PLIN009 – Strojový překlad

Pravidlový strojový překlad

Vít Baisa

jaro 2012

28. března 2013

Úvod

Úvod

Rule-based Machine Translation – RBMT

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

Knowledge-based Machine Translation – KBMT

- systémy využívající znalosti o jazyce
- obecnější pojem

Knowledge-based MT

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

Rozdělení systémů KBMT

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

Do 90. let pouze tyto dva typy systémů.

Systém přímého překladu

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

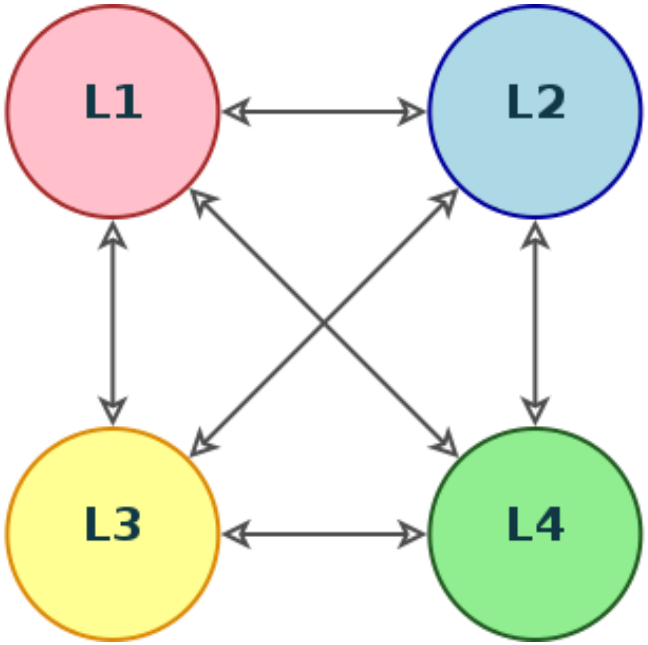
Přístup pomocí interlinguy

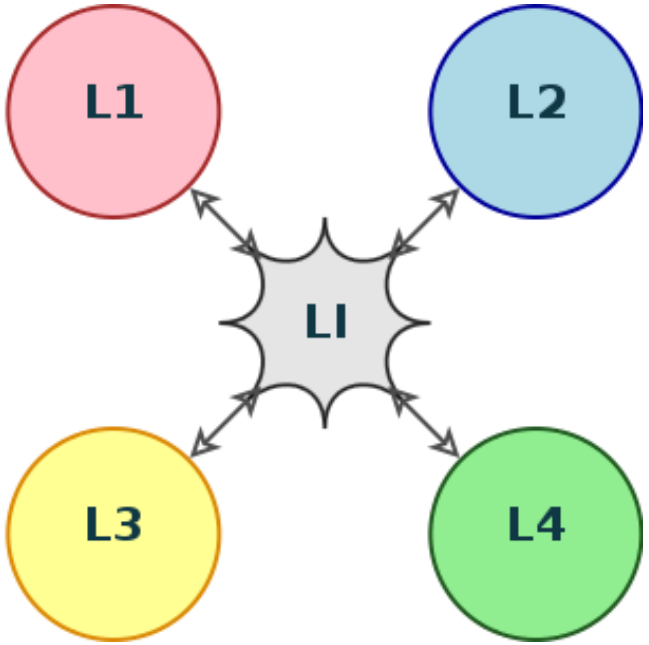
- předpokládá, že je možné SL konvertovat do sémanticko-syntaktické reprezentace, která je (částečně) nezávislá na jazyku
- interlingua musí být jednoznačná (unambiguous)
- z této podoby (interlingua) je generován TL
- analýza SL je jazykově závislá, ale nezávislá na TL
- analogicky syntéza TL
- SL a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

Transferové systémy strojového překladu

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

Interlingua vs. transferové KBMT





Tokenizace

Tokenizace

Co to je?

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

Problémy

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

Tokenizace – jak se to dělá?

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

Dělení na tokeny

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

Dělení na věty

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

Morfologická rovina

Morfologická rovina

- druhé patro v překladovém trojúhelníku
- je nutné eliminovat obrovský počet slovních variant
- převod slovní formy na základní tvar
  give, gives, gave, given, giving $\rightarrow$ give
  dělá, dělám, dělal, dělaje, dělejme, ... $\rightarrow$ dělat
- analýza gramatických kategorií slovních tvarů
  dělali $\rightarrow$ dělat + minulost + průběh + plurál + 3. osoba
  did $\rightarrow$ do + minulost + dokonavost + osoba ? + číslo ?
  Robertovým $\rightarrow$ 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* $\rightarrow$ bych?, *nejpomalejšími* $\rightarrow$ pomalý *neschopný* $\rightarrow$ schopný?
- slovesa: infinitiv
- *neraď* $\rightarrow$ radit?, *bojím se* $\rightarrow$ bát (se)
- Proč infinitiv? nejčastější tvar slovesa
- lemma souvisí s rozsahem/obsahem použitého slovníku

Morfologické značky, tagset

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

Problém s víceznačností

- v mnoha případech: více 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

Morfologická disambiguace

- nutno vybrat *jednu* značku a *jedno* lemma
- ke slovu přichází *morfologická disambiguace*
- nástroj *tagger*
- překladová víceznačnost je něco jiného
  pubblico $\rightarrow$ 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

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

Universal POS tags

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

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

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

Odhadování POS na základě gramémů

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

Problém: *myší, west, fotbal,* . . .

Tomáš Hanák – Sám v lese II

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

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

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

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

Ano, teď teprve snad poprvé sám svět's.

Brillův tagger

- učení z trénovacích dat
- transformation-based, error-driven
- úspěšnost přes 90 %

1. inicializuj značkování (nejčastější značka)
2. porovnej s trénovacími daty
3. vytvoř sadu pravidel pro změnu značek
4. 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

Morfologie – shrnutí

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

Slova a slovníky ve strojovém překladu

Slovníky ve strojovém překladu I

- 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

Slovníky ve strojovém překladu II

- 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

Problém s víceznačností

- 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

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

dát (bez *se*)

- 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).

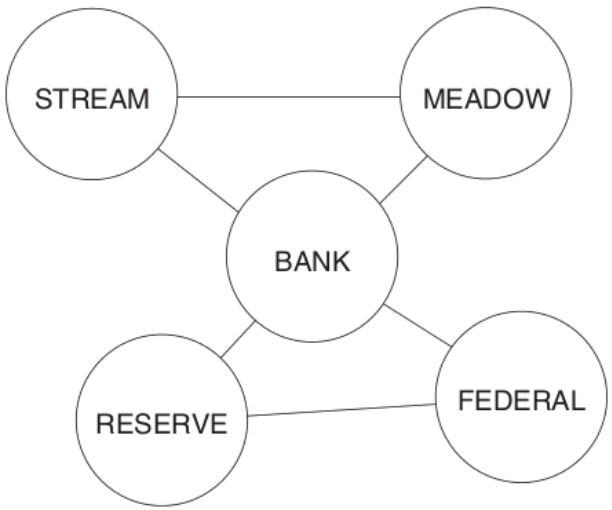
Granularita – malý

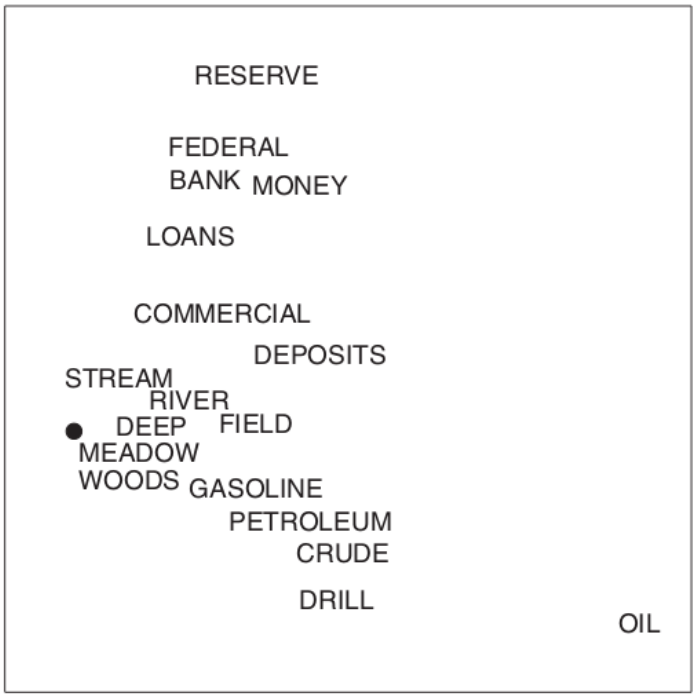
malý, malá

- 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

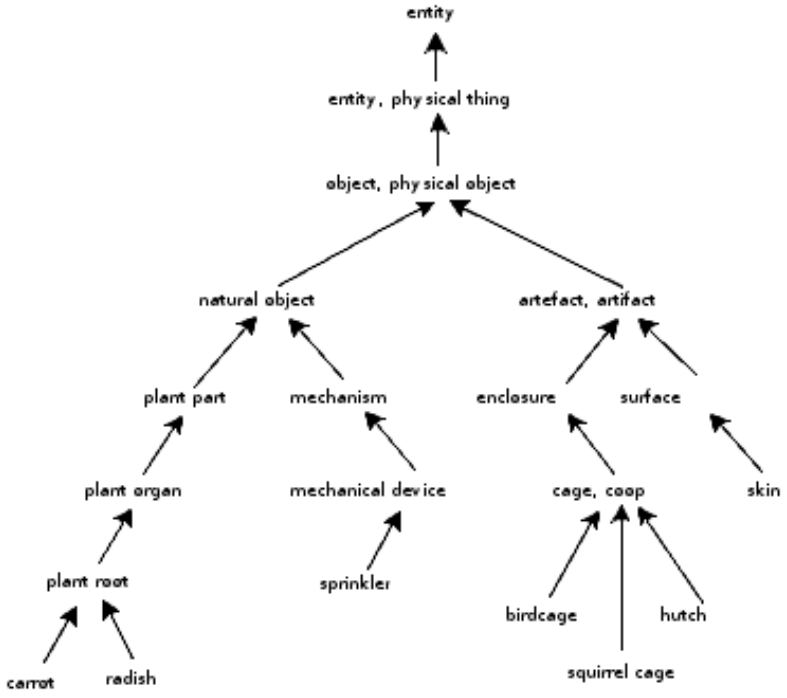
- nejčastější způsob: banka významů
- graf: významy jsou uzly, sémantické relace jsou hrany
- prostor: významy jsou body, podobné významy jsou prostorově blízko





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



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)
  $\mathrm{AG}^{kdo1}_{person:1} + \mathrm{VERB} + \mathrm{OBJ}^{co4}_{object:1} + (\mathrm{PART}^{v\ čem6}_{hand:1})$
- 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 %

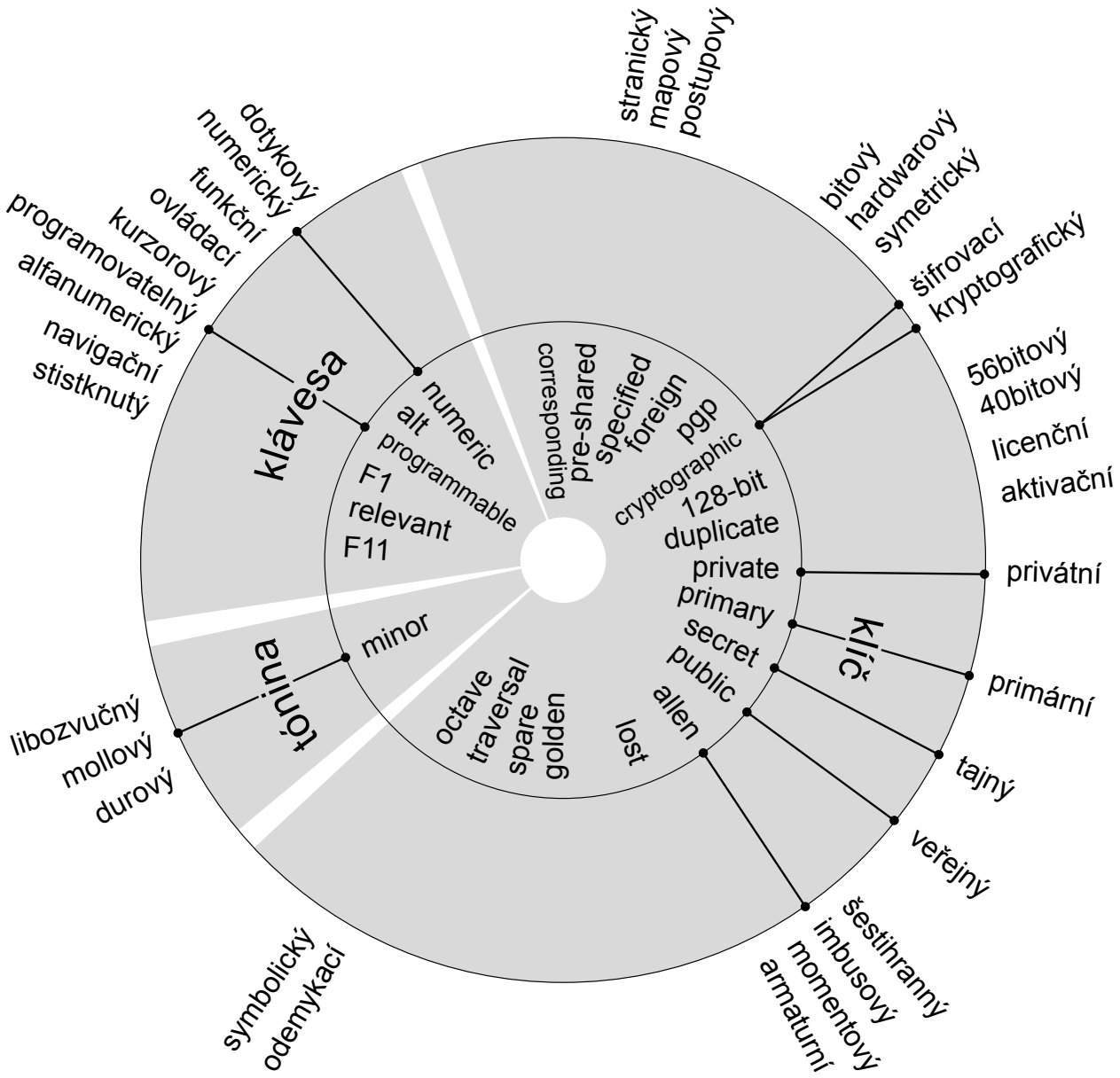
Pro strojový překlad zásadní:
Ludvig dodávka Beethoven, kiss me honey, . . .

WSD – metody

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
  - algoritmy s využitím valenčních slovníků (BP)
- 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

Jak určit vhodnout granularitu?



Lexika – shrnutí

- význam hlavně na úrovni slov (překladové slovníky)
- WSD zcela klíčový pro pravidlové systémy
- počet slov se mezi jazyky řádově liší
- na přesnost WSD má nejvíce vliv požadovaná granularita
- lexikální víceznačnost je bottleneck (RB)MT

Syntaktická analýza

Syntaktická analýza I

- 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

Syntaktická analýza II

- 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

Syntaktická analýza III

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

- *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.

Syntaktická víceznačnost II

- *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.*
- *Neboť každý, kdo prosí, dostává a kdo hledá, nalézá a tomu, kdo tluče, bude otevřeno. (Lk: 11,10)*

Částečná synt. víceznačnost – 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*.