Kapitola 28

Státnice I3: Strojový překlad

28.1 Proč je strojový překlad těžký?

- Strukturální rozdíly mezi jazyky
  - Různé způsoby řazení podmětu, slovesa a předmětu — SVO, SOV, VSO languages
  - Head-marking

```
English: the man's house
Hungarian: az ember háza
           the man house-his
```

- – Pro-drop languages: Některé jazyky umožňují vynechávání zájmen
- Lexikální rozdíly
  - Překlad homonym: slovo ve zdrojovém jazyce může mít více významů, každý význam je potřeba přeložit jiným slovem cílového jazyka (anglické slovo **bass** může označovat jak hudební nástroj tak i druh ryby, tomu odpovídají dva různé překlady do španělštiny)
  - polysémie: koruna v češtině znamená jak část stromu, tak ozdobu hlavy – významy spolu souvisí. V angličtině ale tahle souvislost není vidět a překládá se treetop a crown.
  - distinkce jazykového významu: anglické slovo **know** označuje jak znalost faktu tak i znalost osoby či místa a je to jeden význam. Ve francouzštině odpovídají těmto významům 2 různá slovesa **connaitre** a **savoir**.
  - Slovo/fráze jednoho jazyka nemusí mít ekvivalent v druhém jazyce.

28.2 Úkoly strojového překladu

Zatím se nepodařilo vytvořit plně automatický systém, který by se mohl měřit s živými překladateli co se týče kvality překladu. Přesto je strojový překlad užitečný, používá se pro řešení následujících úkolů:

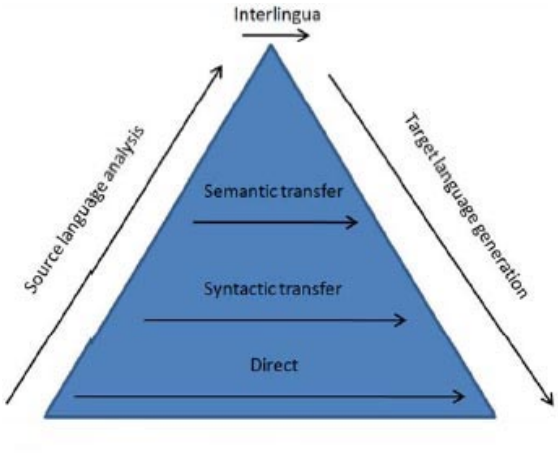
- **rough translation** — orientační překlad nízké kvality (Např. Google translate). Cílem je, aby uživatel neznalý daného jazyka mohl s určitým úsilým porozumět obsahu dokumentu/webové stránky.
- **computer-aided human translation*** — Výstup strojového překladu musí být natolik kvalitní, že je pro překladatele rychlejší provést drobné změny ve výstupu MT než psát vlastní překlad od nuly.
- **subdomain translation** — plně automatický kvalitní překlad na velmi omezené doméně: Např. překlady předpovědí počasí — velmi omezená slovní zásoba, omezená množina jazykových konstrukcí. (Další domény: software manuals, air travel queries, appoitment scheduling, restaurant recommendation).

28.3 Překladový trojúhelník (Vauquois triangle)

Toto schéma popisuje MT jako proces, který lze rozdělit do tří kroků:

- **Analysis** — ligvistická analýza zdrojové věty, může zahrnovat morfologickou, syntaktickou, sémantickou analýzu. Účelem je vytvořit lingvistickou reprezentaci zdrojové věty vhodnou pro překlad.
- **Transfer** — Převod lingvistické reprezentace zdrojové věty do lingvistické reprezentace cílové věty.
- **Generation** — Vytvoření povrchové reprezentace věty v cílovém jazyce.

Trojúhelníkový tvar naznačuje, že při hlubší analýze je transfer jednodušší.



Obrázek 28.1: Vauquois triangle

28.4 Metody Evaluace

28.5 Rule-based strojový překlad

Přímý překlad

- Analysis — pouze morfologická analýza
- Transfer — Překlad jednotlivých slov/frází za pomocí dvojjazyčného slovníku (pravidla ve formě Rozhodovacích stromů), local reordering (záměna pořadí slov/frází)
- Generation — morphological generation

```
//Příklad: Překlad slov much a many z angličtiny do ruštiny
//
if (preceding word is how)
  return skol'ko;
else if (preceding word is as)
  return skol'ko zhe;
else if (word is much)
{
  if (preceding word is very)
    return nil;
  else if (following word is a noun)
    return mnogo;
}
else
{
  if (preceding word is a preposition and following word is a noun)
    return mnogii;
  else
    return mnogo;
}
```

Rule-based překlad se zapojením syntaktické analýzy

V rámci analýzy je proveden syntaktický parsing. Strom zdrojové věty je převeden na strom cílové věty pomocí **contrastive knowledge** — znalosti rozdílů mezi jazyky (Příklad: Při překladu z SVO jazyka do SOV jazyka bude v syntaktickém stromě prohozeno pořadí uzlu odpovídajícího slovesu s uzlem odpovídajícím předmětu)

Combined Approach

- Analysis
  - Morphological analysis and part-of-speech tagging

  - Chunking of NPs, PPs, and larger phrases
  - Shallow dependency parsing (subject, passives, head modifiers)

- transfer
  - Translation of idioms
  - Word sense disambiguation
  - Assignment of prepositions according to governing verbs

- synthesis
  - Lexical translation with a rich bilingual dictionary
  - Reorderings
  - Morphological generation

Překlad pomocí sémantické analýzy — Interlingua

Interlingua označuje jazykově nezávislou reprezentaci významu. Překladové modely pracující na bázi interlingvy nejprve vytvoří jazykově nezávislou semantickou reprezentaci výchozí věty, na jejím základě pak zkonstruují cílovou větu (žádný transfer, pouze analýza a generování). Použití v systémech pro překlad textů z velmi omezené domény — model sémantiky lze vytvořit (předpověď počasí, rezervace letenek atd.)

28.6 Statistický strojový překlad — phrase-based

značení: z historických důvodů E resp. e označuje větu resp. frázi cílového jazyka, F resp. f větu resp. frázi výchozího jazyka (První překladové systémy: francouzština => angličtina).

- **základní model — Noisy-channel model**

$$E^* = \arg\max_E P(E|F) = \arg\max_E \frac{P(F|E)P(E)}{P(F)} = \arg\max_E P(F|E)P(E)$$

- State of the art systémy používají obecnější "'log-linear model'

$$E^* = \arg\max_E exp[\sum_{m=1}^{M} \lambda_m h_m(E,F)]$$

h_i označuje libovolnou **feature function**

Feature functions

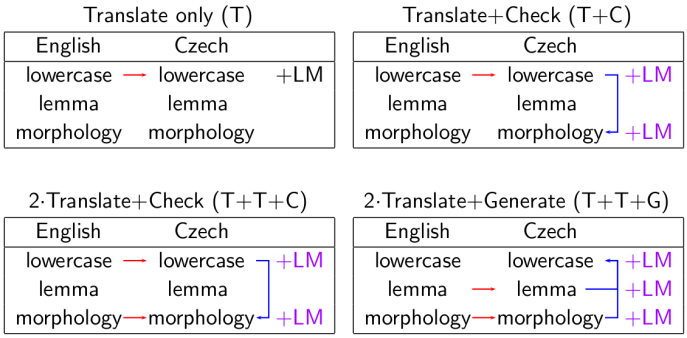
- **Jazykový model** P(E): klasický n-gram language model. Překladové systémy běžně kombinují více jazykových modelů na různých faktorech (slovní formy, lemmata, morfologické tagy...)

$$h_L = \log(P(E))$$

- **Překladový model** P(F|E)

$$h_T = \log(P(F|E))$$

- **Word penalty** — penalizace hypotéz, které obsahují příliš málo/příliš hodně slov.
- **Unknown-word penalty**
- ...atd



Obrázek 28.2: Translation scenarios (picture by Ondřej Bojar)

Vícefaktorový překlad — více různých jazykových modelů

- Both input and output words can have more factors (a ty jsou pak použity jako featury loglineárního modelu).
- Arbitrary number and order of:
  - Mapping steps (red arrows): Translate (phrases of) source factors to target factors.
  - Generation steps (blue arrows)
    - Generate target factors from target factors.

`dvě` → `fem-nom; dva` → `masc-nom`

-
  - ∗ ⇒ Ensures "vertical" coherence.
  - Target-side language models (+LM)
    - Applicable to various target-side factors.
    - ⇒ Ensures "horizontal" coherence.

Překladový model P(F|E)

Úkolem překladového modelu je odhadnout pravděpodobnost toho, že E "generuje" F. Generování se skládá ze tří kroků

- Slova z E jsou rozdělena do frází $e_1, e_2 ... e_I$
- Každá fráze e_i je přeložena jako f_i
- reordering frází f_i

Odhad pravděpodobnosti je založen na

- **translation probability**: $p(f_i, e_i)$
  - pravděpodobnost, že e_i bude přeložena jako f_i. (Bayes tu obrátil směr překladu)
  - pravděpodobnost stanovena na základě alignovaného paralelního korpusu (alignment — mapování mezi frázemi zdrojových a cílových vět v paralelním korpusu): $p(f, e) = \frac{count(f,e)}{count(e)}$
- **distortion probability**: $d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$
  - a_i: počáteční pozice f_i
  - b_{i-1}: počáteční pozice e_{i-1}
  - Neformálně: ...Čím víc reorderingu, tím menší pravděpodobnost...

$$P(F|E) = \prod p(f_i, e_i) d(a_i - b_{i-1})$$

Hledání nejlepší překladové hypotézy

Dekodér

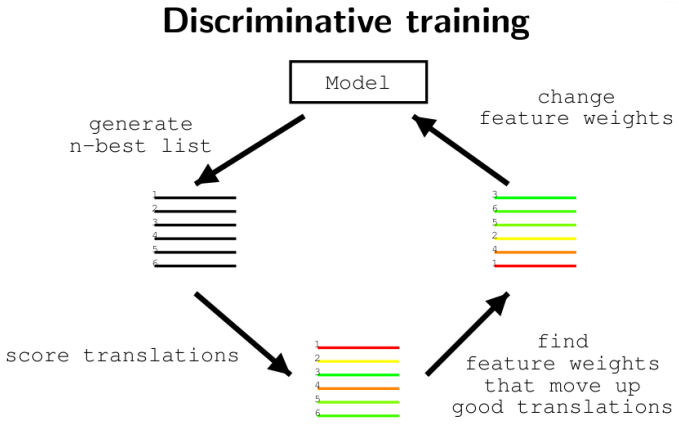
- Úkolem dekodéru je nalézt **N-best list** nejlepších překladových hypotéz na základě pravděpodobnostního modelu
  - Ve fázi dekódování bývá často použit jednodušší pravděpodobnostní model (pouze podmnožina všech feature functions) — tzv. generativní model

  modely na morfologických faktorech atd...)
- algoritmus: A* search
  - jeho varianty používané v MT a speech recognition běžně označované jako **stack decoding**
  - heuristika: best-first search
  - beam search pruning
- Popis dekodéru včetně názorných obrázku je možné nalézt zde.

Rescoring, Discriminative model

- Na výstupu dekodéru je N nejlepších překladových hypotéz podle **generativního modelu**.
- Tyto hypotézy jsou ohodnoceny pomocí podrobnějšího **Diskriminativního modelu** (všechny feature functions), je vybrána ta nejlepší
  - Motivace: Některé feature functions není možné vyhodnocovat pro částečné hypotézy nebo by to bylo moc drahé (časová náročnost)
  - Příklad (můj vlastní):
    * Generativní model vytvoří N-best list na základě jazykového modelu P(E) a překladového modelu P(F|E).
    * Na všech překladových hypotézách bude proveden HMM part-of-speech tagging, pravděpodobnost otagování $\prod_{i=1}^{n} P(w_i|t_i) \prod_{i=1}^{n} P(t_i|t_{i-2}, t_{i-1})$ bude použita jako další feature v diskriminativním modelu.

Minimum Error Rate Training (MERT)



Obrázek 28.3: MERT — discriminative training (picture by Phillip Koehn)

- Nalezení optimálních vah $\lambda_1 ... \lambda_M$ pro jednotlivé feature functions.
- Provádí se na held-out datech — část paralelního korpusu, která nebyla použita při trénování modelu.
- Och's minimum error rate training (MERT):

```
given: sentences with n-best list of translations
iterate n times
  randomize starting feature weights
    iterate until convergences
```

```
      for each feature
        find best feature weight
        update if different from current
return best feature weights found in any iteration
```

28.7 Alignment

Sentence alignment

- Classical algorithm: Gale and Church (1993).
  - Based on similar character length of aligned sentences, no words examined.
  - Dynamic-programming search for the best alignment.
  - Allows 0 to 2 sentences in a group: 0-1, 1-0, 1-1, 2-1, 1-2, 2-2

Word alignment

- Goal: Given a sentence in two languages, align words (tokens).
  - Lexical probabilities: IBM Model 1,
  - Expectation-Maximization Loop for IBM1.
  - Symmetrization techniques.
  - The "standard tool": GIZA++ (Och and Ney, 2000)
  - **Details here!**

28.8 Evaluace

- evaluace MT je subjektivní a netriviální úkol
- výzkum metodologie evaluace hrál ve vývoji MT vždy důležitou úlohu

"Ruční" evaluace

- Kvalitu překladů posuzují dobrovolníci
- Zvlášť se hodnotí:
  - **přesnost** (fidelity) — věrnost zachycení obsahu sdělení
  - **plynulost** (fluency) — zda je věta dobře srozumitelná, jasná; styl

Vyhodnocení plynulosti

- nejjednodušší metoda: dobrovolníci přiřazují skóre každé větě (např. od 1 do 5)
- **cloze** — některá slova na výstupu jsou zakryta, dobrovolník má za úkol uhádnout, jaké slovo je zakryto. Čím větší plynulost, tím je hádání jednodušší
- metoda měření času potřebného pro přečtení věty — čím větší plynulost, tím se věta čte snadněji

Vyhodnocení přesnosti

- nejjednodušší metoda: dobrovolníci přiřazují skóre každé větě (např. od 1 do 5)
- kladení otázek týkajících se obsahu textu — dobrovolník má odpovídat pouze na základě informací obsažených v překladu

Metody měření celkové kvality

- všechny aspekty (plynulost, přesnost) dohromady
- **edit cost of post-editing**
  - kolik slov je potřeba upravit, aby bylo dosaženo překladu rozumné kvality
  - kolik je potřeba stisknutí kláves, aby bylo dosaženo překladu rozumné kvality

Automatická evaluace

- horší kvalita, za to však výrazně levnější a rychlejší než manuální evaluace
- díky tomu, že se dá automatická evaluace provádět rychle a často, může být použita pro optimalizaci parametrů modelu (váhy jednotlivých features jsou nastaveny tak, aby bylo co největší **BLEU** score na heldout datech), nebo k posouzení zlepšení systému při změnách v implementaci.
- automatické metriky: **BLEU, NIST, TER, METEOR**
- všechny tyto metriky jsou založeny na porovnávání výstupu systému s **referenčnímy překlady** — lidmi vytvořené kvalitní překlady. Pro každou větu z testovací množiny je k dispozici více referenčních příkladů (věta se dá obvykle dobře přeložit více způsoby)
- jednotlivé metriky se od sebe navzájem liší tím, jak počítají podobnost mezi výstupem systému a referenčnímy překlady

BLEU

- nejpoužívanější metrika
- založená na **modified n-gram precision**
  - unigram precision: Počítá se poměr, kolik slov z výstupní věty je obsaženo v nějakém referenčním překladu.

```
Output: the the the the the the the
Reference 1: the cat is on the mat
Reference 2: there is a cat on the mat
```

-
  - V uvedeném příkladě je unigram precision $7/7 = 1$, protože všech 7 slov se vyskytlo v nějakém ref. překladu.
  - Uvedený příklad poukazuje na problém — vysoké ohodnocení (maximální hodnota precision), přestože je na výstupu velmi špatný překlad
  - problém odstraní **modified precission**
    - pro každé slovo výstupní věty se spočítá $Count_{clip}$ — upravený počet výskytů
      - Pokud je počet výskytů (Count) slova ve výstupní větě menší nebo roven počtu výskytů v některém z referenčních výskytů, tak $Count_{clip} = Count$
      - V opačném případě bude $Count_{clip}$ rovno maximálnímu počtu výskytů slova v referenčních větách
  - V uvedeném příkladě bude **modified unigram precision** rovna $2/7$, $Count_{clip}(the$ je totiž rovno 2. (Maximální počet výskytů slova the v některém z referenčních překladů je 2).
- modified n-gram precission pro n-gramy vyšších řádů se určuje stejným způsobem. Výpočet modified precision přes celou testovací množinu:

$$p_n = \frac{\sum_{O \in Output} \sum_{n\text{-}gram \in O} Count_{clip}(n\text{-}gram)}{\sum_{O' \in Output} \sum_{n\text{-}gram' \in O'} Count_{clip}(n\text{-}gram')}$$

Output: množina výstupních vět pro celá testovací data

- penalizace příliš krátkých překladů: **brevity penalty (BP)** (nelze použít recall, protože máme více referenčních překladů)

c: součet délek všech výstupních vět

r: součet délek nejpodobnějších referenčních překladů k výstupním větám

$$BP = \begin{cases} 1 & if\ c > r \\ e^{(1-r/c)} & if\ c <= r \end{cases}$$

- BLEU score je definováno jako harmonický průměr modified ungram precision pro n-gramy do řádu N normalizovaného pomocí brevity penalty

$$BLEU = BP \times exp(\frac{1}{N} \sum_{n=1}^{N} \log(p_n))$$

28.9 Historie

- První patenty – už 1930's: G. Artsrouni – automatický slovník na děrné pásce, P. Troyanskii – lidský editor, vyjadřující log. formy a synt. funkce, automatický překlad a editor, který přepíše log. formu cílového jazyka do textu.
- 1946 – A. D. Booth: automatický slovník, překlad slovo od slova, 1949 W. Weaver – informační teorie, desambiguace na zákl. kontextu, kryptografické metody, univerzálie
- 1948 – R.M.Richens – slovník s kořeny, předponami a příponami zvlášť
- 1950 – E. Reifler – **preediting** a **postediting** (zjednodušení textu pro účely překladu, oprava chyb, které udělal stroj)
- 1952 – 1. konference na MIT, L. Dostert – pivotní jazyk pro překlad do více jazyků
- 1954 – Georgetownský experiment: Rusko-anglický text o 250 slovech, 6 synt. pravidel, bez negací, slovesa ve 3. osobě, málo předložek; byl vidět úspěch, zkouší to další
- 1955 – Anglicko-ruský překlad v Moskvě
- 1956 – První mezinárodní konference
- 1957 – Chomsky: Syntactic Structures
- 1960 – Yehoshua Bar Hillel: "Fully automatic high quality machine translation is not feasible.", 1966 ALPAC (Amer. Lang. Processing Advisory committee) – zpráva, která způsobila útlum, mimo USA výzkum pokračoval

Projekty po ALPACu:

- SYSTRAN, Grenoble (GETA), SUSY (Saarbrücken), LOGOS (Texas), TAUM (Montreal), ETAP (Moskva)

TAUM METEO (1976)

- Montreal, překlad meteorologických zpráv z angličtiny do francouzštiny (wen:TAUM system)
- dobře definovaná a správně omezená podmnožina syntaxe a sémantiky
- vhodná implementace (Q-systémy), systém sám rozpozná, že text neumí přeložit
- praktická implementace METEO System fungovala až do 2001 (wen:METEO System)

SYSTRAN

- překlad dokumentů EU, přímý (každý pár zvlášť, cca 20, uspokojivě jen AJ, FJ, NJ, wen:SYSTRAN)
- data oddělena od programu
- řešeno ad-hoc

EUROTRA

- oficiální projekt EU v 80. letech, pokus nahradit Systran (72 jazykových párů, v každé zemi jedno centrum, wen:Eurotra)
- nezvládnutá modularita (každý si měl analyzovat sám, domlouvat se na rozhraní)
- negativní efekt

VERBMOBIL

- Německý nástupce Eurotry, víc jak 30 univerzit; překlad mluvené řeči: domluva obchodníků na příští schůzce
- Patent, prezentace na EXPO 2000, pak ticho

28.10 Systémy podporující překlad

- Využití dříve přeložených textů, princip **překladové paměti** (wen: Computer-assisted translation)
- IBM Translation Manager, Déja Vu, SDL TRADOS – prodává se sám systém, paměť si překladatel zajistí sám
- Hledání shodných úseků, oprava odlišností
- Zejména pro překlady dokumentace k systémům různých verzí
- Dnes se kombinují se statistickým překladem

28.11 České systémy

(převzano z wiki poznámek Úvod do počítačového zpracování přirozeného jazyka)

První překlad 1957 – jedna věta na SAmočinném POčítači: "The consonants have not by far been investigated to the same extent as the vowels." — „Souhlásky zdaleka nebyly prozkoumány do stejné míry jako samohlásky." Později se tu objevily Q-Systémy, takže se začaly psát gramatiky.

APAČ (80. léta)

- Z. Kirschner, slovník pokrýval oblast vodních pump (dokumentace), cca 1500 slov; Q-systémy
- **Transdukční slovník** pro latinské výrazy: -zation -> -zace, -ic -> ický atd., seznam výjimek

Ruslan (1985-1990)

- Překlad manuálů sálových počítačů z češtiny do ruštiny
- Slovník: cca 8500 slov, transdukční slovník (ale příbuznosti jazyků se u něj využít nedalo), Q-systémy
- Použití synt. transferu, očekával se minimální, ale ten stále rostl
- Tehdy na PC 286 trval překlad 1 věty asi 4 minuty, dnes 4 vteřiny
- Spec. kódování: háček = "3" za písmenem, čárka = "2" za písmenem, kroužek = "7" (-1- $n(5) + $gs(1) + < + Z3LUT3OUC3KY2 + KU7N3 + U2PE3L + D3A2BELSKE2 + O2DY + . + > -2-).
- Před operačními zkouškami vývoj ukončen

Česílko (od 1998)

- Překlad příbuzných jazyků, kvůli překladům dokumentací: lidský překlad z angličtiny do češtiny a odtud automaticky do slovenštiny a polštiny. Následně se výsledek opravuje.
- Morfologické slovníky, statistická analýza češtiny
- Využívá (většinou) shodné syntaxe, jsou tu ale odlišné slovníky (ač jistá pravidelnost) a úplně odlišné tvarosloví

PC Translator

- Komerční systém, založený na pravidlech, vyvíjený už hodně dlouho.