Neřízený závislostní parsing

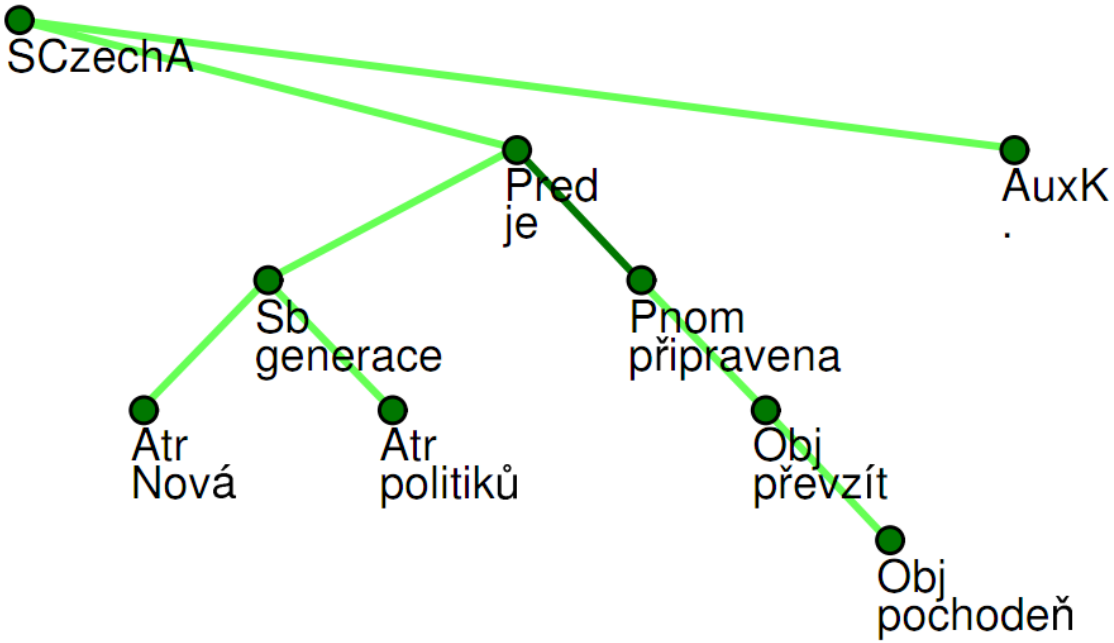
David Mareček

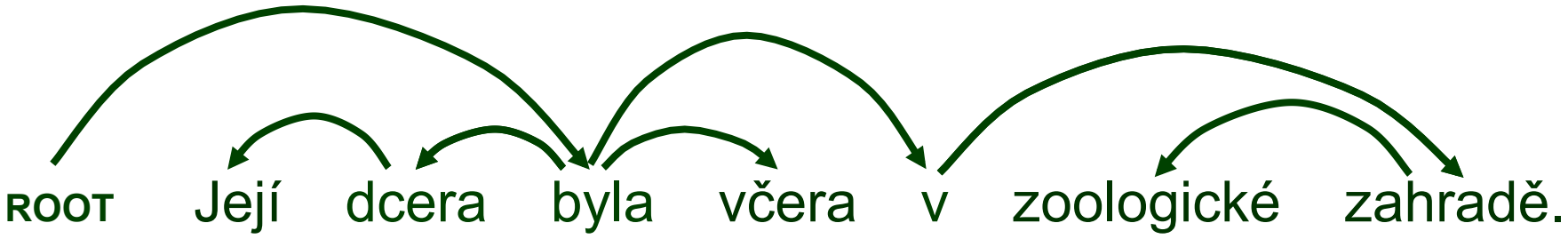
Ústav formální a aplikované lingvistiky
Univerzita Karlova v Praze

Schůzka projektu Res Informatica
15. listopadu 2011, Praha

Závislostní parsing

- Jedna ze základních úloh počítačové lingvistiky
- Analýza zadané věty
  - orientovaný strom
  - hranám mohou být přiřazeny syntaktické funkce
- Různé způsoby zobrazování





Řízený parsing

- Je trénovaný na příkladech – na lidmi anotovaném korpusu
  - Pražský závislostní korpus (PDT) obsahuje asi 90 000 vět (1,5 milionu slov)
- Nejlepší řízené parsery dosahují na češtině úspěšnosti 86%
  - (86% hran je na testovacích datech shodných s lidskou anotací)
- Nevýhody:
  - lidské anotace stojí značné množství času a peněz
  - ručně anotované korpusy jsou dostupné pouze pro některé jazyky
  - dostupné pouze pro některé domény (například pouze novinové články)
  - závislé na dané lingvistické teorii (zachycení koordinací, složených slovesných tvarů)

Neřízený parsing

- Nepotřebuje žádná lidmi anotovaná data
  - Učí se gramatiku sám na základě velkého množství textů
- Nejlepší neřízené parsery dosahují na angličtině úspěšnosti 68%
  - srovnání s ručně anotovanými daty
- Výhody:
  - snadné přizpůsobení jazyku a doméně
  - nevyžaduje nákladné ruční anotace
  - lidmi neanotovaných textů máme k dispozici celý Internet
- Nevýhody:
  - náročné na výpočetní výkon
  - zatím nedosahuje takových výsledků jako řízený parsing

Evaluační metriky

- Evaluace porovnáním s lidmi anotovanými daty je problematická
  - před každou lingvistickou anotací je třeba udělat spoustu rozhodnutí jak anotovat dané jazykové jevy
  - koordinační struktury, pomocná slovesa, členy, předložkové skupiny
- Dvě metriky:
  - UAS (unlabeled attachment score) – kolik orientovaných hran ve stromě je určeno správně
  - UUAS (undirected unlabeled attachment score) – nezávisí na orientaci hrany, vyřeší například problém s rozdílným zachycováním předložkových skupin
- UAS < UUAS
- Ideální by bylo změřit kvalitu parsingu na nějaké cílové aplikaci
  - strojový překlad, question answering,...

Jednoduchý neřízený parser

- Nepracuje se slovy, ale pouze se značkami (tagy)
  - slovní druh + některé morfologické kategorie

Její	dcera	byla	včera	v	zoologické	zahradě.
P1	N1	Vp	Db	R6	A6	N6

- Aproximuje pravděpodobnost stromu jako součin pravděpodobností jednotlivých hran
  - nevyužívá kontext
- Využívá Gibbsův sampling pro generování stromů
  - náhodná inicializace
  - podmínka stromovitosti

Gibbsův sampling

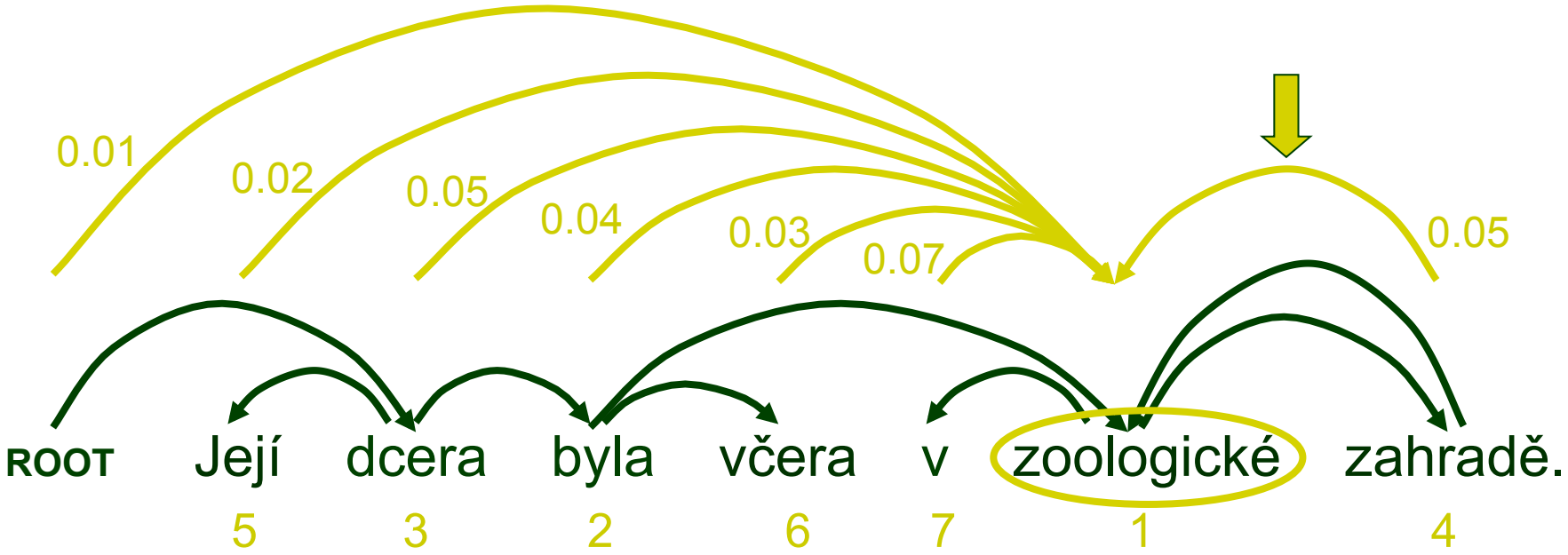
- Každé větě v korpusu na začátku přiřadím náhodný strom
- Iteruji náhoně přes všechna slova v korpusu a náhodně je převěšuji s ohledem na P(T)
- The rich get richer
  - často vyskytující se hrana v korpusu bude ještě častější
- Závislostní a vzdálenostní model
- Dirichletovské hyperparametry α_1 α_2 byly nastaveny experimentálně

$$P(\mathcal{T}) = \prod_{i=1}^{N} P(T_i^g | T_i^d) \cdot P(D_i | T_i^d)$$

$$= \prod_{i=1}^{N} \left(\frac{count^{(-i)}(T_i^g, T_i^d) + \alpha_1}{count^{(-i)}(T_i^d) + \alpha_1 |T|} \cdot \frac{count^{(-i)}(D_i, T_i^d) + \alpha_2}{count^{(-i)}(T_i^d) + \alpha_2 |D|} \right)$$

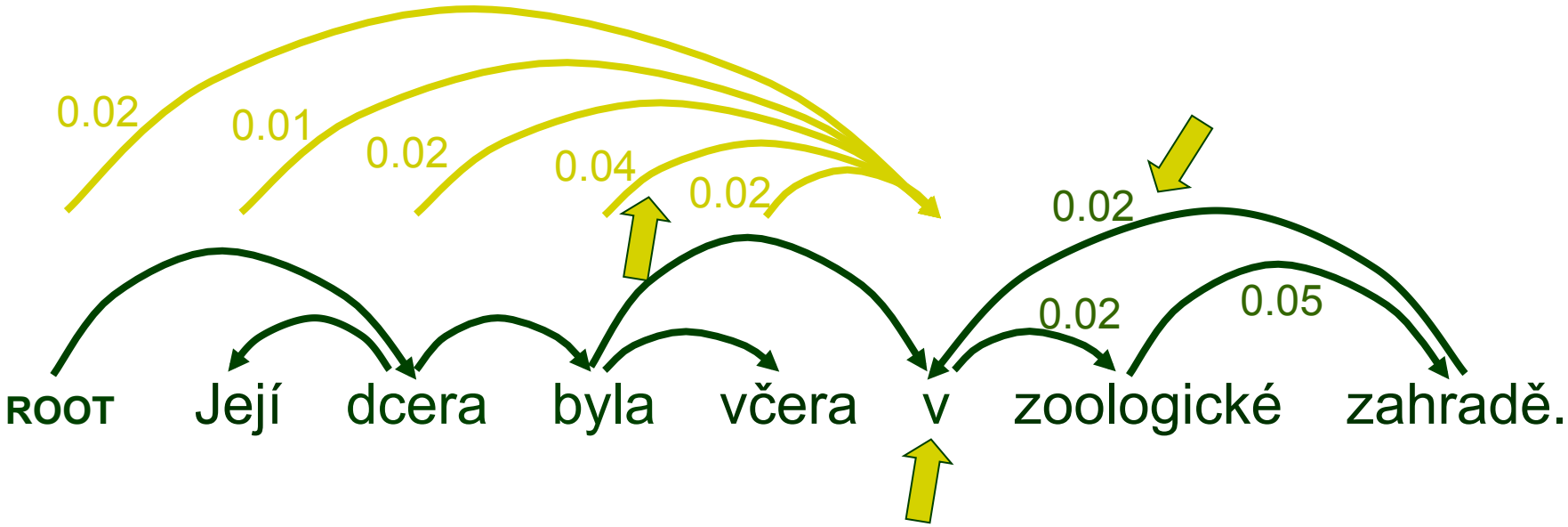
Základní sampling

- Pro každé slovo vyber jeho rodiče s ohledem na P(T)
- Slova v dané větě se vybírají náhodně
- Problém: vytváří i cykly



Podmínka stromovitosti

- Pokud je vytvořen cyklus:
  - vyber hranu v cyklu (s ohledem na P(T)) a vymaž ji
  - osamocený podstrom připoj tak, aby nebyl vytvořen cyklus



Výsledky na češtině

- Evaluace na PDT 2.0

Configuration	UAS	UUAS
Random baseline	12.0	19.9
LeftChain baseline	30.2	53.6
RightChain baseline	25.5	52.0
Base	36.7	50.1
Base+Treeness	41.2	58.6
Base+Treeness+NounRootRepression	49.8	62.6

Výsledky na jiných jazycích z CoNLL

Language			Baselines			Results			
name	code	CoNLL	rand.	left	right	Our-NR	Our	Spi5	Spi6
Arabic	ar	2007	3.9	59.0	6.0	24.8	25.0	22.0	**49.5**
Bulgarian	bg	2006	8.0	38.8	17.9	**51.4**	25.4	44.3	43.9
Catalan	ca	2007	3.9	30.0	24.8	56.3	55.3	**63.8**	59.8
Czech	cs	2007	7.4	29.6	24.2	**33.3**	24.3	31.4	28.4
Danish	da	2006	6.7	47.8	13.1	38.6	30.2	**44.0**	38.3
German	de	2006	7.2	22.0	23.4	21.8	26.7	**33.5**	30.4
Greek	el	2007	4.9	19.7	31.4	33.4	**39.0**	21.4	13.2
English	en	2007	4.4	21.0	29.4	23.8	24.0	34.9	**45.2**
Spanish	es	2006	4.3	29.8	24.7	**54.6**	53.0	33.3	50.6
Basque	eu	2007	11.1	23.0	30.5	34.7	29.1	**43.6**	24.0
Hungarian	hu	2007	6.5	5.5	41.4	**48.1**	48.0	23.0	34.7
Italian	it	2007	4.2	37.4	21.6	**60.6**	57.5	37.6	52.3
Japanese	ja	2006	14.2	13.8	67.2	**53.5**	52.2	**53.5**	50.2
Dutch	nl	2006	7.5	24.5	28.0	**43.4**	32.2	32.5	27.8
Portugese	pt	2006	5.8	31.2	25.8	41.8	**43.2**	34.4	36.7
Slovenian	sl	2006	7.9	26.6	24.3	**34.6**	25.4	33.6	32.2
Swedish	sv	2006	7.8	27.8	25.9	26.9	23.3	42.5	**50.0**
Turkish	tr	2006	6.4	1.5	65.4	32.1	32.2	33.4	**35.9**
Chinese	zh	2007	15.3	13.4	41.3	34.6	21.0	34.5	**43.2**
		Average:	7.2	26.4	29.8	**39.4**	35.1	36.7	39.3

Závěr

- Neřízený parsing má smysl
- I velmi jednoduchý parser pracující pouze s malým množství textů a používající pouze značky místo slovních forem dosahuje relativně dobrých výsledků
- Ve 12 z 19 jazyků z CoNLL dosahuje lepších výsledků dříve publikované parsery

Děkuji za pozornost.