PLIN037 Sémantika a počítače

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

4. prosince 2014

PLIN037 Sémantika a počítače

Předmět PLIN037 Sémantika a počítače je podpořen projektem OP VK Mezi bohemistikou a informatikou. Inovace vysokoškolské výuky češtiny v kontextu počítačového zpracování přirozeného jazyka (INOVA.CZ).

www.projekt-inova.cz

esf evropský sociální fond v ČR
EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost
UNIVERSITAS MASARYKIANA BRUNENSIS
INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Úspěšnost jazykových nástrojů a zdrojů

Jak měřit? Porovnáním se správnými výsledky:

- ve shodě se “zdravým rozumem”
- ve shodě s experty
- rozdělení dat na trénovací a testovací
- zlatý standard

Testování modelů

v kolika případech došlo ke shodě modelu s informace od učitele?

Testování modelů

v kolika případech došlo ke shodě modelu s informace od učitele?

křížová validace (cross validation): můžeme použít pro klasifikační úlohy (kde máme dost dat)

1. rozdělíme data náhodně v poměru 1:10 až 1:6
2. větší část je trénovací $\Rightarrow$ natrénujeme klasifikátor
3. menší část je testovací
4. výsledkem trénování je model
5. změříme chybu trénovacích dat
6. změříme chybu testovacích dat
7. proces je možné opakovat s jiným rozdělením dat

Testování modelů

matice záměn (confusion matrix): můžeme použít pro klasifikační úlohy o dvou třídách

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

Testování modelů: příklad

Program se učí určovat slovní druhy:

pes	1
kost	1
ležet	5
stát	5
rychle	6
hodný	2
jako	6
jak	1

Testování modelů: příklad

Program se učí určovat slovní druhy:

pes	1	1
kost	1	1
ležet	5	5
stát	5	1
rychle	6	6
hodný	2	2
jako	6	8
jak	1	8

Testování modelů: příklad

Program se učí určovat slovní druhy:

pes	1	1	ok
kost	1	1	ok
ležet	5	5	ok
stát	5	1	ne
rychle	6	6	ok
hodný	2	2	ok
jako	6	8	ne
jak	1	8	ne

Testování modelů: příklad

Program se učí určovat slovní druhy:

pes	1	1	ok
kost	1	1	ok
ležet	5	5	ok
stát	5	1	ne
rychle	6	6	ok
hodný	2	2	ok
jako	6	8	ne
jak	1	8	ne

Umí systém určovat správně podstatná jména?

	co určil systém	
správná klasifikace	+	−
+	2	1
−	1	4

Co říká matice záměn?

Umí systém určovat správně podstatná jména?

	co určil systém	
správná klasifikace	+	−
+	2	1
−	1	4

Co říká matice záměn?

Umí systém určovat správně podstatná jména?

	co určil systém	
správná klasifikace	+	−
+	2	1
−	1	4

Otázka ANO/NE $\Rightarrow$ skóre (reálné číslo)

Co říká matice záměn?

Umí systém určovat správně podstatná jména?

	co určil systém	
správná klasifikace	+	−
+	2	1
−	1	4

Otázka ANO/NE ⇒ skóre (reálné číslo):

na 6 otázek z 8 odpověděl systém správně
na 2 otázky z 8 odpověděl systém špatně

Co říká matice záměn?

Umí systém určovat správně podstatná jména?

	co určil systém	
správná klasifikace	+	−
+	2	1
−	1	4

Otázka ANO/NE ⇒ skóre (reálné číslo):

na 6 otázek z 8 odpověděl systém správně
na 2 otázky z 8 odpověděl systém špatně

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

Skóre

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

Skóre

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

přesnost (precision): $\frac{TP}{TP+FP}$

pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

Skóre

	co určil systém	
správná klasifikace	+	−
+	true positive	false negative
−	false positive	true negative

celková správnost (overall accuracy): $Acc = \frac{TP+TN}{TP+TN+FP+FN}$

celková chyba (overall error): $Err = \frac{FP+FN}{TP+TN+FP+FN}$

přesnost (precision): $\frac{TP}{TP+FP}$

pokrytí/úplnost (recall): $\frac{TP}{TP+FN}$

průměr: $\frac{P+R}{2}$ míra F1 (F1 score): $\frac{2PR}{P+R}$

Skóre a vyváženost tříd (skewed classes)

Příklad: rozložení tříd: 3:97

Systém bude vždy klasifikovat 0

Skóre a vyváženost tříd (skewed classes)

Příklad: rozložení tříd: 3:97

Systém bude vždy klasifikovat 0

	co určil systém	
správná klasifikace	1	0
1	0	3
0	0	97

Kde je pravda?

více anotací, shoda (Interannotation Agreement): Cohen Kappa, Fleiss Kappa

Význam promluvy

Psí granule a kafe.

Význam promluvy

Psí granule a kafe.

- Co si přejete?
- Co se to tady vysypalo?
- Co máte nejraději?
- Cos dnes jedl?
- Co po tobě ten pes chtěl?
- Co po tobě ten člověk chtěl?
- Co po tobě ten člověk mrštil?
- ...

Analýza promluvy: krabicový model

A: Už jsi ten motor smontoval?
– Provleč lano tím okem na horní straně motoru.

– Jo, mimochodem, koupils už ten benzín?
B: Jasně, koupil, když jsem sháněl disk do sekačky.

– Zapomněl jsem vzít kanystr, tak jsem koupil nový.
A: Byl drahý?
B: Ne, ale bude se mi hodit do auta.
A: Fajn.

– Už to máš provlečené?

Prostředky koherence

- časová souslednost (jednota času, místa a děje)
- porušení časové souslednosti je vyjádřeno explicitně: „ještě předtím“
- výrazy jako „Nejprve . . . , potom . . . “, „Oproti tomu . . . “, „také“
- elipsa: Koupila jsem si auto a Marie [si koupila auto] taky.

Elipsa, výpustka (ellipsis)

- Petr šel na večírek, kde [Petr] potkal Pavlu.
- Koupila jsem si auto a Marie [si koupila auto] taky.
- Mám zavolat já tobě, nebo ty [máš zavolat] mně?
- [Mám vám dát na ty brambory] máslo?
- Nevím proč [bych měla tuhle knížku číst].

Promluvové objekty

seznam objektů promluvy (promluvový objekt, PO; discourse entity):

- množina prvků znalostní báze (knowledge base, KB), které byly zmíněny a mohou být odkazovány pomocí zájmen
- pokud prvek nebyl zmíněn, a přesto může být odkazován, byl evokován

jmenná fráze typicky vyjadřuje nějaký PO

Promluvové objekty

seznam objektů promluvy (promluvový objekt, PO; discourse entity):

- množina prvků znalostní báze (knowledge base, KB), které byly zmíněny a mohou být odkazovány pomocí zájmen
- pokud prvek nebyl zmíněn, a přesto může být odkazován, byl evokován

jmenná fráze typicky vyjadřuje nějaký PO

Promluvové objekty

seznam objektů promluvy (promluvový objekt, PO; discourse entity):

- množina prvků znalostní báze (knowledge base, KB), které byly zmíněny a mohou být odkazovány pomocí zájmen
- pokud prvek nebyl zmíněn, a přesto může být odkazován, byl evokován

jmenná fráze typicky vyjadřuje nějaký PO

Promluvové objekty

seznam objektů promluvy (promluvový objekt, PO; discourse entity):

- množina prvků znalostní báze (knowledge base, KB), které byly zmíněny a mohou být odkazovány pomocí zájmen
- pokud prvek nebyl zmíněn, a přesto může být odkazován, byl evokován

jmenná fráze typicky vyjadřuje nějaký PO

Karlovi$_i$ někdo ukradl auto$_j$, které$_j$ [on]$_i$ měl zaparkované před domem$_k$. [on]$_i$ Zavolal na policii$_l$, [oni]$_l$ přijeli, [oni]$_l$ sepsali to$_m$. Za měsíc mu$_i$ [oni]$_l$ napsali, že [oni]$_l$ případ$_m$ odkládají.

Odkazy v diskurzu

exofora (odkaz mimo text)
Co je *to*?

endofora (odkaz do textu)
v takovém případě

anafora (zpětný odkaz) – antecedent (dříve evokovaný PO)
Anežka na *sebe* hodila kabát a vyrazila.

katafora (dopředný odkaz)
Protože *[on]* byl chytrý, vydal se David nejprve za svým šéfem.

koreference: Václav Klaus, Klaus, bývalý prezident, on, čórlpero

druhy anafor:

- deixe: Petr si ukrojil chleba a pak *ho* snědl.
- synonymum: Petr si ukrojil chleba a pak *krajíc* snědl.

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

Karel

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

auto
Karel

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

auto
Karel

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

Karel
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

dům
Karel
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

policie
Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

policie
Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

policie
Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

?
policie
Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

? Karel? dům?
policie
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

policie
? Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

policie
? Karel? dům?
auto

Rozpoznání anafor, rezoluce anafor (anaphora resolution, AR): základní algoritmus

1. objekty promluvy (PO): promluvový zásobník (*history list*)
2. při každé zmínce objektu se PO posune na vrchol zásobníku
3. každý odkaz se nahradí PO, který je nejblíž vrcholu zásobníku a obsahuje gram. shodu (číslo, příp. rod)
4. v jedné klauzi se PO vyskytuje jen jednou

Karlovi někdo ukradl auto, které [on] měl zaparkované před domem. [on] Zavolal na policii, [oni] přijeli, [oni] sepsali to. Za měsíc mu [oni] napsali, že [oni] případ odkládají.

případ? Karel? dům?
policie
auto

Promluvové objekty a znalost světa

Karlovi$_i$ někdo ukradl auto$_j$, které$_j$ [on]$_i$ měl zaparkované před domem$_k$. [on]$_i$ Zavolal na policii$_l$, [oni]$_l$ přijeli, [oni]$_l$ sepsali to$_m$. Za měsíc mu$_i$ [oni]$_l$ napsali, že [oni]$_l$ případ$_l$ odkládají.

Jak poznáme, že to$_m$ =případ$_m$? Jak poznáme, že [oni]$_l$ =policie?

Promluvové objekty a znalost světa

Karlovi$_i$ někdo ukradl auto$_j$, které$_j$ [on]$_i$ měl zaparkované před domem$_k$. [on]$_i$ Zavolal na policii$_l$, [oni]$_l$ přijeli, [oni]$_l$ sepsali to$_m$. Za měsíc mu$_i$ [oni]$_l$ napsali, že [oni]$_l$ případ$_l$ odkládají.

Jak poznáme, že to$_m$ =případ$_m$? Jak poznáme, že [oni]$_l$ =policie?

Potřebujeme znalost o světě.