PLIN021 Sémantická analýza v praxi

OP VK Mezi bohemistikou a informatikou
www.projekt-inova.cz

Zuzana Nevěřilová
xpopelk@fi.muni.cz

Centrum zpracování přirozeného jazyka, B203
Fakulta informatiky, Masarykova univerzita

28. března 2012

Word Sense Desambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

minule jsme mluvili o metodách založených na znalostech (Leskův algoritmus pracující se slovníkovými definicemi a příklady užití)

Algoritmy strojového učení

- „pravidlové“
  - rozhodovací seznamy
  - rozhodovací stromy
- „matematické“
  - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
  - maximální entropie: Berger 1996
  - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvové“
  - one sense per discourse (Gale 1992)
  - one sense per collocation (Yarowsky, 1995)
  - redundance atributů

Nebudeme se tu nějak moc věnovat ML, ale přeci jen poskytnu povrchní přehled. S uvedenými termíny se totiž může počítačový lingvista poměrně často setkat, tak ať aspoň ví, na čem je.

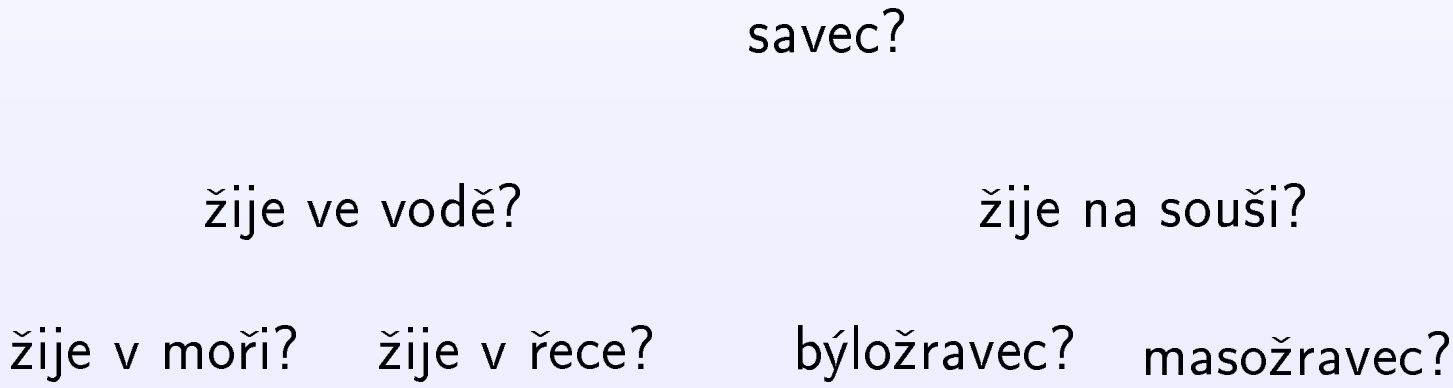
Pro lidi jsou typické (a intuitivní) spíš „pravidlové“ systémy.

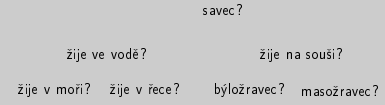
Příkladem je hra Myslím si zvíře (protihráč se snaží zvíře $z \in \mathcal{Z}$ uhádnout pomocí otázek, na které dostává odpovědi ano/ne).

Rozhodovací seznam

```
if (zvíře má chobot) then output(slon)
if (zvíře má pruhy) then output(zebra)
if (zvíře má ploutve & zvíře není ryba) then
output(žralok)
```

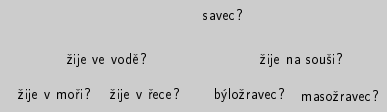
Rozhodovací strom





Seznam je jednodušší na implementaci, ale vidíme, že strom je přehlednější při stejné i vyšší složitosti.

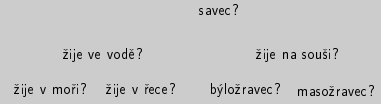
Častěji pracujeme se stromy.



V této hře jsou 2 aspekty:

- Jak poznám z množiny otázek $\mathcal{O} = \{o_1, o_2, \ldots, o_n\}$, kde o_i je např. „Má zvíře srst?“, o jaké zvíře jde? Redukcí. Pokud odpověď na o_i je „ne“, vyloučím ze správných odpovědí všechna zvířata z_j, která mají srst. Podobně pro další otázky, dokud nezůstane (ideálně) 1 zvíře.
- Jaká je strategie kladení otázek? Cílem je minimalizovat n. Prostředkem k dosažení tohoto cíle je neklást otázky, které dělí množinu možných zvířat stejným způsobem. Např. otázky „Má zvíře srst?“ a „Má zvíře 4 nohy?“ dělí $\mathcal{Z}$ na dvě téměř stejné části.

Na celou hru můžeme pohlížet jako na množinu zvířat (které známe) a rozhodovací strom, který nás „dovede“ k vítěznému zvířeti.



„Matematické“ algoritmy zde uvedené jsou každý úplně jiný, spíš jde o reprezentanty různých skupin algoritmů.

Naivní Bayesovský klasifikátor

Naivní Bayesovský alg. předpokládá nezávislost znaků (což nemusí být správně), ale je rychlý.

$$P(C|F_1,\ldots,F_n) = \frac{P(C)\cdot P(F_1,\ldots,F_n|C)}{P(F1,\ldots,F_n)}$$

zvíře	velikost	barva	potrava
slon	velký	šedý	býložravec
slon	střední	šedý	býložravec
kráva	velká	černá	býložravec
kráva	velká	strakatá	býložravec
kráva	malá	strakatá	býložravec
kráva	velká	bílá	býložravec
vlk	velký	černý	masožravec
vlk	malý	šedý	masožravec

Naivní Bayes pro velkého černého býložravce I

zvíře	velikost	barva	potrava
slon	velký	šedý	býložravec
slon	střední	šedý	býložravec
kráva	velká	černá	býložravec
kráva	velká	strakatá	býložravec
kráva	malá	strakatá	býložravec
kráva	velká	bílá	býložravec
vlk	velký	černý	masožravec
vlk	malý	šedý	masožravec

Na základě těchto dat můžeme vypočítat, že zvíře, které vidíme, bude: na 25 % slon, na 50 % kráva a na 25 % vlk, tj. $P(\text{slon}) = \frac{2}{8}$, $P(\text{kráva}) = \frac{4}{8}$ a $P(\text{vlk}) = \frac{2}{8}$.

Podmíněné pravděpodobnosti jsou $P(\text{černá barva}|\text{slon}) = 0$, $P(\text{černá barva}|\text{kráva}) = \frac{1}{4}$, $P(\text{černá barva}|\text{vlk}) = \frac{1}{2}$.

Naivní Bayes pro velkého černého býložravce II

Podmíněné pravděpodobnosti jsou $P(\text{velký}|\text{slon}) = \frac{1}{2}$, $P(\text{velký}|\text{kráva}) = \frac{3}{4}$, $P(\text{velký}|\text{vlk}) = \frac{1}{2}$.

Podmíněné pravděpodobnosti jsou $P(\text{býložravec}|\text{slon}) = \frac{2}{2}$, $P(\text{býložravec}|\text{kráva}) = \frac{4}{4}$, $P(\text{býložravec}|\text{vlk}) = 0$.

Naivní Bayes pro velkého černého býložravce I

$$P(C|F_1,\ldots,F_n) = \frac{P(C)\cdot P(F_1,\ldots,F_n|C)}{P(F1,\ldots,F_n)}$$

$P(\text{slon}) = \frac{2}{8}$, $P(\text{kráva}) = \frac{4}{8}$, $P(\text{vlk}) = \frac{2}{8}$, $P(\text{černý}|\text{slon}) = 0$,
$P(\text{černý}|\text{kráva}) = \frac{1}{4}$, $P(\text{černý}|\text{vlk}) = \frac{1}{2}$, $P(\text{velký}|\text{slon}) = \frac{1}{2}$,
$P(\text{velký}|\text{kráva}) = \frac{3}{4}$, $P(\text{velký}|\text{vlk}) = \frac{1}{2}$, $P(\text{býložravec}|\text{slon}) = \frac{2}{2}$,
$P(\text{býložravec}|\text{kráva}) = \frac{4}{4}$, $P(\text{býložravec}|\text{vlk}) = 0$

$P(\text{slon}|\text{velký, černý, býložravec}) = P(\text{slon})P(\text{velký}|\text{slon}) \cdot$
$P(\text{černý}|\text{slon}) \cdot P(\text{býložravý}|\text{slon}) = 0.25 \cdot 0.25 \cdot 0 \cdot 1 = 0$

$P(\text{kráva}|\text{velký, černý, býložravec}) = P(\text{kráva})P(\text{velký}|\text{kráva}) \cdot$
$P(\text{černý}|\text{kráva}) \cdot P(\text{býložravý}|\text{kráva}) = 0.5 \cdot 0.75 \cdot 0.25 \cdot 1 = 0.09375$

Naivní Bayes pro velkého černého býložravce II

$$P(\text{kráva}|\text{velký, černý, býložravec}) = P(\text{vlk})P(\text{velký}|\text{vlk}) \cdot P(\text{černý}|\text{vlk}) \cdot P(\text{býložravý}|\text{vlk}) = 0.25 \cdot 0.5 \cdot 0.5 \cdot 0 = 0$$

Algoritmy strojového učení

- „pravidlové“
  - rozhodovací seznamy
  - rozhodovací stromy
- „matematické“
  - pravděpodobnostní: naivní Bayesovský (Duda et Hart, 1973)
  - maximální entropie: Berger 1996
  - podobnostní: k-NN ve vektorovém prostoru (Ng et Lee, 1996)
- „promluvové“
  - one sense per discourse (Gale 1992)
  - one sense per collocation (Yarowsky, 1995)
  - redundance atributů

Algoritmus strojového učení [Yarowsky, 1995] I

hledáme význam *s* slova *w*

- 1. vezmi všechny výskyty slova *w* z korpusu včetně jejich kontextů
- 2. pro každý možný význam slova, vytvoř malou sadu příkladů (buď ručně, nebo pomocí kolokací)
- 3. vytvoř rozhodovací seznam s pravděpodobnostmi pro další slova, která se vyskytují v kontextech
- 4. aplikuj tento seznam na celý korpus (s prahem pro pravděpodobnost)
- 5. nově zařazená slova obsahují další slova v kontextech
- 6. algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
- 7. opakuj kroky 3–6

Algoritmus strojového učení [Yarowsky, 1995] II

- 8. jakmile množiny přestanou narůstat, zastav
- 9. systém je nyní natrénovaný i na jiný korpus!

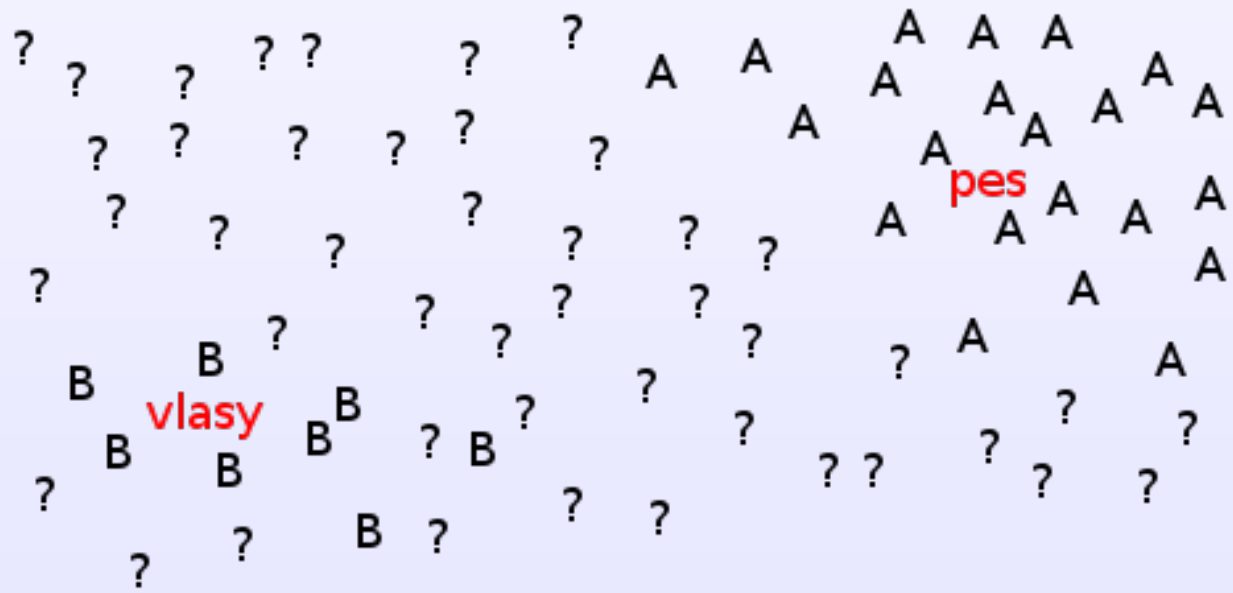
Algoritmus strojového učení [Yarowsky, 1995] III

kočičím emblémem. Na hlavu nespadla žádná	kočka	na fintění, protože to by mě nepustili
Vyjížděli jsme pro jistotu brzy, aby si náhodou	kočky	v poledne neusmyslely, že se kolem nich
šťastně a já nechápala proč, protože vždy na	kočky	nadával. Říkal, že nepatří ani tak k falešným
Pak za mříže a zase na nás. "Taky jdete na	kočky	?" vyzvídala jsem. "Taky. Vypadá to mrtvě
vyťukávání čísla, aby se dozvěděla, co se těm	kočkám	přihodilo. Dozvěděli jsme se akorát bolestivý
místo - místní benzinovou pumpu, jestli ty	kočky	nepřemístili jinam. "Nevíte, kde jsou ty
nepřemístili jinam. "Nevíte, kde jsou ty	kočky	, co měly být na výstavišti?" "Kočky? Zeptejte
ty kočky, co měly být na výstavišti?" "	Kočky	? Zeptejte se kolegyně, já nedávno nastoupila
ale nevím. U nás žádný plakát nevisel." "	Kočky	se nevedou!" povzdechla jsem a zavrtěla
zaslechl a přidal se k hovoru. "Chcete vidět	kočky	?" "Jste místní? Ano, chceme! Tam nahoře
plešatých, mouratých, černých, bílých a jiných	koček	. "Chudinky, trápí je," ozvala jsem se.
se suďte! A ostatní se přidávali, brali	kočky	pro sebe i pro sousedy, pro známé a příbuzné
No a pan z posledního auta si vzal místo	kočky	překvapenou Jarunu a odvezl ji k sobě do
" Jindy se s ní zkouším domluvit: "Hele,	Kočko	, dneska by se mi vážně hodilo, abys zůstala
aby to dneska všechno dobře dopadlo..."	Kočka	je Osobnost a kašle na můj osud. Trénuje
Postupem času zjišťuji, že jsem se naučila brát	Kočku	jako zvířecího kamaráda, který nemá s mým
tlapkách moji budoucnost ani věci příští.	Kočka	je zkrátka Kočka. Ulevilo se mi tímto zjištěním
budoucnost ani věci příští. Kočka je zkrátka	Kočka	. Ulevilo se mi tímto zjištěním. Cítím se
V některých domácnostech spolu kamarádí	kočka	a pes. V jiných zase spolu žijí masochistka
volný čas, energii a peníze na záchranu psů,	koček	a dalších zvířat. Proto tak obdivuji a

```
('pes' in context(w)) then s(w,A)=1
```

Algoritmus strojového učení [Yarowsky, 1995] IV

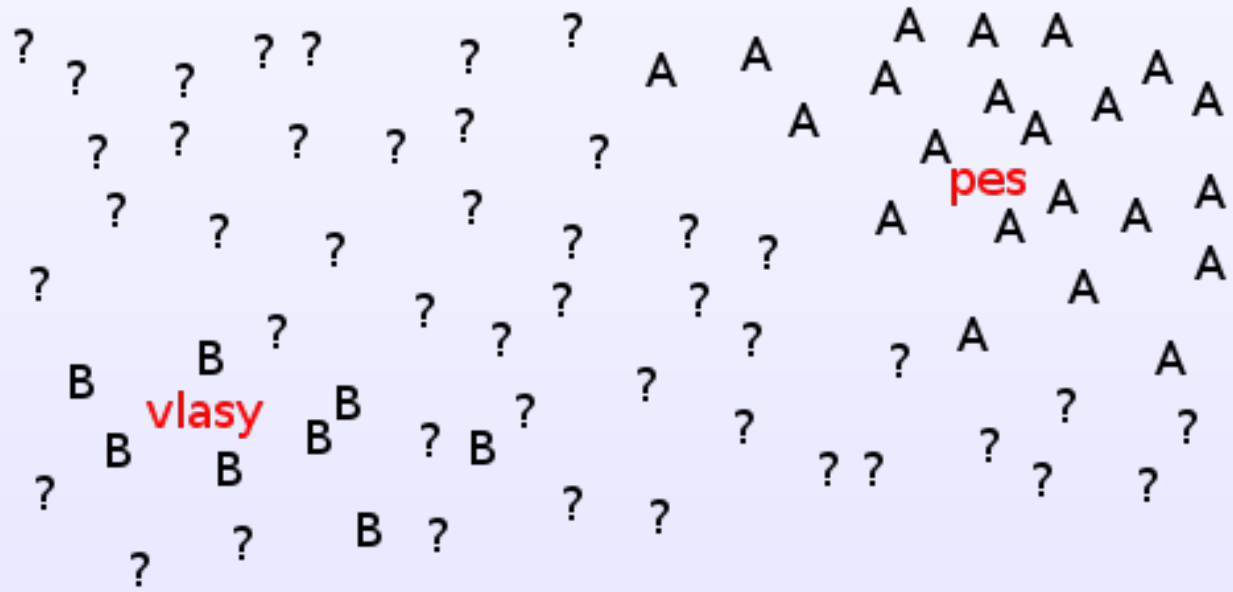
```
if ('vlasy' in context(w)) then s(w,B)=1
if ('prase' in context(w)) then s(w,A)=0.9
if ('prsa' in context(w)) then s(w,B)=1
if ('oči' in context(w)) then s(w,B)=0.6
if ('člověk' in context(w)) then s(w,A)=0.8
```




```
if ('mazlit' in context(w)) then s(w,A)=0.7
```

Algoritmus strojového učení [Yarowsky, 1995] V

```
if ('šaty' in context(w)) then s(w,B)=0.8
if ('boty' in context(w)) then s(w,B)=0.6
if ('kotě' in context(w)) then s(w,A)=0.8
```



Algoritmus strojového učení: shrnutí

1. vezmi všechny výskyty slova *w* z korpusu včetně jejich kontextů
2. pro každý možný význam slova, vytvoř malou sadu příkladů (buď ručně, nebo pomocí kolokací)
3. vytvoř rozhodovací seznam s pravděpodobnostmi pro další slova, která se vyskytují v kontextech
4. aplikuj tento seznam na celý korpus (s prahem pro pravděpodobnost)
5. nově zařazená slova obsahují další slova v kontextech
6. algoritmus můžeme upravit pomocí zařazení předpokladu one-sense-per-discourse
7. opakuj kroky 3–6
8. jakmile množiny přestanou narůstat, zastav
9. systém je nyní natrénovaný i na jiný korpus!

Algoritmus strojového učení

závisí na:

- první volbě kolokací
- způsobu určení pravděpodobnosti: typicky log likelihood $\log \frac{P(senseA,collocateA)}{P(senseB,collocateA)}$
- prahu pro pravděpodobnost
- správnosti předpokladu one-sense-per-discourse

Word Sense Disambiguation

úkolem WSD je zjistit, jaký význam (z inventáře významů) má slovo ve vstupním textu

ukázali jsme si dva reprezentanty metod pro WSD: Leskův algoritmus pracující se slovníkovými definicemi a příklady užití a Yarowského algoritmus strojového učení

Word Sense Disambiguation: slabiny

největší slabinou je inventář významů

proto existují jednak snahy vytvořit dobré inventáře, jednak snahy úplně se inventářím vyhnout (Hyperlex)

největší slabinou je inventář významů

proto existují jednak snahy vytvořit dobré inventáře, jednak snahy úplně se inventářům vyhnout (Hyperlex)

projekt HyperLex je dobrá inpirace pro BP nebo referát

Word Sense Disambiguation: shrnutí

- všechny algoritmy pro WSD pracují s kolokacemi
- všechny pracují s určitým oknem, ve kterém kolokace sledují

- všechny algoritmy pro WSD pracují s kolokacemi
- všechny pracují s určitým oknem, ve kterém kolokace sledují

Ono okno může zásadně ovlivňovat průběhy algoritmů. Není žádná „doporučená velikost“ okna. Hlavním důvodem je to, co možná tušíme: různá slova mají různý „dopad“ na význam promluvy. Sledováním velikosti a kvality tohoto okna (tj. kontextu) se budeme zabývat o něco později, až budeme znát také přístupy z úplně opačného konce.

Word Sense Disambiguation: měření kvality

soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např. přiřazení emoce ke krátkému textu, detekce metonymie . . .)
- čeština (zatím) chybí
- data z proběhlých kol jsou k dispozici

Word Sense Disambiguation: měření kvality

soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např. přiřazení emoce ke krátkému textu, detekce metonymie . . .)
- čeština (zatím) chybí
- data z proběhlých kol jsou k dispozici

Ne, že by bylo třeba se Senseval/Semeval účastnit. Je dobré podívat se na ručně anotovaná data (málokdy je máme). Mnoho prací se také na soutěže odvolává.

soutěž SENSEVAL (www.senseval.org)

- vyhodnocení systémů pro WSD
- od roku 1998 (Senseval-1, -2, -3, Semeval-2007, -2010)
- od Semeval-1 jsou úkoly různé (např. přiřazení emoce ke krátkému textu, detekce metonymie . . .)
- čeština (zatím) chybí
- data z proběhlých kol jsou k dispozici

Cokoli ze Senseval/Semeval je inpirací pro BP nebo referát.


Yarowsky, D. (1995).
Unsupervised word sense disambiguation rivaling supervised methods.
In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.