Sémantika přirozeného jazyka

Popis syntaxe umožňuje rozlišit gramaticky správně a nesprávně utvořené věty.

Gramatická správnost matematických výrazů:
3+2; 3*2=6; 3/0 správné
3+*2; =7/2 nesprávné

Správné (syntakticky) české věty:
Kulatý obdélník zeleně masíroval kouř vedle minuty.
Před čtvrtkem následuje pátek.
Pes Alík je veselý a smutný.
Premiér Klaus má dlouhé černé vlasy.
Jestli mi bude oběd chutnat, přidám si.

Sémantika přirozeného jazyka

Význam a **pravdivost** sdělení v přirozeném jazyce jsou dvě naprosto odlišné záležitosti! I nepravdivá sdělení mají svůj význam, u jiných zase není možné ověřit pravdivost.

Jde o věty se stejným významem?

Hlavní budova FEL ČVUT je hned vedle Vítězného náměstí.
Hlavní budova FEL ČVUT je hned vedle Kulaťáku.

Studenti se stravují v menze.
Studenti se smějí stravovat v menze.
V menze se stravují studenti
V menze se smějí stravovat studenti.

Tuto knihu vydalo nakladatelství Paseka.
Tato kniha byla vydána nakladatelstvím Paseka.

Pozorovali ho dobrovolně.
Byl jimi pozorován dobrovolně.

Sémantika přirozeného jazyka

Vyplývání – to, že je věta pravdivá, mívá důsledky, věta nese víc informací:
Karel prodal auto sousedovi => Karel měl auto, už ho nemá, soused je od něj koupil a teď ho má

Pozor!

Tučňáci jsou ptáci ?=>? Tučňáci mají křídla a létají?

Sémantika formálních jazyků často spojuje pravdivost s významem, pro přirozené jazyky je nutné zvolit jiné teorie.

Fregeho princip kompozicionality (Gottlob Frege, 1848-1925)
Význam složeného výrazu je jednoznačně určen významy jeho částí a způsobem jejich kombinace.

Lexikální sémantika

Význam slov můžeme popisovat zase pouze pomocí nějakého (meta)jazyka:

- formálního (vhodný matematický nebo logický kalkul či soustava sémantických rysů, sémů)

- přirozeného (téhož nebo jiného)

či v reálném světě kombinací jazyka a situace (předmětu):
Toto je křída.

Význam ale závisí i na kontextu:

Střílení poslanců *ohrožuje naši křehkou demokracii.* (Pala)

Lexikální sémantika

Význam slov můžeme nezávisle na kontextu popisovat pomocí významových (sémantických) tříd (rysů).

Ontologie = množina tříd objektů, která představuje klasifikaci objektů universa U, např.:
- fyzické objekty
- kvantity
- vztahy
- vlastnosti
- akce atd.

Tyto třídy lze dále zjemňovat: slovesa pohybu, péče o tělo, změny, komunikace apod.

Existují doménové (domain) a vrcholové (upper) ontologie.

Význam slov ale není jednoznačný: *kohoutek, štěně, hlava*

Popis významu slov

Ve slovnících
- pomocí synonym, např. OALD, SSJČ,
- pomocí definic,
- pomocí množiny vybraných primitivních výrazů daného přir. jazyka,
- pomocí speciálního metajazyka: sémantických rysů, např.
muž = HUM, MASK, ADU
dívka = HUM, FEM, -ADU

Pomocí sémantické sítě
- forma sémantické sítě lépe zachycuje víceznačnosti
- je možné pracovat s hierarchií pojmů
- výhodnější pro počítačové zpracování

Sémantická síť

WordNet
1993 George A. Miller z Princetonu:
http://wordnet.princeton.edu

- WordNet® je rozsáhlá lexikální databáze angličtiny obsahující podstatná a přídavná jména, slovesa a příslovce seskupená do množin synonym (synsetů). Každý synset vyjadřuje určitý koncept, jsou mezi sebou navzájem propojeny sémantickými a lexikálními relacemi.
- Celou sítí je možno procházet pomocí prohlížeče.
- WordNet je veřejný, je možné si jej stáhnout.
- Ve verzi 3.0 obsahuje téměř 155 000 hesel a 117 000 synsetů.

Příklad z WordNetu

Chair

Noun

•S: (n) **chair** (a seat for one person, with a support for the back) *"he put his coat over the back of the chair and sat down"*

•S: (n) professorship, **chair** (the position of professor) *"he was awarded an endowed chair in economics"*

•S: (n) president, chairman, chairwoman, **chair**, chairperson (the officer who presides at the meetings of an organization) *"address your remarks to the chairperson"*

•S: (n) electric chair, **chair**, death chair, hot seat (an instrument of execution by electrocution; resembles an ordinary seat for one person) *"the murderer was sentenced to die in the chair"*

Verb

•S: (v) **chair**, chairman (act or preside as chair, as of an academic department in a university) *"She chaired the department for many years"*

•S: (v) moderate, **chair**, lead (preside over) *"John moderated the discussion"*

Sémantická síť

EuroWordNet
V roce 1997 prof.Vossen z Amsterdamu založil EuroWordnet 1 obsahující holandštinu, italštinu a španělštinu. K němu přibyl v roce 1998 EuroWordNet 2 s francouzštinou, němčinou, češtinou a estonštinou.

Proti původnímu WordNetu zavedeny změny:
- vrcholové ontologie (63 nejdůležitějších jazykově nezávislých konceptů
- množiny základních konceptů (1000 základních konceptů tvořících jádra sítí slov, jazykově závislé)
- jazykově nezávislý soubor indexů (interlingual index - ILI)
- vztahy ekvivalence (EQ-relations)

Aplikace WordNetu

- Automatický překlad – může fungovat jako slovník
- IE – extrakce informací
  jednak umožňuje pracovat se sémantickými vztahy (zejména synonymie), jednak může sloužit při vícejazyčném vyhledávání
- Určování jednotlivých významů slov (Word Sense Disambiguation) – zdroj dat pro rozpoznávání jednotlivých významů
- Reprezentace znalostí, odvozování využívající významů slov, vztah k sémantickému Webu
- Vyhodnocování kvality překladu (zlepšení automatických metrik typu BLEU)

Reprezentace významu věty

Predikátová logika 1.řádu
Konstruuje logické formule z jednotlivých výrazů věty na základě principu kompozicionality – jednotlivým složkám věty náleží odpovídající části sémantického zápisu

Alík skáče jump(Alík), ∃x x=Alík & jump(Alík)

Všichni psi skáčou ∀x dog(x) → jump(x)

Každý student podepsal ∀x student(x) → ∃y petition(y) & sign(x,y)
petici

Petici podepsal každý ∃y petition(y) & ∀x student(x) → sign(x,y)
student

Reprezentace významu věty

Hranice predikátové logiky 1.řádu
Modalita, čas a postoj – nové operátory, které mají jako argumenty formule
possible(F), necessary(F)
believe(x,F)
true_at_some_time_in _the_future(F)

Presupozice
Předpoklad, který musí být pravdivý, aby celá věta vůbec měla pravdivostní hodnotu
Jupiterův měsíc má oranžové pruhy. – Jupiter musí mít právě jeden měsíc.

Neurčitost (Fuzziness)
Nevystačíme s T/F hodnotami, potřebujeme jemnější dělení
Pavel je mladý. Většina špičkových sportovců dopuje.

Extenze a intenze

Jakmile začneme predikátovou logiku rozšiřovat, musíme rozlišovat mezi funkcí a její hodnotou

Cena Big Macu je 20 Kč.

Nahraďme výraz „Cena Big Macu" jeho hodnotou 90 Kč:

90 Kč je 20 Kč.

a dostaneme NEPRAVDU.

Pokud se však začneme pohybovat v oblasti postoje mluvčího:

Myslím, že cena Big Macu je 20 Kč. – nemůžeme už nahrazení provést, není to ekvivalentní tvrzení *Myslím, že 90 Kč je 20 Kč.*

Intenze výrazu – samotný popis, charakteristika - intenzí pojmu čtverec je pravoúhlost a stejná délka stran

Extenze výrazu – souhrn věcí, které pod pojem spadají

Základní přístupy k sémantice

Modelově-teoretická sémantika
Pracuje s pravdivostními podmínkami vztaženými k určitému modelu. Reprezentantem je **montagueovská gramatika**:
- syntaktické kategorie odpovídají sémantickým typům
- základní (lexikální) výrazy a jejich interpretace
- syntaktická a sémantická pravidla

Kompozicionální sémantika
Vychází z principu kompozicionality, používá různé reprezentace
- sémantické rysy a jejich skládání
- koncepty a převod (překlad) ze syntaktické reprezentace
- logickou reprezentaci a zjišťování pravdivosti

Montagueovská gramatika

původně nazývaná **Universal Grammar**

- teorie sémantiky přirozeného jazyka vytvořená americkým logikem Richardem Montague (1930-1971) a srozumitelněji vyložená Barbarou H. Partee.

- teorie je založena na formální logice, zvláště na lambda kalkulu a teorii množin, a používá pojmy intenzionální logiky a teorie typů.

- Montague byl přesvědčen, že neexistuje žádný zvláštní rozdíl mezi sémantikou přirozených a formálních jazyků. Základní zásady jeho teorie vyšly v článku "*The Proper Treatment of Quantification in Ordinary English*" (1973).

- byl to první pokus aplikovat formální sémantiku na přirozený jazyk. Logici před Montaguem považovali přirozený jazyk za příliš mnohoznačný a nestrukturovaný pro formální logickou analýzu, zatímco lingvisté měli pocit, že formální jazyky nejsou schopny zachytit struktury jazyků přirozených.

Syntaktické kategorie v MG

Category	Abbreviation	PTQ Name	Nearest linguistic equivalent
t	(primitive)	Truth-value expression; or declarative sentence	Sentence
e	(primitive)	Entity expression; or individual expression	(noun phrase)
t/e	IV	Intransitive verb phrase	transitive verb, transitive verb and its object, or other verb phrases
t/IV	T	Term	Noun phrase
IV/T	TV	Transitive verb phrase	Transitive verb
IV/IV	IAV	IV-modifying adverb	VP-adverb and prepositional phrases containing in and about.
t//e	CN	Common noun phrase	Noun or NOM
t/t	None	Sentence-modifying adverb	Sentence-modifying adverb
IAV/T	None	IAV-making preposition	Locative, etc., preposition
IV/t	None	Sentence-taking verb phrase	V which takes that-COMP
IV/IV	None	IV-taking verb phrase	V which takes infinitive COMP

Category Definitions/Generation

- Categories are of form: X/Y
  - semantics of Y into the truth value of X
- Abbreviations for first 5 categories
  - IV = t/e
  - T = t/IV
  - TV = IV/T
  - IAV = IV/IV
  - CN = t//e
- Infinite number of possible categories
  - May use as many slashes as needed for new categories

Example Expressions

	Category	Basic Expression
1	B_{IV}	{run, walk, talk, rise, change}
2	B_{T}	{John, Mary, Bill, ninety, he0, he1, he2, …}
3	B_{TV}	{find lose, eat, love, date, be, seek, conceive}
4	B_{IAV}	{rapidly, slowly, voluntarily, allegedly}
5	B_{CN}	{man, woman, park, fish, pen, unicorn, price, temperature}
6	$B_{t/t}$	{necessarily}
7	$B_{IAV/T}$	{in, about}
8	$B_{IV/t}$	{believe that, assert that}
9	$B_{IV/IV}$	{try to , wish to}

Partee 1973

Example Rule F_3 For B_{TV}

$F_3(\alpha, \beta)$ = If first word of α is a TV:

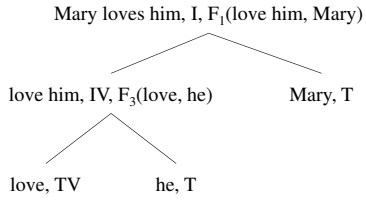
$\alpha\ \beta$	if β is not a variable
α him$_i$	if β is he$_i$

If α is $\alpha_1\ \alpha_2$ where α_1 is a TV/T:

$\alpha_1\ \beta\ \alpha_2$	if β is not a variable
α him$_i\ \alpha_2$	if β is he$_i$

F_3(shave, a fish) = shave a fish
F_3(seek, he$_1$) = seek him$_1$
F_3(read a large book, Mary) = read Mary a large book

Syntactic Rules



If $\alpha \in X/Y$ and $\beta \in Y$ then $F_i(\alpha,\beta) \in X$

TIL

Transparentní intenzionální logika (TIL) – Pavel Tichý (1936-94)
Reaguje na fakt, že predikátový kalkul 1.řádu, který stále mnoho teorií používá k popisu významu jazykových výrazů, nedostačuje. Intenzionální logika je vhodnější.

TIL
- je založen na modifikaci typovaného lambda kalkulu
- je transparentní systém, tj. pro TIL není formální aparát reprezentující způsoby, jakými jsou konstruovány objekty, předmětem studia, nýbrž pouze prostředkem ke studiu těchto konstrukcí.
- nepreferuje jistá vybraná slova jako tzv. logická slova (logické spojky, kvantifikátory apod.), jež by určovala charakter logiky.
- TIL aplikována na analýzu přirozeného jazyka se stává sémantikou založenou na pojmu **možných světů**
- Univerzum je v TIL chápáno jako množina společná všem možným světům, kromě možných světů se neuvažuje o tzv. možných individuích.

Použití TILu – příklad (Pala)

Věta:

Studentka Alena si myslí, že ministr financí je hezčí než ministr zahraničí.

Jednotlivé atomy (pro jednoduchost tvořené složenými výrazy)
studentka Alena: **A**/ι – nálepka individua
myslet si: **M** /$(o\iota o_{\tau\omega})_{\tau\omega}$ – vztah mezi individuem a propozicí
ministr financí: **F**/ $\iota_{\tau\omega}$ – individuální koncept
hezčí než: **Hn**/$(o\iota\iota_{\tau\omega})_{\tau\omega}$ – vztah mezi dvěma individui
ministr zahraničí: **Z**/$\iota_{\tau\omega}$ – individuální koncept.

Propozice popisující vedlejší větu:
$\lambda w\, \lambda t\, (\mathrm{Ht}_{wt}\, \mathrm{F}_{wt}\, \mathrm{M}_{wt})$..

Po přidání atomů M a A
$\lambda\, w\, \lambda t\, (\mathrm{M}_{wt}\, (\mathrm{A}\, (\lambda w\, \lambda t\, (\mathrm{Hn}_{wt}\, \mathrm{F}_{wt}\, \mathrm{Z}_{wt}))))$

Celý výraz konstruuje objekt $o_{\tau\omega}$ - tedy propozici, což je funkce, která každému možnému světu W v okamžiku S přiřadí nejvýše jednu pravdivostní hodnotu.

Rozpoznávání vztahů v textu

Anafora

Výraz, jehož interpretace závisí na kontextu. Obecně rozlišujeme:

- Exofora (odkazování mimo text)
  Vidíš ho? Dejte mi, prosím, tyhle tři.
- Endofora (odkazování v rámci textu)
  - Anafora (zpětně)
    Petr se seznámil se sympatickou dívkou. Pozval ji do kina.
    Petr vyzradil tajemství. To neměl dělat.
  - Katafora (dopředu)
    Když se zlobí, není s Petrem žádná řeč.
    Věřte tomu nebo ne, máme schodkový rozpočet.
    Vyšel jsem z domu. Věděl jsem, že jsem sledován. Když jsem se zastavil, zastavil se i on. Když jsem se ohlédl, dělal, že lelkuje. Měl na sobě stejný šedý kabát jako vždycky. Už ho důvěrně znám, estébáka Jiřího.

Anafora

Anaforický vztah **předchůdce – následník**

Typy anaforických vztahů:

Zájmena a „nulové výrazy" (nevyjádřený podmět nebo jiný větný člen)
Petr si koupil vstupenku. Vsunul ji do kapsy. Byla děravá.

Určité jmenné skupiny (*Elektronický zesilovač Tesla* vs. *Toto zařízení…*)

Elipsa (vypuštěné části výrazů na základě paralelismu s předchůdcem)
jmenná vs. slovesná
Včera jsem šel pěšky. Kam? Domů.

Petr přinesl dva stoly. Dřevěný a kovový.

Petra půjde do kina. Jirka taky.

Textové spojovací výrazy (výrazy vyjadřující mezivětné souvislosti v textu) – např. souřadicí a podřadicí spojky a výrazy jako *například, na jedné straně – na druhé straně, jednak –jednak, nejdříve – potom* apod.

Anafora

Důležitost pro aplikace

Získávání informací z textu
Škoda představila nový model Octavie. Jde o pětidveřové kombi, které má ….

Automatický překlad
Otevřenou tabulku upravte podle potřeby. Uložte ji pomocí ikony v panelu nástrojů.

Dialogové systémy
Kdy jede nejbližší vlak do Ostravy? Má jídelní vůz?

Uživatel: *Které vzorky obsahují magnézium?*
Systém: *ID123, ID147, ID159, ID369*
Uživatel: *Které obsahují (také) křemík?*

Řešení anafory

Je nutné využít celou řadu informací:

• morfologické značky
- např. u zájmen musí být shoda v rodě

• syntaktická struktura věty
- pomůže určit vhodné kandidáty na předchůdce
- valenční informace umožní doplnit elipsu

• statistické přístupy
- pravděpodobnost výběru některého z určených kandidátů

• aktuální členění
- témata zmíněná v základu a v ohnisku věty jsou odkazována různými způsoby – využito v algoritmu Zásoby sdílených znalostí

• rozsáhlé pomocné znalosti
- ontologie, sémantické sítě, tezaury apod.

Zásoba sdílených znalostí

Modeluje zásobu znalostí, o které mluvčí předpokládá, že ji sdílí s posluchačem. Tato zásoba se mění v souladu s tím, co je „v centru pozornosti "v daném časovém okamžiku.

Každá věta má vliv na tuto „hierarchii sdílení," avšak ne každý zmíněný objekt má stejný účinek.

Jednoduchá pravidla:

Objekt **a**, jenž má stupeň sdílení **n**, je označen jako $\mathbf{a^n}$

Pokud k **a** referujeme slabou formou zájmene nebo pokud je vynecháno, zůstává **n** stejné

Pokud je jmenná skupina **a** zmíněna v **ohnisku**, potom a^n -> a^0

Pokud je jmenná skupina **a** zmíněna v **jádru**, a^n -> a^1

Pokud a^n -> a^m potom všechny objekty asociované s **a** obdrží hodnotu a^{m+2}

Výrazy "As for a", "Concerning a" atd. zvýší aktivaci na a^1

Pokud není **a** ani zmíněno, ani asociováno, potom se jeho stupeň aktivace zvýší o 2

Zásoba sdílených znalostí

Příklad

[1] The school garden was full of children.
[2a] They talked noisily,
[2b] but the teachers did not reprove them,
[2c] because they were so excited.
[3] Outside parents were waiting.
[4] A group of about five parents grouped around a microphone.
[5] One of them should probably speak.
[6] The teachers were very serious.

[1] $children^0$ $school\ garden^1$ $parents^2$ $school^3$ $pupils^3$ ….
[2a] $children^0$ $parents^2$ $school\ garden^3$ $school^5$
[2b]
[2c]
[3]
[4]
[5]
[6]

Základy teorie informace

Základy položil Claude Shannon v roce 1948 v práci
A Mathematical Theory of Communication
Zavedl BIT (BInary digiT) jako jednotku informace

Množství informace
Podle Shannona je množství informace obsažené v události (zprávě) x rovno

$$h(x) = \log_2 \frac{1}{P(x)}$$

Měříme ji v bitech.

Množství informace obsažené ve zprávě X souvisí s pravděpodobnosti P(X) s jakou může příjemce uhodnout obsah zprávy X, neboli jaká je pravděpodobnost výskytu dané zprávy u příjemce před jejím přijetím.

Příklady

Množství informace
Zpráva „V ruletě padlo číslo 17" přináší větší množství informace než zpráva „V ruletě padlo liché číslo".
Pravděpodobnost padnutí lichého čísla je 0.5, čímž příjemce mohl tento stav rulety snadněji uhodnout, než padnutí čísla 17, neboť tento stav má mnohem menší pravděpodobnost. První zpráva přináší tedy větší množství informace.

Příkladem dvou stavů systému je hod korunovou minci. Pravděpodobnost, že padne hlava je P(X)=1/2 a množství informace, kterou nese zpráva „Padla hlava" je:

$h(x) = \log_2 \frac{1}{P(x)} = \log_2 2 = 1bit$

Základy teorie informace

Entropie

Míra neurčitosti v nějaké zprávě X o daném systému. Tato míra neurčitosti se po příjmu zprávy odstraňuje a tím se vyjádří míra získané informace. Při růstu informace klesá entropie a naopak.

Entropie souboru X je definována jako průměrné Shannonovo množství informace přes všechny události x:

$$H(X) = \sum_{x \in X} P(x) \log \frac{1}{P(x)}$$

(používáme konvenci, že pokud P(x)=0, potom se celý výraz za sumou také rovná nule)

Jednotkou entropie je také bit.

Entropie

Vlastnosti entropie

1) Entropie je spojitá a nezáporná funkce - všechny náhodné procesy mají nezápornou neurčitost.
2) Není jednoznačnou funkcí svých argumentů, tzn. může nabývat téže hodnoty pro různé hodnoty svých argumentů
3) Entropie je rovna nule tehdy a jenom tehdy, když pravděpodobnost výskytu některého znaku x_i je
   $p(x_i) = 1$ a $p(x_j) = 0$ pro všechny $i \neq j$.

Jsou-li pravděpodobnosti výskytu jednotlivých stavů stejné, má zpráva maximální entropii.

Další vzorce

Spojená entropie

$$H(X,Y) = \sum_{xy \in XY} P(xy) \log \frac{1}{P(xy)}$$

Podmíněná entropie

$$H(X|Y) = \sum_{xy \in XY} P(x,y) \log \frac{1}{P(x|y)}$$

Vzájemná informace mezi X a Y

I(X;Y) ≡ H(X) + H(X|Y) ≡ H(Y) + H(Y|X) ≡ I(X;Y),
I(X;Y) ≥ 0

Měří průměrnou redukci neurčitosti X, která je důsledkem informací obsažených v Y nebo jinak, průměrné množství informací, které X sděluje o Y.