

Sémantika

Syntaxe a sémantika

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.

Význam a pravdivost

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.* (příklad převzat od prof.Paly)

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
- kvantify
- 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 přír. 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 síť 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 $\text{jump}(\text{Alík}), \exists x x=\text{Alík} \ \& \ \text{jump}(\text{Alík})$

Všichni psi skáčou $\forall x \text{dog}(x) \rightarrow \text{jump}(x)$

Každý student podepsal petici $\forall x \text{student}(x) \rightarrow \exists y \text{petition}(y) \ \& \ \text{sign}(x,y)$

Petici podepsal každý student $\exists y \text{petition}(y) \ \& \ \forall x \text{student}(x) \rightarrow \text{sign}(x,y)$

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.

Estenze a intenze

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

Cena Big Macu je 20 Kč.

Nahradíme 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

Estenze 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

Definice kategorií/Generování

- Kategorie mají tvar: X/Y
 - sémantika Y vyjádřená pravdivostní hodnotou X
- Zkratky prvních 5 kategorií
 - $IV = t/e$
 - $T = t/IV$
 - $TV = IV/T$
 - $IAV = IV/IV$
 - $CN = t//e$
- Nekonečný počet možných kategorií
 - Pro nové kategorie je možné použít libovolně velký počet lomítek

Příklady výrazů

	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

Příklad pravidla F_3 pro B_{TV}

$F_3(\alpha, \beta) =$ If first word of α is a TV:
 $\alpha \beta$ if β is not a variable
 $\alpha \text{ him}_i$ 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
 $\alpha \text{ him}_i \alpha_2$ if β is he_i

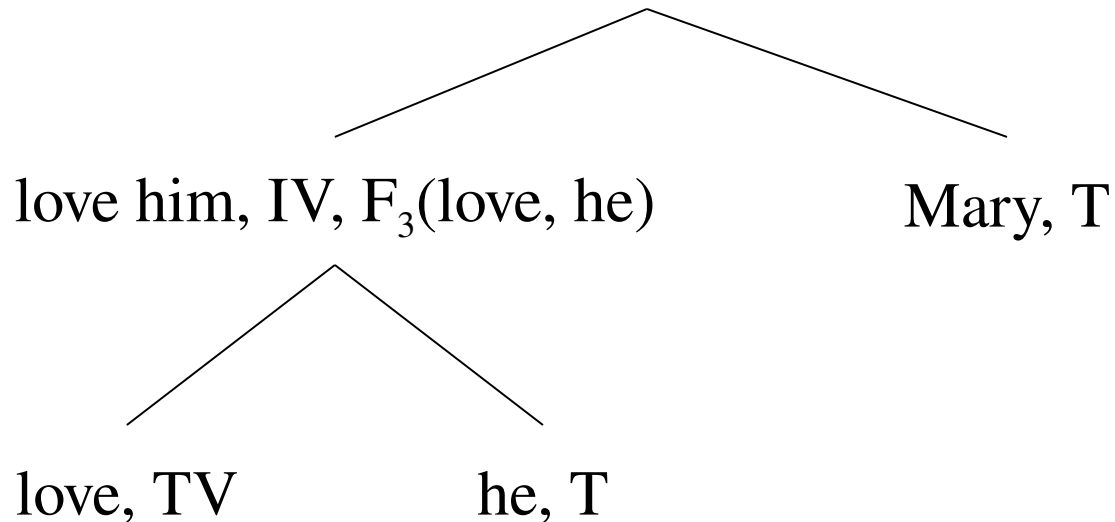
$F_3(\text{shave}, \text{a fish}) = \text{shave a fish}$

$F_3(\text{seek}, \text{he}_1) = \text{seek him}_1$

$F_3(\text{read a large book}, \text{Mary}) = \text{read Mary a large book}$

Syntaktická pravidla

Mary loves him, I, $F_1(\text{love him}, \text{Mary})$



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 uvažuje 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** /($\text{O}\iota\text{O}_{\tau\omega}$) $_{\tau\omega}$ – vztah mezi individuem a propozicí

ministr financí: **F**/ $\iota_{\tau\omega}$ – individuální koncept

hezčí než: **Hn**/($\text{O}\iota\iota_{\tau\omega}$) $_{\tau\omega}$ – vztah mezi dvěma individui

ministr zahraničí: **Z**/ $\iota_{\tau\omega}$ – individuální koncept.

Propozice popisující vedlejší větu:

$\Delta w \Delta t (Ht_{wt} F_{wt} M_{wt})..$

Po přidání atomů M a A

$\Delta w \Delta t (M_{wt} (A (\Delta w \Delta t (Hn_{wt} F_{wt} Z_{wt}))))$

Celý výraz konstruuje objekt $\text{o}_{\tau\omega}$ - 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:

- Exoforma (odkazování mimo text)
Vidíš ho? Dejte mi, prosím, tyhle tři.
- Endoforma (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 dřeváková.

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 **aⁿ**

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

Pokud je jmenná skupina **a** zmíněna v **ohnisku**, potom $a^n \rightarrow a^0$

Pokud je jmenná skupina **a** zmíněna v **jádro**, $a^n \rightarrow a^1$

Pokud $a^n \rightarrow a^m$ 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] children⁰ school garden¹ parents² school³ pupils³

[2a] children⁰ parents² school garden³ school⁵

[2b]

[2c]

[3]

[4]

[5]

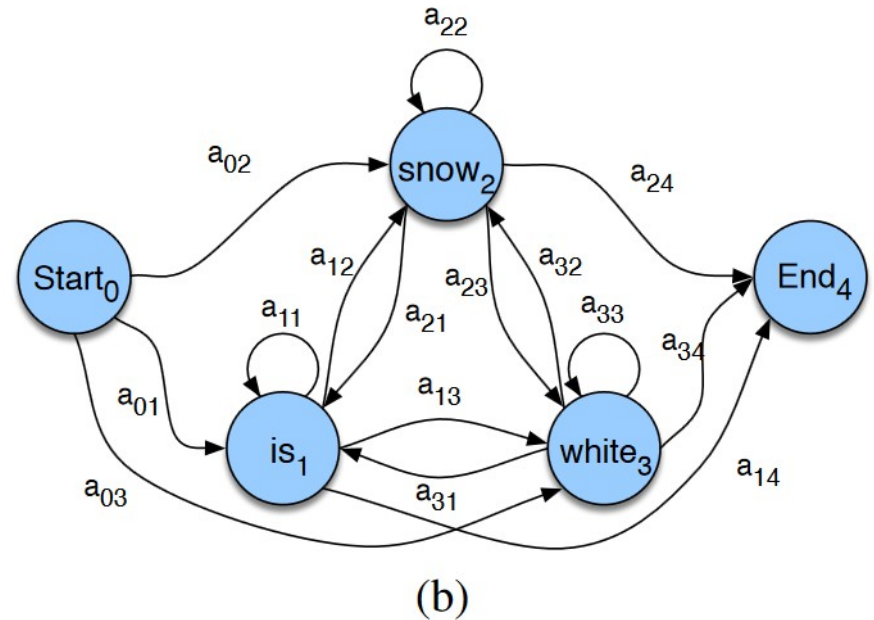
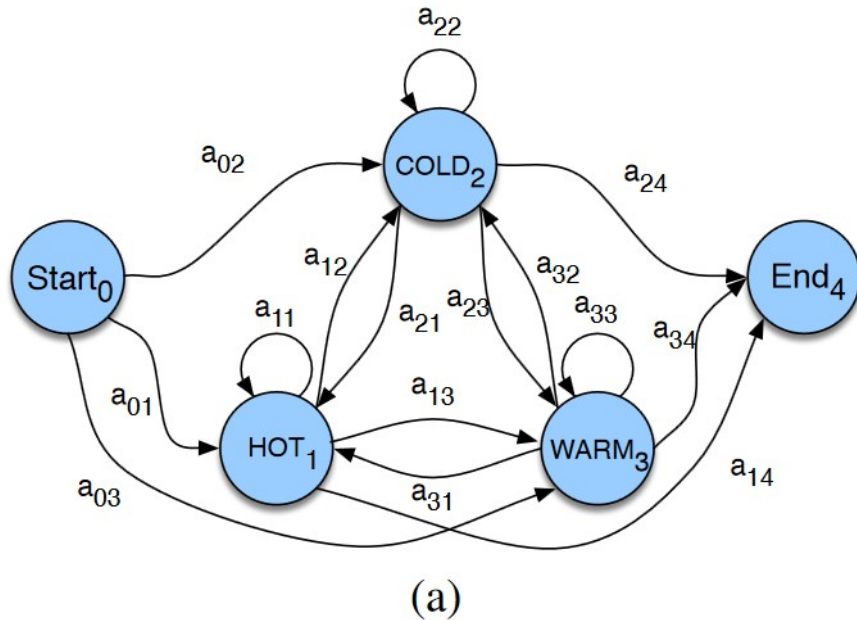
[6]

Skryté markovské modely

Hidden Markov Models - HMM

- Standardní metoda analýzy řad (posloupností událostí v čase) na základě kontextu (řečový signál, posloupnost morfologických značek v textu apod.)
- Posloupnost rozhodnutí, která jsou na sobě závislá
- Markovova hypotéza – kontext je možno zkrátit na délku, která je spočítatelná (bigramy, trigramy)
- Slovo „skryté“ reprezentuje fakt, že některé vlastnosti posloupnosti nejsou pozorovatelné (ve větě vidíme slova, nikoli morfologické značky, které jim přiřazujeme)
- Jedná se v podstatě o stochastický konečný automat.

Příklad Markovského řetězce



Markovův řetězec pro počasí (a) a pro slova (b)

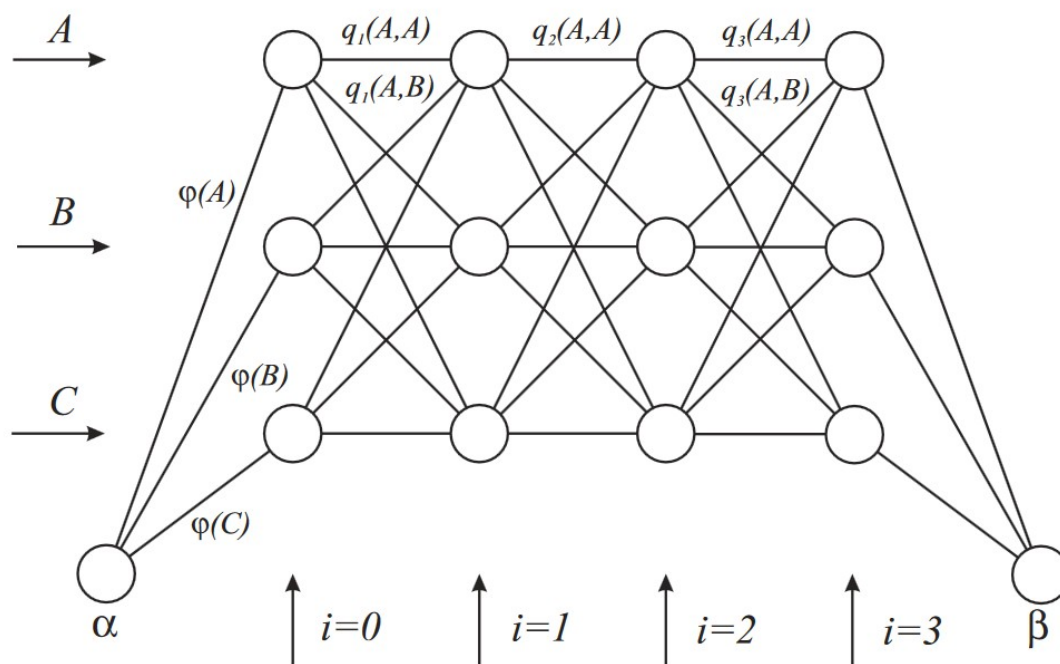
Markovův řetězec je určen strukturou, přechody mezi stavy a počátečním a koncovým stavem.

Tři základní úlohy s HMM

1. **Rozpoznávání** (ohodnocení) statistického modelu (dynamické programování, v angl. literatuře Forward-backward algoritmus). Jsou dány parametry HMM, cílem je spočítat pravděpodobnost, že je pozorována posloupnost X . Použití: např. rozpoznávání obrázků (reg. značek aut)
2. **Dekódování**, neboli hledání nejpravděpodobnější posloupnosti skrytých stavů (Viterbiho algoritmus, dynamické programování). Je dán statistický model a posloupnost pozorování, cílem je najít nejpravděpodobnější posloupnost skrytých stavů.
3. **Učení** statistického modelu (Baum-Welshův algoritmus). Je dána struktura modelu (počet skrytých stavů) a trénovací množina. Cílem je najít parametry modelu, tedy pravděpodobnosti přechodů mezi stavy a pravděpodobnosti jednotlivých prvků posloupnosti.

Dekódování

- Viterbiho algoritmus – v podstatě hledání nejkratší cesty v grafu
- Orientovaný graf s hranami orientovanými zleva doprava. Počátečním vrcholem je α a cílovým β , ohodnocení hran odpovídá pravděpodobnosti přechodu
- Schéma pro skryté stavy A,B a C a pozorování 0,1,2 a 3



Viterbiho algoritmus pro značkování

- Konkrétní příklad pro větu „Janet will back the bill“

