaků
- výstup *tokenizace* = seznam tokenů
- slouží jako vstup pro další zpracování
- označení hranic vět

Problémy

- don't: do_n't, do_n_'t, don_'t, ?
- červeno-černý: červeno-_-černý, červeno-černý, červeno-_-černý
- Zeleninu jako rajče, mrkev atd. ¶Petr nemá rád.
- Složil zkoušku a získal titul Mgr. ¶Petr mu dost záviděl.

Tokenizace – jak se to dělá?

V drtivé většině případů heuristika. (`unitok.py`)

Dělení na tokeny

- pro jazyky používající hlásková písmata: dělení podle mezer
- a podle dalších interpunkčních znamének
- ?! . , - () / : ;

Dělení na věty

- MT v naprosté většině případů pro věty
- u plaintextu: podle seznamu interpunkčních znamének
- problém: Měl jsem 5 (sic!) poznámek.
- výjimky: zkratky (aj., atd., etc.), tituly (RNDr., prof.)
- někdy (HTML) lze využít strukturní značky

Morfologická rovina

- 1 Úvod
- 2 Tokenizace
- 3 Morfologická rovina**

Morfologická rovina

- druhé patro v překladovém trojúhelníku
- je nutné eliminovat obrovský počet slovních variant
- převod slovní formy na základní tvar
give, gives, gave, given, giving → give
dělá, dělám, dělal, děláje, dělejme, ... → dělat
- analýza gramatických kategorií slovních tvarů
dělali → dělat + minulost + průběh + plurál + 3. osoba
did → do + minulost + dokonavost + osoba ? + číslo ?
Robertovým → Robert + pád ? + adjektivum + číslo ?

Morfologická analýza

- pro každé slovo získáme základní tvar, gramatické kategorie, případně segmentaci
- Co je to základní slovní tvar? Lemma.
- jména: singulár, nominativ, pozitiv, maskulinum
- *bycha* → bych?, *nejpomalejšími* → pomalý
neschopný → schopný?
- slovesa: infinitiv
- *nerad'* → radit?, *bojím se* → bát (se)
- Proč infinitiv? nejčastější tvar slovesa
- lemma souvisí s rozsahem/obsahem použitého slovníku

Morfologické značky, tagset

- silně závislé na jazyce (různé morfologické kategorie)
- brněnský atributový systém: dvojice kategorie-hodnota
maminkou → k1gFnSc7
udělány → k5eAaPmNgFnP
- pražský poziční systém: 16 pevných pozic
kontury → NNFP1-----A-----
zdají → VB-P---3P-AA---
- Treebank tagset (angličtina): omezená množina značek
faster → RBR
doing → VBG
- a další (němčina)
gigantische → ADJA . ADJA . Pos . Acc . Sg . Fem
erreicht → VVPP . VPP . Full . Psp

Problém s víceznačností

- v mnoha případech: více morfológických značek
- víceznačnost mezi slovními druhy (více lemmat)
jednou → k4gFnSc7, k6eAd1, k9
ženu → k1gFnSc4, k5eAaImIp1nS
- víceznačnost v rámci slovního druhu
- typicky (čeština): nominativ = akuzativ
víno → k1gNnSc1, k1gNnSc4, ...
odhalení → 10 značek

Morfologická disambiguace

- nutno vybrat *jednu* značku a *jedno* lemma
- ke slovu přichází *morfologická disambiguace*
- nástroj *tagger*
- překladová víceznačnost je něco jiného
pubblico → Öffentlichkeit, Publikum, Zuschauer
- drtivá většina metod využívá kontext
- okolní slova a jejich značky

Statistická disambiguace

- nejpravděpodobnější posloupnost značek
Ženu je domů.
k5 | k1, k3 | k5, k6 | k1
Mladé muže
gF | gM, nS | nP
- těžká situace: *dítě škádlí lvíče*
- strojové učení na ručně značkovaných datech
- různé metody: Brill, TreeTagger
- pro češtinu: Desamb (hybridní)
- je nutné mít k dispozici trénovací data (korpus)

Pravidlová disambiguace

- pokud není k dispozici anotovaný korpus – nutné
- pravidla vyžadují dobrou znalost jazyka
- většinou se používá jako filtr před použitím statistického taggeru
- pravidla mohou zachytit širší kontext
- typicky: shoda v pádu, čísle a rodu ve jmenných frázích
malému (c3, gIMN) *chlapci* (nPc157, nSc36, gM)
- sofistikovanější: valenční struktura věty
valence: *vidět koho/co*
vidím stůl → c4
- systémy DIS, VaDIS

Morfologická segmentace

- proč místo lemmatu (např. infinitiv) nepoužít kořen slova?
- existují i systémy, které provádí segmentaci automaticky na základě seznamu slov pro daný jazyk
- problém: *mít, měj, mám, měl, mívá, ...* – různé podoby téhož morfému
- problém: *i, ové, a, y* – stejná gramatická funkce, různé morfémy
- *bychom* → bych?
- gramatické kategorie mají konkrétní formu (gramémy)
nad-měr-ný, ne-patr(n)-ně, vid-ím, ne-chci, čtyř-i-cet, po-po-sun-out, u-děl-al-i
- nutné pokud nemáme morfologický analyzátor k dispozici

slovo	analýzy	disambiguace
Pravidelné	k2eAgMnPc4d1, k2eAgInPc1d1, k2eAgInPc4d1, k2eAgInPc5d1, k2eAgFnSc2d1, k2eAgFnSc3d1, k2eAgFnSc6d1, k2eAgFnPc1d1, k2eAgFnPc4d1, k2eAgFnPc5d1, k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1, ... (+ 5)	k2eAgNnSc1d1
krmení	k2eAgMnPc1d1, k2eAgMnPc5d1, k1gNnSc1, k1gNnSc4, k1gNnSc5, k1gNnSc6, k1gNnSc3, k1gNnSc2, k1gNnPc2, k1gNnPc1, k1gNnPc4, k1gNnPc5	k1gNnSc1
je	k5eAalmlp3nS, k3p3gMnPc4, k3p3gInPc4, k3p3gNnSc4, k3p3gNnPc4, k3p3gFnPc4, k0	k5eAalmlp3nS
pro	k7c4	k7c4
správný	k2eAgMnSc1d1, k2eAgMnSc5d1, k2eAgInSc1d1, k2eAgInSc4d1, k2eAgInSc5d1, ... (+ 18)	k2eAgInSc4d1
růst	k5eAalmlF, k1gInSc1, k1gInSc4	k1gInSc4
důležité	k2eAgMnPc4d1, k2eAgInPc1d1, k2eAgInPc4d1, k2eAgInPc5d1, k2eAgFnSc2d1, k2eAgFnSc3d1, k2eAgFnSc6d1, k2eAgFnPc1d1, k2eAgFnPc4d1, k2eAgFnPc5d1, k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1, ... (+ 5)	k2eAgNnSc1d1

Universal POS tags

Počet značek se v různých jazycích značně liší → snaha o zjednodušení.

TAG	význam
VERB	verbs (all tenses and modes)
NOUN	nouns (common and proper)
PRON	pronouns
ADJ	adjectives
ADV	adverbs
ADP	adpositions (prepositions and postpositions)
CONJ	conjunctions
DET	determiners
NUM	cardinal numbers
PRT	particles or other function words
X	other: foreign words, typos, abbreviations
.	punctuation

Vytvořeno mapování pro cca 25 jazyků s *tree banks*.

Odhadování POS na základě gramémů

EN	CZ	význam
-s	-á	3. os., j. č., přít.
-ed	-al, -l, -en.	minulý čas
-ing	-(ov)ání	průběhový čas
-en	-en(.)	příčestí minulé
-s	-y, -i, -ové, -a	množné číslo
-’s	ov(o, a, y)	přivlastňování
-er	-ší	komparativ
-est	nej-, -ší	superlativ

Problém: *myší, west, fotbal, . . .*

Tomáš Hanák – Sám v lese II

Když jsi sám v lese,
 ano, sám-li v lese's,
 však skutečně, v lese sám's-li.
 Zkrátka v lese sám-li's.

Však kde vlastně vzal ty tu's?
 Z meze-li v les's vlez?
 Či z nebes v les se snesl's?

Pověz, ach, tvář tvá perlí přívalem se slz.
 Teď rud's, zas bled's, co pivoňka's
 Snad tedy autem's tu, či kolmo's?

Mlčíš a slza tvá dál
 sama malá padá v mechu číš.

Ano, teď teprve snad poprvé sám svět's.

Brillův tagger

- učení z trénovacích dat
 - transformation-based, error-driven
 - úspěšnost přes 90 %
- 1 inicializuj značkování (nejčastější značka)
 - 2 porovnej s trénovacími daty
 - 3 vytvoř sadu pravidel pro změnu značek
 - 4 ohodnot' pravidla
 - 5 aplikuj pravidlo a opakuj od 2. dokud je co zlepšovat

Problémy s POS

- kvalita MA ovlivňuje všechny další roviny zpracování
- kvalita se liší pro různé jazyky (angličtina vs. maďarština)
- **chončaam** (tj) – můj malý dům (domek) (tádžičtina)
- **kahramoni** (tj) – jsi hrdina
- **legeslegmagasabb** (hu) – úplně nejvyšší
- **raněný** – SUBS / ADJ
- the big red **fire** truck – SUBS / ADJ?
- The Duchess was **entertaining** last night.
- Pokojem se neslo tiché **pšššš**

Co s neznámými slovy?

- jde nám o *pokrytí*: analýza co nejvíce slov
- nová, přejatá slova
- řeší *guesser*
- sedm **dunhillek**
- bez **facebooku** strádám
- **třítisícdevětědesát pět** znaků

Morfologie – shrnutí

- první rovina, která zanáší do analýzy významné chyby
- snaha omezit počet slovních tvarů
- nahrazení slovního tvaru za dvojici **lemma + značka**
- pro angličtinu s 36 značkami snadné
- POS tagging dosahuje pro různé jazyky různé kvality
- typicky kolem 95 %