5 Různé koncepce strojového překladu a jejich porovnání

5.1 Dvoujazyčný nebo vícejazyčný systém

Dvoujazyčný systém

jednosměrný nebo obousměrný
obousměrný systém: nereversibilní nebo reversibilní

Vícejazyčný systém

velké rozpětí co do počtu jazyků a možností překladu mezi nimi: obousměrný překlad mezi všemi jazyky systému (Eurotra) na straně jedné a jednosměrný překlad z jednoho jazyka do 2–3 jiných na straně druhé; mezi tím řada možností — např. japonština ⟷ angličtina ⟷ francouzština, němčina, španělština
vícejazyčný systém má tím větší smysl, čím je modulárnější, čím víckrát se dají jednotlivé moduly využít a čím je koncepce pro všechny jazyky jednotnější, pak:

- analýza a syntéza jednoho jazyka může zůstat stejná bez ohledu na to, z/do jakého jazyka se překládá
- lingvistické metody a prostředky mohou zůstat stejné pro všechny zúčastněné jazyky
- lze použít stejných softwarových prostředků

Avšak ve dvoujazyčném systému lze využít případů, kdy slova a konstrukce jsou v obou jazycích stejně nejednoznačné (zvláště u jazyků příbuzných).

5.2 Přímá metoda, interlingva, transfer

Přímá metoda

základem je doslovný překlad s lokálními úpravami slovosledu, první pokusy koncem 40. let, celý proces překladu se odehrává v jediné fázi, chybí syntaktická a sémantická analýza, programování ve strojovém kódu
typický postup:

1. morfologická analýza
2. hledání v dvoujazyčném slovníku
3. pravidla pro lokální slovosledné úpravy

výsledek: často chybné lexikální ekvivalenty, syntaktická struktura kopíruje vstup
např.:
My trebuem mira. — We require world.
Ona navarila ščei na neskolko dnej. — It welded on cabbage soups on several days.
Přímá metoda může být výhodná pro jednosměrný překlad do příbuzného jazyka, kde je možné využít podobností v lexiku a syntaxi a soustředit se na rozdíly. (Systran, částečně METEO, APAČ, Ruslan)
Neúspěch přímé metody pro překlad mezi odlišnými jazyky vedl k vývoji systémů, v nichž existuje abstraktní representace (syntakticky, popř. sémanticky) analyzovaného vstupu a/nebo výstupu, a v nichž jsou odděleny fází překladu podle jazyků.

Interlingva

analýza vstupního textu vede k reprezentaci nezávislé na jazyku vstupu a výstupu, z níž se přímo syntetizuje výstupní text; jazyk takové representace se nazývá interlingva; původně se usilovalo o universální reprezentaci, která by mohla být prostředníkem mezi libovolnými dvěma přirozenými jazyky; dnes jsou cíle skromnější
Koncepce s interlingvou je výhodná zvláště pro vícejazyčné systémy: pro jeden jazyk stačí modul pro analýzu a syntézu. Problémy: (a) jak má interlingva vypadat (slovník, typy vztahů mezi větnými členy), (b) interlingvou je obtížné explicitně vyjádřit všechny informace přítomné v "povrchových" jazykových strukturách, v průběhu analýzy se tyto informace ztrácejí a v syntéze pak chybějí (problém lexikální a strukturní synonymie).

Transfer

“překlad” reprezentace vstupního textu do reprezentace textu výstupního; na rozdíl od interlingvy jsou tyto reprezentace závislé na konkrétním jazyku; ani reprezentace vstupu a výstupu v jednom jazyce nemusí být totožná
U vícejazyčných systémů je třeba počítat s modulem transferu pro každou dvojici jazyků, pro nereversibilní transfer se dvěma moduly na dvojici. (Pro všechny kombinace n jazyků to znamená $n(n-1)$ modulů transferu.)
Na druhé straně (a) odpadá problém specifikace jazykově nezávislé representace, (b) analýza a syntéza může jít jen do takové hloubky, jak to překlad vyžaduje, (c) v transferu lze využít podobností mezi jazyky.

Flexibilní nebo víceúrovňový překlad

Pro některé dvojice nebo i skupiny jazyků může stačit nižší rovina (čeština – slovenština, švédština – norština), takže je výhodné, když úroveň (obecného) systému, na které probíhá transfer, je možné přizpůsobit konkrétním jazykům.
Překládat na různých úrovních může být výhodné i v rámci jedné dvojice nebo skupiny jazyků, a to v závislosti na typu výrazu. Tak např. syntaktický transfer s interlingvou lze někdy vhodně zkombinovat, aby moduly transferu mohly být co nejmenší — pro terminologickou oblast lexika je možné zavést jazykově neutrální ekvivalenty.

5.3 Dávkový nebo interaktivní systém

Dávkový systém

lidský zásah je možný až po skončení procesu (překladu věty, odstavce, celého textu), to nevylučuje použití specializovaných prostředků pro editaci přeloženého textu (paralelní zobrazení vstupu a výstupu, přístup do lexikálních databází), dávkové zpracování vstupu může být výhodné pro překlad velkého množství textů bez vysokých nároků na kvalitu

Interaktivní systém

systém vyžaduje od uživatele spolupráci (HAMT) v průběhu překladu za účelem

- doplnění lingvistických informací (chybějící slovníkové heslo, chybějící kategoriální interpretace, rámec, neznámá gramatická konstrukce, chybějící lexikální nebo strukturní ekvivalent)
- schválení zvolené možnosti (víceznačnost v analýze, překladová víceznačnost v transferu, souznačnost v syntéze, řešená výběrem nejpravděpodobnější varianty)
- volby z více možností (lexikální a strukturní víceznačnost a souznačnost)

Dobře navržený systém se neptá na každý případ, ale jen tehdy, nemůže-li věc rozhodnout ani po analýze celé věty. Systém se také může učit, může si např. doplňovat chybějící lingvistické znalosti nebo dělat statistiku odpovědí uživatele a po překročení určitého limitu se neptat vůbec nebo jen žádat potvrzení správnosti.
Interaktivní systémy lze dále rozdělit na ty, které předpokládají, že

a) uživatel ovládá jazyk vstupního i výstupního textu,

b) uživatel ovládá jen jazyk vstupu,

c) uživatel ovládá jen jazyk výstupu. Podle toho je třeba formulovat dotazy na uživatele.

Příklady:

1) *she put on a lot of weight, she put on an additional train, she put me on*

2) *she put on her clothes, she put on her shoes, she put on her hat, she put on her glasses*

Otázky uživateli:

a) *put on* zde znamená *přibrat, přidat, 'tahat za nos,' obléknout si, obout si, nasadit si, vzít si* ?

b) (pro Angličana:) co je (nevyjádřeným) předmětem slovesa *put on*: *weight, a service, a person, clothes, shoes, hat, glasses*?

c) (pro Čecha, věta *she put on her hat*:) z následujících možností vyberte nejvhodnější: *přibrala svůj klobouk, přidala svůj klobouk, tahala svůj klobouk za nos, oblékla si klobouk, obula si klobouk, nasadila si klobouk, vzala si klobouk*.

Nebo z češtiny do japonštiny:

a) *nosit* $\longrightarrow$ *hakobu, haoru, haku, kaburu, hameru, šimeru, cukeru, kakeru, hajasu, kiru*?

b) (pro Čecha:) co je (nevyjádřeným) předmětem slovesa *nosit*: *břemeno*; *kabát* nebo *sako*; *boty* nebo *kalhoty*; *klobouk*; *prsten* nebo *rukavice*; *opasek*, *vázanka* nebo *šátek*; *brož* nebo *spona*; *brýle* nebo *náhrdelník*; *knír*; ostatní nebo blíže neurčené případy

c) (pro Japonce:) ...

5.4 Pravidla nebo příklady

Možnosti systému vždy pokulhávají za realitou, a to přinejmenším z těchto důvodů:
(a) přirozený jazyk se neustále vyvíjí (nová slova, nové významy, nová spojení slov)
(b) jazyk je obtížné popsat beze zbytku, i kdyby se nevyvíjel
(c) popis jazyka je obtížné formalizovat
Tradičně se lingvistické znalosti v systémech strojového překladu vyjadřují "ručně" sestavenými formalizovanými pravidly a slovníkem. Takové systémy jsou náchylné k zaostávání ze všech výše uvedených důvodů. U praktických systémů se tedy hlavním problémem stává nedostatečnost zdrojů informací o gramatice a slovníku příslušných jazyků.
'Ručně' zkompilované zdroje nepostačují minimálně v těchto třech ohledech: (a) rozsahem (kvantitou), (b) problémy s údržbou a rozšiřováním, (c) kvalitou.
Existují pokusy tato omezení a z toho vyplývající problémy překonat. Systémy EBMT (*Example-Based Machine Translation*) částečně překonávají omezení formalisovatelnosti znalostí o jazyce, systémy založené na statistickém zpracování částečně překonávají omezení nepopsatelnosti jazyka v jeho úplnosti. Jinou cestou k částečnému překonání bariér nedostatečnosti zdrojů je snaha o vícenásobnou využitelnost a s tím spojenou standardizaci.
Ve výkladových a překladových slovnících pro lidi bývá kromě ekvivalentu jeden nebo více příkladů užití. Lidé často rozumějí a překládají na základě podobnosti s jinými výrazy. Také některé systémy s úspěchem používají databáze příkladů při analýze, syntéze, ale zejména v transferu jako alternativu nebo doplnění systému pravidel. Příklady je možné sestavovat ručně, přejímat ze slovníků, manuálně nebo automaticky excerpovat z (paralelních) textů, doplňovat na základě zpracovaných textů. Místo (přepisovacích) pravidel, která v tradičním pojetí strojového překladu řeší jazykové jevy a korespondence mezi výrazy různých jazyků explicitně, se v 'překladu podle příkladu' (*example-based translation*) nebo 'překladu podle analogie' (*translation by analogy*) používá procedura, která využívá databáze ekvivalencí mezi konkrétními výrazy příslušných jazyků — příklady překladů. K tomu je třeba:

1. zkompilovat (nejlépe automaticky) dvoujazyčnou databázi ekvivalencí konkrétních výrazů;

2. implementovat algoritmus, který ke každému výrazu na vstupu vyhledá v databázi ekvivalencí nejbližší výraz příslušného jazyka; při hledání nejbližšího výrazu se může uplatnit tezaurus s hierarchií výrazů, v níž se hledá nejspecifičtější výraz nadřízený oběma porovnávaným

3. vytvořit nástroj, který po porovnání překládaného výrazu a výrazu v databázi sestaví dostatečně abstraktní schéma, které bude zaplněno tím, čím se vstup od příkladu v databázi liší; většinou se počítá s lexikálními rozdíly

Typické využití: určení ekvivalentního výrazu (ekvivalentní konstrukce) v transferu, řešení lexikální i strukturní víceznačnosti v analýze. Místo příkladů jako takových se obvykle používají jejich representace (obvykle závislostní, např. DLT/BKB, IBM Japan, UMIST-ATR). Tradiční přístup (pravidla) a koncepce s příklady se tedy dá vhodně kombinovat: např. tradiční syntaktická analýza a sytéza s příklady pro transfer, nebo tradiční transferová pravidla jako nouzové řešení v případě nenalezení vhodného příkladu.

Databáze příkladů:

wildlife protection – ochrana volně žijících zvířat
radiation protection – ochrana před radiací
police protection – policejní ochrana
Tourists eat hamburgers. – Turisté jedí hamburgery.
Acid eats metal. – Kyselina ničí kov.

Vstup:

endangered species protection, tropical forest protection, ozone layer protection, protection of inhabitants
noise protection, drugs-related hazards protection
government protection, neighbourhood watch protection
She eats a lot of vegetables.
Exhaust fumes eat the marble statues.

Výhody EBMT:

- inkrementální vývoj, je-li něco přeloženo špatně, stačí přidat nový příklad, odpadá modul překladových pravidel, který se obtížně sestavuje a doplňuje
- lze bezprostředně využít zkušenosti překladatelů
- modul příkladů je málo závislý na konkrétním systému
- systém je odolný proti selhání v důsledku chybějící informace, vždy vydá nějaký výsledek
- lze určit míru spolehlivosti výsledku, chtít druhý a další nejlepší výsledek

Problémy:

- kolik příkladů je třeba? 1000000? (ATR: 10000)
- jak dlouhé mají příklady být? slova i věty jsou víceznačné, fráze (syntagmata) již méně
- v některých případech (idiomy, slovosled) systém není schopen najít správné řešení i za předpokladu přítomnosti ideálních příkladů v databázi (vliv širšího kontextu)
- vstupu může odpovídat více různých příkladů, se vzájemnými přesahy

5.5 Překladová paměť (Translation Memory)

– databáze ekvivalentů, většinou vět a (terminologických) výrazů

5.5.1 Využití:

- opakování vět nebo výrazů uvnitř dokumentu
- opakování vět nebo výrazů v různých dokumentech, různé verze téhož dokumentu
- stejná nebo příbuzná témata, ne nutně technická ("birdwatching")
- originál v elektronické podobě, překlad ve stejném formátu
- čím víc a déle, tím lépe

5.5.2 Výhody:

- využití minulé práce (i cizí)
- dodržení stejné terminologie
- stejné prostředí pro různé formáty

5.5.3 Komponenty:

- program pro párování vět (alignment)
- program pro údržbu databáze vět
- program pro údržbu (a využívání) databáze termínů
- editor překladu
- filtry (konverzní programy)

5.5.4 Pracovní postup:

1. nastavení segmentace textu
2. párování hotových překladů s originály
3. zadání údajů o typu textu (jazyky, formáty, téma, zákazník)
4. naplnění databáze paralelními texty
5. naplnění terminologické databáze
6. vytvoření "projektu", zadání údajů o typu textu
7. import textu, který se má přeložit:
    (a) předběžný překlad celého textu nebo každé věty:
    (b) jen přesně odpovídající věty v databázi
    (c) "fuzzy" matching
    (d) kombinace z úseků vět v databázi vět a z terminologické databáze
    (e) zachování neměnných kousků z originálu (číselné výrazy, formátovací kódy)
    (f) formální úpravy (čísla)
8. revize, editování nebo vytvoření překladu

9. uložení přeložené věty do databáze
10. využití přeložené věty při předběžném překladu zbytku dokumentu
11. kontrola pravopisu
12. export, revize, import revidovaného překladu
13. uložení definitivního znění do databáze
14. uložení terminologických výrazů do databáze
15. export výsledného textu

5.5.5 Další možnosti:

- úprava segmentace v průběhu překladu
- paralelní konkordance
- extrakce ekvivalentů z textu
- export a import databáze
- kontrola terminologie
- distribuce částí projektu mezi více překladatelů
- vytvoření obrácené databáze
- více databází pro jeden projekt
- práce v běžném editoru
- nastavitelný SGML filtr

WordFast: http://champollion.net/
TRADOS TRANSLATOR'S WORKBENCH: http://www.trados.com/
Déjà Vu: http://www.atril.com/
SDL SDLX: http://www.sdlintl.com/products/sdlx/nav/main.htm
STAR TRANSIT: http://www.star-ag.ch/products/
IBM TRANSLATION MANAGER: http://www-4.ibm.com/software/ad/translat/tm/index.html
Translation Memory:
http://dmoz.org/Computers/Software/Globalization/Computer_Aided_Translation/Translation_Memory/

More Translation Memory Tools (not many more, but good ones)
by Suzanne Assénat-Falcone
http://www.accurapid.com/journal/12TM.htm

How To Select the Right CAT Tool Solution
http://www.languagepartners.com/reference-center/whitepapers/howto.htm

What you need to know about Translation Memories
http://www.multilingualwebmaster.com/library/trmemories.html

5.6 Statistické a symbolické metody

Zjišťování frekvence výskytu různých jevů (morfémů, slov, kolokací, syntaktických konstrukcí) v jednom jazyce a kontrastivní zkoumání založená na frekvenci výskytu jejich ekvivalentů v paralelních textech se často užívají jako podklad pro rozhodnutí o různých variantách řešení systému a jako doplnění symbolických lingvistických informací. Existují však také pokusy využít statistické metody jako jediný nástroj pro strojový překlad — systém IBM založený na zpracování korpusu paralelních textů. Současný trend se vyznačuje snahou kombinovat oba přístupy.

Stručný popis statistické metody

- Překlad z francouzštiny do angličtiny, Brown et al., 1989
- Inspirace z oboru kódování signálu: na začátku byly anglické věty, které byly 'zakódovány' (nebo 'zkresleny šumovým kanálem') do vět francouzských. Naším úkolem je najít původní věty jazyka anglické, t.j. dekódovat věty francouzské. Francouzština se považuje za jazyk cílový a angličtina za jazyk zdrojový.
- Překladem určité anglické věty S (source) může být kterákoli francouzská věta T (target). Každé dvojici S a T přisoudíme podmíněnou pravděpodobnost $P(T|S)$, kterou interpretujeme jako pravděpodobnost, že překladatel přeloží větu S větou T.
- Problém strojového překladu pak převedeme na problém následující: máme-li větu T, hledáme takovou větu S, která byla (překladatelem) přeložena jako věta T. Nejlepšího možného výsledku pak dosáhneme v případě, kdy zvolíme nejpravděpodobnější větu S, za předpokladu, že známe T.
- Chceme tedy takovou větu S, která maximalizuje pravděpodobnost $P(S|T)$. Podle Bayesovy věty pak můžeme napsat:

$$P(S|T) = \frac{P(S)P(T|S)}{P(T)}$$

.

Jmenovatel nezávisí na S, a tak stačí najít takové S, které maximalizuje součin $P(S)P(T|S)$.

  - $P(S)$ .... pravděpodobnost S v modelu zdrojového jazyka (volba a pořadí slov ve větě S)
  - $P(T|S)$ .... pravděpodobnost překladu věty S větou T (jaká slova z S vedla ke slovům v T).

- Pro systém strojového překladu je tedy třeba:

  1. spočítat pravděpodobnosti jazykového modelu
  2. spočítat pravděpodobnosti překladového modelu
  3. najít takovou větu S, která maximalizuje součin obou pravděpodobností

Jazykový model

Pravděpodobnost výskytu určitého slova závisí na slovech předcházejících. Pravděpodobnost výskytu určitého řetězce slov lze převést na součin pravděpodobností výskytu všech slov v řetězci:
$P(s_1 s_2 ... s_n) = P(s_1)P(s_2|s_1)...P(s_n|s_1 s_2 ... s_{n-1})$
I pro krátké věty je však takový výpočet kapacitně velmi náročný, proto se běžně počítá s jedním až dvěma předcházejícími slovy, tedy s tzv. bigramy nebo trigramy.
Jazykový model lze ověřit např. pokusem najít správný slovosled, máme-li k dispozici slova původní věty.

Překladový model

Předpoklad: věty T a S spolu korespondují po jednotlivých slovech, přičemž jedno slovo z S se většinou překládá jedním slovem z T, ale nemusí být také přeloženo vůbec, nebo může být přeloženo více slovy. $P(T|S)$ je pak součin pravděpodobností, že:

1. slovo s z S je přeloženo jako (též prázdný) řetězec slov z T, pro jednoslovný řetězec: $P(t|s) = P(Jean|John)$
2. slovo s je přeloženo jako n slov, $n \geq 0$ – tzv. plodnost (fertility): P(fertility=1|John)
3. došlo k nějakému 'zkreslení' (distortion), t.j. změně pozice překladu slova s v T: $P(i|j,l)$; i je pozice v T, j je pozice v S, l je délka T

Parametry překladového modelu:

- množina pravděpodobností $P(n|e)$ pro každé anglické slovo e a pro plodnost n od 0 až do nějaké rozumné hranice (zde: 25)
- množina pravděpodobností překladu $P(f|e)$ pro každé francouzské slovo f a každé anglické slovo e
- množina pravděpodobností zkreslení $P(i|j,l)$ pro každou pozici i v T, j v S a délku l věty T. Hodnoty i,j,l jsou v rozsahu od 1 do 25.

Hledání optima

Věta S, která maximalizuje součin $P(S)P(T|S)$, se hledá tak, že k větě T se hledá nejpravděpodobnější S postupným přidáváním nejpravděpodobnějších slov.

Odhad parametrů

Pro jazykový model stačí anglický text, pro překladový model jsou nutné texty paralelní. Hansard corpus má v anglické i francouzské části asi 100 mil. slov. Z nich byly vybrány 3 mil. párů vět, z nichž 99 procent je přeloženo 1:1.

První experiment: Z korpusu bylo vybráno 9000 nejčastějších anglických a 9000 nejčastějších francouzských slov, ostatní byla označena jako neznámá anglická/francouzská slova. Ze 40 000 párů vět (asi 800000 slov v každém jazyce) bylo odhadnuto asi 81 milionů parametrů. Na začátku iterace se předpokládalo, že každé francouzské slovo je stejně pravděpodobným překladem každého anglického slova, že každá plodnost od 0 do 25 je stejně pravděpodobná pro každé anglické slovo a že pro každou pozici v S a každou délku T je každá pozice v T stejně pravděpodobná.

Druhý experiment: 1000 anglických slov, 1700 francouzských z překladů vět, které neobsahovaly jiná anglická slova. Ze 117 000 párů vět bylo odhadnuto 17 mil. parametrů. Věty neobsahovaly žádná neznámá slova. Z 570 000 anglických vět byly odhadnuty parametry anglického jazykového modelu. 73 francouzských vět z jiné části korpusu bylo 'dekódováno' takto: 5% přesně, 25% alternativně, 18% odlišně, 15% chybně, 37% negramaticky.