Potřeby strojového učení pro počítačovou lingvistiku

Co se počítačoví lingvisté potřebují strojově učit

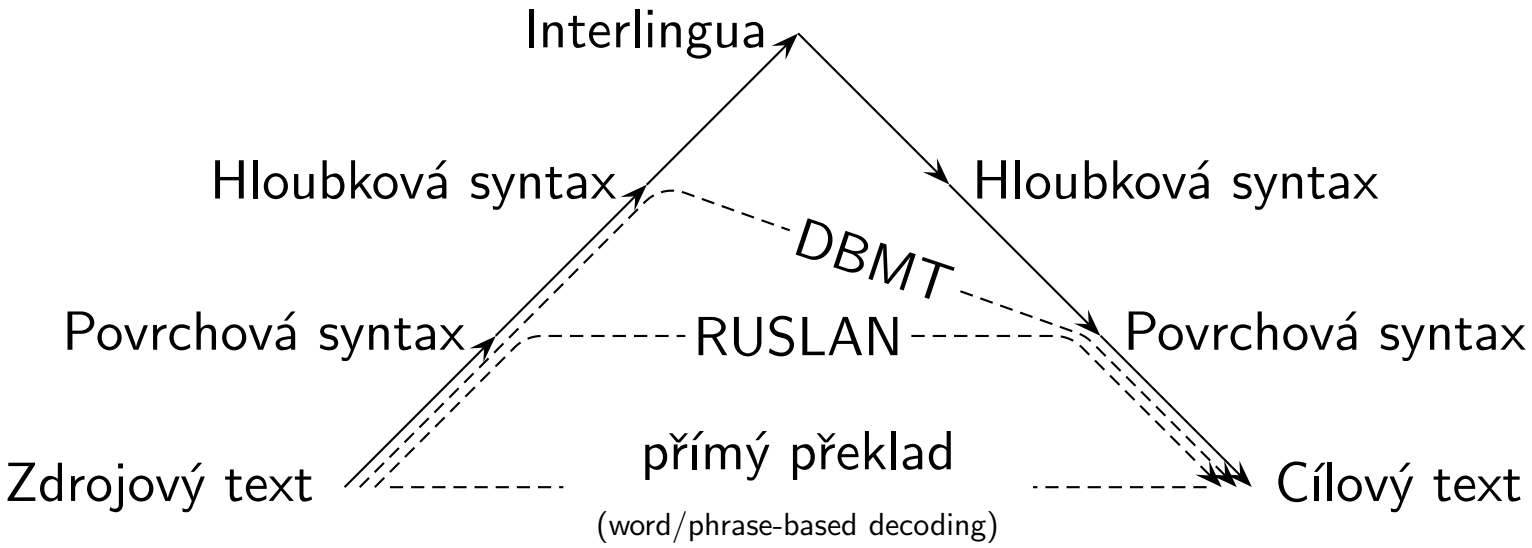
Ondřej Bojar
obo@cuni.cz

5. duben, 2006

Osnova

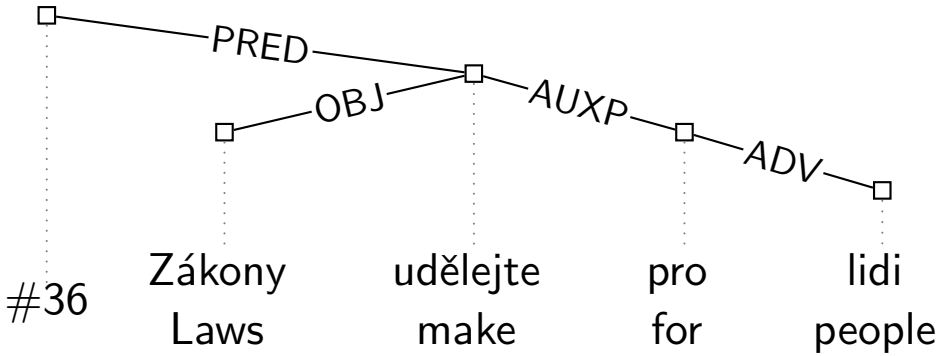
- Motivace a pojmy z (pražské tradice) (počítačové) lingvistiky
- Přehled úloh strojového učení
- Shrnutí charakteru úloh
- Co se potřebuji strojově učit já (dnes, příští měsíc, příští rok)
- Závěr: máme pro vás mají řadu úloh a dat k experimentům

Jedna z motivací: strojový překlad (MT)

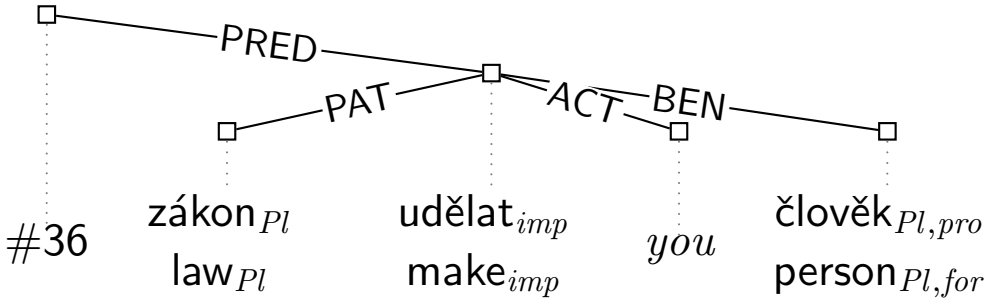


Rozbor českých vět (pražská tradice)

Analytický strom (povrchová syntax):



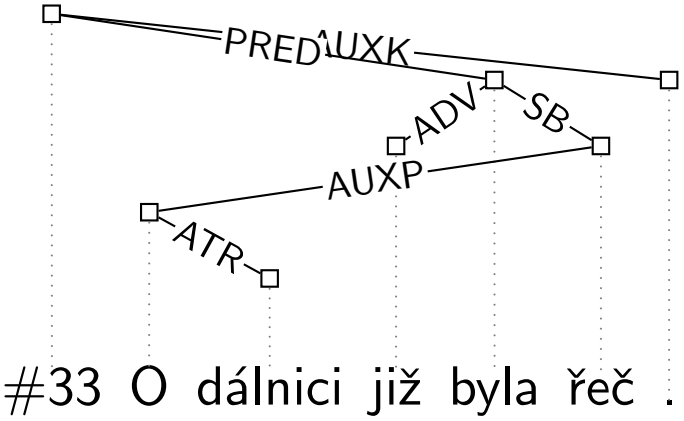
Tectogramatický strom (hloubková syntax):



Morfologická analýza:

Forma	Lema	Morphologická značka
zákony	zákon	NNIP1-----A----
zákony	zákon	NNIP4-----A----
zákony	zákon	NNIP5-----A----
zákony	zákon	NNIP7-----A----
udělejte	udělat	Vi-P---2--A----
udělejte	udělat	Vi-P---3--A---4
pro	pro-1	RR--4----------
lidi	člověk	NNMP1-----A----
lidi	člověk	NNMP4-----A----
lidi	člověk	NNMP5-----A----

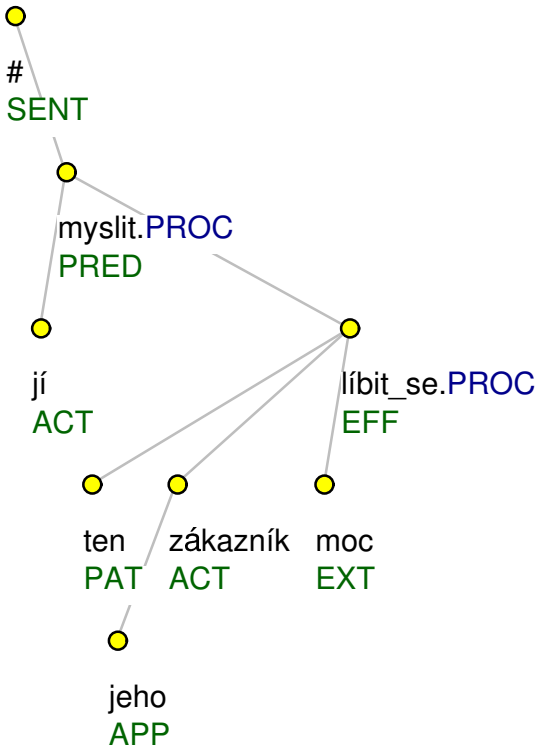
Neprojektivita

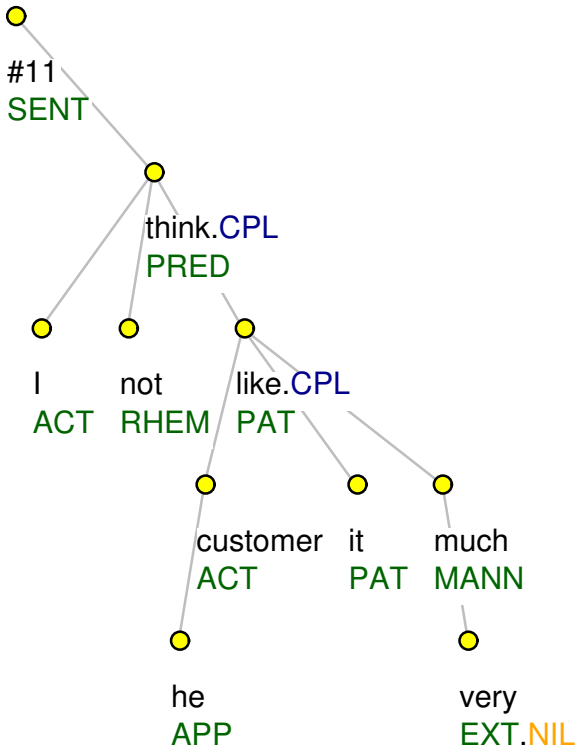


Neprojektivita:

- nezdá se, že by způsobovala prodlevy ve čtení (Bojar et al., 2004)
- na hloubkové syntaktické rovině mizí (Veselá, Havelka, and Hajičová, 2004)
- automatická syntaktická analýza s neprojektivitou vyřešena teprve nedávno (navíc $O(n^2)$) (McDonald et al., 2005)

Ilustrace: hloubková syntax sbližuje

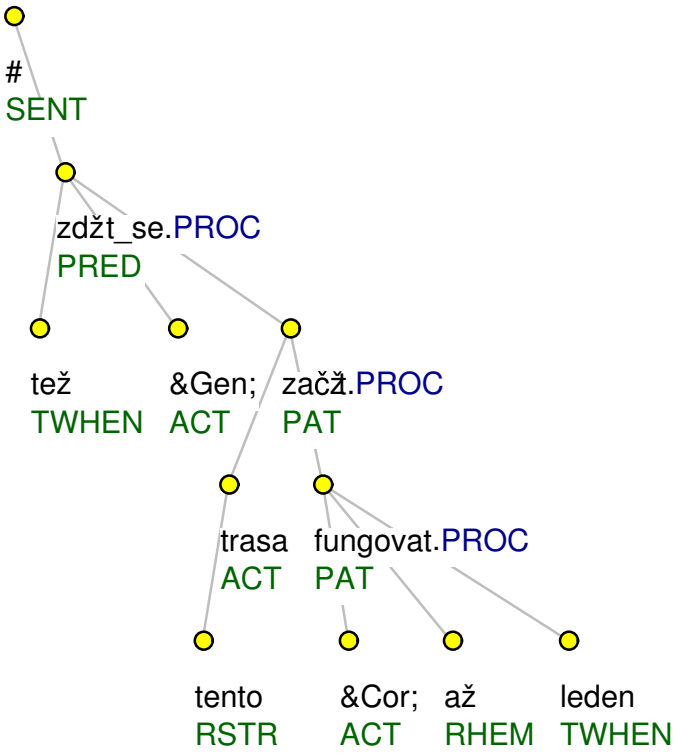




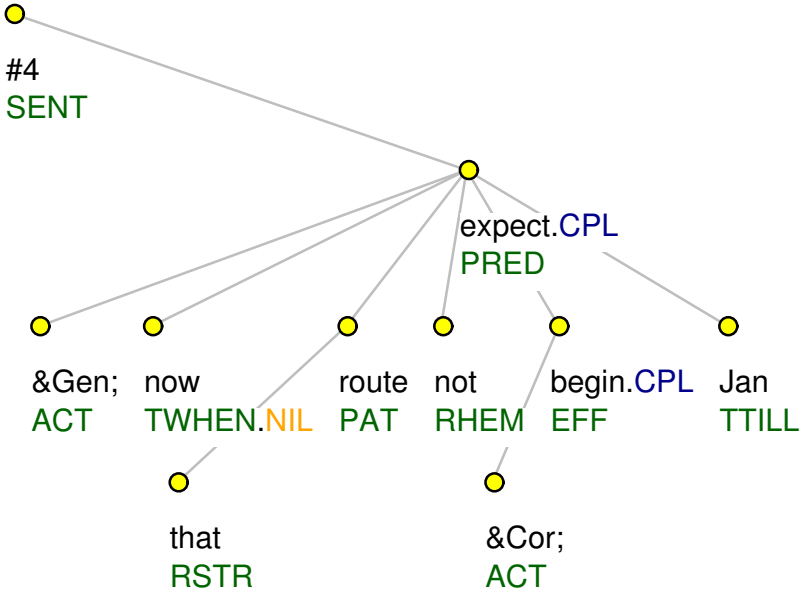
" Nemyslím , že by se to jejich zákazníkům moc líbilo . "

" I do n't think their customers would like it very much . "

Ilustrace: sbližuje hloubková syntax?

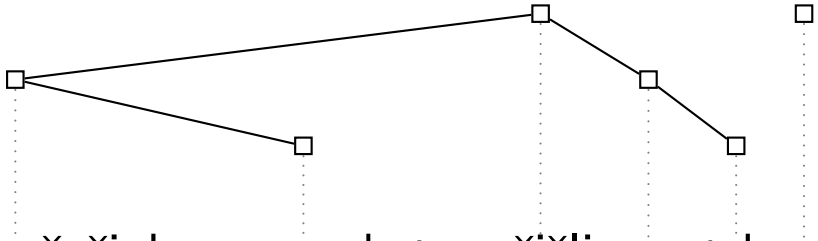


Teď se zdá , že tyto trasy začnou fungovat až v lednu .

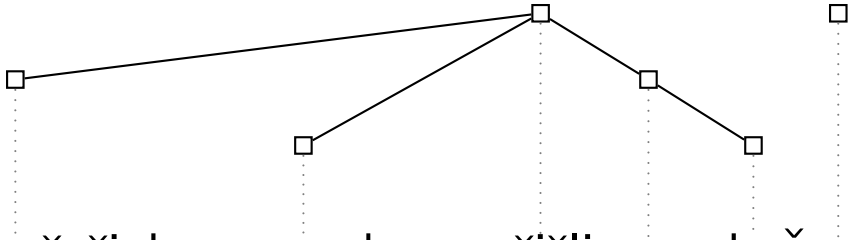


Now , those routes are n't expected to begin until Jan .

Valence: pomáhá zjednoznačnit strukturu



Manažeři hypermarketu přišli na trh .
přijít: kdo kam



Manažeři hypermarketu přišli na chuť .
přijít na chuť: kdo čemu

- Povrchová: co musím ve větě uvést, aby "jí nic nechybělo"
- Hloubková: co všechno musí být v situaci myšleno, aby se dalo užít dané slovo (zejm. sloveso)
- Povrchová valence pomáhá syntaktickému rozboru (Zeman, 2002).

Několik čísel o češtině

	čeština	angličtina
Bohatá morfologie	$\geq$ 4 000 možných značek, $\geq$ 1 400 viděno	užívá se 50
Pořádek slov	volný, vč. neprojektivity	pevný

Po nasbírání slovníku z	20 000	75 000	vět
nové lema (tj. slovo) přichází každých	1.6	1.8	vět
nová morf. značka přichází každých	110	290	vět
nová zjednodušená morf. značka každých	280	870	vět

Neprojektivních vět	16,920	23.3%
Neprojektivních hran	23,691	1.9%

Automatická (povrchová) syntaktická analýza	čeština	angličtina
Podíl správných hran	69.2–82.5–84%	91%
Podíl správných vět	15.0–30.9%	43%

Jazyková data: Anotované korpusy (říjen 2005)

Korpus a verze	Vět	Tokenů	Slovník	Lemat	Pozn.
Český národní korpus (SYN2000d)	6.8M	114M	1.7M	775k	jen morf.
Prague Dep Tbk (PDT 1.0 & 2.0)	82k	1.3M	130k	55k	3 roviny

Paralelní česko-anglické

Korpus a verze	Vět	Tokenů	Slovník	Lemat	Pozn.
Prague Cz-En Dep Tbk (PCEDT 1.0)	22k/49k	0.5M/1.2M	57k/30k	28k/25k	
Reader's Digest (PCEDT 1.0)	44k/44k	658k/755k	84k/36k	?	beletrie
Kačenka	128k/105k	1.5M/1.5M	102k/47k	39k/22k	beletrie
OPUS EU Constitution	11k/10k	127k/164k	?	?	špatná tok.
Kolovratník	107k/107k	1.3M/1.5M	190k/92k	?	netokeniz.
. . . a další					

Shrnutí základů z (pražské) poč. lingv.

- morfologická rovina: lemata, značky
- analytická (povrchově-syntaktická) rovina: 1 slovo $\sim$ 1 uzel
- tektogramatická (hloubkově-syntaktická): věta $\sim$ term predikátové logiky

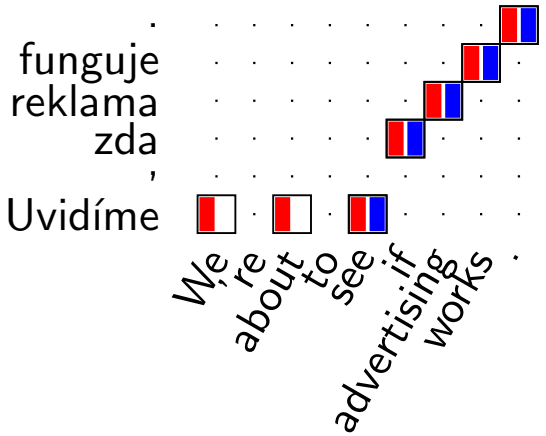
- neprojektivita a bohatost morfologie ztěžují syntaktickou analýzu

- valence: uzly stromu "vyžadují" potomky, povrchová pomáhá automatickému rozboru

- na experimenty je dostatek dat

Ilustrace nelingvistického přístupu k překladu

- trénovací soubor **paralelních textů**
- zarovnání po slovech
- extrakce slovníku (překlady slov či frází)
- decoding (překlad) = hledání "nejhladší formulace"
  nejhladší $\sim$ 3-gramy v mé hypotéze ať jsou v průměru (součin pstí) co nejběžnější (často spatřeny korpusu cílového jazyka, tzv. **jazykovém modelu**)



Skóre	Zdrojová fráze	Cílová fráze
2.30	že bude	it would
2.79	že bude	he would
3.08	že bude	he will
3.08	že bude	it will
3.48	že bude	it will be
3.77	že bude	it would be
4.17	že bude	be
4.17	že bude	it is
...		

Příklady strojového učení v počítačové lingvistice

Typický vývoj: formulace úlohy, ruční pravidla, ruční anotace + strojové učení

Úlohy řešené s učitelem:

- tagging = volba správné morfologické značky + lematu (HMM, MBL, TBL)
- parsing = nalezení nejpravděpodobnějšího povrchového stromu pro danou větu (PCFG, MST, MBL)
- tektogramatický parsing = hloubkový strom z povrchového (zatím ve fázi 3 pevných kroků, každý krok řešen jako klasif. úloha)
- identifikace významů slov (word-sense disambiguation) = rozhodni, které ze slovníkových hesel je v daném příkladu realizováno (DecTrees, SVM)

Úlohy řešené bez učitele: (ruční data jen pro vyhodnocení)

- zarovnání paralelního korpusu po větách
- zarovnání paralelního korpusu po slovech

Charakter úloh

Při lingvisticky motivovaném postupu: (zajímá náš ústav)

- bohatá anotace
- bohatě strukturovaná data, diskrétní veličiny, minimum spojitých veličin
- řídká data

Při postupu cíleném na řešení úlohy, např. strojového překladu:

- velké objemy dat (na dnešní stolní počítače)
- řídká data
- neadekvátní zjednodušení modelu (n-tice po sobě jdoucích slov místo závislostí), nicméně větší data to zachraňují

Obecně:

- víc dat $\Rightarrow$ lepší výsledek; techniky pro back-off zatím jen ad-hoc

Aktuální témata (můj velmi zúžený pohled)

Aktuální (disertace):

- výroba slovníkových hesel pro nová slovesa

Blízká budoucnost:

- Trénování konverze stromu na strom
  Hloubkový transfer čj↔aj.
  Převod povrchový↔hloubkový strom.

Výhled do dáli:

- Unsupervised (paralelní) závislostní větný rozbor a generalizace:
  Možná lingvisticky motivované stromy v čj a aj vykazují víc odlišností než je nutné.
  Možná kategorie definované lingvisty nejsou optimální pro danou úlohu (překlad ap.).

VALLEX = tektogramatický valenční slovník

Základní součásti: rámce∼významy, reflexivita, funktory, obligatornost, povrchové realizace

Word entry — Frame entry

odpovídat (imperfective)

1 odpovídat$_1$ ∼ odvětit [answer; respond]

- frame: $\mathrm{ACT}^{obl}{}_1\ \mathrm{ADDR}^{obl}{}_3\ \mathrm{PAT}^{opt}{}_{na+4,4}\ \mathrm{EFF}^{obl}{}_{4,aby,ať,zda,že}\ \mathrm{MANN}^{typ}$
- example: *odpovídal mu na jeho dotaz pravdu / že ...* [he responded to his question truthfully / that . . .]
- asp.counterpart: odpovědět$_1$ pf.
- class: communication

2 odpovídat$_2$ ∼ reagovat [react]

- frame: $\mathrm{ACT}^{obl}{}_1\ \mathrm{PAT}^{obl}{}_{na+4}\ \mathrm{MEANS}^{typ}\ \gamma$
- example: *pokožka odpovídala na včelí bodnutí zarudnutím* [the skin reacted to a bee sting by turning red]
- asp.counterpart: odpovědět$_2$ pf.

. . .

odpovídat se (imperfective)

1 odpovídat se$_1$ ∼ být zodpovědný [be responsible]

- frame: $\mathrm{ACT}^{obl}{}_1\mathrm{ADDR}^{obl}{}_3\mathrm{PAT}^{obl}{}_{z+2}$
- example: *odpovídá se ze ztrát* [he answers for the losses]

Zkrácený příklad pro sloveso "odpovídat".

Strojové učení pro hesla ve VALLEXu

Přestává být ekonomické doplňovat nová slovesa do slovníku ručně.

Vybudovat hesla pro nové sloveso (zadané lematem) znamená:

- poznat, co jsou doplnění slovesa (parsing, "hotovo", ∼83%, ale jen v ∼55 % byla správně pozorována celá mža doplnění pod slovesem)
- poznat funktory doplnění (klasif. úloha, "hotovo", ∼80%)
- rozdělit výskyty na reflexivní a nereflexivní
- rozdělit výskyty slovesa do skupin odpovídajících jednotlivým rámcům
- sdružit doplnění vyjadřující tutéž funkci (tentýž funktor) napříč výskyty
- rozhodnout, která doplnění patří do popisu rámce a která jsou "volná"
- rozhodnout obligatornost doplnění
- vyjmenovat povolené formy u jednotlivých doplnění

Lze se přitom inspirovat hotovými rámci pro známá slovesa.

Problematičnost přiřazení významu

Při přiřazování jednoho z předem definovaných rámců danému výskytu slovesa se dva lidé shodnou v 70–76 % případů, tři lidé se shodnou v 61–68 % případů.

Na datech, kde se tři lidé shodli nebo supervizor shodu vynutil (oprava jasných překlepů ap.), stroj umí rámec poznat v 67–71 % případů.

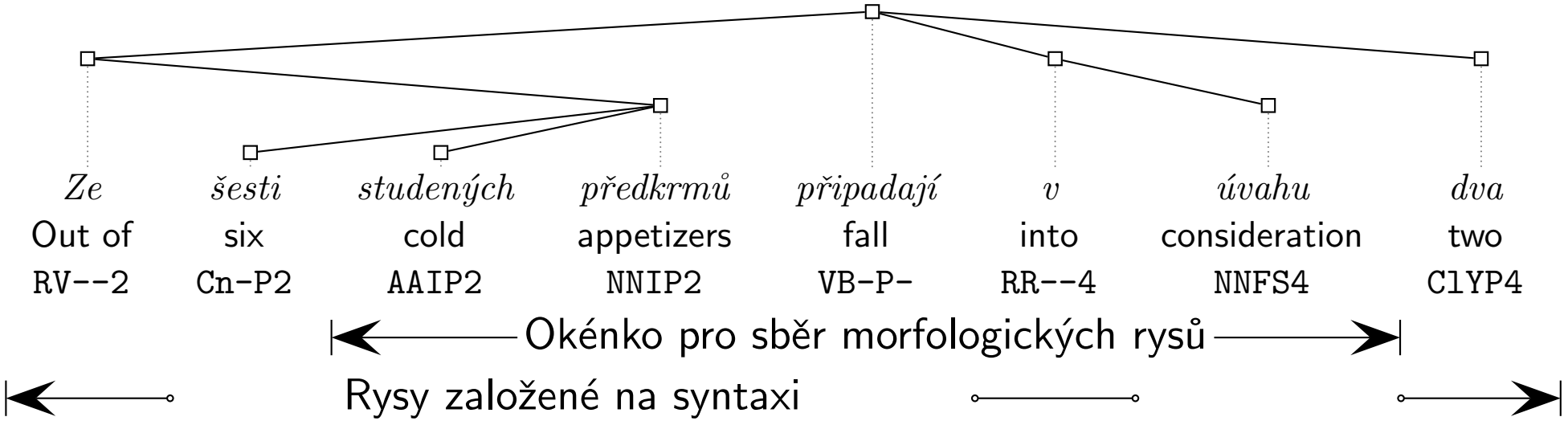
Zásadnější problémy:

- množinu významů slova nelze předem definovat
- hranice mezi významy slova není ostrá, podle situace přihlížíme k rozdílům různé jemnosti

. . . ale to zatím ignorují skoro všichni.

Jaké rysy ve strojovém učení užíváme

Věty s anotovanými slovesy → parsing → extrakce rysů → klasifikace pomocí C4.5



Morfologické rysy: `AAIP2 NNIP2 VB-P- RR--4 NNFS4`

Rysy založené na syntaxi: *ze+2*, *v+4*, *4*

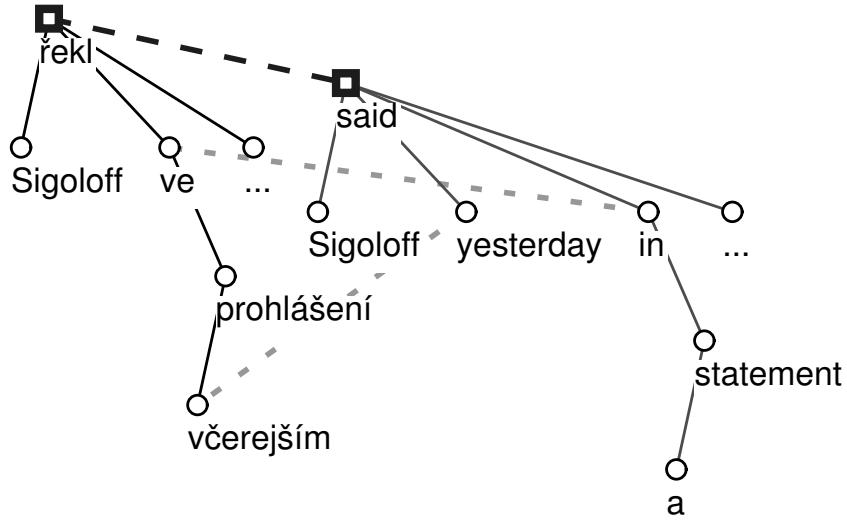
Booleovské rysy popisují (ne)přítomnost určitých typů doplnění slovesa.

Mé aktuální úvahy a experimenty

- Jak definovat metriku syntaktické podobnosti mezi dvěma výskyty slovesa?
- Jak definovat metriku sémantické podobnosti mezi dvěma výskyty slovesa?
- Která z těchto metrik (nebo kombinace?) bude nejlepším podkladem pro clustering?
  - Aby se clustering nejvíc podobal ručnímu označení užitých rámců.
  - Abychom z clusterů automaticky odvodili rámce co nejpodobnější těm ve slovníku. (Vyhodnocení end-to-end totiž může být jiné než vyhodnocení po komponentách.)
- Jak vlastně měřit podobnost dvou clusteringů?
- Jak automaticky poznat, kolik clusterů definovat?

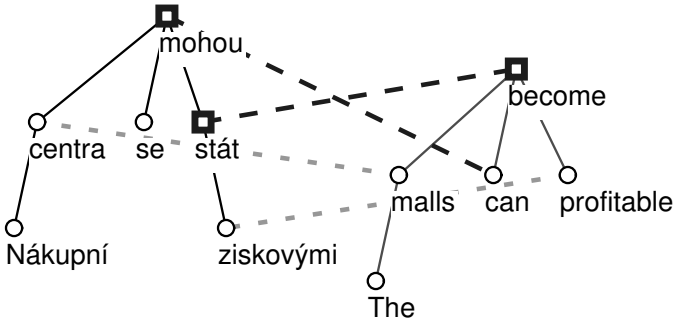
. . . hledám rady a zkušenosti.

K transformacím stromů: Posun doplnění



Sigoloff řekl ve včerejším prohlášení . . .
Sigoloff said yesterday in a statement . . .

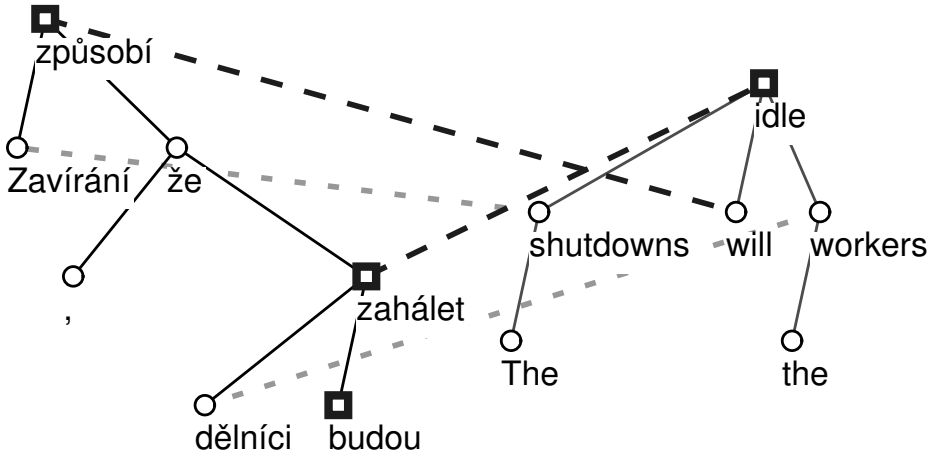
K transformacím stromů: Prohazování hlav (1)



Nákupní centra se mohou stát ziskovými . . .
The malls can become profitable . . .

. . . na vině jsou jen jiná pravidla pro anotaci závislostí struktury

K transformacím stromů: Prohazování hlav (2)



Zavírání způsobí , že dělníci budou zahálet . . .
The shutdowns will idle the workers . . .

. . . skutečné prohození řídicího a závislého uzlu (navíc sloveso→vedlejší věta)

Úvahy do daleka

Lingvistické abstrakce se při praktickém nasazení ukazují:

- jako vynikající zdroj inspirace, na jaké rysy hledět
- většinou příliš složité (co do stupňů volnosti) s ohledem na množství dat
- často příliš obecně zaměřené, a tedy suboptimální pro každou konkrétní úlohu

Konkrétně např. pro úlohu strojového překladu přes hloubkovou syntax:

- je množina funktorů definována účelně?
- je množina rámců definována účelně?

Účelně $\sim$ kategorie v datech rozděleny "rozumně rovnoměrně", prostor potenciálních řešení prohledatelný v reálném čase, dostatek trénovacích dat pro spolehlivé odhady řešení.

. . . budu vděčný za formalizaci, zatím jsem neměl čas ani na rešerši.

Závěrem

Co si zapamatovat:

- Počítačoví lingvisté mají velkou zásobu dat a řadu úloh, jejichž řešení potřebuje vylepšit.
- Metody statistiky a strojového učení často fungují lépe než ruční pravidla.
- Počítačoví lingvisté tady v Praze statistiku zatím moc neumějí.
- Z metod strojového učení umíme používat vlastně jen klasifikátory.

- Lingvistická data mají zipfovské rozdělení (je jich moc a málo současně)

Co bych potřeboval:

- odkazy na metody (implementace) strojového učení s daty strukturovanými do různé hloubky a zipfovským rozdělením na každé úrovni.

Literatura

Bojar, Ondřej, Jiří Semecký, Shravan Vasishth, and Ivana Kruijff-Korbayová. 2004. Processing noncanonical word order in Czech. In *Proceedings of Architectures and Mechanisms for Language Processing, AMLaP 2004*, pages 91–91, Université de Provence, September 16-18.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-Projective Dependency Parsing using Spanning Tree Algorithms. In *Proceedings of HLT/EMNLP 2005*, October.

Veselá, Kateřina, Jiří Havelka, and Eva Hajičová. 2004. Condition of Projectivity in the Underlying Dependency Structures. In *Proceedings of Coling 2004*, pages 289–295, Geneva, Switzerland, August. COLING.

Zeman, Daniel. 2002. Can Subcategorization Help a Statistical Parser? In *Proceedings of the 19th International Conference on Computational Linguistics (Coling 2002)*, Taibei, Tchaj-wan. Zhongyang Yanjiuyuan (Academia Sinica).

Analytic vs. Tectogrammatical (2)