1 Reprezentace znalostí jako jeden ze zákl. problémů AI

Historicky vyvíjeny soustavy rep. znalostí jako podklad pro systémy členění problémů, hledání informací, rozbor dotazů.

V 60. letech vznikají **databázové systémy**. Rozdíly RZ od databáze:

1. Repr. znalostí má mít víc možností *inference* (odvozování), to v DB nejde; + klasifikace dat, ověření platnosti, dekompozice
2. Tehdy DB chápány jako "úplný popis světa", RZ ne
3. Systém musí sám vědět, k čemu je, co dělá

Výchozí teze: "AI bez přir. jazyka neexistuje." Spojení AI s přir jazykem – uplatnění:

- Univerzálnost (+): je možné mluvit o všem
- Víceznačnost (-)
  - např. *bank* = břeh, banka, existují kontexty, kde i člověk těžko rozezná
  - ještě horší – slovní spojení, struktura: *řízení našeho ústavu* (ústav řídí nebo je řízen?)
- Pragmatická vágnost (-): *Chce se oženit s američankou* (jednou určitou?, čistě kvůli získání pasu?)
- Reference (-): *Děti měly radost. Dostaly vláček* (kdo dostal?)
- Bohatství jazyka (+): korpusová lingvistika, slovníky nabízí materiál k řešení, čím víc ho je, tím líp.

Nilsson (Stanford, US 1980) v definici AI zahrnuje jazyk (ač sám nebyl lingvista), dává porozumění jazyku na 1. místo. Jeho definice ukazuje vztah lingvistiky a AI – v aplikacích, v zachycení světa (pamatovat si podobně jako člověk).

2 Taxonomie reprezentačních schémat

Jaké jsou typy reprezentací znalostí? Co může být využito v systémech komunikace s AI?

Ivan Havel (1977) rozlišuje 2 typy:

1. S použitím predikátové logiky
2. Relační struktury (grafy, sítě, rámce). Řadí rámce a sítě dohromady, ač se v lecčems liší.

Mylopoulos (1980) – teoretičtější, představuje si reprezentaci jako *individiua*, vztahy mezi nimi; stav systému konstituuje soubor individuí a vztahů. Změny stavu jsou transformace vztahů. Podle toho rozlišuje sémantické sítě (kde jsou hlavní individua a vztahy), logická schémata (kde jsou hlavní pravdivé výroky o stavech), procedurální schémata (kde jsou hlavní transformace stavů) a jako zvláštní typ rámcová schémata.

2.1 Sémantické sítě

Hlavní – **individua a vztahy**, tj. uzly a ohodnocené hrany. Vztahy – různé typy: *abstraktum x konkrétum, část x celek, typ x instance*. Můžu podle toho definovat různé organizační principy, což u logických schémat moc nemám. Díky asociacím se informace hledají snadno, problém je málo formalismu.

2.2 Logická schémata

Soubor logických pravdivýh formulí o stavech. Nejsou uzly a hrany, ale formule. To je výhoda, první systémy na tom stály, pro logiku máme odvozovací pravidla, takže nemusíme vymýšlet vlastní. Problém: chybí org. princip (to v sítích je), těžko se zachycují procedurální znalosti (znalosti postupu práce).

PROLOG vychází z logických schémat, ale přidává i procedurálno, proto může být docela dobře použitelný.

2.3 Procedurální schémata

Svého času velmi populární. **Stavové transformace**, zásoba znalostí jako **soubor procedur v programovacím jazyce** (např. LISP). Proti logice výhoda – reprezentace procedur ("vědět jak"), efektivita sytému. *Default reasoning* – na základě "default" situací se může rozhodovat i s neúplnou informací o světě. Může automaticky akutalizovat data, se kterými pracuje.

Nevýhody – chybí nějaký univerzální podklad (to v logice je), chybí konzistence při default reasoning, obtíže při změnách: když se někde něco změní, změna se propaguje všude, nekontrolovatelně.

2.4 Rámcová schémata

Kombinace předchozích 3, využití i logiky, i procedur, i sítí.

3 Porozumění a počítače, ELIZA

A. M. Turing (1947, 50): Tvrdí, že překlad pomocí počítačů je možný, chce odpovědět na otázku, zda je možné porozumění, definuje *Turingův test* na základě lidského poznání, zda reaguje člověk nebo počítač. Test byl sporný, vyvrátil ho **Weizenbaum**(1966) s programem ELIZA (název podle Pygmalionu – reakce na úrovni společenské konverzace). ELIZA prošla testem, i když inteligentní vůbec není, jen simuluje inteligenci různými triky. Pracuje v omezeném prostředí (psychiatr). Zahrnuje:

- Soubor klíč. slov, usp. podle důležitosti – reaguje se na 1 prioritní klíč. slovo
- Pro každé klíč. slovo – seznam reakcí, uspořádaný, co se už použilo, jde na konec (a cyklí se)
  - Pro obecné reakce (neurčitá slova) výzva, aby jmenoval zvláštní případ
  - Otázky na "rozpovídání se", typu "In what way?"
  - Pro popisy špatné duševního stavu se ptá, jak může pomoct
  - Pokud se vyčerpají možnosti, použije se "Go on.", nebo "Say me more about ...", příp. "That is interesting."
- Není-li v reakci klíčové slovo, vrací se k poslednímu použitému, případně opět neurčitá reakce, zopakování pacientovy výpovědi.

Trvá to, než si člověk uvědomí, že se počítač opakuje a neví, o čem mluví.

4 Winogradův systém ovládání robota (SHRDLU)

Terry Winograd (Stanford 1972, pozd. Edinbourgh) – první opravdu vážný pokus o komunikaci s počítačem. Na omezeném prostoru zkoušel vlastnosti porozumění: simulace robota v omezeném světě s několika objekty, které mohl přesouvat (nešlo o fyzicky existujícího robota, jen počítačovou simulaci jeho akcí).

4.1 Složení systému

- Obsahuje *autoreference* – robot ví, v jakém je stavu, co zrovna dělá.
- *Procedurální reprezentace znalostí* (jazyk PLANNER) – seznam použitelných údajů a uskutečnitelných instrukcí + dodané údaje
- Rozložil systém na několik modelů, ale šlo mu i o celek
  - Hlavní ovládání: MONITOR
  - Pod ním INPUT (pro čtení vstupu, morfologická analýza), GRAMMAR (zpracování vstupu), SEMANTICS (interpretace), ANSWER (zajištění výstupu: vytvoření odpovědi a její provedení, tj posouvání beden)
  - Pro čtení vstupu používán DICTIONARY, PROGRAMMAR (pomoc gramatice – syntaktické stromy), SEMANTIC FEATURES, jazyk PLANNER, BLOCKS (v PLANNERU napsaná robotova znalost světa) + DATA, MOVER (visualizace)

4.2 Výsledek pokusu

Winogradovi se pokus zdařil, založil celé odvětví **box systems**. Nešlo ale o nijak příliš vyspělý systém, dost profitoval z omezení světa – např. lexikální analýza měla hodně malou slovní zásobu, v syntaxi šlo větš. o jednoduchý imperativ apod. Systém navíc neformuloval moc odpovědi (negeneroval vlastní), v ANSWER měl zabudované matice, kam vkládal slova od operátora (tj. nebylo to moc daleko od ELIZY ;-)).

Název **SHRDLU** byl libovolně vybrán Winogradem, aby se nepletl s žádným existujícím slovem (není to zkratka). Winogradův systém chtěli mít hned na všech univerzitách – i na MFF 1980 systém NALCOM (Natural Language Communicator, autor Hajič).

5 Sémantické sítě, Quillianův model paměti, rozčlenění sítí (Hendrix)

Někteří nedělají mezi nimi a rámci rozdíl, ale Hajičová ho v tom vidí.

5.1 Sémantické sítě

Jiné názvy: kognitivní, konceptuální síť. Snaha reprezentovat znalosti podobně jako v lidském mozku. *Motivace: centralizace faktů z urč. oboru na jednom místě a jejich logické propojení.*

Reprezentace: deklarativní zápis, procedura k vybavení znalosti. Závislosti *paradigmatické* (záměny např. konkrétnějšího pojmu za nadřízený) i *syntagmatické* (kombinace, spojování).

Možné interpretace sítě:

1. Hrany pointry, uzly jako místa, kam směřují
2. Uzly jako predikáty, hrany jako logické vztahy
3. Uzly jako významy slov, hrany jako konceptuální vztahy mezi nimi

Problémy:

- Překřížení vztahů (dá se tomu zabránit duplikací uzlů)
  - Např.: *Bakterie odolávají nízkým teplotám.* + *Kaktusy odolávají nedostatku vody* = *Bakterie odolávají nedostatku vody* atp.
- Ztráta kvantifikátorů při ztrátě pořadí slov ve větě, aktuální členění věty
  - Např.: *Nepřišel, protože pršelo.* x *Protože pršelo, nepřišel*

Zastánce sítí **Hermann Helbig** – do svých modelů začleňoval i jazykové vztahy typu "nositel vlastnosti", "konatel", řeší i aktuální členění větné, jeho přístup je brán jako jeden z nejlepších.

Příklad sítě: Wordnet – zobrazení slovníku (odpovídá Quillianovu modelu).

Dnešní sítě se většinou nazývají *kognitivní* a mají klasifikaci hran, vztahů.

5.1.1 Quillianův model

První koncepce na téma sémantické sítě, **Quillian** (1968). Šlo o modelování uspořádání znalostí v lid. mozku, ne přímo o komunikaci s počítačem.

- Uzly jako pojmy, vyjádřitelné slovem nebo souslovím

1. *Typové* uzly (hlavní uzly, z nich vedou hrany k ostatním uzlům) – představují jednu "typovou oblast" (znalosti o tom jednom pojmu)
2. *Zástupné* uzly (zastupují jiný typový uzel s jiným okruhem hran) – z nich vedou "asociační hrany" do zastupovaného uzlu

- Pro každý význam daného slova je jiný typový uzel
- Význam pojmu = množina všech uzlů, které lze získat průchodem po hranách z typového uzlu
- Dá se použít na porovnávání významů slov

5.1.2 Rozčleněné sém. sítě (partitioned networks)

Hendrix (Stanford, 1978) vymyslel pro sém. sítě řešení proti ztrátě vztahů, amorfnosti sítí. Vylepšení:

- Hranice mezi výroky, syntagmaty – *spaces*,
- *ohodnocení hran* (v Quillianově modelu není, dnes má podobné věci Helbig) podle druhu vztahu (část-celek apod.),
- *Průhledy (vistas)* – spojování několika spaces, přístupná je jen informace obsažená v některém ze spojených spaces.

Šlo o *uspořádání* znalostí, rozčlenění i podle dosahu kvantifikátorů, správné připojení k jiným vztahům s kvantifikátory. Zachovávají se pořadí slov ve větě, hranice vět.

Cílem byla komunikace s počítačem, Hendrix chtěl udělat reálně použitelný systém. Založil na tom později komerční projekt Q/A, který pomáhal s vyplněním daňového přiznání (a měl s ním docela úspěch, ač byl taky "zadrátovaný natvrdo").

6 Rámce, systém GUS, konceptuální závislosti

V AI přišel s nápadem rámců první **Minsky** (Stanford, 1975), kterému šlo o prostorové vidění. V lingvistice existují de facto od **Fillmora** a jeho práce o argumentech slovesa (valenčních rámcích, 1968). Ty se dodnes používají při anotacích korpusu atd. Fillmore vytvořil i projekt FrameNet, který zahrnuje sémantický přístup k doplněním slovesa, což je dnes běžný přístup.

Minsky uvažoval představu vstupu člověka do místnosti: člověk má předem představu, jak bude místnost vypadat. Volné rámce pro stěny, strop apod., ty se zaplňují: místo abstrakta si doplňuji konkrétní vjem z místnosti. *V dlouhodobé paměti člověka jsou stereotypy: defaultní hodnoty, člověk má v paměti rámce.*

- *Rámec* = pevný "horní" uzel, obsahující slots
- *Slot* = volné rubriky v rámci, které se mohou zaplnit

Pro konkrétní děj / věc musím mít předem rámec se sloty, které pak z informací o světě (z textu na vstupu) zaplňuju. Některé mohou zůstat prázdné. Oproti sítím jsou tedy *předem organizované.*

Výtky rámcům: neschopnost dynamického zpracování (jen deklarativní).

6.1 Systém GUS (Genial – "duchaplný" Understanding System)

První systém, který použil rámce: Stanford (Bobrow, Kaplan, Kay, Norman, Thompson, 1977). Mělo fungovat jako rezervace letenek, modelovat cestovní kancelář. Měl podávat rozumné doplňující otázky a další nabídky. Pokusili se *spojit rámce a dynamičnost* (nová myšlenka). Zahrnovalo:

- Významový zápis pomocí rámců
- Báze znalostí (letový řád) ve vnější paměti
- Matice odpovědí (negeneruje se inteligentní odpověď)

Rámce v GUS:

- každá rubrika má jméno, hodnotu, příp. soubor přiřazených procedur
  - hodnotou může být i jiný rámec
  - procedury slouží jako spojení procedurální a deklarativní sémantiky
- procedury - dva druhy (chytré):
  - *sluhové* – aktivují se, když jsou potřeba; tj. jen když klient specifikuje netypický požadavek, přepíše se default
  - *démoni* – aktivují se, jakmile je vložena hodnota; např. za účelem dopočítání další informace z vloženého (den v týdnu)

6.2 Konceptuální závislosti

Rámce měly hlavně v teorii velkou odezvu, i ve spojení s lingvistikou. Začal s tím **Roger Schank** (USA, 1969), v něčem vychází ze závislostního popisu P. Sgalla z MFF. Prezentoval na COLINGu systém *konceptuálních závislostí* – vztahy (hrany) mezi *koncepty*, podobné závislostním vztahům mezi větnými členy (čerpá z Fillmora).

Shankova koncepce:

- Máme nominální koncepty a dějové koncepty (primitivní děje), pro nominální koncepty máme závislostní vztahy stavové, pro dějové závislostní vztahy specifikační.
- 5 konceptuálních pádů: *agens, patiens, objekt, směr, nástroj*
- Primitivní děje (celkem 11): *jít, vidět, dát* (přenos něčeho někam, někomu), *pohyb, zažívání, vlastnění, obsahování* (specifikační vztahy, vlastnické vztahy)
- To vše se dá spojovat do *kauzálních řetězců*: děje mohou měnit stavy, stavy mohou ovlivňovat děje atd.

Pomocí toho chtěl reprezentovat všechny věty jazyka (vliv sémantičky **Věržbické**, která propagovala koncepci složení všeho z několika málo prvků). Vztahy se daly rekurzivně do sebe zapouštět.

Rámce pojmenoval obecné informace o skutečnostech, byly 2 typů:

- *scénáře* (standardní situace, např. "restaurace")
- *plány* (standardní posloupnosti lidských akcí, příčinné vztahy mezi scénáři, např. "host odejde bez placení")

7 Lingvistická koncepce, závislostní stromy

Předností přirozeného jazyka je možnost popsat cokoliv (s případným vytvořením nových pojmenování). Při využití zápisu hloubkové struktury lze odstranit i homonymie a synonymie. Lze se domnívat, že uspořádání informací v lidské paměti je podobné.

Zatím převládají systémy s omezeným pohledem na svět, pro typické situace – nevychází se tedy z povahy jazyka, ale ze situací. Lingvisticky ale lze dojít k obecnému inventáři znaků i operací nad nimi. Pro každý způsob reprezentace (sítě, rámce atp.) je třeba zachytit vztahy děje (slovesa) a jeho účastníků (jména) – např. **Schank** rozpoznává na 70 různých vztahů, rozlišitelných přímo podle lex. významu slova.

7.1 Významový zápis věty

Významový zápis věty může být podkladem pro síť znalostí. *Význam* lze vztahovat nejen na jednotlivá slova ve větě, ale i na celé věty (lze zjišťovat, jestli tvrdí to samé). Lze říct, že dva výrazy mají ten samý význam, pokud se dají zaměnit kdekoliv kromě metajazykových kontextů (tedy při použití jak v intenzionálním, tak v extenzionálním smyslu). Věty potom mají stejný význam, pokud v nich jsou na stejných pozicích výrazy o stejném významu.

Takovýto význam odpovídá *tektogramatickému zápisu* věty. Je ale vázán jen na jazyk, bez vztahu k objektům mimojazykové reality, pokud význam vztáhneme k realitě, dostaneme *smysl* věty. Kromě smyslu si posluchač z věty odvodí další důsledky, *inference*, jazykové ekvivalenty věty.

Významový zápis vychází ze závislostního popisu, jde o čtveřici M, D, W, T, kde:

- M je množina uzlů (výrazů použitých ve větě v jednotl. pozicích),
- D je relace, vytvářející graf – strom nad objekty z M,
- W je relace uspořádání uzlů, splňující podmínku projektivity; hloubkový slovosled,
- T je zobrazení množiny M do množiny A (slovníku tektogramatické roviny).

Prvky slovníku tektogramatické roviny A potom sestávají z:

- lexikálního významu výrazu;
- *gramatémů* – morfologických údajů ("významový" slovní druh, tj. např. "použít" a "použití" jsou sémantickými slovesy), rozlišuje se tu i aktuální členění;
- *funktoru* – druhu závislosti na nadřízeném členu (aktor, patiens, původ apod.)

Tektogramatickým slovům přiřazujeme valenční rámce a doplňujeme další slova jako doplnění. K tomuto je nutné přidat ještě vztah souřadného spojení, aby bylo možné zapsat všechny věty.

Problémem je také zachycení mezivětných vztahů – můžeme vycházet ze zápisu jednotlivých vět, je ale nutné spojit odkazování se na stejné objekty, případně vztahy mezi nimi (část-celek, druh-instance). Nejsložitější je postarat se o odkazovací zájmena – odkazování dopředu, zpět, zvratnost, odlišnosti na základě akt. členění.

7.2 Metoda TIBAQ: Text and Inference Based Answering of Questions

Projekt vypracovaný na KAM MFF s pomocí významových zápisů vět a porozumění jazyku s inferenčními pravidly. Porozumění se ověřuje schopností vyhledat odpověď na otázku v přirozeném jazyce. Nejde až tolik o zodpovídání otázek, důležitý je způsob práce s jazykem.

Systém obsahuje:

- Morfologickou a syntakticko-sémantickou analýzu věty (převedení do významového zápisu)
- Soubor znalostí (kam jsou ukládány významové zápisy oznamovacích vět)
- Systém výběru relevantních znalostí, aplikace inferenčních pravidel (a vytvoření zásoby relevantních / rozšířených znalostí) pro otázky na vstupu
- Systém výběru a syntézy odpovědi (vybere se z rozšířených znalostí taková, která má stejnou lex. hodnotu vrcholu a cestu ve stromě až k tázacímu slovu jako otázka)

Zkoušelo se na článku o operačním zesilovači, bylo zjednodušení (např. nemělo schopnost instanciace typu). Mělo to inferenční pravidla typu "může dělat" –¿ "dělá", odvozování typu "X je Y" a "X dělá Z" –¿ "X je Y, které dělá Z".

8 Problémy syntaktické analýzy

8.1 Druhy syntaktické analýzy

Rozlišují se dvě základní strategie:

- **Shora dolů** – pravou stranou pravidla se přepisuje levá strana, vychází se od symbolu S pro větu. Podobá se generativní gramatice.
- **Zdola nahoru** – začíná se od slov, postupně se nahrazují synt. kategoriemi až do S. Tj. najdu-li pravidla, která dojdou až k S, pak je všechno v pořádku, jinak backtrackuju.

Dívám se, ke kterým kategoriím patří slova ve slovníku. U obou strategií je běžný způsob zápisu **strom**.

Kromě shora dolů a zdola nahoru existuje i způsob průchodu **middle-out** – najdu sloveso a zjišťuju doplnění. Jde vlastně o závislostní přístup k problému, de facto tak dnes pracují analýzy češtiny.

Jinou možností zápisu během hledání je **přechodová síť (transition network)**. To je vynález **Woodse** (1978), idea **Kaye** (chart parser). Jde o přechod zleva doprava větou s pomocí přepisovacích pravidel, na základě konečného automatu se schopností backtrackingu (přechody představují hrany). V 80. letech se opravdu používalo, bylo to považováno za úspěch. Kay tvrdil, že kromě frázových to bude fungovat i pro závislostní gramatiky. Hlavní problém ale nastal pro jazyky s volným slovosledem (např. na začátku věty musí být pravidla úplně pro všechno, takže se to stane neúnosně náročným na výpočet), což ukázal Sgall. Ani morfologická analýza tomu nepomohla. Podobné problémy se slovosledem měli i při používání v Polsku.

Člověk většinou při porozumění věty používá známé informace, nečemu dává přednost, jde směrem k determinismu.

8.2 Víceznačnost a její řešení

Problém – syntaktická víceznačnost (např. Majitel držel psa v domě: "držel v domě", nebo "pes v domě"?). Řešení:

8.2.1 Strategie "Minimal attachment"

Chci co nejmenší počet uzlů v bezprostředních složkách, nebo co nejmenší počet úrovní v závislostním stromě (tam nemám meziuzly). Tato strategie byla odůvodňovaná podobnými pochody v mysli člověka ("We painted all the walls with cracks." – většinou člověk řekne "painted with cracks", i když je to blbost).

8.2.2 Strategie připojení zprava

Jiná strategie, zastávaná jinými lingvisty. Nové složky se přednostně připojí k nejbližší (právě objevené) složce, ne k nějaké objevené dříve. Náhodně se může shodovat s předchozí. U závislostních stromů se vybírá spíš konstrukce, která je projektivní (ale ne vždy to platí).

8.2.3 Lexikální preference

Kontrola pro spojení, tj. pokud se rozhodnu pro nějaké spojení podle libovolné z předchozích strategií, zkontroluju, jestli to dává smysl podle lexikálních spojení (např. "*vyndat* knížku *z poličky*" x "číst *knížku z poličky*"). Někdy ale stejně nestačí ("kreslit tužkou na obočí").

8.3 Garden path sentences

Jsou věty, které je těžké zanalyzovat, které klamou. Nevím hned, že mám různé možnosti syntaktické analýzy. Např. "The horse raced past the barn fell." Slovo raced vypadá jako časované sloveso, ale je particip. V angličtině je problém kategoriální víceznačnosti, větš. se dělají heuristiky na zákl. shody podmětu s přísudkem apod.

Člověk je zřejmě neřeší zkoušením všech možností, má nějaký *look-ahead*, ale ne až do konce věty. Neví se kolik, 1-2-3 jednotky? Slova nebo struktury? Za magickou hranici se považuje 7 slov / struktur dopředu. Pokus o řešení – **Mitch Markus** (Pennsylvania Univ., žák Chomského, 1980) – analyzátor Parsifal se zásobníkovým look-aheadem (na 3 struktury), nějak to fungovalo, ale stejně mělo problémy.

8.4 Problémy zpracování přir. jazyka

- Jazykové výrazy se vždy vztahují k situaci, mluvčí vyjádřením rozumí více, než řeknou:
  - Význam slova nebo slovního spojení je závislý na kontextu, který předchází nebo následuje (a hlavně následnost je problém).
  - Význam textu není jen složením slov nebo vět, ze kterých se skládá, ale i situace, ve které je užitý. Tedy při porozumění textu člověk hledá i ve vlastní hlavě, umí interpretovat. Počítač potřebuje "vědět", nebo mít "model světa".

  - Slovní spojení v různých jazycích si neodpovídají, specifická použití slov si neodpovídají (např. "loose ice" - "offenes Eis" - "roztátý led", "validate a ticket" - "Karte entwerten")

- Výběr adekvátního ekvivalentu v jiném jazyce je problém (v předch. se dá ještě najít):
  - některé jazyky rozlišují, jiné ne. Např. "fish" -¿ "pescado" / "pez", "go" -¿ "jít" / "jet" apod.
  - V některých jazycích jsou někt. gramatické kategorie, jinde ne, musí se hádat.
- Co je ještě překlad a co už výklad? Kdy si člověk musí text při překladu vyložit?
- Nejednoznačnosti v jazyce – nejlepší je je zachovat i ve výstupu:
  - Lexikální ("All of the expressions are typed.")
  - Syntaktické ("Attach the amplifier to the output terminal with the red dot." x "Attach the amplifier to the output terminal with the red wire").

9 Problémy sémantické interpretace, reference

9.1 Sémantická interpretace

Je nutné odlišovat **jazykový význam** a **kognitivní obsah** (ten je dán mimojazykově, "význam je obsah v zrcadlení formy, vyjádřením obsahu formou se projeví rozdíly ve významu"). Pro význam stačí odstranění lexikální a syntaktické víceznačnosti. Pro obsah je nutné mít znalost světa a situaci.

Přirozené inference ze slov, které člověk udělá, se dají vyjádřit ve slovníku. Slovník pak obsahuje:

- **sémantické rysy** – to, co slovo samo o sobě obsahuje (např.: prezident: muž, člověk, státník)
- **selekční omezení** – do jakého kontextu se může slovo vložit (např.: žrát: subjekt - životné, nelidské + předmět - jedlý; platí jen pro jeden význam)

Sloveso a jeho doplnění se dá taky interpretovat na rovině významu nebo obsahu:

- Z hlediska významu je "stroj běžel" a "Honza běžel" a "potok běžel" to samé, jde o agens
- Z hlediska obsahu závisí sémantické role na konkrétním slovese, nelze dělat moc přesunů ("koupit" nelze beze všeho převést na "prodat")

Při interpretaci by se neměla vynechávat gram. struktura, např. kvůli negacím, kvantifikaci apod., kvůli aktuálnímu členění.

9.2 Problémy referencí

Reference jsou záležitostí roviny obsahu. Druhy:

- **anaforická (koreference)** – odkazy k předchozímu textu
  - v rámci jedné věty, mezivětná – jak daleko se lze odkazovat? 98
- **neanaforická** – odkazy úplně mimo text (to co se ještě v textu neobjevilo).

Problémy:

- Odkazování *k příbuzným prvkům* – dvě možnosti řešení: buď zavádět do báze znalostí asociované prvky hned, nebo až budou potřeba, je hledat. Hajičová doporučuje něco mezi tím, hádat, jestli bude asociovaný prvek potřeba.
- Odkazování *k prvkům množiny*, odkazování *k dějům* – podobné.
- Rozdíl odkazování v *atributivním* a *referenčním smyslu* ("Potkal jsem předsedu." x "Chci být zvolen předsedou.")
- Odkazování vpřed.

10 Historický pohled na strojový překlad

První myšlenka strojového překladu – **Warren Weaver** (1946, v dopise N. Wienerovi) na základě dešifrace Enigmy (myslí, že je to stejný proces). V roce 1949 pak v USA vyšlo memorandum *Translation*, které odstartovalo práce, hl. v podnikových ústavech. První sympozium se uskutečnilo 1952 (ještě před Chomského disertací, Chomsky sám popírá souvislost své práce s počítači).

1954 první projekt – Geoergetown Univ. (Washington) + IBM: Rusko-anglický překlad 250-slovného textu (vybraných vět z jednoho konkrétního článku) podle 6 syntaktických pravidel. Rusko-anglický – na základě studené války. Vzniká časopis *Machine Translation*.

V SSSR se tím taky začli zabývat – překlad anglicko-ruský, skupiny v Moskvě a Leningradě; neměli počítače, šlo o teoretický popis, ale ten byl dost dobrý. Další centra překladu – Cambridge (**Margaret Masterman, Martin Kay**), Grenoble (**Bernard Vaquois**) – tam spolupracovali s Moskvou. Kanada – **Alain Colmerauer, Richard Kittridge** – vytvořili systém Q-jazyk (využívající unifikace nad větami reprezentovanými acyklickými grafy) a na něm postavené systémy TAUM METEO (1978, dodnes se používá).

V Praze 1959 první pokus o překlad 4 vět z angličtiny, taky 6-8 pravidel, dělalo se na 1. českém počítači Prof. Svobody. Založeno oddělení algebraické lingvistiky a teorie překladu na FF.

Susumu Kuno vytvořil na Harvardu *prediktivní analýzu* – parsování věty odleva s backtrackingem, vybíral si v každém kroku nějaké predikce. Proti tomu Rusové shlukují dohromady věci, co jsou vedle sebe – originální přístup, je na tom vidět charakter ruštiny proti angličtině. Hajičová, Jelínek v Praze 1962 chtěli prediktivní analýzu využít, ale pak došlo k organizačním obtížím.

1960 prohlásil **Yehoshua Bar-Hillel**, že "fully automatic high quality" strojový překlad je nemožný, což zvedlo skepsi. Na základě jeho prohlášení vytvořen výbor ALPAC (Automatic Language Processing Advisory Committee), složený z ekonomů, techniků, informatiků, jednoho lingvisty. Ten vydal 1965 zprávu – tzv. *Černou knihu*, kde postuluje:

1. důraz na rychlost a ekonomičnost ručního překladu,
2. podporu projektů komputační lingvistiky,
3. podporu lingvistického výzkumu, bez ohledu na aplikace.

Zpráva nezněla úplně pesimisticky, ale bez ohledu na doporučování se spousta ambicí, projektů zavřela. Lidi na tom dál nepracovali, skončila finanční podpora. Některá pracoviště (hl. mimo USA) přežila.

Se vznikem EU – potřeba mnohojazyčnosti, použití nemožného SYSTRANu (1973), vytvoření projektu EUROTRAN (1983) pro překlad ze všech jazyků. Ten byl úspěšný jen papírově, vzbudil ale zájem o překlady, další výzkum. Důvod neúspěchu EUROTRANU – decentralizace výzkumu, nepoužití převodního jazyka. Udělalo se ale dost práce na dvoujazyčných překladech.

10.1 Mnohojazykový překlad

Různé přístupy:

- Běžně pro každé dva jazyky přímo ze zdroje k cíli
- Přes *interlinguu* – převodní jazyk, který je univerzální. Pak existuje pro každý jazyk jen jedna analýza a jedna syntéza.
- Něco mezi tím – přes *transfer* – malý, ale jazykově závislý modul

Metoda přes transfer je obvyklý přístup, i Vaquoisův. V analýze se přestane, když už dál nemůžu abstrahovat bez ztráty významu.

11 Různé přístupy ke strojovému překladu

11.1 Example-based

Založený na dvojjazyčných korpusech (**Nagao**), počítá pravděpodobnosti, které části korpusu použít (exact match, partial match), problém je udělat správnou metriku podobnosti vět, vytvořit dost velký korpus. Dodnes se v Japonsku používá.

11.2 Knowledge-based

Žádalo se po tom spíše znalost vnějšího světa než znalost překladů frází, mělo obsahovat odvozování. Nedošlo se s tím daleko, hodně práce a málo výsledků. Problém: nutné znát "všechno".

11.3 Translation by negotiation

V mnohojazyčném překladu, vychází z myšlenky přiblížení se interlingue, vyráběl **M. Kay**. Jednotlivé moduly by se měly mezi sebou domluvit, když analýza dojde k maximální možné abstrakci, hlavně na nejednoznačnostech ve vstupu (a jejich interpretaci nebo ponechání). Nejde tedy jen o převod do *transferu*. Jde jen o teoretické myšlenky, podle Hajičové slibné.

11.4 Pražský přístup

Plně automatický: **Z. Kirschner, A. Rosen**. Závislostní, mnohovrstevný, lingvisticky založený překlad z angličtiny do češtiny, s jazykem Q (TAUM METEO). Neměl žádný transfer, problémy hlavně v rozdílech obou jazyků. Veselá historka, kdy málem selhalo předvedení kontrole z ministerstva kvůli nedoplněné tečce za větou.

Poloautomatický: **P. Strossa** – spec. typ text. editoru.

12 Statistický strojový překlad

Předpokládá se *jazykový model* $P(E)$ a *překladový model* $P(F|E)$, kde F symbolizuje francouzský text a E anglický. Potom cílem je najít co nejlepší F pro E s pomocí jazykového a překladového modelu, o což se stará *search procedure*. Chyby mohou nastat v obou modelech i při hledání (hlavně pro příliš složité search procedury). Jazykový model se typicky dělá s pomocí trigramů nebo rozhodovacích stromů, čehokoliv, co jde pak nasadit na hledání. Mnohem lepší model se dostane s pomocí morfologické analýzy, teoreticky i se syntaktickou analýzou, ale ta v praxi většinou zatím moc nepomáhá.

Lze udělat drastické zjednodušení modelu na odpovídání si slovo od slova. Je to hodně přibližné, ale pořád ještě relevantní – překlad je (mimo jiné) dán slovem. Teoreticky by šlo dělat závislosti na slovních spojeních, ale na to bývá málo dat. Aby se vyrovnal počet slov v obou větách, provádí se *alignment (párování)*. Tak můžou třeba některá slova i vypadnout, spárují se s *nulovým slovem*.

```
. And the program has been implemented
/   /     /     /   /     / |  \
.  Le programme a  été  mis en application
```

Kvůli search proceduře může z francouzštiny pro každé slovo jen jeden spoj (opačně to nevadí) – jinak by bylo pro $|F| = m$ a $|E| = l$ až $2^{l \cdot m}$ spojení, takhle získáme jenom $(l+1)^m$, což je o dost méně. Tohle se dá ale řešit předzpracováním, protože pokud by mělo z francouzštiny vést více spojů, půjde většinou o odborné termíny. Ty se dají potom považovat za jednolitý celek.

Alignment je potřeba natrénovat, nejlépe ručně, což ale vzhledem k objemům dat není moc proveditelné. Používá se na to strojové učení.

Kdybychom přidali alignmenty do modelu, museli bychom ho celý přepočítat přes pravděpodobnost všech možných alignmentů, což by bylo moc složité. Proto se volí jeden alignment a považuje za správný. Potom platí

$P(F, A|E) = P(m|E) \prod_{j=1}^{m} P(a_j|a_1^{j-1}, f_1^{j-1}, m, E) P(f_j|a_1^j, f_1^{j-1}, m, E)$

Alignment je funkce, má tedy jen jednu hodnotu, takže jde spočítat. Každé slovo v cílovém jazyce je tak určeno z alignmentů od něj vlevo a ze slova v původním jazyce.

Tohle je pořád moc složité – v praxi se udělá jeden hodně jednoduchý model a podle něj se řídí vyšší a vyšší modely. Celkově dostávám pětistupňovou hierarchii (podle prvního použití v IBM, stupně jsou ustálené):

- **Model 1** počítá pravděpodobnost délky angl. věty v závislosti na franc. větě (aby nedovoloval přeložit dlouhou angl. jako krátkou franc. a naopak). Slova jsou na sobě nezávislá.
- **Model 2** podporuje svislý alignment (ve směru poměru délek vět), tj. preferuje se shodný slovosled.
- **Model 3** se dívá, kolik spojů vede z angličtiny k francouzštině (tj. počet spojů z jednoho slova) a zjišťuje pravděpodobnosti počtu v závislosti na slově. Snižují se pravděpodobnosti pro alignmenty, které se hodně kříží ve větě (*distortion*).
- **Model 4** opět zvýší pravděpodobnosti pro distortion celých frází (aby se daly přesouvat ve větě).
- **Model 5** odstraní *deficiency*, tj. zařídí, aby nenulovou pravděpodobnost nedostalo něco, co není řetězec (má mít dvě slova na stejné pozici). Může ale být vypuštěn, výsledek většinou už moc neovlivní.

Po natrénování všech modelů (postupně vyšší nad nižším) už mám jen nejvyšší a ten používám.

Search procedura pro daný výstup ve francouzštině najde odpovídající vstup v angličtině (tj. hledá "obráceně"). Prochází větu postupně, alignuje a podle pravděpodobností z modelu si pamatuje n nejlepších a zbytek zahazuje (A*-algoritmus). Je tu nutné porovnávat pravděpodobnosti jen pro prefixy stejné délky!

Předem se mohou udělat (nejlépe pro oba jazyky) některá vylepšení – tagování, lematizace a disambiguizace významů slov (na základě kontextu). Mohou se lepit i některé termíny, aby se překládaly přesně. Výsledek přípravy pak považuju za "slova" a search procedura s nimi normálně pracuje.

Hlavním problémem těchto typů překladů jsou "sparse data" – paralelních dat je málo a typicky nepotkám dvakrát stejnou větu. Nejvíc dat existuje právě pro angličtinu a francouzštinu díky kanadské dvojjazyčnosti. Pro češtinu a angličtinu je dat málo a navíc jsou problémy s právy je používat.

13 Textově lingvistický přístup – aktivovanost prvků v textu

Zajímáme se o návaznost věcí na sebe, o aktuální členění. Předpokládáme při začátku komunikace (textu):

- *Sdílenou zásobu znalostí (stock of shared knowledge)* – zásoby znalostí producenta i adresáta se částečně překrývají, cílem je překryv zvýšit.
- Ve sdílené zásobě znalostí je od začátku hierarchie *aktivovanosti* – některé objekty jsou více v popředí než jiné.

V průběhu komunikace se aktivovanost může měnit, hovor ji ovlivňuje. Při stanovení pravidel změny aktivovanosti vycházíme z akt. členění věty. Dostáváme pět heuristik (přibližných):

1. Při vstupu do textu je pořadí *topic-focus* (říkám o něčem něco), tedy předpokládáme, že aktivovanost prvků v základu je vysoká, ale ne tak vysoká, aby *focus* nepřekryl *topic*.
2. Potom *focus* překryje *topic* – je ještě aktivovanější.
3. To, o čem nemluvím, aktivovanost ztrácí.
4. Rychlost ztrácení aktivovanosti závisí na tom, jak moc a jak dlouho byl objekt aktivovaný předtím.
5. Když má něco urč. stupeň aktivovanosti, ovlivňuje to i "asociované" prvky (Tenhle bod je dost složitý, je obtížné se jím zabývat).

Aktivovanost se dá vyjádřit čísly (stupeň aktivovanosti je číslo), závisí ale hlavně na jejich poměru, ne na konkrétních hodnotách. De facto se aktivovanost chová jako zásobník, nahoře jsou aktivované koncepty, neaktivované klesají.

Dá se nakreslit graf aktivovanosti, vytváří zvláštní patterny – čáry "kooperují", dlouho zůstává v popředí jeden koncept, čímž se dají nalézt segmenty textu. Můžu podle aktivovaných prvků říct, "o čem to je", tedy domyslet si téma článku, ač není v nadpisu – např. podle rámců situací. Podle toho je možné hodnotit i kvalitu překladu textu.

Text se může teď segmentovat nejen "horizontálně" na odstavce, ale i "vertikálně" – jak daleko klesl prvek v aktivovanosti? To je důležité pro zájmenné reference. Hranice toho, co ještě může být odkázáno zájmenem, se asi nedá číselně vyjádřit přesně, ale odhadnout na konkrétním textu jde. Podle toho by se možná dalo vyrobit i pravidla odkazování.

Mnohem složitější je zaznamenat takhle dialog – nevíme, zda je lepší brát mluvčí odděleně nebo dohromady. Pro automatické vyrábění analýzy aktivovanosti by bylo potřeba vyřešit koreference, odkazování na část celku apod.