Strojový překlad

učební text pro PLIN019

Brno, květen 2013 Vít Baisa

Tento text je určen výhradně pro studenty semináře PLIN019 *Strojový překlad*.

Text vznikl transformací slajdů promítaných na semináři. Veškeré zdroje, které jsem při přípravě použil, řádně *necituji*. Hlavní předlohou pro kapitolu *Statistické systémy* byla kniha *Statistical Machine Translation* od Philippa Koehna.

Vznik tohoto textu byl podpořen v rámci OPVK projektu INOVA.CZ.

Vít Baisa

Obsah

1. Překlad

1.1 Obecný překlad

Překlad je převod textu ze zdrojového jazyka do jazyka cílového.

Tlumočení je ústní překlad mluveného jazyka.

- ▷ odborný překlad × literární překlad
- ▷ přesná reprodukce × volná převodová parafráze

Pro překlad slova je rozhodující kontext.

—Maimonidés, 12. stol.

Každé slovo je element vytržený z celkového jazykového systému a jeho vztahy k jiným segmentům systému jsou v jednotlivých jazycích rozdílné.

Každý význam je element z celého systému segmentů, v něž mluvčí rozčleňuje skutečnost.
V jazyce Mohave: otec ženy ≠ otec muže

—Werner Winter

Překlad je jako žena: buď věrný, nebo hezký.

—poučený anonym

	odborný styl	publicistická a rétorická próza	umělecká próza a drama	volný verš	pravidelný verš	hudební text (libreto)	dabing
denotativní význam	i	i	i	i	i	i-v	i-v
konotativní význam	v	i-v	i	i	i	i	i
stylistické zařazení slova	i-v	i	i	i	i	i	i
větná stavba	v	i-v	i	i	i	i	i
opakování (rytmus, rým)	v	v	v	i-v	i	i	i-v
délka a výška samohlásek	v	v	v	i-v	i-v	i	i
způsob artikulace	v	v	v	i-v	i-v	i-v	i

Obrázek 1.1: Jaké vlastnosti zdroje mají být zachovány? – *J. Levý* , **i**nvariabilní, **v**ariabilní

Teorie překladu, Jiří Levý

- ▷ musí reprodukovat
  - slova originálu
  - ideje originálu
- ▷ se má dát číst jako originál
- ▷ má být čten jako překlad
- ▷ by měl
  - obrážet styl originálu
  - ukazovat styl překladatelův
  - být čten jako text náležející do doby
    - originálu
    - překladatelovy
- ▷ může k originálu něco přidávat nebo z něho vynechávat
- ▷ by neměl nikdy k originálu nic přidávat a vynechávat

Translatologie

- ▷ vědní obor zabývající se překladem textů mezi jazyky a sémiotickými systémy
- ▷ otázky přesnosti (věrnosti), přeložitelnosti
- ▷ překlad mezi kulturními oblastmi, obdobími
- ▷ větev deskriptivní (kritika a dějiny) × aplikovaná (praxe)
- ▷ 60.–70. léta vznik, lingvistická orientace
- ▷ 80. léta přiblížení literární teorii
- ▷ 90. léta obrat k překladateli jako jedinci

Co by měl překladatel znát (Levý)

- ▷ zdrojový jazyk
- ▷ cílový jazyk
- ▷ věcný obsah textu: dobové reálie, obor (u odborného překladu)

Překlad má působit jako umělecké dílo.

– Jiří Levý

Strojovému překladu jde nutně o atomizování věty na nejjednodušší srovnatelné jednotky; uměleckému naopak o převádění co nejvyšších celků.

—Jiří Levý o strojovém překladu

Typy překladu podle Romana Jakobsona

- ▷ **mezijazykový** – převod mezi různými jazyky
- ▷ **vnitrojazykový** – převod v rámci jazyka, např. do jiného nářečí, do spisovné podoby apod.
- ▷ **meziznakový** – převod mezi různými znakovými systémy

Otázky překladu

- ▷ Je vůbec přesný překlad mezi jazyky možný?
- ▷ Jak se pozná, že w_1 je překladový ekvivalent slova w_2?
- ▷ anglické typy větru: airstream, breeze, crosswind, dust devil, easterly, gale, gust, headwind, jet stream, mistral, monsoon, prevailing wind, sandstorm, sea breeze, sirocco, southwester, tailwind, tornado, trade wind, turbulence, twister, typhoon, whirlwind, wind, windstorm, zephyr
- ▷ jak přeložit slova jako *alkáč*, *večerníček*, *telka*, *čoklbuřt*, *knížečka*, *ČSSD* . . . ?
- ▷ film Kód Navajo – neznámý jazyk pomáhá utajit informace před nepřítelem = šifra

1.2 Jazykový relativismus

- ▷ vlastnosti jazyka podstatně ovlivňují naše vnímání světa
- ▷ vlastnosti různých jazyků se výrazně liší
- ▷ jejich mluvčí tudíž žijí v různých, nepřevoditelných světech

Hranice mého jazyka znamenají hranice mého světa.

—Ludwig Wittgenstein

Kdyby byl Aristoteles z kmene Dakotů, jeho logika by nabyla zcela odlišné podoby.

—Fritz Mauthner

- ▷ **teorie matrice** (mould theories): jazyk a myšlení jsou totožné, myslíme jazykem
- ▷ **teorie pláště** (cloak theories): jazyk je na povrchu, za ním je složitá spleť myšlenek

Kam patří *jazykový relativismus*?

Sapir-Whorfova hypotéza

- historicky významná teorie psycholingvistiky
- 30. léta 20. století, Edward Sapir, původ v jazykovém relativismu
- srovnání pojmů v indiánských a indoevropských jazycích
- teorie rozpracována Benjaminem Lee Whorfem
- později kritizována
- testovatelná podoba hypotézy (pojmy pro barvy) prokázala spíše opak

2. Úvod do strojového překladu

> *Strojový překlad je obor počítačové lingvistiky zabývající se návrhem, implementací a aplikací automatických systémů (programů) pro překlad textů s minimálním zásahem člověka.*

Např. používání elektronických slovníků při překladu nepatří do strojového překladu.

Předmět zájmu

Uvažujeme pouze odborné texty:

- webové stránky
- technické manuály
- vědecké dokumenty
- prospekty, katalogy
- právnické texty
- obecně texty z omezených domén

Nuance na různých jazykových vrstvách v umělecké literatuře jsou mimo schopnosti současných nástrojů NLP.

Ve skutečnosti je výstup z SP vždy revidován. Mluví se o *před-překladu* resp. o nutné *post-editaci*. Ta je někdy nutná i u člověka, ovšem systémy SP dělají zcela rozdílné chyby.

Chyby člověka a stroje

Pro člověka jsou typické chyby:

- špatné předložky (*I am in school*)
- chybějící členy (*I saw man*)
- špatný čas (*Uviděl jsem – I was seeing*), . . .

Pro počítač jsou typické zejména chyby významové:

- *Kiss me, honey.*
- *Ludvig dodávka Beethoven*

Přímé metody zlepšení kvality strojového překladu

- omezení vstupu na:
  - podjazyk (krátké věty, oznamovací věty)
  - doménu (právnické texty)
  - typ dokumentu (patentové dokumenty)
- pre-processing textu (např. ruční syntaktická analýza)

2.1 Základní pojmy

- **přesnost** (accuracy, precision)
- **srozumitelnost** (intelligibility)
- **plynulost** (fluency)
- **zdrojový** (výchozí) jazyk (source language, SL)
- **cílový jazyk** (target language, TL)
- **korpus** (corpus, corpora)
- **víceznačnost** (ambiguity)

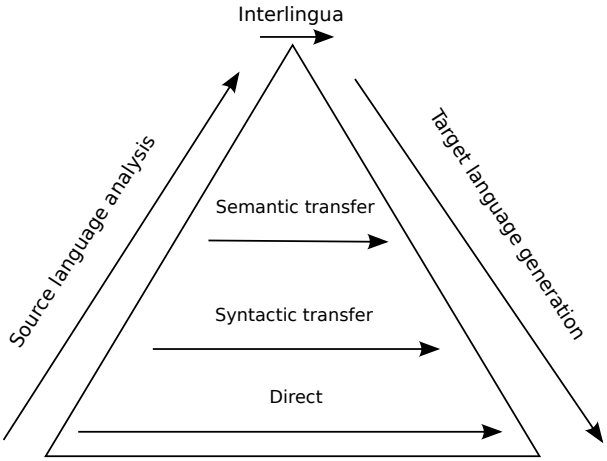
2.2 Rozdělení systémů strojového překladu

Klasifikace podle přístupu (approach)

- pravidlový (znalostní) strojový překlad
  rule-based, knowledge-based – RBMT, KBMT
  - transferový
  - interlingua
- statistický strojový překlad
  statistical machine translation – SMT
- hybridní strojový překlad
  hybrid machine translation – HMT, HyTran

Klasifikace podle interakce s uživatelem

- (ruční překlad)
- ruční překlad s pomocí počítače
  machine-aided human translation – MAHT
- automatický překlad s interagujícím překladatelem
  human-aided machine translation – HAMT



Obrázek 2.1: Vauquoisův trojúhelník

- plně automatický překlad
  fully automated high-quality (M)T – FAHQMT

HAMT a MAHT někdy souhrnně označovány jako CAT – computer-aided translation.

Klasifikace podle směru a četnosti překladu

Podle četnosti:

- dvojjazyčné systémy (bilingual)
- vícejazyčné systémy (multilingual)

Podle směru:

- jednosměrné (unidirectional)
- obousměrné (bidirectional)

2.3 Reálie z oblasti strojového překladu

Systémy strojového překladu

Apertium (RBMT, open-source), **Babelfish** (Yahoo), **Caitra** (CAT systém), **ČESILKO** (česko-slovenský překlad), **EuroTra** (ambiciózní projekt EK), **Google Translate**, **Logos** (OpenLogos, jeden z nejstarších MT systémů), **METEO** (překlad předpovědí, angličtina, francouzština), **Moses** (open-source MT systém), **Pangloss** (example-based MT), **Rosetta** (obsahuje logickou analýzu), **Systran** (jeden z nejstarších MT systémů), **Trados** (překladová paměť, CAT systém), **Verbmobil** (překlad řeč↔řeč mezi němčinou, angličtinou a japonštinou), . . .

Konference, workshopy

- ACL – Annual meetings of the Association for Computational Linguistics

- NIST – National Institute of Standards and Technology
- Translating and the Computer (Londýn)
- RANLP – Recent Advances in Natural Language Processing
- MT Summit
- The Xth Conference of the Association for Machine Translation in the Americas
- LREC – Language Resources and Evaluation Conferences
- `www.wikicfp.com`

(Elektronické) informační zdroje

- odkazy na stránkách předmětu
- MT Archive
- `www.statmt.org`
- ACL Anthology
- Translation Journal

Instituce

- IAMT – International Association for Machine Translation:
  - EAMT – European Association for Machine Translation
  - AMTA – The Association for MT in the Americas
  - AAMT – The Asian-Pacific Association for MT
- META-NET – sdružuje evropská MT pracoviště
- British Computer Society Natural Language Translation Group
- UK MFF ÚFAL
- Obec překladatelů (překlady krásné literatury)
- Jednota tlumočníků a překladatelů
- Ústav translatologie, FF UK

2.4 Nástin vývoje SP

Počátky, 40. léta 20. století

Motivace pro strojový překlad po 2. světové válce

- období informačního boomu
  - 1922 – pravidelné rozhlasové vysílání BBC
  - 1923 – rozhlasové vysílání v ČR
  - 1936 – pravidelné televizní vysílání BBC
  - 1953 – začíná TV vysílání v ČR
- rozvoj počítačů
  - nultá generace – Z1–3, Colossus, ABC, Mark I,II
  - první generace – ENIAC, MANIAC

V roce 1947 měla RAM kapacitu 100 čísel a sčítání dvou čísel trvalo 1/8 sekundy!

Ranné názory na strojový překlad

- ▷ překlad je často opakovaná činnost – věřilo se, že bude tuto proceduru možné počítačem napodobit
- ▷ úspěchy použití počítačů v kryptografii: vhodné i pro strojový překlad?

> *When I look at an article in Russian, I say: This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.*
>
> —Warren Weaver

50. léta, MT boom

První impulsy

V roce 1950 rozesílá Weaver memorandum 200 adresátům, ve kterém nastiňuje některé problémy strojového překladu.

- ▷ víceznačnost jako častý jev
- ▷ průnik logiky a jazyka
- ▷ souvislosti s kryptografií
- ▷ univerzální vlastnosti jazyka

Zájem o strojový překlad podnícen na několika pracovištích. Do té doby pouze na University of London vedené A. Boothem. Zejména na MIT, University of Washington, University of California, Harvard, Georgetown, . . .

Témata a první výměny zkušeností

- ▷ morfologická, syntaktická analýza
- ▷ reprezentace významu a znalostí
- ▷ tvorba a práce se slovníky
- ▷ 1952 – první veřejná konference na MIT
- ▷ 1954 – předvedení systému pro strojový překlad

Georgetown experiment

První funkční prototyp strojového překladu.

- ▷ 50 vět (zřejmě pečlivě vybraných)
- ▷ spolupráce s IBM
- ▷ slovník obsahoval 250 slov
- ▷ překlad z ruštiny do angličtiny
- ▷ gramatika pro ruštinu obsahovala 6 pravidel

Demonstrace systému vyvolala nadšení. MT bylo očividně možné. Následně odstartovalo mnoho nových projektů, hlavně v USA a Rusku.

Vývoj v 50. letech

- MT oblast podnítila rozvoj a výzkum na poli
  - teoretické lingvistiky (Chomsky)
  - počítačové lingvistiky
  - umělé inteligence (60. léta)
- s větším pokrytím kvalita strojového překladu klesala
- i nejlepší systémy (GAT, Georgetown, RE→EN) poskytovaly nepoužitelný výstup

60. léta, zklamání ze slabých výsledků

- i přes nevalné výsledky přetrvával optimismus
- Yehoshua Bar-Hillel píše v roce 1959 kritiku stavu strojového překladu
- tvrdí, že počítače nejsou schopné provádět lexikální desambiguaci
- fully automated high-quality translation (FAHQT) podle Bar-Hillela stěží dosažitelné

> *Little John was looking for his toy box. Finally, he found it. The box was in the pen. John was very happy.*
>
> —Yehoshua Bar-Hillel, příklad pro desambiguaci

Výdaje na projekty strojového překladu se začaly snižovat.

ALPAC report

- Automatic Language Processing Advisory Commitee
- organizace pod U.S. National Academy of Science
- analýzy a vyhodnocení kvality a použitelnosti systémů SP
- doporučila omezit výdaje na podporu strojového překladu
- negativní dopad na strojový překlad jako vědeckou oblast
- chyba spočívala zřejmě v silném podceňování složitosti porozumění přirozenému jazyku
- vývoj strojového překladu v Evropě a Japonsku pokračoval nepřerušeně dál
- celých 15 let trvalo než SP v USA znovu získal vážnost a původní postavení

Renesance strojového překladu

70. léta

TAUM-METEO

- překlad z angličtiny do francouzštiny
- od r. 1977 používán pro překlad předpovědí počasí
- vyvinut na University of Montreal

Systran

- velmi populární překladový systém
- využíván v projektu Apollo a Sojuz (od r. 1975)
- od r. 1976 oficiální MT systém používaný Evropským hospodářským společenstvím

80. léta

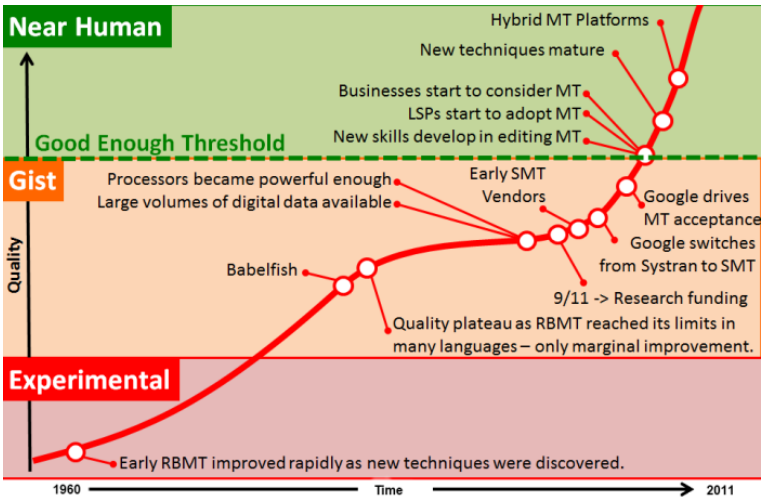
- vývoj zejména pravidlových systémů s použitím interlinguy
- první daty řízené systémy (Example-based MT)
- rozmach komerčních MT systémů

90. léta

- výzkum statistického překladu (IBM)
- pravidlové systémy stále dominují

po roce 2000

- statistické systémy převládají
- kvalita pravidlových systémů je zvyšována statistickými metodami (hybridní metody)
- přidávání dalších jazykových párů



Obrázek 2.2: Příliš positizní prognóza pro vývoj SP

2.5 Strojový překlad dnes

- výpočetní technika a datové struktury dovolují práci s miliardami slovy
- Google 1PB sort, rok 2008
  - bilión 100bytových záznamů
  - 6 hodin
  - 4 000 počítačů
  - 48 000 disků
- vývoj MT systému dostupné komukoli
- roste počet paralelních korpusů
- přibývají jazykové zdroje pro minoritní jazyky
- kvalita překladu neustále (byť pomalu) stoupá

- SMT rulezz
- intenzivní sběr paralelních dat
- vývoj systémů vzhledem k hodnotícím metrikám
- USA: zájem o angličtinu jako TL
- EU: překlad mezi úředními jazyky EU (EuroMatrix)
- korporace (Microsoft) zaměřeny na En jako SL
- velké páry (En↔Sp, En↔Fr): velmi dobrý překlad
- SMT obohacována syntaxí
- Google Translate jako gold standard
- morfologicky bohaté jazyky jsou opomíjeny
- En-* a *-En páry převažují

Motivace pro strojový překlad ve 21. století

- překlad webových stránek pro pochopení obsahu
- metody pro výrazné urychlení překladatelské práce (překladové paměti)
- extrakce a vyhledávání informací mezi jazyky (cross-lingual IR)
- instantní překlad elektronické komunikace (ICQ)
- překlad na mobilních zařízeních

2.6 Výzvy pro strojový překlad

Lexikální výběr

Výběr správného překladového ekvivalentu:

- homonymie: *slaď, pila, baby, ženu*
- polysémie: *run, bank, klíč, kohout*
- synonymie: *kluk, chlapec, hoch; dívka, holka, děvče*

Word order	English equivalent	Proportion of languages	Example languages
SOV	"I you love."	45%	Hindi, Japanese, Latin
SVO	"I love you."	42%	English, Mandarin, Russian
VSO	"Love I you."	9%	Hebrew, Irish, Zapotec
VOS	"Love you I."	3%	Baure, Fijian, Malagasy
OVS	"You love I."	1%	Apalai, Hixkaryana, Tamil
OSV	"You I love."	0%	Jamamadi, Warao, Xavante

Obrázek 2.3: Slovosled

Volný slovosled

Čím více morfologicky bohatší, tím volnější slovosled. Katka **snědla** kousek koláče.

- ▷ Kati megevett egy szelet tortát → Katie eating a piece of cake
- ▷ Egy szelet tortát Kati evett meg → Katie ate a piece of cake
- ▷ Kati egy szelet tortát evett meg → Katie ate a piece of cake
- ▷ Egy szelet tortát evett meg Kati → Katie ate a piece of cake
- ▷ Megevett egy szelet tortát Kati → Katie eating a piece of cake
- ▷ Megevett Kati egy szelet tortát → Katie ate a piece of cake

2.7 Shrnutí

- ▷ strojový překlad patří mezi AI-complete problémy
- ▷ máme k dispozici obrovskou výpočetní sílu
- ▷ tržní potenciál je větší než kdy dřív
- ▷ je stále co zlepšovat
- ▷ statistické metody se zdají vhodnější (rychlé, levné)

3. Pravidlové systémy

3.1 Rozdělení systémů

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)
- ▷ k tomu, abychom přeložili *vrána na větvi* nemusíme vědět, že vrána je pták a létá
- ▷ 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 (PC Translator)
  - tři kroky (+ transfer)

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

Přístup pomocí interlinguy

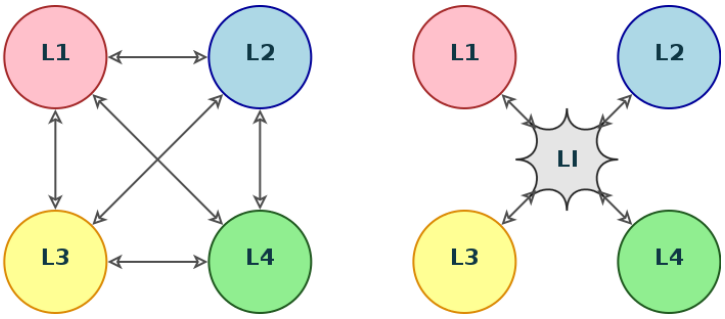
- ▷ předpokládá, že je možné SL konvertovat do reprezentace, která je nezávislá na jazyku
- ▷ interlingua musí být jednoznačná (unambiguous)
- ▷ z této podoby (interlingua) je generován TL
- ▷ analýza SL je jazykově závislá, ale nezávislá na TL
- ▷ analogicky syntéza TL
- ▷ SL a TL nepřijdou do styku
- ▷ 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

3.2 Tokenizace

- ▷ rozdělení vstupního řetězce do tokenů



Obrázek 3.1: Interlingua vs. transferové KBMT

- 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

3.3 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, dělaje, 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, positiv, 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`

Morfologická disambiguace

- ▷ v mnoha případech: více morfologických 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
- ▷ 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 $\rightarrow$ `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* $\rightarrow$ 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

Universal POS tags

- ▷ počet značek se v různých jazycích značně liší
- ▷ $\rightarrow$ snaha o zjednodušení.
- ▷ vytvořeno mapování pro cca 25 jazyků s *tree banky*.

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	k5eAaImIp3nS, k3p3gMnPc4, k3p3gInPc4, k3p3gNnSc4, k3p3gNnPc4, k3p3gFnPc4, k0	k5eAaImIp3nS
pro	k7c4	k7c4
správný	k2eAgMnSc1d1, k2eAgMnSc5d1, k2eAgInSc1d1, k2eAgInSc4d1, k2eAgInSc5d1, . . . (+ 18)	k2eAgInSc4d1
růst	k5eAaImF, k1gInSc1, k1gInSc4	k1gInSc4
důležité	k2eAgMnPc4d1, k2eAgInPc1d1, k2eAgInPc4d1, k2eAgInPc5d1, k2eAgFnSc2d1, k2eAgFnSc3d1, k2eAgFnSc6d1, k2eAgFnPc1d1, k2eAgFnPc4d1, k2eAgFnPc5d1, k2eAgNnSc1d1, k2eAgNnSc4d1, k2eAgNnSc5d1, . . . (+ 5)	k2eAgNnSc1d1

Tabulka 3.1: Morfologická disambiguace

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)

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

Tabulka 3.2: Universal POS tags

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

Tabulka 3.3: Odhalování POS na základě gramémů, možné problémy: *myší, west, fotbal, . . .*

2. porovnej s trénovacími daty
3. vytvoř sadu pravidel pro změnu značek
4. ohodnoť 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
- **třitisícedvěstědevadesátpět** znaků

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
- pro některé jazyky těžké až nemožné
- POS tagging dosahuje pro různé jazyky různé kvality
- typicky kolem 95 %

3.4 Lexikální rovina

Slova a slovníky ve strojovém překladu

- propojení mezi jazyky typicky na úrovni slov (slovníky)
- u transferových systémů i na úrovni syntaktických struktur
- pro KBMT systémy jsou slovníky nezbytné
- typicky 10k a více položek
- GNU-FDL slovník
- kolik položek ve slovníku potřebujeme / chceme?
  → pojmenované entity, slang
  listém – jazyková položka, kterou nelze odvodit na základě principu kompozicionality (*slaměný vdovec*)
- v jakém tvaru mají být slova ve slovníku?
  → lemmatizace
- jak odlišit jednotlivé významy pro potřeby strojového překladu?
  → budování slovníků pro strojový překlad
- kolik různých významů má smysl rozlišovat?
  → granularita

Víceznačnost

- slovům odpovídají významy
- co je ale význam? pro počítač potřebujeme formální popis
- počítač je diskrétní, význam je zřejmě spojitý
- *muž* – dospělý člověk mužského pohlaví
- co 17letý člověk mužského pohlaví?

Spojitost významu

špalek

?

židle

Typy víceznačnosti

Víceznačnost se projevuje na více úrovních:

- morfologie (-s, viz výše)
- slova (oko)
- slovní spojení (bílá vrána)
- věty (I saw a man with a telescope.)

- homonymie: náhodný jev
  - úplná homonymie: líčit, kolej
  - částečná h.: los, stát
- polysémie je přirozená: oko, táhnout, . . .

Granularita

Kolik významů má slovo *kočka*?

- malá kočkovitá šelma chovaná v domácnostech
- malá nebo středně velká šelma s hustým kožichem
- samice kočkovité šelmy
- kožešina na límci, kolem krku nebo ramen
- kocovina
- věc připomínající vlastnost kočky
- druh důtek

Pro strojový překlad může stačit granularita překladového slovníku: slovo x má tolik významů jako má překladových ekvivalentů ve slovníku.

Granularita – oko

- zrakový orgán
- klička, smyčka, kroužek z různého materiálu
- věc připomínající tvarem oko (morské oko)
- jednotka v kartách, loterii
- druh karetní hry

Granularita – dát (SSJČ)

- ▷ odevzdat do vlastictví, darovat, prodat
- ▷ vyžádat, způsobit (dá to mnoho práce)
- ▷ umístění něčeho
- ▷ dopřát, dovolit, připustit (nedej pane)
- ▷ projevit nedostatek odporu (dát se ošidit)
- ▷ přikázat (dát něco udělat)

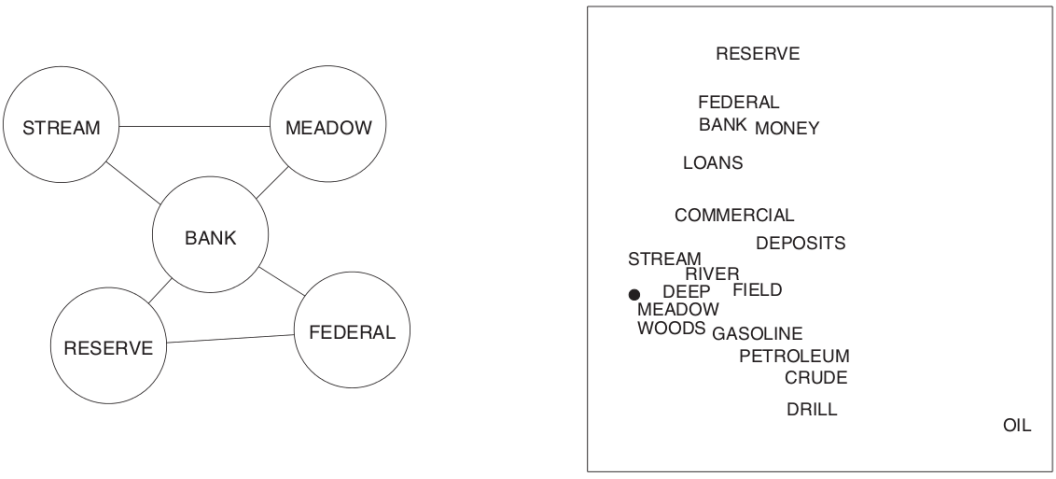
VerbaLex uvádí 32 (!) významů (nezvratné varianty).

Granularita – malý

- ▷ neveliký rozměry, počtem, časovým rozsahem
- ▷ nedospělý
- ▷ slabý, nevydatný (malý rozhled)
- ▷ nevýznamný (malý pán)
- ▷ téměř (malý zázrak)
- ▷ děvčátko (malá)
- ▷ přihrávka vlastnímu brankáři (malá domů)

Reprezentace významu

- ▷ nejčastější způsob: banka významů
- ▷ graf: významy jsou uzly, sémantické relace jsou hrany
- ▷ prostor: významy jsou body, podobné významy jsou prostorově blízko

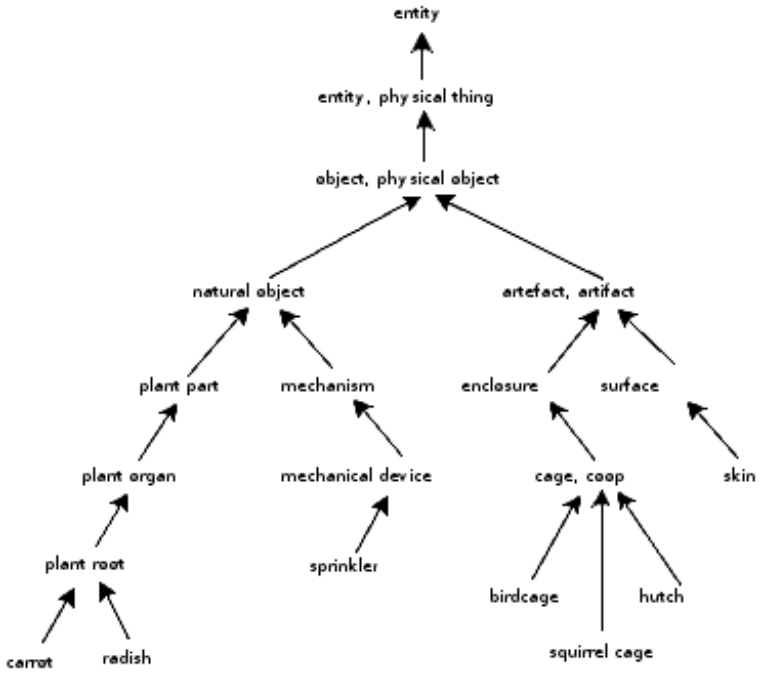


Obrázek 3.2: Typy reprezentace významů

Sémantická síť – WordNet

- ▷ **literál** dát:8, **synset** louže:1, kaluž:1, tratoliště:1

- ▷ sémantické relace: hypero-, hypo-, holo-, meronymum
- ▷ 150k slov, 117k synsetů: n, adj, v a adv
- ▷ WN používán jako referenční banka významů



Obrázek 3.3: Ukázka sémantické sítě

VerbaLex

- ▷ WordNet neobsahuje syntaktické vazby, morfosyntaktické omezení
- ▷ synsety (6 256)
  atakovat:1, útočit:2, dorážet:3, napadnout:6
- ▷ valenční rámce (mačkat:1) a sloty (19 247)
  $\mathsf{AG}^{kdo1}_{person:1} + \mathsf{VERB} + \mathsf{OBJ}^{co4}_{object:1} + (\mathsf{PART}^{v\ čem6}_{hand:1})$
- ▷ sémantické role
  I: ABS, ISUB, AG, KNOW, PAT, VERB, ... (29)
  II: abstraction:1, person:1, artifact:1, body part:1, ... (10^3)
- ▷ další omezení:
  předložkové pády, životnost, slovní druhy, obligatornost
- ▷ synsety napojeny na WordNet

Word Sense Disambiguation

- ▷ nalezení významu slova v daném kontextu
- ▷ pro člověka triviální, pro PC těžké
- ▷ jde o klasifikační úlohu
- ▷ potřebujeme konečný inventář významů
- ▷ při použití WN: pro dané slovo určit konkrétní synset

- ▷ kvalita se těžko vyhodnocuje (SensEval, SemEval)
- ▷ přesnost kolem 90 %

Metody WSD

Problém: jak přeložit **box in the pen** (Bar-Hillel).

- ▷ hloubkové (deep)
  - využívají znalosti o světě (common sense)
  - nejsou vhodné pro obecný jazyk (spíše omezené domény)
  - znalosti typu: ptáci umí létat, jablka rostou na stromě, . . .
  - metody založené na reprezentaci znalostí, na slovníku
  - Leskův algoritmus: shoda slov z okolí se slovy ze slovníku patřícími ke konkrétnímu významu
- ▷ povrchové (shallow)
  - využívají slova z kontextu
  - levnější, rychlejší implementace
  - různé metody strojového učení (klasifikační problémy)
  - učení s učitelem (supervised), bez učitele (unsupervised)
  - možné použít varianty Brillova algoritmu

Shrnutí

- ▷ význam hlavně na úrovni slov (překladové slovníky)
- ▷ WSD zcela klíčový pro pravidlové systémy
- ▷ počet slov se mezi jazyky řádově liší
- ▷ lexikální víceznačnost je bottleneck strojového překladu

3.5 Syntaktická rovina

Syntaktická analýza

- ▷ další patro v MT trojúhelníku
- ▷ snaha o konečný popis nekonečného množství frází, vět
- ▷ konečným způsobem = gramatikou
- ▷ vstup (většinou): morfologicky označkovaná data
- ▷ výstup: syntaktický strom, les, graf
- ▷ úkol SA: pro danou gramatiku a vstupní větu vrať všechny možné derivační stromy
- ▷ potenciálně milióny různých analýz (viz Synt)
- ▷ pro analýzu je potřeba:
  - výběr formalismu
  - napsání gramatiky
  - implementace algoritmu analýzy
- ▷ v současnosti většina **parserů** využívá statistiky

Gramatické formalismy

- bezkontextová gramatika: na levé straně mohou být pouze jednoduché **neterminály**
- regulární gramatika: bezkontextová + pravidla pouze typu
  `N -> epsilon | A | bB`
- tree-adjoining: podobné bezkontextovým, přepisují se stromy nikoli znaky (řetězce)

Typy analýz

- top-down analýza (shora): hledá se taková nejlevější derivace, která generuje analyzovaný řetězec
- bottom-up analýza (zdola): hledají se pravidla, která přepíší vstupní řetězec na výslednou posloupnost pravidel

K čemu je syntaktická analýza?

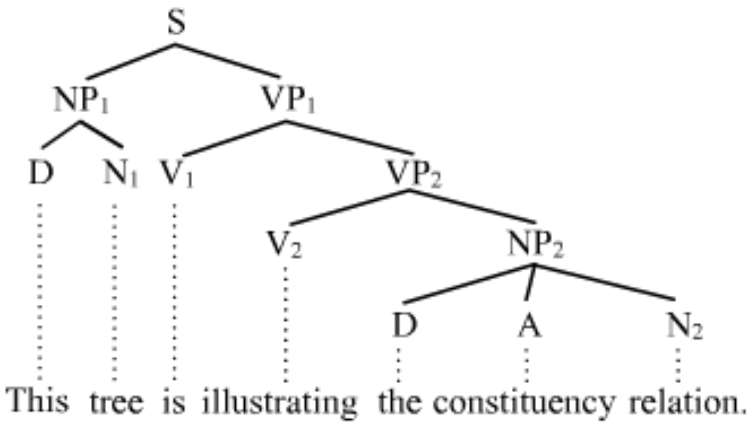
- sémantická interpretace zdrojového kódu (informatika)
- mezistupeň k sémantické reprezentaci věty
- transferové systémy: konečný počet transferových pravidel pro nekonečný počet možných frází
- WSD: zachycení vztahů na větší vzdálenosti (širší kontext)
- jaká slova k sobě patří a jaká ne

Syntaktická víceznačnost

- *I saw a man with a telescope.*
  Uzřel jsem muže (s) dalekohledem.
- *I'm glad I'm a man, and so is Lola.*
  Jsem rád, že jsem muž a Lola také.
- *Someone ate every tomato.*
  Někdo snědl všechna rajčata.
  Každé rajče bylo někým sněženo.
- *Lvíče škádlí dítě.*
  A child teases a lion cub.
  A lion cub teases a child.
- *Letadlo spadlo do pole za lesem.*
- *Ženu holí stroj.*
  Ženu holý stroj.
- *Zabít ne propustit.*
  Ibis, redibis nunquam per bella peribis.
- *Rodiče by mu mohli závidět.*

Garden path

- *The man returned to his house ... was happy.*
- *The man whistling tunes ... pianos.*
- *Time flies like an arrow; fruit flies like a banana.*
- *Ženu krávy ... nezajímají.*



Obrázek 3.4: Frázový strom

Vyhodnocení kvality syntaktické analýzy

- ▷ jaká analýza je nejlepší? (viz experiment)
- ▷ vyhodnocení kvality je obtížné a interpretace je sporná
- ▷ nejlepší analyzátory dosahují přesnosti cca 85 %

Frázová struktura jazyka

- ▷ jeden z nejstarších formalismů
- ▷ gramatika obsahuje přepisovací pravidla
- ▷ nejčastěji bezkontextová gramatika
- ▷ zachycuje, jak se skládají fráze: **konstituenty**

```
S   -> NP VP
VP  -> ADV V | V ADV
NP  -> DET N
DET -> the | a | an
N   -> cat | dog
...
```

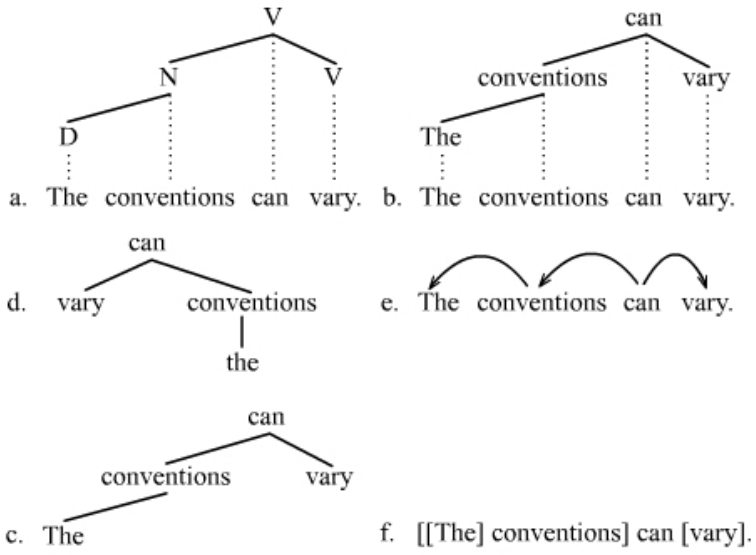
Analýza: **the dog runs fast** (shora a zdola)

Závislostní struktura

- ▷ zachycuje závislosti mezi slovy
- ▷ strom neobsahuje **neterminály**
- ▷ hlava a závislá slova
- ▷ vhodné pro jazyky s volným slovosledem (čeština)

Constituency vs. Dependency

- ▷ každé paradigma vhodné pro něco jiného



Obrázek 3.5: Závislostní strom

- ▷ složky: pevný slovosled, koordinace
- ▷ nevýhoda: neschopnost zachytit neprojektivitu souvislým složkovým stromem
  neprojektivní závislost = závislost mezi dvěma slovy oddělenými ve větě třetím slovem, které nezávisí na žádném z nich
  I saw a man with a dog yesterday which was a yorkshire terrier.
- ▷ závislosti: volný slovosled, morfosyntaktická shoda
- ▷ nevýhoda: neschopnost zachytit doplněk (dvojí závislost)
  Babička seděla u stolu shrbená. (doplněk)
  Babička seděla u stolu shrbeně. (PUZ)
- ▷ lze převádět mezi sebou nebo kombinovat: hybridní stromy

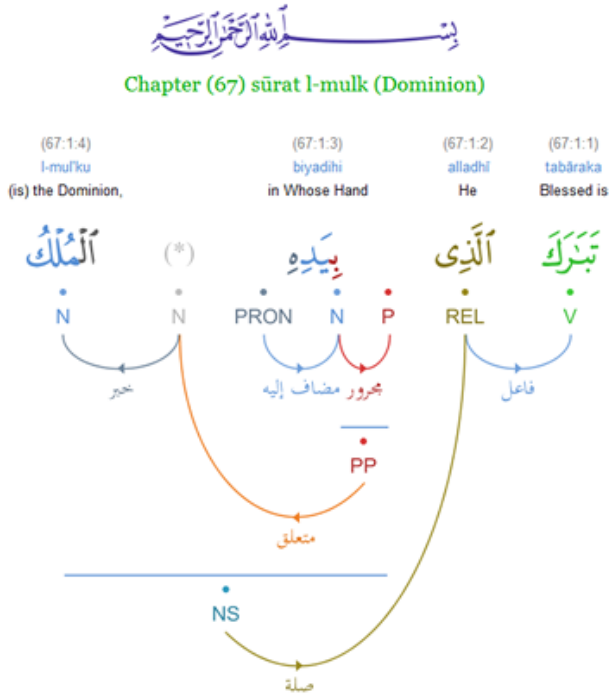
Intermezzo – hledání slov a vět splňujících podmínku

Slovní tvary jako ve scrabble.

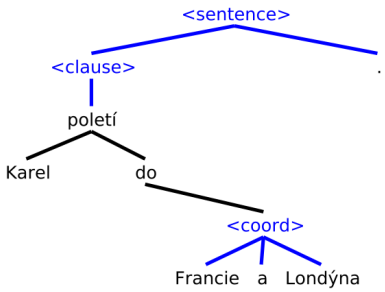
- ▷ slovo obsahující 3x „r“ reproduktor
- ▷ slovo obsahující 3 po sobě jdoucí diakritická znaménka jednodušší
- ▷ věta obsahující 4x po sobě jdoucí „se“ nesnese se se sestrou
- ▷ slovo, 5 písmen, význam i retrográdně tokej, jelen
- ▷ slovo, které má význam i v češtině i v angličtině mat, user
- ▷ slovo, které obsahuje dvě zvířata (nepřekrývají se) rusalka, sobeckost

3.6 Sémantika a logika

- ▷ reprezentace totálního významu nemožná: znalosti světa, smyslové vnímání, mezilidské vztahy, neverbální komunikace, ...
- ▷ některé transferové systémy nevyžadují sémantickou analýzu



Obrázek 3.6: Hybridní strom



Obrázek 3.7: Hybridní strom II

- ▷ hranice mezi syntaxí a sémantikou často zastřená (deep analysis)
- ▷ další úroveň jazyka: pragmatika (řečové akty)
- ▷ logika: jak velký je průnik s jazykem? Je logika pro MT nezbytná?
- ▷ argumenty proti IL: význam je subjektivní, významy jsou často jazykově, kulturně, historicky závislé

Sémantické role

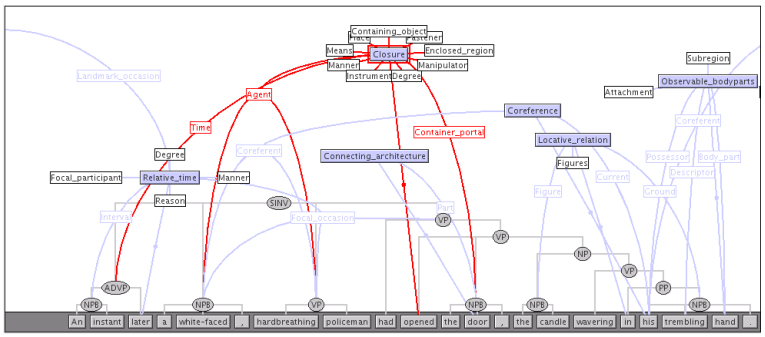
- ▷ syntaxe umožňuje odhalit sémantické vztahy
- ▷ konstituenty vět odpovídají **sémantickým rolím**
- ▷ vztah **predikátu** a ostatních větných členů
- ▷ také **semantic case**, **thematic role**, **theta role**
- ▷ agent, causer, instrument, manner, patient, result, time, source
- ▷ různé množiny rolí, viz např. VerbaLex (29 rolí)

Dítě **škádlí** **lvíče.**
AG/SUBJ PRED/V PAT/OBJ

A child (SUBJ) **teases** (PRED/V) **a lion cub** (PAT/OBJ).
A lion cub (SUBJ) **teases** (V) **a child** (OBJ).

FrameNet

- ▷ elektronický „slovník“ **sémantických rámců**
- ▷ rámec popisuje věc, stav či děj a jeho účastníky
- ▷ situace: děj *vaření* zahrnuje kuchaře, jídlo, nádobu na vaření, zdroj tepla atd.
- ▷ rámec `Apply_heat`, role `Cook, Food, Heating_instrument`, ...
- ▷ 800 rámců, 10k lex. jednotek, 120k anotovaných vět
- ▷ zdroj pro automatické přiřazování sémantických rolí



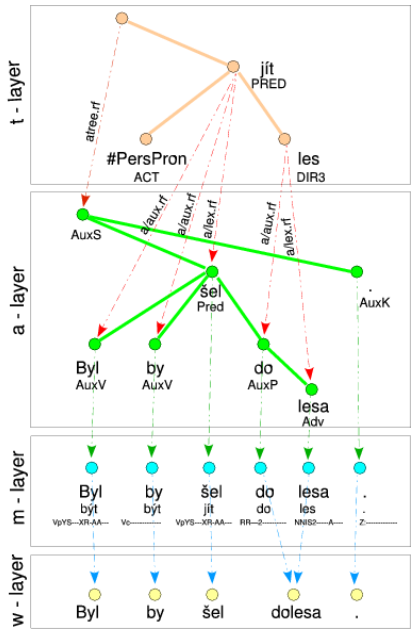
Obrázek 3.8: Ukázka anotací dat z FrameNet

Prague Dependency TreeBank 2.0

- ▷ aplikace teorií Pražského lingvistického kroužku
- ▷ funkční generativní popis jazyka
- ▷ rovina: fonologická a fonetická, morfonologická, morfematická, povrchová syntax a
- ▷ tektogramatická rovina – rovina významu jazyka
- ▷ *nižší rovina je formou vyšší a vyšší rovina funkcí nižší*
- ▷ 2M morfologicky, 1,5M syntakticky a 800k sémanticky označkovaných slov z novinových článků v ČNK
- ▷ koreference a aktuální členění větné
  Petr dal Petře kytici. Pak ji vzal a dal do vázy.
- ▷ uzly pro nevyjádřená slova
- ▷ vazby mezi uzly na různých úrovních

Transferový systém TectoMT

- ▷ vysoká modularita
- ▷ maximální rozložení úkolů do série bloků – scénáře
- ▷ bloky jsou Perl moduly, komunikují přes API
- ▷ struktura systému odpovídá struktuře PDT
- ▷ vnitřní reprezentace jazyka: stromy v tmt formátu odvozeném od PML pro PDT



Obrázek 3.9: Jazykové úrovně v PDT

- ▷ bloky umožňují masivní zpracování dat, paralelizace
- ▷ bloky mohou implementovat pravidlové, stochastické či hybridní metody
- ▷ zpracování:
  1. konverze do formátu tmt
  2. aplikace scénáře
  3. konverze do výstupního formátu

TectoMT – jednoduchý blok

Převod anglických negativních částic na příznaky sloves.

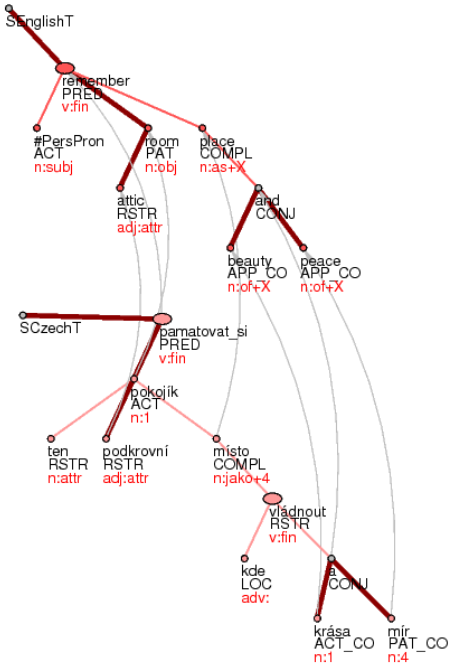
```
sub process_document {
  my ($self,$document) = @_;

  foreach my $bundle ($document->get_bundles()) {
    my $a_root = $bundle->get_tree('SEnglishA');

    foreach my $a_node ($a_root->get_descendants) {
      my ($eff_parent) = $a_node->get_eff_parents;
      if ($a_node->get_attr('m/lemma')=~/^(not|n\'t)$/
          and $eff_parent->get_attr('m/tag')=~/^V/ ) {
        $a_node->set_attr('is_aux_to_parent',1);
      }
    }
  }
}
```

Interlingua systém KBMT

- ▷ morfologická: získání základních slovních tvarů



Obrázek 3.10

- syntaktická: na úrovni vět, využívá nějaký formalismus a odpovídající parser
- sémantická: zachycení významu lexikálních jednotek, vztahů mezi slovy, většinou na úrovni vět; většinou omezená na doménu (ontologie)
- pragmatická, analýza diskurzu: nad úrovní vět; anafory, záměr, řečové akty

Syntéza

- vyčlenění obsahu: co je výstup, co má čtenář domyslet
  Koupil jsem si nový mobil. Nový mobil má velký display. Nový mobil má velká tlačítka.
- pořadí propozic
  Nový mobil má velký display. Koupil jsme si nový mobil.
- lexikální výběr (odpovídá WSD)
- syntaktický výběr
  Uvařil jsem guláš. Guláš byl mnou uvařen.
- uspořádání konstituent
  Uvařil jsem guláš. Guláš jsem uvařil.
- koreference: např. vložení anafor
  Koupil jsem nový mobil. Má velký display.
- generování povrchových struktur (řetězce znaků)

Shrnutí

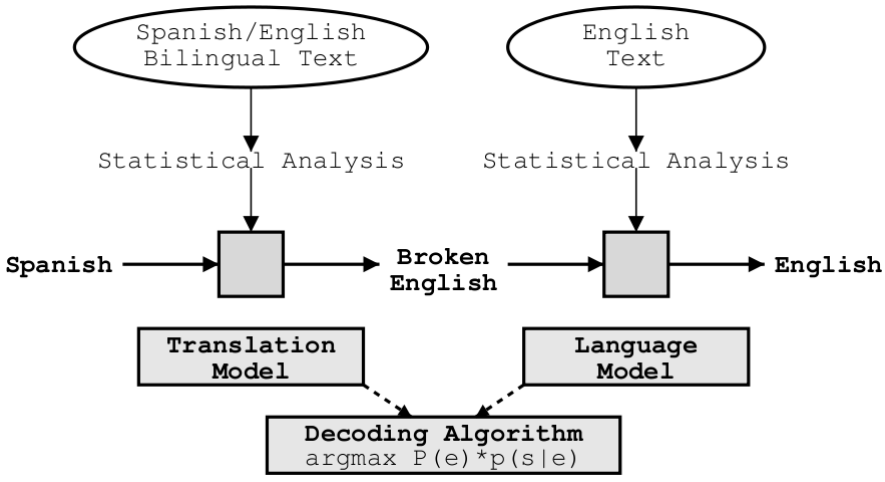
- pravidlové systémy na ústupu
- statistické systémy dosahují lepších výsledků
- mnoho lingvistických jevů je těžké rozlišit i pro člověka (mezianotátorská shoda)
- mnoho metod z pravidlových systémů vylepšují výkon statistickému MT

- ▷ vývoj RBMT je spíš pomalejší
- ▷ v mnohých oblastech se vedou dlouholeté spory

4. Statistické systémy

4.1 Úvod

- pravidlové systémy motivovány lingvistikou
- SMT inspirován teorií informace a statistikou
- v současnosti mnoho společností se zaměřením na SMT: Google, IBM, Microsoft, Language Weaver (2002)
- 50 miliónů stránek denně přeložených pomocí SMT
- **gisting**: stačí, má-li překlad nějaký užitek, nepotřebujeme přesný význam; nejčastější užití MT na internetu



Obrázek 4.1: Schéma statistického strojového překladu

Nástroje SMT

- GIZA++: IBM modely, zarovnávání na úrovni slov

- ▷ SRILM: trénování jazykových modelů
- ▷ IRST: trénování velkých jazykových modelů
- ▷ Moses: frázový dekodér, trénování modelů
- ▷ Pharaoh: předchůdce Mosese
- ▷ Thot: trénování frázových modelů
- ▷ SAMT: tree-based modely

Data pro SMT – (paralelní) korpusy

- ▷ Linguistics Data Consorcium (LDC): paralelní korpusy pro páry arabština-angličtina, čínština-angličtina atd.
  Gigaword korpus (angličtina, 7 mld slov)
- ▷ Europarl: kolekce textů Evropského parlamentu (11 jazyků, 40 M slov)
- ▷ OPUS: paralelní texty různého původu (lokalizace software)
- ▷ Acquis Communautaire: právní dokumenty Evropské únie (20 jazyků)

Pravidelné události v oblasti SMT, soutěže

Většinou roční vyhodnocování kvality SMT. Tvorba testovacích sad, manuální vyhodnocování dat, referenční systémy.

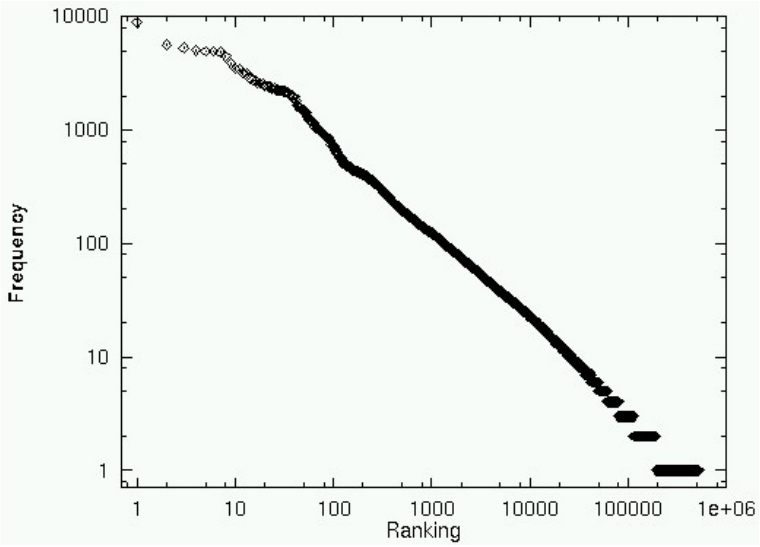
- ▷ NIST: National Institute of Standards and Technology; nejstarší, prestižní; hodnocení překladu arabštiny, čínštiny
- ▷ IWSLT: mezinárodní workshop překladu mluveného jazyka; překlad řeči; asijské jazyky
- ▷ WMT: Workshop on SMT; překlady mezi evropskými jazyky

Základy SMT

Slova

- ▷ pro SMT v drtivé většině případů základní jednotka = slovo
- ▷ v mluvené řeči slova neoddělujeme: jak je od sebe oddělíme?
- ▷ SMT systémy provádí de-tokenizaci
- ▷ překlad samotný je většinou s lowercase textem
- ▷ jaká slova má angličtina $\rightarrow$ jaká slova jsou v anglických korpusech
- ▷ *the* tvoří 7 % anglického textu
- ▷ 10 nejčastějších slov (tokenů) tvoří 30 % textu (!)
- ▷ *Zipfův zákon*: r rank (pořadí ve frekvenčním seznamu slov), f = frekvence výskytu slova, c = konstanta; platí $r \times f = c$
- ▷ překlepy, čísla, vlastní jména, názvy a cizí slova

Zipfův zákon



Věty

- ▷ syntaktická struktura se v jazycích liší
- ▷ vkládání funkčních slov, která jsou typická pro daný jazyk (*the*, interpunkce)
- ▷ přerovnávání: *er wird mit uns gehen* $\rightarrow$ *he will go with us*
- ▷ některé jevy nelze přeložit na úrovni věty: anafory
- ▷ úroveň celého dokumentu: téma (topic) může pomoci při volbě vhodného překladového ekvivalentu
- ▷ v textu o jeskynních živočiších zřejmě nebude překládat *bat* jako *pálka*

Paralelní korpusy

- ▷ základní datový zdroj pro SMT
- ▷ volně dostupné jsou řádově 10 a 100 miliónů slov veliké
- ▷ je možné stáhnout paralelní texty z internetu
- ▷ vícejazyčné stránky (BBC, Wikipedie)
- ▷ problém se zarovnáním dokumentů, odstavců, . . .
- ▷ srovnatelné korpusy (comparable corpora): texty ze stejné domény, ne přímé překlady: New York Times – Le Monde
- ▷ Kapradí – korpus překladů Shakespearových dramat (FI)
- ▷ InterCorp – ručně zarovnané beletr. texty (ČNK, FFUK)

Zarovnávání vět

- ▷ věty si neodpovídají 1:1
- ▷ některé jazyky explicitně nenaznačují hranice vět (thajština)
- ▷ *It is small, but cozy. – Es is klein. Aber es ist gemütlich.*
- ▷ pro věty $e_1, \cdots e_{n_e}$ a $f_1, \cdots f_{n_f}$
- ▷ hledáme páry $s_1, \ldots s_n$
- ▷ $s_i = (\{f_{\text{start}-f(i)}, \cdots f_{\text{end}-f(i)}\}, \{e_{\text{start}-e(i)}, \cdots e_{\text{end}-e(i)}\})$

P	typ zarovnání
0.98	1–1
0.0099	1–0 nebo 0–1
0.089	2–1 nebo 1–2
0.011	2–2

Základy pravděpodobnosti pro SMT

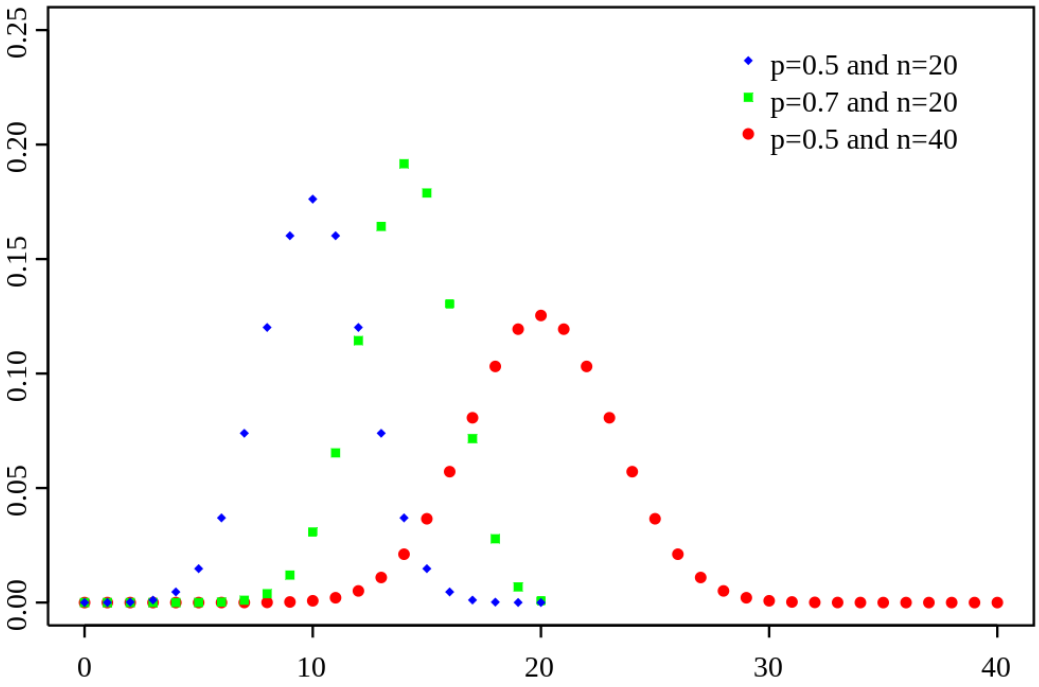
Pravděpodobnostní rozložení

- graf hodnot pravděpodobnosti pro elementární jevy náhodné veličiny
- **rovnoměrné**: hod kostkou, mincí (diskrétní veličina)
- **binomické**: vícenásobný hod

$$b(n,k;p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- **normální, Gaussovo**: spojité, dobře aproximuje ostatní rozložení; zahrnuje rozptyl

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



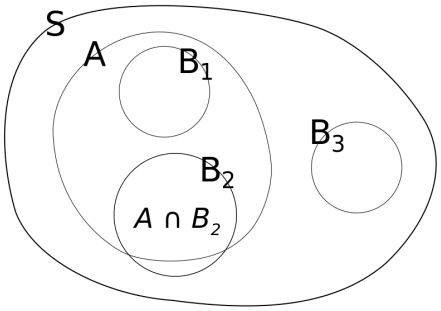
Obrázek 4.2: Binomické rozložení

Základní pojmy

- náhodná proměnná, pravděpodobnostní funkce, ...
- máme data, chceme spočítat rozložení, které nejlépe tato data vystihuje

- ▷ **zákon velkých čísel**: čím víc máme dat, tím lépe jsme schopni odhadnout pravděpodobnostní rozložení
- ▷ např.: hod falešnou kostkou; výpočet π
- ▷ nezávislé proměnné: $\forall x, y : p(x, y) = p(x).p(y)$
- ▷ **spojená (joint) pravděpodobnost**: hod mincí a kostkou
- ▷ **podmíněná pravděpodobnost**: $p(y|x) = \frac{p(x,y)}{p(x)}$
  pro nez. proměnné platí: $p(y|x) = p(y)$

Podmíněná pravděpodobnost



Shannonova hra

Pravděpodobnostní rozložení pro následující znak v textu se liší v závislosti na předchozích znacích.

Doplňujeme postupně znaky (malá abeceda a mezera). Některé znaky nesou více informace (jsou uhádnuty později).

Bayesovo pravidlo

$$p(x|y) = \frac{p(y|x).p(x)}{p(y)}$$

- ▷ příklad s kostkou
- ▷ $p(x)$ – prior
- ▷ $p(y|x)$ – posterior

Další pojmy ze statistiky

- ▷ střední hodnota (diskrétní): $\mathrm{E}\, X = \sum_I s_i.p_i$
- ▷ rozptyl: $\sigma^2 = \sum_{i=1}^{n} [x_i - \mathrm{E}(X)]^2 p_i$
- ▷ očekávaná hodnota: $\mathrm{E}[X] = \sum_{x \in X} x.p(x)$

4.2 Princip noisy channel

Vyvinut Shannonem (1948) pro potřeby samoopravujících se kódů, pro korekce kódovaných signálů přenášených po zašuměných kanálech na základě informace o původní zprávě a typu chyb vznikajících v kanálu.

Příklad s OCR. Rozpoznávání textu z obrázků je chybové, ale dokážeme odhadnout, co by mohlo být v textu (jazykový model) a jaké chyby často vznikají: záměna l-1-I, rn-m apod.

$$\begin{aligned} e^* &= \arg\max_e p(e|f) \\ &= \arg\max_e \frac{p(e)p(f|e)}{p(f)} \\ &= \arg\max_e p(e)p(f|e). \end{aligned}$$

SMT – komponenty noisy channel principu

- ▷ jazykový model:
  - jak zjistit $p(e)$ pro libovolný řetěz e
  - čím víc vypadá e správně utvořené, tím vyší je $p(e)$
  - problém: co přiřadit řetězci, který nebyl v trénovacích datech?
- ▷ překladový model:
  - pro e a f vypočítej $p(f|e)$
  - čím víc vypadá e jako správný překlad f, tím vyšší p
- ▷ dekódovací algoritmus
  - na základě předchozího najdi pro větu f nejlepší překlad e
  - co nejrychleji, za použití co nejmenší paměti

4.3 Jazykové modely

Jak pravděpodobné je pronesení české věty s?

- ▷ LM pomáhají zajistit **plynulý výstup** (správný slovosled)
- ▷ p_{LM}(včera jsem jel do Brna) $\geq p_{LM}$(včera jel do Brna jsem)
- ▷ co však s p_{LM}(jel jsem včera do Brna)?
- ▷ LM pomáhají s **WSD v obecných případech**
- ▷ pokud má slovo více významů, můžeme vybrat nejčastější překlad (*pen* $\rightarrow$ pero)
- ▷ ve speciálních textech nelze použít, ale
- ▷ LM pomáhají s **WSD pomocí kontextu**
- ▷ p_{LM}(i go home) $\geq p_{LM}$(i go house)

N-gramové modely

Využití statistického pozorování dat. Některé slova se vyskytují často v určitých dvojicích (*chudý student, vážený pane, pracující lid*), po slovech *I go* je častější *home* než *house* apod.

$$W = w_1, w_2, \cdots, w_n$$

Jak vypočítat $p(W)$? Spočítáme výskyty všech W v datech a normalizujeme je velikostí dat. Pro většinu velkých W však nebudeme mít v datech ani jeden výskyt. Úkolem je zobecnit pozorované vlastnosti trénovacích dat, která jsou většinou řídká (**sparse data**).

Markovův řetězec a Markovův předpoklad

$p(W)$, kde W je posloupnost slov, budeme modelovat postupně, slovo po slovu, užitím tzv. **pravidla řetězu**:

$$p(w_1, w_2, \ldots w_n) = p(w_1)p(w_2|w_1)p(w_3|w_1, w_2)\ldots p(w_n|w_1 \ldots w_{n-1})$$

Jelikož nemáme k dispozici pravděpodobnosti pro dlouhé řetězce slov, omezíme historii na m slov použitím **Markovova předpokladu**:

$$p(w_n|w_1, w_2, \ldots w_{n-1}) \simeq p(w_n|w_{n-m}, \ldots w_{n-2}, w_{n-1})$$

Číslo m nazýváme řádem odpovídajícího modelu. Nejčastěji se používají **trigramové** modely.

Výpočet, odhad pravděpodobností LM

Trigramový model používá pro určení pravděpodobnosti slova dvě slova předcházející. Použitím tzv. **odhadu maximální věrohodnosti** (maximum likelihood estimation):

$$p(w_3|w_1, w_2) = \frac{count(w_1, w_2, w_3)}{\sum_w count(w_1, w_2, w)}$$

w	**počet**	$p(w)$
paper	801	0.458
group	640	0.367
light	110	0.063
party	27	0.015
ecu	21	0.012

Tabulka 4.1: trigram: *(the, green, w)* (1748)

Kvalita a srovnání jazykových modelů

Chceme být schopni porovnávat kvalitu různých jazykových modelů (trénovány na různých datech, pomocí jakých n-gramů, jak vyhlazených apod.).

Dobrý model by měl přiřadit dobrému textu vyšší pravděpodobnost než špatnému textu.

Pokud máme nějaký testovací text, můžeme spočítat pravděpodobnost, jakou mu přiřazuje zkoumaný LM. Lepší LM by mu měl přiřadit vyšší pravděpodobnost.

Cross-entropy (křížová entropie)

$$H(p_{LM}) = -\frac{1}{n}\log p_{LM}(w_1, w_2, \ldots w_n)$$
$$= -\frac{1}{n}\sum_{i=1}^{n}\log p_{LM}(w_i|w_1, \ldots w_{i-1})$$

Křížová entropie je průměrná hodnota záporných logaritmů pravděpodobností slov v testovacím textu. Odpovídá míře nejistoty pravděpodobnostního rozložení (zde LM). Čím menší, tím lepší.

Dobrý LM by měl dosahovat entropie blízké skutečné entropii jazyka. Tu nelze změřit, ale existují relativně spolehlivé odhady (např. **Shannonova hádací hra**). Pro angličtinu je entropie na znak rovna cca 1.3 bitu.

Perplexita

$$PP = 2^{H(p_{LM})}$$

Perplexita je jednoduchá transformace křížové entropie.

Dobrý model by neměl plýtvat p na nepravděpodobné jevy a naopak.

Čím nižší entropie, tím lépe $\rightarrow$ čím nižší perplexita, tím lépe.

Vyhlazování jazykových modelů

Problém: pokud není v datech určitý n-gram, který se vyskytne v řetězci w, pro který hledáme pravděpodobnost, bude $p(w) = 0$.

Potřebujeme rozlišovat p i pro *neviděná data*. Musí platit

$$\forall w.p(w) > 0$$

Ještě větší je problém u modelů vyšších řádů.

Snaha o úpravení reálných počtů n-gramů na očekávané počty těchto n-gramů v libovolných datech (jiných korpusech).

Add-one vyhlazování

Maximum likelihood estimation přiřazuje pravděpodobnost na základě vzorce

$$p = \frac{c}{n}$$

Add-one vyhlazování používá upravený vzorec

$$p = \frac{c+1}{n+v}$$

kde v je počet všech možných n-gramů. To je však velmi nepřesné, neboť všech možných kombinací je většinou řádově víc než ve skutečnosti (Europarl korpus má 86,700 tokenů, tedy víc jak 7,5 mld možných bigramů. Ve skutečnosti má korpus 30 mil. slov, tedy maximálně 30 mil. bigramů.) Vyhlazování nadhodnocuje neviděné n-gramy.

r	**FF**	r^*
0	7 514 941 065	0,00015
1	1 132 844	0,46539
2	263 611	1,40679
3	123 615	2.38767
4	73 788	3,33753
5	49 254	4,36967
6	35 869	5,32929
8	21 693	7,43798
10	14 880	9,31304
20	4 546	19.54487

Tabulka 4.2: Ukázka Good–Turing vyhlazování (Europarl)

Add-α vyhlazování

Nebudeme přidávat 1, ale koeficient α. Ten lze odhadnout tak, aby add-α vyhlazování bylo spravedlivější.

$$p = \frac{c + \alpha}{n + \alpha v}$$

α můžeme experimentálně zjistit: zvolit více různých a hledat pomocí perplexity nejlepší z nich. Typicky bude spíše malé (0.000X).

Deleted estimation

Neviděné n-gramy můžeme vytvořit uměle tak, že použijeme druhý korpus, případně část trénovacího korpusu. N-gramy obsažené v jednom a ne v druhém nám pomohou odhadnout množství neviděných n-gramů obecně.

Např. bigramy, které se nevyskytují v trénovacím korpusu, ale vyskytují se v druhém korpusu milionkrát (a všech možných bigramů je cca 7,5 mld), se vyskytnou cca

$$\frac{10^6}{7.5 \times 10^9} = 0.00013\times$$

Good–Turing vyhlazování

Potřebujeme upravit počet výskytů v korpusu tak, aby odpovídal obecnému výskytu v textu. Použijeme *frekvenci frekvencí*: počet různých n-gramů, které se vyskytují n-krát.

$$r^* = (r + 1)\frac{N_{r+1}}{N_r}$$

Speciálně pro n-gramy, které nejsou v korpusu máme

$$r_0^* = (0 + 1)\frac{N_1}{N_0} = 0.00015$$

kde $N_1 = 1.1 \times 10^6$ a $N_0 = 7.5 \times 10^9$ (Europarl korpus).

metoda	perplexita
add-one	382,2
add-α	113,2
deleted est.	113,4
Good–Turing	112,9

Tabulka 4.3: Srovnání metod vyhlazování (Europarl)

řád	unikátní	singletony
unigram	86 700	33 447 (38,6 %)
bigram	1 948 935	1 132 844 (58,1 %)
trigram	8 092 798	6 022 286 (74,4 %)
4-gram	15 303 847	13 081 621 (85,5 %)
5-gram	19 882 175	18 324 577 (92,2 %)

Tabulka 4.4: Velké jazykové modely – počet n-gramů

Interpolace a back-off

Předchozí metody zacházely se všemi neviděnými n-gramy stejně. Předpokládejme 3-gramy:

nádherná červená řepa
nádherná červená mrkev
I když ani jeden nemáme v trénovacích datech, první 3-gram by měl být pravděpodobnější.
Budeme využívat pravděpodobnosti n-gramů nižších řádů, u kterých máme k dispozici více dat:
červená řepa
červená mrkev

Interpolace

Použijeme interpolaci:

$$p_I(w_3|w1, w_2) = \lambda_1 p(w_3) \times \lambda_2 p(w_3|w_2) \times \lambda_3 p(w_3|w_1, w_2)$$

Pokud máme hodně dat, můžeme věřit modelům vyšších řádů a přiřadit odpovídajícím pravděpodobnostem větší váhu.
p_I je pravděpodobnostní rozložení, proto musí platit:

$$\forall \lambda_n : 0 \leq \lambda_n \leq 1$$
$$\sum_n \lambda_n = 1$$

Kolik je různých n-gramů v korpusu? Europarl, 30 miliónů tokenů.

i:	1	2	3	4	5
	the	castle	is	very	old
	ten	hrad	je	velmi	starý
j:	1	2	3	4	5

4.4 Překladové modely

Lexikální překlad

Standardní slovník neobsahuje informace o tom, jak často se překládá dané slovo na své různé překladové ekvivalenty.

key $\rightarrow$ *klíč, tónina, klávesa*

Jak často jsou zastoupeny jednotlivé překlady v překladech?

key $\rightarrow$ *klíč* (0.7), *tónina* (0.18), *klávesa* (0.12)

Potřebujeme lexikální překladové pravděpodobnostní rozložení p_f s vlastností

$$\sum_e p_f(e) = 1$$
$$\forall e : 0 \leq p_f(e) \leq 1$$

S jakou pravděpodobností se přeloží *babička* $\rightarrow$ *appropriate*?

Zarovnání slov, zarovnávací funkce

Překlady si často neodpovídají v počtu slov ani ve slovosledu. Zavádí se *alignment function*

$$a : j \rightarrow i$$

kde j je pozice odpovídajícího slovo v cílové větě (čeština), i je pozice ve zdrojové větě (angličtina).

a je funkce, tedy pro každé slovo w_e z cílové věty existuje právě jedno slovo w_f ze zdrojové věty.

Zarovnání slov – další případy

- jiný slovosled:
  it was written here
  bylo to zde napsané
  $a : 1 \rightarrow 2, 2 \rightarrow 1, 3 \rightarrow 4, 4 \rightarrow 3$
- jiný počet slov:
  jsem maličký
  i am very small
  $a : 1 \rightarrow 1, 2 \rightarrow 1, 3 \rightarrow 2, 4 \rightarrow 2$
- slova bez překladových ekvivalentů:
  have you got it ?
  máš to ?
  $a : 1 \rightarrow 1, 2 \rightarrow 4, 3 \rightarrow 5$

▷ opačný případ, přidáme nové slovo NULL, pozice 0:
NULL laugh
smát se
$a : 1 \to 1, 2 \to 0$

IBM modely

IBM model 1

Nemůžeme hledat p_f pro jednotlivé věty. Překlad rozložíme do menších kroků, budeme používat p_f pro slova. Tomuto přístupu se říká *generative modeling*.

Překladový model IBM-1 je definován jako

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

kde $\mathbf{e} = (e_1, \ldots e_{l_e})$ je cílová věta, $\mathbf{f} = (f_1, \ldots f_{l_f})$ zdrojová věta, l_e je délka cílové věty, l_f délka zdrojové věty, ϵ je normalizující konstanta, aby byl výsledný součin pravděpodobnostní rozložení. $(l_f + 1)^{l_e}$ je počet všech možných zarovnání mezi **e** a **f**, přičemž k l_f přičítáme 1 kvůli speciálnímu slovu NULL, t je pravděpodobnostní překladová funkce.

Výpočet překladové pravděpodobnosti

Pro výpočet $p(\mathbf{e}, a|\mathbf{f})$ potřebujeme znát hodnotu funkce t pro všechna slova (věty). K tomu budeme využívat základní zdroj pro SMT: **paralelní korpus** se zarovnanými větami. Bohužel nemáme zarovnání slov mezi sebou. To je úkol tzv. **word-alignment**. Ke slovu přichází **expectation-maximization (EM)** algoritmus.

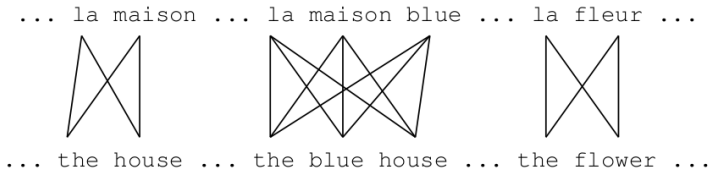
EM algoritmus

1. inicializuj model (typicky uniformní p. rozložení)
2. aplikuj model na data (krok expectation)
   hledáme $p(a|e, f) = \frac{p(e,a|f)}{p(e|f)}$
   kde $p(e|f) = \sum_a p(e, a|f)$
3. uprav model podle dat (krok maximization)
   upravíme počty zarovnání slova w_e na w_f (funkce c) pomocí předchozího
   $c(w_e|w_f; e, f) = \sum_a p(a|e, f) \sum_{j=1}^{l_e} \delta(e, e_j)\delta(f, f_{a(j)})$
   kde $\delta(x, y) = 1 \iff x == y$, jinak 0
4. opakuj E-M kroky dokud je co zlepšovat

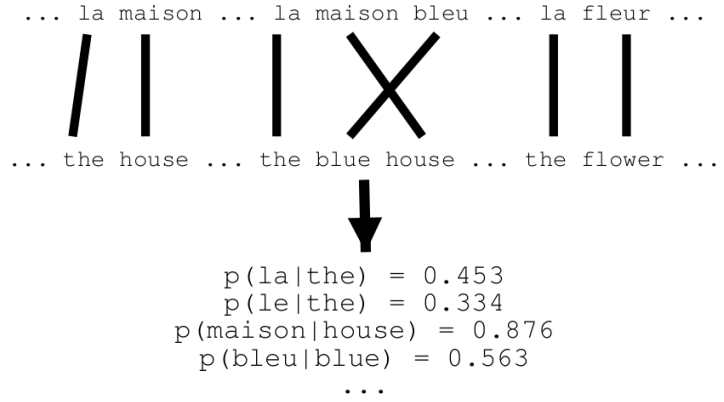
Překladová pravděpodobnost z EM algoritmu

Výsledná překladová pravděpodobnost se vypočítá pomocí c:

$$t(w_e|w_f) = \frac{\sum_{(e,f)} c(w_e|w_f; e, f)}{\sum_{w_e} \sum_{(e,f)} c(w_e|w_f; e, f)}$$



Obrázek 4.3: Ilustrace EM algoritmu – inicializace



Obrázek 4.4: Ilustrace EM algoritmu – výsledná fáze

IBM modely

IBM model 1 je značně jednoduchý. Neuvažuje kontext, neumí přidávat a vypouštět slova. Všechna různá zarovnání považuje za stejně pravděpodobné. Ostatní modely vždy přidávají něco navíc.

- ▷ IBM-1: lexikální překlad
- ▷ IBM-2: přidává model absolutního zarovnání
- ▷ IBM-3: přidává model **fertility**
- ▷ IBM-4: přidává model relativního zarovnání
- ▷ IBM-5: ošetřuje nedostatečnosti předchozích modelů

IBM-2

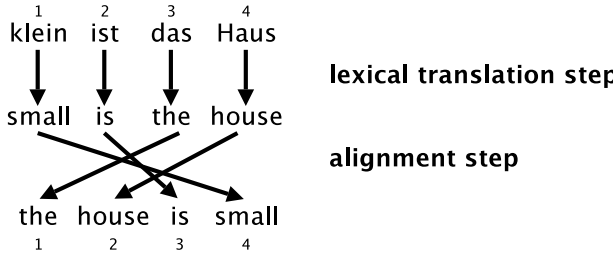
Pro IBM-1 jsou všechny možné překlady s různým uspořádáním slov stejně pravděpodobné. IBM-2 přidává explicitní model pro zarovnání, tzv. **alignment probability distribution**:

$$a(i|j, l_w, l_f)$$

kde i je pozice zdrojového slova, j pozice cílového slova.

IBM-2 – 2 kroky překladu

Překlad se tedy rozdělí na dva kroky. V prvním se přeloží lexikální jednotky, v druhém se podle modelu zarovnání přeskupí přeložená slova.



Obrázek 4.5: Kroky překladu modelu IBM-2

IBM-2

První krok je stejný jako u IBM-1, používá se $t(e|f)$. Funkce a i pravděpodobnostní rozložení a je v opačném směru než je překlad. Obě rozložení se kombinují do vzorce pro IBM-2:

$$p(e, a|f) = \epsilon \prod_{j=1}^{l_e} t(e_j|f_{a(j)}) a(a(j)|j, l_e, l_f)$$

$$\begin{aligned} p(e|f) &= \sum_a p(e, a|f) \\ &= \epsilon \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i) a(i|j, l_e, l_f) \end{aligned}$$

IBM-3

Modely IBM-1,2 neuvažují vlastnost, kdy se jedno slovo přeloží na více slov, případně se nepřeloží vůbec. IBM-3 řeší tento problém zavedením **fertility**, které je modelována pravd. rozložením

$$n(\phi|f)$$

Pro každé zdrojové slovo f rozložení n říká, na kolik cílových slovo se obvykle f přeloží.

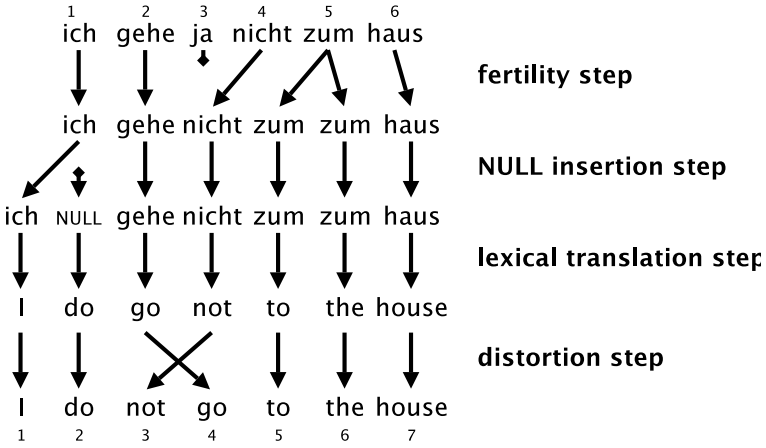
$$\begin{aligned} n(0|\text{a}) &= 0.999 \\ n(1|\text{king}) &= 0.997 \\ n(2|\text{steep}) &= 0.25 \\ &\dots \end{aligned}$$

Vložení tokenu NULL

Pokud chceme správně překládat do cílového jazyka, který používá slova, jež nemají ve zdrojovém jazyce překladové ekvivalenty, musíme řešit vkládání pomocného tokenu NULL.

Nepoužívá se $n(x|NULL)$, protože vložení NULL záleží na délce věty.

Přidáme tedy další krok **vložení NULL** do procesu překladu. Používají se p_1 a $p_0 = 1 - p_1$, kde p_1 znamená pravděpodobnost vložení tokenu NULL za libovolné slovo ve větě.



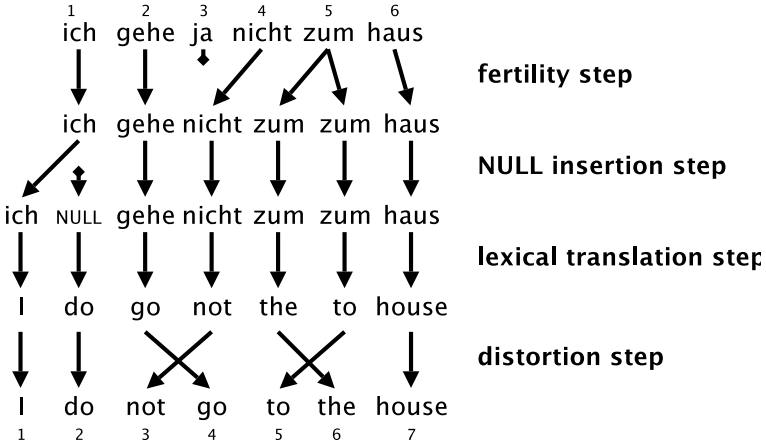
Obrázek 4.6: Kroky překladu modelu IBM-3

IBM-3 – distortion

Poslední krok je téměř shodný s 2. krokem překladového procesu IBM-2 a je modelován tzv. **distortion probability distribution**:

$$d(j|i, l_e, l_f),$$

která modeluje pozice v opačném pořadí: pro zdrojové slovo na pozici i modeluje pozici j cílového slova. Proces překladu z předchozího obrázku se může drobně lišit (viz 4.7).



Obrázek 4.7: Kroky překladu modelu IBM-3, alternativní

IBM-4, IBM-5

IBM-4

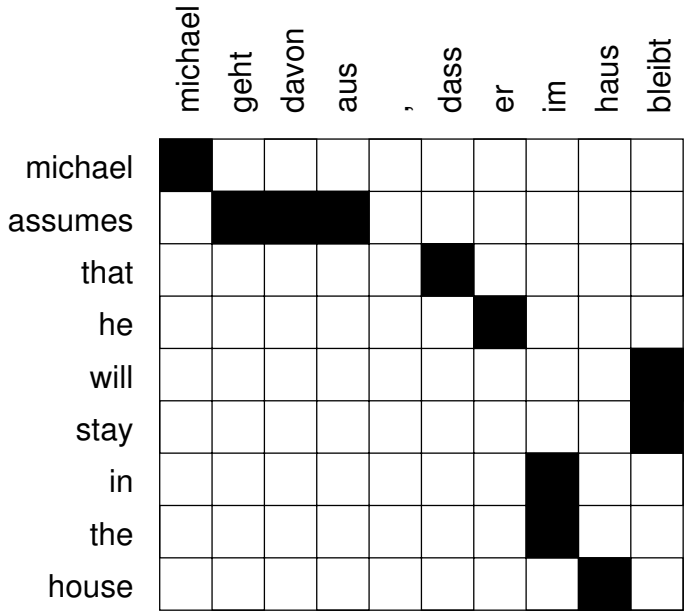
Problém distorze tkví v řídkých datech pro dlouhé věty. IBM-4 zavádí tzv. **relativní distorzi**, kde změny pozic slov závisí na předcházejících slovech. Vychází z předpokladu, že se překládá po frázích, které se přesunují vcelku, případně že některé přesuny jsou více časté (angličtina: ADJ SUB, francouzština SUB ADJ apod.).

IBM-5

Tento model řeší další nedostatky předchozích modelů. Např. hlídá, aby se dvě různá zdrojová slova

nedostala na jednu pozici v cílové větě atd.

Word-based metody – zarovnání slov



Obrázek 4.8: Matice zarovnání slov

Frázový překladový model

State-of-the-art statistického strojového překladu. Nepřekládají se pouze samostatná slova. Když to jde, tak i celé sekvence slov.

Fráze nejsou lingvisticky motivované, pouze statisticky. Německé *am* se zřídka překládá jedním slovem *with*. Statisticky významný kontext *spass am* pomáhá správnému překladu. Klasické fráze by se dělily jinak: *(fun (with (the game)))*.

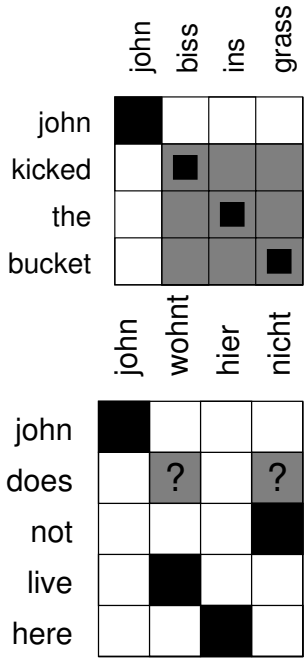
Výhody

- ▷ často překládáme $n : m$ slov, slovo je tedy nevhodný atomický prvek
- ▷ překlad skupin slov pomáhá řešit překladové víceznačnosti
- ▷ můžeme se učit překládat delší a delší fráze
- ▷ jednodušší model: neuvažujeme fertilitu, NULL token atd.

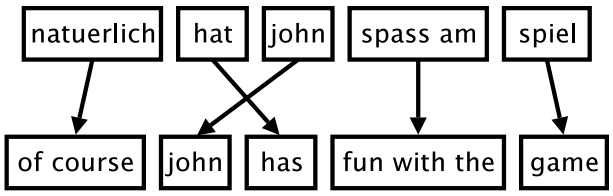
Překladová pravděpodobnost $p(f|e)$ se rozloží na fráze

$$p(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(\text{start}_i - \text{end}_{i-1} - 1)$$

Věta f se rozloží na I frází $\bar{f}_i$, všechna dělení jsou stejně pravděpodobná. Funkce ϕ je překladová pravděpodobnost pro fráze. Funkce d je přerovnávací model založený na vzdálenosti (**distance-based reordering model**), modelujeme pomocí předchozí fráze. start_i je pozice prvního slova ve frázi věty f, které se překládá na $i$tou frázi věty e.



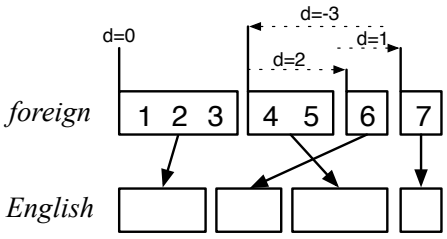
Obrázek 4.9: Problémy se zarovnáním slov



Obrázek 4.10: Frázový překladový model

Distance-based reordering model

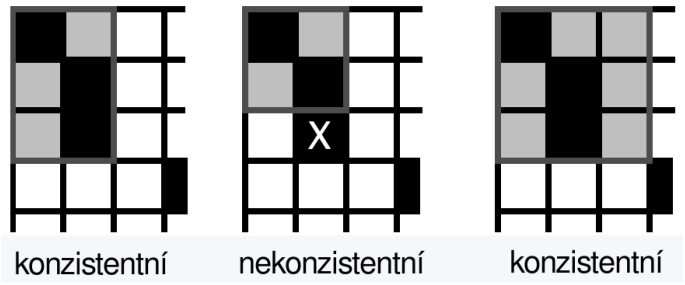
Preferuje se minimální přesun frází. Čím větší přesun (měří se na straně výchozího jazyka), tím dražší tato operace je.



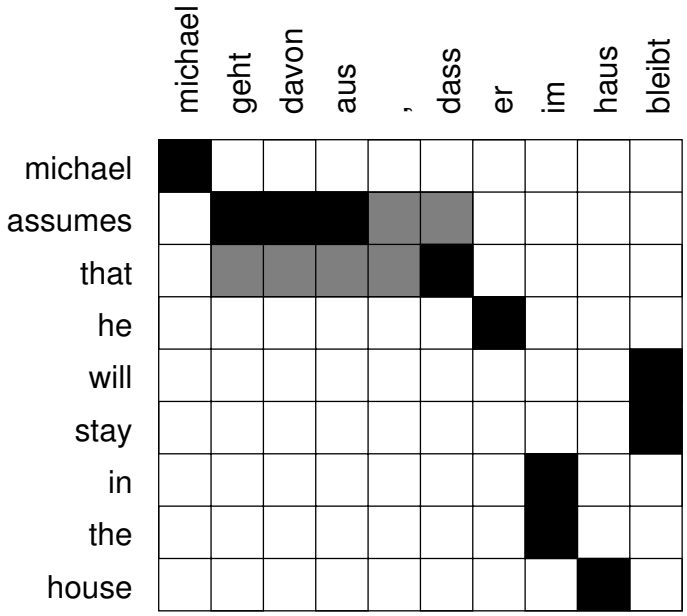
Obrázek 4.11: Distance-based reordering model

Budování překladové tabulky frází

Použijeme zarovnání slov (získané pomocí EM algoritmu pro IBM-1) a pak hledáme **konzistentní fráze**. Fráze $\bar{f}$ a $\bar{e}$ jsou konzistentní se zarovnáním A, pokud všechna slova $f_1, \ldots f_n$ ve frázi $\bar{f}$, která mají zarovnání v A, jsou zarovnaná se slovy $e_1, \ldots e_n$ ve frázi $\bar{e}$ a naopak.



Obrázek 4.12: Konzistentní a nekonzistentní fráze



Obrázek 4.13: Extrahování frází

Odhad pravděpodobnosti frází

$$\phi(\bar{f}|\bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} \text{count}(\bar{e}, \bar{f}_i)}$$

Model statistického překladu založený na frázích

$$e^* = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)\ d(start_i - end_{i-1} - 1) \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1...e_{i-1})$$

Vážený frázový model

$$e^* = \text{argmax}_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i)^{\lambda_\phi}\ d(start_i - end_{i-1} - 1)^{\lambda_d} \prod_{i=1}^{|\mathbf{e}|} p_{LM}(e_i|e_1...e_{i-1})^{\lambda_{LM}}$$

michael	michael
assumes	geht davon aus / geht davon aus ,
that	dass / , dass
he	er
will stay	bleibt
in the	im
house	haus
michael assumes	michael geht davon aus / michael geht davon aus ,
assumes that	geht davon aus , dass
assumes that he	geht davon aus , dass er
that he	dass er / , dass er
in the house	im haus
michael assumes that	michael geht davon aus , dass
. . .	. . .

Tabulka 4.5: Extrahované fráze

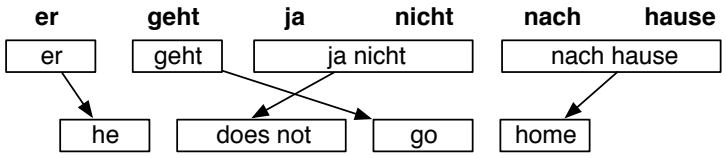
4.5 Dekódování

Máme jazykový model p_{LM} a překladový model $p(f|e)$. Potřebujeme vyhledat z exponenciálního množství všech překladů ten, kterému modely přiřazují nejvyšší pravděpodobnost.

Používá se **heuristické** prohledávání. Nemáme tedy garantováno, že nalezneme nejpravděpodobnější překlad.

Chyby překladu jsou způsobeny 1) chybou v prohledávání, kdy není nalezen nejlepší překlad v celém prohledávacím prostoru a 2) chybou v modelech, kdy i nejlepší překlad podle pravděpodobnostních funkcí není ten správný.

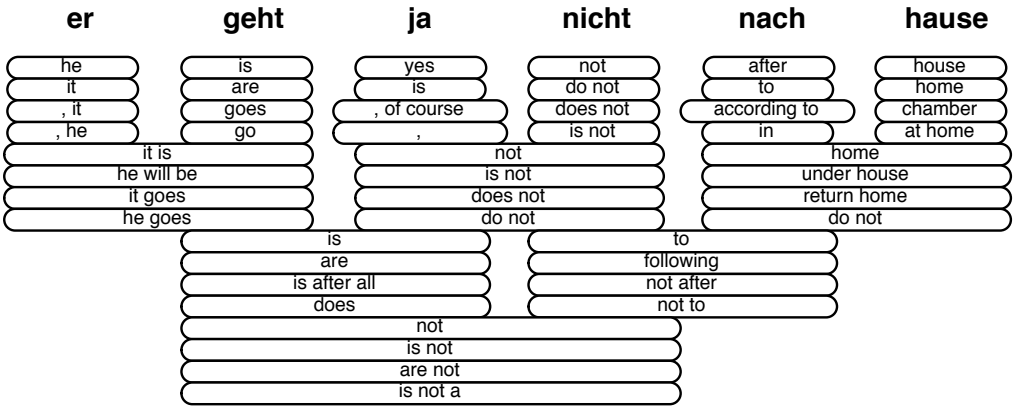
Překlad věty po frázích



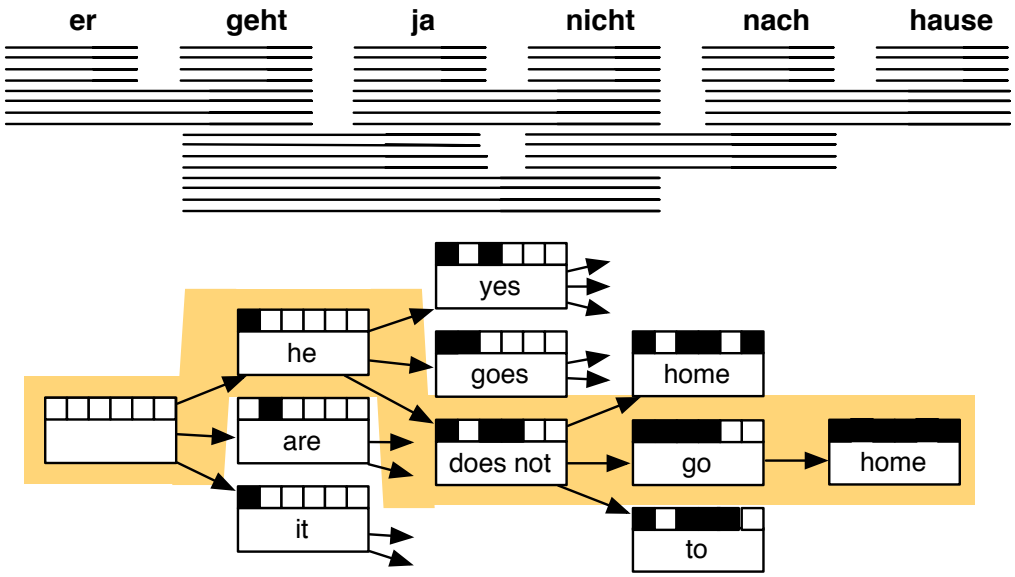
Obrázek 4.14: Kroky překladu

V každém kroku překladu počítáme předběžné hodnoty pravděpodobností z překladového modelu, přerovnávacího modelu a jazykového modelu.

Rozšiřujeme hypotézy v exponenciálním prostoru všech možných překladů. Různými metodami se snažíme prostor zmenšit.



Obrázek 4.15: Prohledávací prostor překladových hypotéz



Obrázek 4.16: Budování hypotéz, beam search

5. Hodnocení kvality překladu

Základní pojmy

- ▷ **plynulost** (fluency) – je překlad plynulý, má přirozený slovosled?
- ▷ **adekvátnost** (adequacy) – zachovává překlad význam, nebo je změněn, nekompletní?
- ▷ **srozumitelnost** (intelligibility)
- ▷ neplést s **přesností** (precision) a **pokrytím** (recall)

Nevýhody ručního hodnocení

- ▷ ruční hodnocení je pomalé, drahé, subjektivní
- ▷ mezianotátorská shoda (MAS) ukazuje, že se lidé shodnou více na plynulosti než na adekvátnosti
- ▷ jiné hodnocení: je X lepší překlad než Y?
- ▷ $\rightarrow$ ještě větší MAS

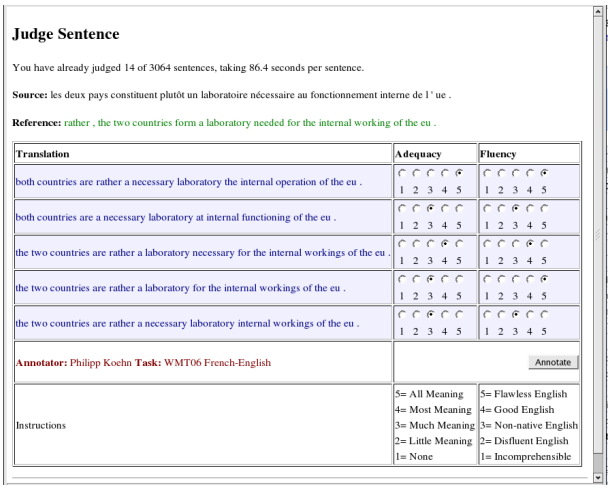
Automatické hodnocení kvality

- ▷ výhody: rychlost, cena; nevýhody: měříme opravdu kvalitu?
- ▷ gold standard: množina ručně připravených referenčních překladů
- ▷ kandidát c se srovnává s n referenčními překlady r_i

adekvátnost	
5	veškerý význam
4	většina významu
3	dostatečně významu
2	málo z původního významu
1	žádný význam

plynulost	
5	bezchybný jazyk
4	dobrý jazyk
3	nepřirozený
2	neplynulý jazyk
1	nesrozumitelný

Tabulka 5.1: Stupnice hodnocení



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

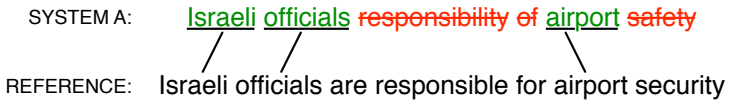
Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Obrázek 5.1: Anotační nástroj

- ▷ paradox automatického hodnocení: úkol AHKSP odpovídá situaci, kdy má student hodnotit svou vlastní písemnou práci: jak pozná, v čem udělal chybu?
- ▷ různé přístupy: n-gramová shoda mezi c a r_i, editační vzdálenost, . . .

Pokrytí a přesnost na slovech

Nejjednodušší způsob automatického hodnocení



- ▷ přesnost

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

- ▷ pokrytí

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

- ▷ f-score

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Pokrytí a přesnost – nedostatky

metrika	systém A	systém B
přesnost	50%	100%
pokrytí	43%	100%
f-score	46%	100%

Nepostihuje se nesprávný slovosled.

BLEU

- ▷ nejznámější (standard), nejpoužívanější, nejstarší (2001)
- ▷ IBM, Papineni
- ▷ n-gramová shoda mezi referencí a kandidáty
- ▷ počítá se přesnost pro 1 až 4-gramy
- ▷ extra postih za krátkost (**brevity penalty**)

$$\text{BLEU} = \min\left(1, \frac{\textit{output-length}}{\textit{reference-length}}\right) \left(\prod_{i=1}^{4} \textit{precision}_i\right)^{\frac{1}{4}}$$

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

metrika	systém A	systém B
přesnost (1gram)	3/6	6/6
přesnost (2gram)	1/5	4/5
přesnost (3gram)	0/4	2/4
přesnost (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0 %	52 %

Další metriky

- ▷ NIST
  - NIST: National Institute of Standards and Technology
  - vážení shod n-gramů podle informační hodnoty
  - velmi podobné výsledky jako BLEU (varianta)
- ▷ NEVA
  - Ngram EVAluation
  - úprava BLEU skóre pro kratší věty
  - bere v potaz i synonyma (kladně hodnotí použití synonyma ve smyslu stylistické bohatosti)
- ▷ WAFT
  - Word Accuracy for Translation
  - editační vzdálenost mezi c a r
  - WAFT $= 1 - \frac{d+s+i}{max(l_r, l_c)}$
- ▷ TER
  - Translation Edit Rate

- nejmenší počet kroků (smazání, přidání, prohození, změna)
- TER $= \frac{\text{počet editací}}{\text{prům. počet ref. slov}}$
- $r =$ dnes jsem si při fotbalu zlomil kotník
- $c =$ při fotbalu jsem si dnes zlomil kotník
- TER $= 4/7$

▷ HTER

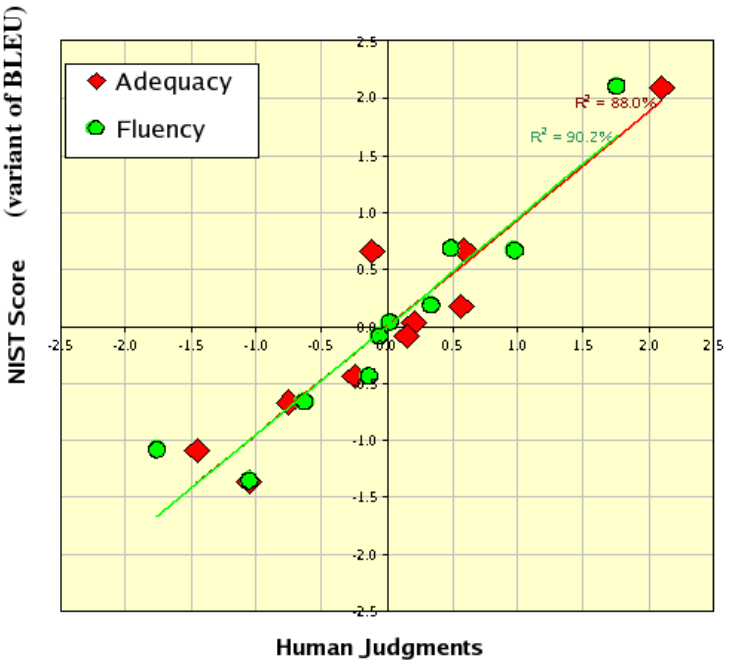
- Human TER
- nejdříve ručně vytvořena r a na ni aplikováno TER

▷ METEOR

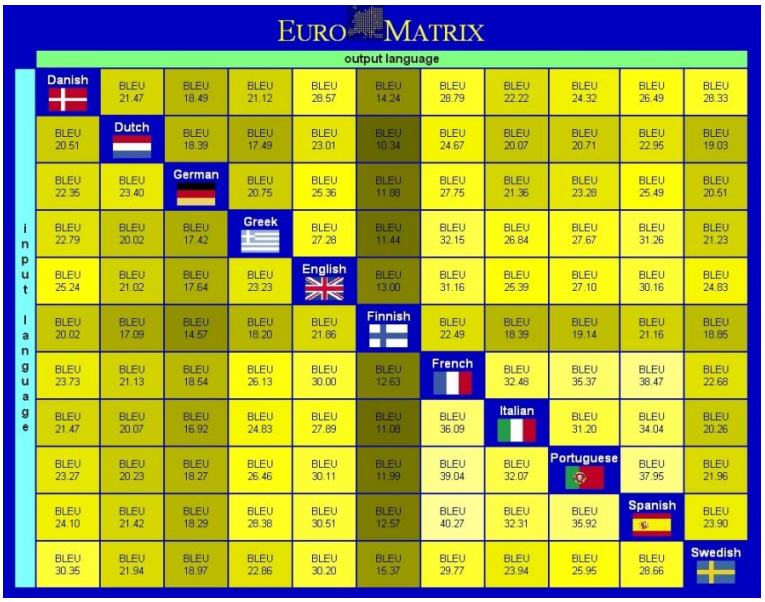
- uvažuje synonyma (WordNet), morfologické varianty slov
- vyšší korelace s ručním hodnocením

Hodnocení hodnotících metrik

Jak zjistit, která metrika je nejlepší?
Změřit, jak koreluje s manuálním hodnocením.



Obrázek 5.2: Korelace automatického a ručního hodnocení



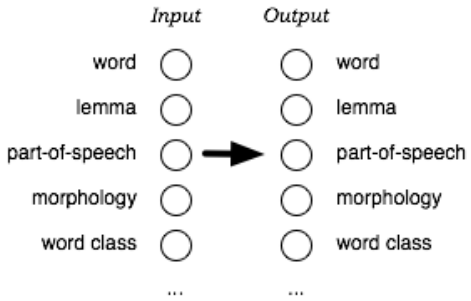
Obrázek 5.3: Hodnocení kvality překladu v projektu EuroMatrix

Obrázek 5.4: Hodnocení překladu podle jazykových párů

6. Další témata

6.1 Faktorované překladové modely

- ▷ běžné SMT modely nevyužívají lingvistickou znalost
- ▷ využití lemmat, POS, kmenů překonává řídkost dat
- ▷ pomocí těchto dat lze lépe a přirozeněji modelovat překlad

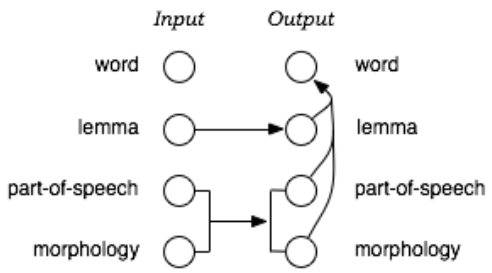


Obrázek 6.1: Překlad mezi vektory namísto tokenů

- ▷ v SMT jsou *domov* a *domovem* nezávislé tokeny
- ▷ ve FPM sdílí lemma, POS a část morf. informace
- ▷ mezi morf. bohatými jazyky lze překládat na úrovni lemat
- ▷ lemma a morfologická informace se přeloží nezávisle
- ▷ v cílovém jazyce se vygeneruje odpovídající slovní tvar

6.2 Tree-based překladové modely

- ▷ SMT překládá sekvence slov

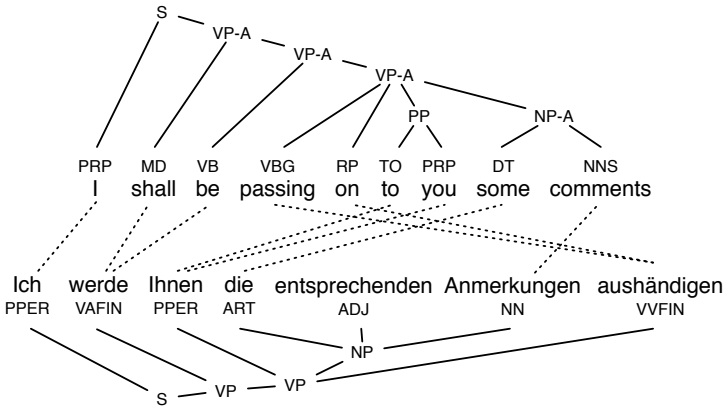


Obrázek 6.2: Schéma překladu faktorovaným modelem

- ▷ mnoho situací lze lépe vysvětlit pomocí syntaxe:
  přesun slovesa ve větě, gramatická shoda na velkou vzdálenost, . . .
- ▷ → překladové modely založené na syntaktických stromech
- ▷ aktuální téma, pro některé jazykové páry dává nejlepší výsledky

Synchronní frázová gramatika

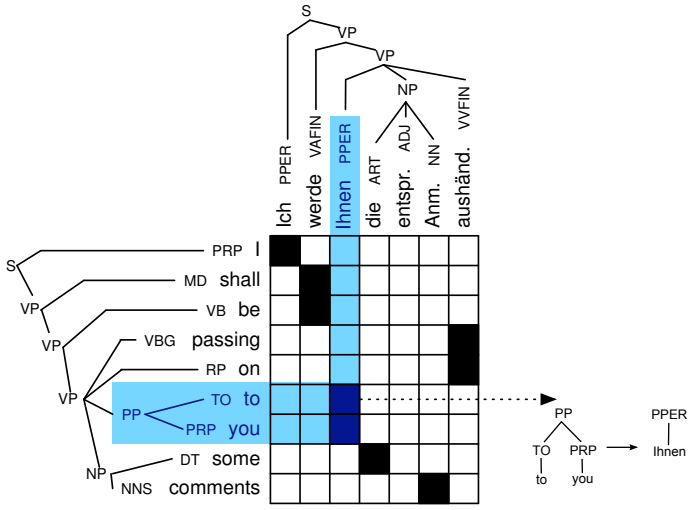
- ▷ EN pravidlo NP → DET JJ NN
- ▷ DE pravidlo NP → DET NN JJ
- ▷ synchronní pravidlo NP → DET_1 NN_2 JJ_3 | DET_1 JJ_3 NN_2
- ▷ koncové pravidlo N → dům | house
- ▷ smíšené pravidlo N → la maison JJ_1 | the JJ_1 house



Obrázek 6.3: Paralelní korpus se syntaktickou anotací

6.3 Hybridní systémy strojového překladu

- ▷ kombinace pravidlových a statistických systémů
- ▷ pravidlový překlad s post-editací statistickým systémem (např. vyhlazení jazykovým modelem)
- ▷ příprava dat pro SMT na základě pravidel, upravení výstupu SMT na základě pravidel



Obrázek 6.4: Extrakce syntaktických překladových pravidel

6.4 CAT – Computer-aided Translation

- CAT – computer-assisted (aided) translation
- mimo rámec strojového překladu
- využití počítače v procesu ručního překladu
- nástroje spadající pod CAT:
  - kontrolory pravopisu (překlepy): *hunspell*
  - kontrolory gramatiky: *Lingea Grammaticon*
  - správa terminologie
  - elektronické překladové slovníky: *Metatrans*
  - korpusové manažery: *Manatee/Bonito*
  - překladové paměti →

Překladová paměť

- databáze segmentů: nadpisy, fráze, věty, termíny
- které byly již dříve přeloženy → překladové jednotky
- výhody:
  - vše se překládá pouze jednou
  - snížení nákladů (opakované překlady mírně změněných manuálů)
- nevýhody:
  - většina systémů je komerčních
  - překladové jednotky nelze jednoduše získat
  - chyba v překladu se opakuje
- systém navrhuje překlad na základě přesné shody
- nebo shody na základě stejného kontextu
- systém může automaticky nahradit shodné segmenty

7. Příklady zkouškových otázek

- ▷ Popište princip *noisy channel* (vzorec, co je co).
- ▷ Uveďte alespoň 3 systémy hodnocení kvality SP; typy překladu podle R. Jakobsona.
- ▷ Co tvrdí Sapir-Whorfova hypotéza?
- ▷ Co víte o Georgetownském experimentu?
- ▷ Uveďte alespoň 2 příklady morfologicky bohatých jazyků.
- ▷ Jaká je výhoda systému s interlinguou oproti transferovému systému? Načrtněte diagram překladu mezi 5 jazyky pro tyto 2 typy překladových systémů.
- ▷ Uveďte příklad problematického řetězce znaků pro tokenizace češtiny.
- ▷ Co je to tagset, treebank, POS tagging, WSD, gisting, FrameNet, granularita významu, FAHQMT?
- ▷ Jakou výhodu má prostorová reprezentace významu?
- ▷ Do jakých dvou skupin se dělí metody WSD?
- ▷ Načrtněte Vauqoisův trojúhelník a načrtněte do něj statistický SP typu IBM-1.
- ▷ Vysvětlete pojem garden path a vymyslete příklad pro češtinu (ne ze slajdu).
- ▷ Načrtněte závislostní strukturu pro větu *Máma mele malou Emu.*; schéma statistického SP.
- ▷ Uveďte alespoň 2 příklady zdrojů paralelních textů.
- ▷ Vysvětlete Zipfův zákon.
- ▷ Máme dvě kostky – modrou a zelenou a házeme jimi zároveň. Jedna náhodná proměnná odpovídá číslu, které padne na zelené, druhá náhodná proměnná, co padne na modré kostce. Jde o závislé nebo nezávislé proměnné?
- ▷ Vysvětlete na příkladu Bayesovo pravidlo (uveďte vzorec).
- ▷ Co dělá *dekódovací algoritmus*?
- ▷ Napište vzorec nebo popište slovy *Markovův předpoklad*.
- ▷ Uveďte 3 příklady častých trigramů (slovních nebo znakových) pro češtinu nebo angličtinu.
- ▷ Pro kvalitu jazykového modelu chceme nízkou nebo vysokou perplexitu?
- ▷ Napište zarovnávací funkci pro dvojici frází *very small house* a *velmi malý dům*.
- ▷ Vysvětlete princip a kroky EM algoritmu, popište stručně IBM modely 1–5.
- ▷ Načrtněte matici zarovnání slov pro věty *I am very hungry.* a *Jsem velmi hladový*.