PLIN009 – Strojový překlad

Pravidlový strojový překlad

Vít Baisa

jaro 2012

7. března 2013

Úvod

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.

Systém 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

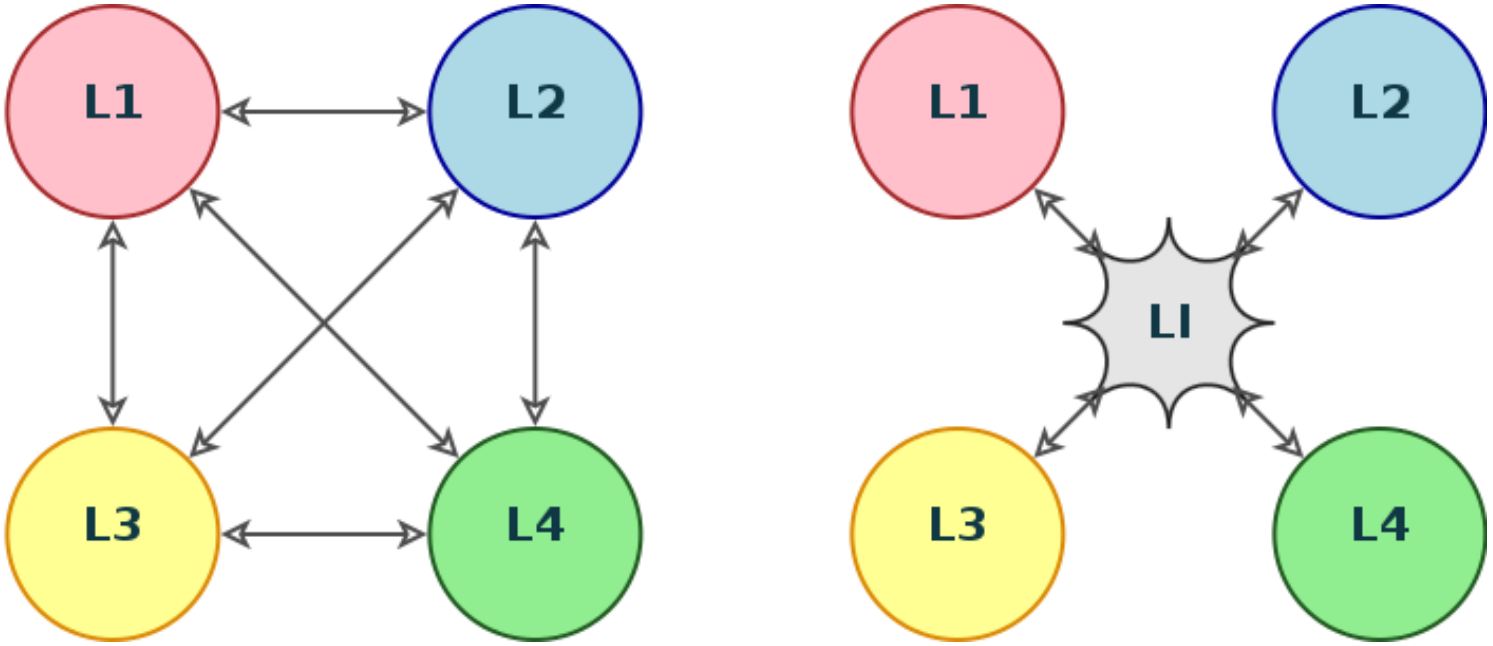
Přístup pomocí interlinguy

- 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



Od řetězce ke slovu

Tokenizace

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 písma: 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