

Vybrané aktuální projekty Laboratoře NLP

Vašek Němčík, Zuzana Nevěřilová

E-mail: xnemcik@fi.muni.cz, xpopelk@fi.muni.cz
http://nlp.fi.muni.cz/poc_lingv/

Obsah:

- Saara – systém na určování anafor
- X-plain – hra, která sbírá Common Sense

Gracie: Minulý týden byl můj bratr vyšetřovat vraždu a představ si, našel toho chlapa za hodinu.

George: Našel za hodinu vraha?

Gracie: Ne, *toho chlapa, co ho zabili.*

George: Tvůj bratr je nejen vysoký, ale i rychlý.

Gracie: No, před časem měli pan a paní Jonesovi manželskou krizi a můj bratr byl najat, aby sledoval paní Jonesovou.

George: No, je to nepochybně moc atraktivní žena.

Gracie: To je – a můj bratr ji sledoval ve dne v noci, půl roku.

George: A jak to skončilo?

Gracie: Požádala o rozvod.

George: Paní Jonesová?

Gracie: Ne, *žena mého bratra.*

(George Burns a Gracie Allen: "The Salesgirl")

- text/**diskurs** – jednotka jazykové komunikace větší než:
- věta/**výpověď** – minimální obsahově úplná jednotka

věta

langue

competence

*produkt**struktura**nedůležité kdy/kde/jak***výpověď**

parole (de Saussure)

performance (Chomsky)

*proces**chování**podmínky/okolnosti/způsob*

- referenční výrazy
- reference (odkazování)

jazykový výraz \mapsto mimojazyková entita

Reference

- **exofora** (vnější reference)

výraz odkazuje k entitě ve světě přímo

“Slunce”, “Alpy”, “Václav Havel”, “ty schody před FI”

- deixis – odkazování k entitám v rámci komunikační situace (gesta, “tady”, “ted”, “tamto”, ...)

- **endofora** (vnitřní reference)

entita je určena na základě vztahu k jinému výrazu v diskursu (nejen mimojazykový, ale i jazykový kontext ...)

- anafora – výraz se vztahuje k výrazu dříve v textu
- katafora – výraz se vztahuje k výrazu dále v textu méně častá; vysktuje se v beletrii (zvyšuje napětí):
“Ranní světlo ho probudilo už v pět. Rychle se oblékl a nasnídal. Detektiv Jones věděl, že nemůže ztrácet čas.”

Anafora

- **anafora** (anaphor) – anaforický výraz (× Chomsky)
 - zejména zájmena, ale i “ten muž”, ...
- **antecedent** – předcházející výraz, ke kterému se anafora vztahuje
- **anafora** (anaphora) – anaforická reference (jev)
- **anaphora resolution** – určování anaforických vztahů (hledání vztahů mezi anaforami a antecedenty)

Příklady:

- **[Petr]_i**; snědl **[koláč]_j**.
[(on)]_i; Byl hladový a **[ten koláč]_j**; vypadal lahodně.
- **[Venus]_i**; rose at 0930, but I didn't see **[the thing]_i**.
- **[Jones]_i**; offered **[[his]_i; furniture]_j** for sale,
but nobody wanted **[the stuff]_j**.

Lze udělat úkrok stranou?

- Můžeme se tomu všemu vyhnout, třeba používáním jen přímé reference?
- Nemuseli bychom se zabývat kontextem ...

NE. Z mnoha vážných důvodů:

- Lidé jsou líní.
 - anafory jsou krátké a snadno se používají
 - patrně vlastní lidské komunikaci (ve všech jazycích!)
 - diskurs není libovolná sekvence výpovědí
 - koherence – sémantická návaznost
 - kohese – gramatické a lexikální vztahy
- ↪ anaforické vztahy drží text pohromadě
(umožňují nám se držet zamýšleného toku myšlenek)

Ilustrační příklad

[Jarda]; si koupil Porsche. (On); Rád jezdí rychle.

[Jarda]; si koupil Porsche. [Jarda]_{*i,j} rád jezdí rychle.

→ delší/složitější věta zní divně (nutí k zamyšlení)

- **Kooperační princip** (Grice)

Komunikační maximy:

- kvality
- relevance
- kvantity
- způsobu
- Posluchač předpokládá, že se jimi mluvčí řídí.
- Když ne, má to hlubší důvody.
- více o pragmatice v “IA091 Sémantika a komunikace”

Proč to učit počítače?

- zásadní úzké hrdlo mnoha NLP aplikací

- **Information Extraction**

- **[Václav Havel]** is a Czech writer and dramatist.
[He] was the ninth and last President of Czechoslovakia and the first President of the Czech Republic. (*Wikipedia*)

- “the best doctor in Europe” → Google

Letters from Asia addressed loosely to The Best Doctor in Europe arrived on **[his]** doorstep.

[His] own reputation as the best doctor in Europe couldn't save **[him]** from the tragedies of **[his]** life.

- Bez AR nenajdeme to, co hledáme.

Pouze anaforické výrazy (které jsou samy o sobě prázdné).

Proč to učit počítače?

- **Strojový překlad**

- CZ \mapsto EN

[Sestřička] mu dala **[pilulku]**. Spolkl **[ji]** a do minuty usnul.

[The nurse] gave him a pill. He swallowed **[her]** and fell asleep in a minute.

- DE \mapsto EN

Ich suche **[meine Uhr]**. Ich kann **[sie]** nirgendwo finden.

I am looking for **[my watch]**. I can't find **[her]** anywhere.

- nelze překládat přímo (různé gramatické kategorie)
- navíc: různé vlastnosti anafor

Definice úlohy

- nalézt anaforické výrazy v textu
- určit k nim antecedenty
- určit typ vztahu
 - koreference
(dva výrazy se odkazují ke stejnému promluvoému objektu)
 - bridging (asociativní/nepřímá anaphora)
(jakákoliv sémantická relace)
 - hyperonymie/hyponymie
“Nábytek je drahý. Židle jsou nejdražší.”
 - část/celek
“Každý majitel bytu se snaží zabezpečit vchodové dveře.”
 - entita/vlastnost
“Pepa má nové auto. Barvu určitě vybírala jeho žena.”
 - příčina/následek
“Včera tu byl požár. Kouř je tu stále cítit.”

Typy anafor

- **textová vs. gramatická**

[Ben] takes a photo of [himself] every day.

- **pronominální** (pro NLP asi nejrelevantnější)

- **nominální**

Od září bude do [Brna] létat nová letecká linka. Očekává se, že přinese [druhému největšímu městu ČR] nové turisty.

- **slovesná**

John likes cats. So does Bill.

- **one-anaphora**

John has a black Porsche. I would like one too.

- **nulová (zero) anafora**

anafora není povrchově realizována

v češtině (a ostatních pro-drop jazycích) nevyjádřené podmínky

Typy pronominálních anafor

- osobní zájmena
 - silná: “jemu”, “on”, “ona”
 - slabá: “mu”, “ho” (klitika)
 - nulová: ∅
- demonstrativní zájmena: “ten”, “ta”, “tomu”
- reflexivní zájmena: “se”, “sebe”, “svůj”
- posesivní zájmena: “jeho”, “jejího”
- relativní zájmena: “který”, “jenž”

ALE jsou i neanaforická zájmena:

- deixe: “to”
- expletivní/pleonastická zájmena:
It's raining. / Es regnet.
It is the first chapter, I enjoy the most.
Zdá se, že tu někdo byl.

Znalosti potřebné pro AR

• morfologie

- shoda v Φ -atributech (závislé na jazyce)
- čeština: osoba, číslo, rod
- angličina: pouze sémantický rod
⇒ nutnost mít informaci jméno \mapsto rod

• syntax

- posice anafory/antecedentu v syntaktické struktuře věty
- paralelismus
tendence k zachování stejných syntaktických rolí:
[Mary] met [Lucy] at the bus station.
[She] asked [her] about the new neighbour.

• pragmatika

- Griceův kooperační princip ...
- komunikační situace + kontext
- scénáře

Sémantika a znalosti o světě

- hraje při interpretaci anafor často rozhodující roli
- sémantická plausibilita zvyšuje/snižuje pravděpodobnost některé interpretace, některé lze zcela vyloučit

After the [bartender] served [the patron], [he] got a big tip. After the [bartender] served [the patron], [he] left a big tip.

- iniciální interpretace (hned)
- pokud pozdější informace vedou ke sporu:
reinterpretace (backtracking)
- **garden-path effect**
- význam slov
- znalosti o světě
- inference

Sémantika a znalosti o světě

- If the baby does not thrive on raw milk, boil it.
- The FBI's role is to ensure our country's freedom and be ever watchful of those who threaten it.
- Stehlíková ustoupila od sbírky. Romové o ni nestojí.
- Klaus dostal dopis podepsaný Aničkou. Má ho policie.
- A: I ve Veselé vačici by mohla být volná místa.
B: Jé, tam jsem ještě nebyla. Slyšela jsem, že tam chodí studenti. A že prý dobře vaří.
- 'I said disarm only!' Lockhart shouted in alarm over the heads of the battling crowd, as Malfoy sank to his knees; Harry had hit him with a Tickling Charm, and he could barely move for laughing.
(*J. Rowling: Harry Potter and the Chamber of Secrets*)

Sémantika a znalosti o světě

- Genau so sei es ihm vorgekommen, sagte Gauss, schief ein und wachte bis zum abendlichen Pferdewechsel an der Grenzstation nicht mehr auf. Während die alten Pferde ab- und neue angeschirrt wurden, assen sie Kartoffelsuppe in einer Gastwirtschaft.
(Daniel Kehlmann: "Die Vermessung der Welt: Die Reise")

- všechny tyto znalosti je obtížné shromáždit
- i kdyby byly k dispozici, bylo by obtížné v nich hledat
- AR je považováno za **"AI-úplný problém"**
AR je stejně obtížný problém jako naučit počítače myslet.
⇒ nutno si úkol zúžit

Teoretické problémy

- John loves his wife. So does Bill.
- The man who gave his **[paycheque]** to his wife was wiser than the man who gave **[it]** to his mistress.
- If any man owns **[a donkey]**, he beats **[it]**.
- **[No one]** will be admitted to the examination, unless **[he]** has registered four weeks in advance.
- **[The man who shows he deserves [it]]** will get **[the prize [he] desires]**.

AR algoritmy

- heuristická pravidla (70. léta)
 - SHRDLU – “block world” Terryho Windograda
 - [Hobbsovo syntaktické hledání](#)
 - jednoduchá pravidla, vzory, časté instance
- sématické teorie
 - centering, focusing – modelování lokální koherence
 - [BFP algoritmus](#)
 - výpočetně problematické
- knowledge-poor (90. léta)
 - kacířství motivované praktickými potřebami
 - založené na datech, která lze dostatečně úspěšně spočítat (morfologie, povrchová syntax, jednoduché sémantické třídy)
 - [RAP](#) – váhování
 - CoGNIAC (pouze 6 pravidel – vysoká přesnost, malé pokrytí)
 - MARS – váhování

AR a strojové učení

- statistika a strojové učení dnes v NLP převažují
- AR není klasifikační problém

předefinování umožňující použití std. ML metod:

- **1 instance**: dvojice anafora-antecedent
- **atributy**: knowledge-poor informace
- **cílový atribut**: 1 pro koreferentní dvojici, jinak 0
- velký nepoměr negativních a pozitivních instancí
- nutno část negativních instancí odstranit z trénovacích dat

AR a čeština

- mnoho teoretických prací (FGP: Sgall, Hajičová)
- PDT 2.0 – velký ručně anotovaný korpus, 3 roviny
- anotace pronominání koreference
- implementace:
Zdeněk Žabokrtský, Nguy Giang Linh
- pouze v rámci formalismu PDT 2.0

- **Saara**
- modulární
- různé algoritmy, formáty dat, pre-processing
- možnost férového porovnání algoritmů

- roviny abstrakce:
 - technická rovina
různé formalismy/formáty dat \mapsto několik obecných metod
 - “markable” rovina
definuje AR algoritmy
metody + “markables” jejich vlastnosti
 - “supervisor”
definuje, který preprocessing/algorithmus se použije
- “markable”
 - jakákoliv jednotka složená z jednodušších jednotek
 - možno definovat různé roviny
 - slova \mapsto klause
 - slova \mapsto referenční výrazy
 - atributy
 - vztahy mezi markables \rightsquigarrow koreferenční třídy
 - MMAX 2

Saara

- načtení dokumentu
- pre-processing
 - rozdělení vět do klauzí
 - detekce nevyjádřených subjektů
 - model diskursu – detekce markables
- AR \rightsquigarrow koreferenční třídy
- výstup: MMAX2 XML

Hobbs syntactic search

- jako syntaktickou strukturu předpokládá frázové stromy
- X-bar theory (Chomsky, Jackendoff)
X – complement – X' – adjunct – X' – specifier – XP
- algoritmus je definován jako procházení stromu
- začíná se v listu dané anafory
- podle kategorie aktuálního uzlu se volí další cesta
- prominentnější posice jsou procházeny dříve
- lze adaptovat na jiné formalismy
- jednoduché, ale nefunguje špatně

BFP algoritmus

- každá výpověď:
 - forward-looking centers (setříděné)
 - preferred center (ten nejdříve postavený)
 - backward-looking center
- formulována 2 jednoduchá pravidla, neformálně:
- preferováno je odkazování zájmeny
- preferováno je zachovávání backward-looking center
- počítají se různé kombinace a filtrují se ty, které nevyhovují pravidlům

RAP

- identifikace NP, filtrování nereferenčních, reflexiva atd.
- přidělí se iniciální váhy kandidátům (součet)
- při hledání antecedentu ke konkrétní anafoře se pro danou kombinaci váhy dále upravují (katafora, paralelismus, ...)
- antecedentem je kandidát s nejvyšší vahou
- při zpracovávání nové věty se všechny váhy podělí dvěma

<i>Factor type</i>	<i>Initial weight</i>
Sentence recency	100
Subject emphasis	80
Existential emphasis	70
Accusative emphasis	50
Indirect object and oblique complement emphasis	40
Head noun emphasis	80
Non-adverbial emphasis	50

Pražské algoritmy

- hned několik algoritmů
- formulovány “na papíře”
- vyhodnocovány ručně
- jako RAP také váhovací princip
- modeluje aktivaci objektu v mysli posluchače
- zohledňuje se informace o AČV
- teoreticky logické, ale prakticky nepotvrzené

Gracie: Last week my brother went out on a murder case, and you know, he found that man in an hour.

George: He found the murderer in an hour?

Gracie: No, *the man who was killed*.

George: Not only is your brother tall, but he's fast.

Gracie: And then Mr. & Mrs. Jones were having matrimonial trouble, and my brother was hired to watch Mrs. Jones.

George: Well, I imagine she was a very attractive woman.

Gracie: She was, and my brother watched her day and night for six months.

George: Well, what happened?

Gracie: She finally got a divorce.

George: Mrs. Jones?

Gracie: No, *my brother's wife*.

(George Burns and Gracie Allen in "The Salesgirl")

Obsah

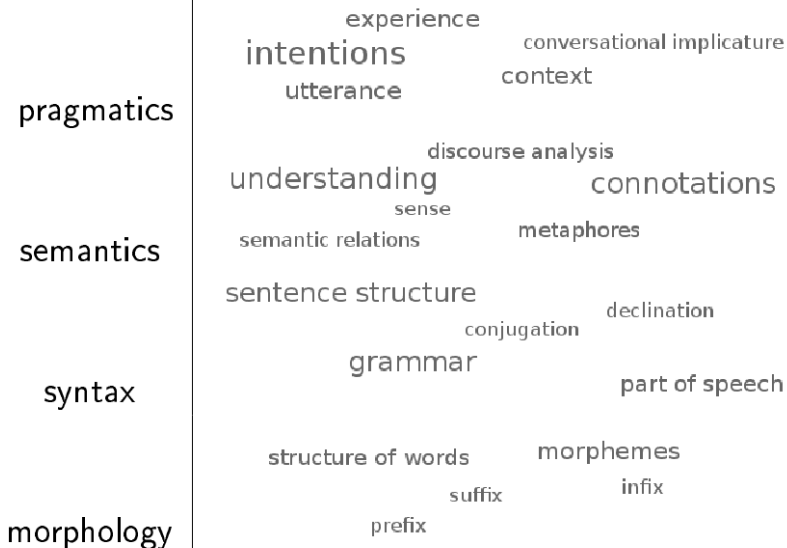
- 1 Saara – systém na určování anafor
 - Anafora
- 2 X-plain – hra, která sbírá Common Sense
 - Úvod
 - Common sense
 - X-plain
 - Výsledky

Úvodem

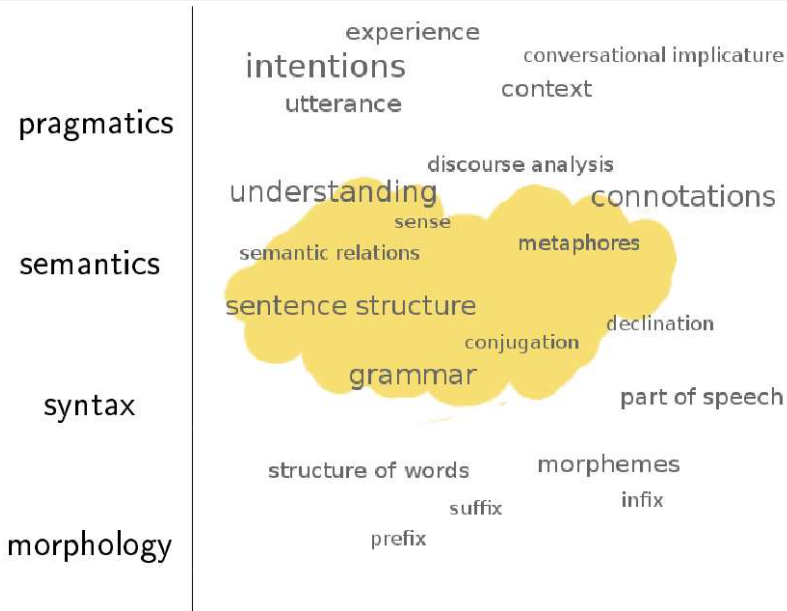
"People have silly reasons why computers don't really think. The answer is we haven't programmed them right; they just don't have much common sense"

– Marvin Minsky

Zpracování přirozeného jazyka



Zpracování přirozeného jazyka



Common sense

„znalosti a zkušenosti, které většina lidí má, nebo se věří, že má, nebo by měla mít“

česky: přirozený lidský rozum, prostý, zdravý rozum, selský rozum, praktické uvažování

„obrovská množina procesů přirozeného poznání a systém domněnek, které (všichni) lidé sdílejí“

česky: common sense ;-)

Common sense

„znalosti a zkušenosti, které většina lidí má, nebo se věří, že má, nebo by měla mít“

česky: přirozený lidský rozum, prostý, zdravý rozum, selský rozum, praktické uvažování

„obrovská množina procesů přirozeného poznání a systém domněnek, které (všichni) lidé sdílejí“

česky: common sense ;-)

Common sense – kde najít?

V encyklopediích?

matted or felted sheet, usually made of cellulose fibres, formed on a wire screen from water suspension.

– Encyclopædia Britannica

a thin substance made from any of a variety of fibrous materials, used for writing, wrapping, decorating walls, and the like.

– Wordsmyth

Common sense – v korpusu?

investičních fondech, o burze cenných
 vynikajícím tiskem na kvalitním křídovém
 zareagoval, se slovem v éteru (s popsaným
 pohybovat ve světě makléřů a cenných
 ulicích v Praze, pět domů ve Vídni a cenné
 tisku, o privatizaci ČTK, o nové ceny
 své reklamní předměty z recyklovaného
 vstupu do sídla firmy, přes hlavičkové
 dodávky kopírek, dodává kancelářský
 třeba že jsem dlouho odvykal tužce a
 zbastlená po domácku, na kancelářském
 nevyklučuje kýč, (viz obr. d) - a tužší
 se však barvitosti dosahuje tónováním
 480X360 pixelů) Půdorysy, pohledy, řezy na
 spíše elementárním pojmům, jako je
 ochrannou mříž spatříme záplavu odpadků,
 ledání. **</p><p>** Notes je něčím mezi listem
 Je to prostě poznámkový blok nebo list
 schůze. Podobně jako konceptní papír i
 počítače. Jednou budou elektronické listy

papírů , zákon směnečný a šekový, o dluhopisech
papíře . Stal se ideálním prostředkem pro prezentaci
papírem , s obrazovkou) se těžko bojuje. A navíc
papírů . Umělecky založené návštěvníčky naláká
papíry v hodnotě asi 2 miliony korun). Hybnou
papíru atd. **</p><p>** ...k dneškuStejně jako se
papíru . Stejně dobře zapůsobí, koresponduje
papíru , vizitky, tištěné materiály až po
papír všeho druhu, tonery pro kopírky a laserové
papíru a psacímu stroji, je dnes počítač pro
papíru nebo fopapíru. Pérovky, fotografie,
papír : možná proto, že z 18 kousků jde až
papíru), 3 kousky jsou tištěny dvoubarevně,
papíře ve formátu max. A3 Aspoň 5, avšak ne
papír , tužka, štětec, hlína a podobně. Dobře
papíru , slupky od banánů... Rudolfova štola
papíru a dnešním laptopem. Příkladem je prototyp
papíru položený na stole, kde je k dispozici
papíry elektronické můžete rozprostřít na
papíru tenké téměř jako listy skutečného

Common sense – specializované kolekce

- CyC – tvořeno experty od r. 1984
(`#$isa #$BillClinton #$UnitedStatesPresident`)
- Thought Treasure – od r. 1993, v r. 2000 vývoj ukončen
[`product-of Intel-8080 Intel`]
(An Intel 8080 is a product of Intel.)
- OpenMind – tvořeno průběžně uživateli, od r. 1999
"The sun is very hot"

X-plain – další způsob, jak sbírat common sense

hra, podobná Kufru nebo Aktivitám

hráči = přispěvatelé do databáze common sense

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač

vypravěč, hadač a tajné slovo

vypravěč vysvětluje tajné slovo, hadač zkouší hádat

časový limit je 3 minuty

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač

vypravěč, hadač a tajné slovo

vypravěč vysvětluje tajné slovo, hadač zkouší hádat

časový limit je 3 minuty

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač

vypravěč, hadač a tajné slovo

vypravěč vysvětluje tajné slovo, hadač zkouší hádat

časový limit je 3 minuty

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač
vypravěč, hadač a tajné slovo

vypravěč vysvětluje tajné slovo, hadač zkouší hádat
časový limit je 3 minuty

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač
vypravěč, hadač a tajné slovo
vypravěč vysvětluje tajné slovo, hadač zkouší hádat
časový limit je 3 minuty

X-plain – pravidla

kooperativní hra pro 1 člověka a počítač
vypravěč, hadač a tajné slovo
vypravěč vysvětluje tajné slovo, hadač zkouší hádat
časový limit je 3 minuty

X-plain – ukázka

X-plain a common sense

trojice: subjekt – relace – objekt,
kde

- subjekt je tajné slovo
- relace je použitá relace
- objekt je nápověda

X-plain a common sense

trojice: subjekt – relace – objekt,
kde

- subjekt je tajné slovo
- relace je použitá relace
- objekt je nápověda

X-plain a common sense

sport	can_be_likely_found	hřiště
sport	can_have_property	bojový
sport	is_related_to	pohyb
sport	is_the_opposite_of	lenošení
sport	is_used_for	udržení kondice
fanoušek	is_part_of	sport
hra	can_be_used_for	sport
sportovec	has_part	sport
trofej	is_used_for	sport
fanoušek	is_related_to	sport
fanoušek	is_for	sport
areál	is_for	sport
fanoušek	can_be_likely_found	tribuna

X-plain – jak je to udělané?

- trojice se ukládají do MySQL, počítač používá data z MySQL + Sketch Engine
- aplikační logika: PHP + JavaScript + AJAX¹
- běží ve webovém prohlížeči

¹Asynchronous JavaScript and XML

X-plain – no, dobře, ale jak je to udělané?

papír preloaded/czes freq = 93440

a_modifier	59603	2.8	prec_s	7769	9.9	prec_pro	3484	6.3	prec_na	5508	3.1	post_pro	1548	2.8
cenný	<u>33859</u>	13.18	obchodník	<u>3419</u>	10.92	komise	<u>3113</u>	8.86	kapička	<u>158</u>	9.19	velk	<u>507</u>	12.79
balící	<u>1483</u>	9.61	obchodování	<u>1090</u>	9.94	Komis	<u>15</u>	7.12	toner	<u>160</u>	9.11	laser	<u>575</u>	10.19
toaletní	<u>1029</u>	8.96	manipulace	<u>390</u>	8.14	černé	<u>14</u>	5.59	páska	<u>168</u>	7.96	snížení	<u>67</u>	4.7
fotografický	<u>580</u>	8.0	potápěč	<u>105</u>	7.99	úřad	<u>222</u>	3.87	zásobník	<u>165</u>	7.73	tisk	<u>100</u>	4.44
křídový	<u>438</u>	7.89	obchodovat	<u>100</u>	7.45	středisko	<u>7</u>	0.23	tisknout	<u>165</u>	7.67	podávání	<u>8</u>	4.15
běžný	<u>1883</u>	7.74	transakce	<u>131</u>	6.67				kontejner	<u>99</u>	7.29	tiskárna	<u>22</u>	3.32
hlavičkový	<u>383</u>	7.68	obchod	<u>1101</u>	6.54	prec_pomocí	61	3.2	tužka	<u>42</u>	6.85	médium	<u>20</u>	2.37
depilační	<u>361</u>	7.61	operace	<u>207</u>	5.93	zbroušení	<u>48</u>	13.05	natisknout	<u>19</u>	6.71	potřeba	<u>12</u>	0.68
ruční	<u>612</u>	7.57	plech	<u>31</u>	5.53				vystřelovat	<u>19</u>	6.56	kancelář	<u>8</u>	0.14
zaknihovaný	<u>331</u>	7.49	velkoobchod	<u>14</u>	5.03				inkoust	<u>27</u>	6.34	post_se	303	2.8
dluhový	<u>328</u>	7.46	trh	<u>521</u>	4.97				tisk	<u>329</u>	6.12	znaménko	<u>73</u>	9.1
recyklovaný	<u>291</u>	7.29	kombinace	<u>70</u>	4.6				rozmnoženina	<u>14</u>	6.09	splatnost	<u>72</u>	7.2
brusný	<u>274</u>	7.18	podvod	<u>21</u>	4.48				psát	<u>356</u>	6.07	sídlo	<u>23</u>	3.05
účastnický	<u>351</u>	7.08	kontrakt	<u>20</u>	4.16				nakreslit	<u>21</u>	6.03	záruka	<u>7</u>	1.14
novinový	<u>263</u>	7.07	tužka	<u>7</u>	4.02				fax	<u>31</u>	5.86	jméno	<u>12</u>	0.56
smirkový	<u>242</u>	7.05	nakládání	<u>11</u>	3.75				kreslení	<u>17</u>	5.8			
kancelářský	<u>424</u>	7.01	spekulace	<u>9</u>	3.34				přenést	<u>50</u>	5.73			
listinný	<u>232</u>	6.94	zacházení	<u>7</u>	3.29				vytisknout	<u>108</u>	5.65			
obchodovatelný	<u>202</u>	6.77	burza	<u>9</u>	2.48				čmárat	<u>8</u>	5.49			
pečlivý	<u>187</u>	6.66	držák	<u>10</u>	2.1				kopírování	<u>24</u>	5.47			
čistý	<u>438</u>	6.61	vzor	<u>7</u>	2.1				poklepávat	<u>7</u>	5.31			
lesklý	<u>190</u>	6.54	práce	<u>73</u>	1.67				potisk	<u>13</u>	5.23			
balící	<u>170</u>	6.53	pracovat	<u>24</u>	1.45				akreditace	<u>14</u>	5.11			
obchodovaný	<u>168</u>	6.48							kresba	<u>25</u>	5.01			
sběrový	<u>162</u>	6.47							hodit	<u>46</u>	4.97			

X-plain – no, dobře, ale jak je to udělané?

=coord

*SYMMETRIC

```
1:[] [word = "a" | word = "nebo"] 2:[] & 1.k=2.k & 1.c=2.c [word
= "ani"] 1:[] [word = ",,"]? [word = "ani"] 2:[] & 1.k=2.k &
1.c=2.c [word = "bud'"] 1:[] [word = ",,"] [word = "nebo"] 2:[]
& 1.k=2.k & 1.c=2.c
```

*DUAL

=a_modifier/modifies

```
2:[tag="k2.*"] [tag="k2.*"|word=", "|word="se"|word="si"]{0,5}
1:[tag="k1.*"] & 1.c=2.c 2:[tag="k2.*"] [word="a" | word="nebo"]
[tag="k2.*"|word=", "|word="se"|word="si"]{0,5} 1:[tag="k1.*"]
& 1.c=2.c [tag="k2.*"|word=", "|word="se"|word="si"]{0,5}
[word="a" | word="nebo"] 2:[tag="k2.*"] 1:[tag="k1.*"] & 1.c=2.c
1:[tag="k1.*"] 2:[tag="k2.*"] & 1.c=2.c
```

*TRINARY

=post_%s

```
1:[tag="k1.*"] 3:[tag="k7.*"] [tag="k2.*"|word=", "|word="se"|word="si"
2:[tag="k1.*"] & 3.c = 2.c 1:[tag="k5.*"] 3:[tag="k7.*"]
[tag="k2.*"|word=", "|word="se"|word="si"]{0,5} 2:[tag="k1.*"]
& 3.c = 2.c
```

X-plain – od gramatiky k sémantice

Pravidla od gramatiky k sémantice:

```
is_similar_to=["coord"]  
can_have_property=["a_modifier"]  
is_the_opposite_of=["prec_oproti", "prec_vůči", "post_vůči", "post_oproti"]  
can_be_likely_found=["post_blízko", "post_nedaleko", "post_na"]
```

Reverzní pravidla od gramatiky k sémantice:

```
is_similar_to=["coord"]  
can_have_property=["modifies"]  
is_the_opposite_of=["prec_oproti", "prec_vůči", "post_vůči", "post_oproti"]  
can_be_likely_found=["prec_blízko", "prec_nedaleko", "prec_na"]
```

X-plain – od gramatiky k sémantice

Program v roli vypravěče:

- hledá kandidáty na náповědu v databázi (pro každou relaci i získá množinu R_i)
- vytvoří word sketch pro tajné slovo
- na základě pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro každou relaci i získá množinu S_i)
- vybere náhodné prvky z T_i , kde $T_i = R_i \cup S_i$ pro každou relaci i

X-plain – od gramatiky k sémantice

Program v roli vypravěče:

- hledá kandidáty na náповědu v databázi (pro každou relaci i získá množinu R_i)
- vytvoří word sketch pro tajné slovo
- na základě pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro každou relaci i získá množinu S_i)
- vybere náhodné prvky z T_i , kde $T_i = R_i \cup S_i$ pro každou relaci i

X-plain – od gramatiky k sémantice

Program v roli vypravěče:

- hledá kandidáty na náповědu v databázi (pro každou relaci i získá množinu R_i)
- vytvoří word sketch pro tajné slovo
- na základě pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro každou relaci i získá množinu S_i)
- vybere náhodné prvky z T_i , kde $T_i = R_i \cup S_i$ pro každou relaci i

X-plain – od gramatiky k sémantice

Program v roli vypravěče:

- hledá kandidáty na nápovědu v databázi (pro každou relaci i získá množinu R_i)
- vytvoří word sketch pro tajné slovo
- na základě pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro každou relaci i získá množinu S_i)
- vybere náhodné prvky z T_i , kde $T_i = R_i \cup S_i$ pro každou relaci i

X-plain – od gramatiky k sémantice

Program v roli hadače:

- hledá kandidáty na odpověď v databázi (pro danou relaci získá množinu R)
- vytvoří word sketch pro objekt
- na základě reverzních pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro danou relaci a objekt získá množinu S_i)
- vybere náhodné prvky z T , kde $T = R \cup S$ pro danou relaci

X-plain – od gramatiky k sémantice

Program v roli hadače:

- hledá kandidáty na odpověď v databázi (pro danou relaci získá množinu R)
- vytvoří word sketch pro objekt
- na základě reverzních pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro danou relaci a objekt získá množinu S_i)
- vybere náhodné prvky z T , kde $T = R \cup S$ pro danou relaci

X-plain – od gramatiky k sémantice

Program v roli hadače:

- hledá kandidáty na odpověď v databázi (pro danou relaci získá množinu R)
- vytvoří word sketch pro objekt
- na základě reverzních pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro danou relaci a objekt získá množinu S_i)
- vybere náhodné prvky z T , kde $T = R \cup S$ pro danou relaci

X-plain – od gramatiky k sémantice

Program v roli hadače:

- hledá kandidáty na odpověď v databázi (pro danou relaci získá množinu R)
- vytvoří word sketch pro objekt
- na základě reverzních pravidel od gramatiky k sémantice hledá kandidáty z word sketch (pro danou relaci a objekt získá množinu S_i)
- vybere náhodné prvky z T , kde $T = R \cup S$ pro danou relaci

Příklad

Tajné slovo je „kometa“

- vypravěč (člověk) vyplní šablonu: ... se skládá z ohonu
- hadač (program) získá $R = \{„liška“\}$ z databáze a $S = \{„kůň“, „kráva“, \dots\}$ z word sketch, hádá „kůň“
- vypravěč (člověk) vyplní šablonu: ... může mít vlastnost Halleyova
- hadač (program) získá $R = \emptyset$ z databáze a $S = \{„kometa“\}$, hádá „kometa“
- bingo!

Příklad

Tajné slovo je „kometa“

- vypravěč (člověk) vyplní šablonu: ... se skládá z ohonu
- hadač (program) získá $R = \{„liška“\}$ z databáze a $S = \{„kůň“, „kráva“, \dots\}$ z word sketch, hádá „kůň“
- vypravěč (člověk) vyplní šablonu: ... může mít vlastnost Halleyova
- hadač (program) získá $R = \emptyset$ z databáze a $S = \{„kometa“\}$, hádá „kometa“
- bingo!

Příklad

Tajné slovo je „kometa“

- vypravěč (člověk) vyplní šablonu: ... se skládá z ohonu
- hadač (program) získá $R = \{„liška“\}$ z databáze a $S = \{„kůň“, „kráva“, \dots\}$ z word sketch, hádá „kůň“
- vypravěč (člověk) vyplní šablonu: ... může mít vlastnost Halleyova
- hadač (program) získá $R = \emptyset$ z databáze a $S = \{„kometa“\}$, hádá „kometa“
- bingo!

Příklad

Tajné slovo je „kometa“

- vypravěč (člověk) vyplní šablonu: ... se skládá z ohonu
- hadač (program) získá $R = \{„liška“\}$ z databáze a $S = \{„kůň“, „kráva“, \dots\}$ z word sketch, hádá „kůň“
- vypravěč (člověk) vyplní šablonu: ... může mít vlastnost Halleyova
- hadač (program) získá $R = \emptyset$ z databáze a $S = \{„kometa“\}$, hádá „kometa“
- bingo!

Příklad

Tajné slovo je „kometa“

- vypravěč (člověk) vyplní šablonu: ... se skládá z ohonu
- hadač (program) získá $R = \{„liška“\}$ z databáze a $S = \{„kůň“, „kráva“, \dots\}$ z word sketch, hádá „kůň“
- vypravěč (člověk) vyplní šablonu: ... může mít vlastnost Halleyova
- hadač (program) získá $R = \emptyset$ z databáze a $S = \{„kometa“\}$, hádá „kometa“
- bingo!

Ach ten češtin

vypravěč píše krátké, ale gramaticky správné věty:

kometa se skládá z ohonu

do databáze se ukládají trojice: kometa -- has_part -- ohon

morfologický analyzátor majka

```
case["has_part"]="c2";
```

Ach ten češtin

vypravěč píše krátké, ale gramaticky správné věty:

kometa se skládá z ohonu

do databáze se ukládají trojice: kometa -- has_part -- ohon

morfologický analyzátor majka

```
case["has_part"]="c2";
```

Ach ten češtin

vypravěč píše krátké, ale gramaticky správné věty:

kometa se skládá z ohonu

do databáze se ukládají trojice: kometa -- has_part -- ohon

morfologický analyzátor majka

```
case["has_part"]="c2";
```

Výsledky

v současnosti necelých 500 her, přes 3100 trojic
trojice se opakují
v databázi jsou překlepy a chyby

Výsledky

v současnosti necelých 500 her, přes 3100 trojic
trojice se opakují
v databázi jsou překlepy a chyby

Výsledky

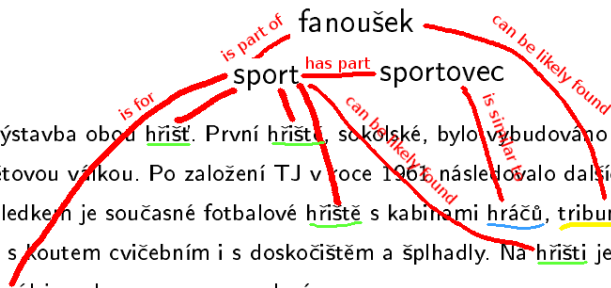
v současnosti necelých 500 her, přes 3100 trojic
trojice se opakují
v databázi jsou překlepy a chyby

Výsledky a jejich použití

sport	can_be_likely_found	hřiště
sport	can_have_property	bojový
sport	is_related_to	pohyb
sport	is_the_opposite_of	lenošení
sport	is_used_for	udržení kondice
fanoušek	is_part_of	sport
hra	can_be_used_for	sport
sportovec	has_part	sport
trofej	is_used_for	sport
fanoušek	is_related_to	sport
fanoušek	is_for	sport
areál	is_for	sport
fanoušek	can_be_likely_found	tribuna

Výsledky a jejich použití

Velkým počinem byla výstavba obou hřišť. První hřiště, sokolské, bylo vybudováno dlouho před druhou světovou válkou. Po založení TJ v roce 1961 následovalo dalších 40 let práce, jejichž výsledkem je současné fotbalové hřiště s kabinami hráčů, tribunou, odpočinkovým koutem, s koutem cvičebním i s doskočištěm a šplhadly. Na hřišti je i osvětlení obou ploch, areál je oplocen a osazen zelení.



Další práce

- lepší grafika
- registrace, hráčské úrovně, odměny apod.
- odstupňování tajných slov od „lehkých“ po „těžká“
- měření a porovnávání kvality získaných dat
- aplikace
- další hry

náměty na BP, DP :-)

Další práce

- lepší grafika
- registrace, hráčské úrovně, odměny apod.
- odstupňování tajných slov od „lehkých“ po „těžká“
- měření a porovnávání kvality získaných dat
- aplikace
- další hry

náměty na BP, DP :-)

<http://nlp.fi.muni.cz/~xpopelk/xplain/>
<mailto:xpopelk@fi.muni.cz>