Úvod do počítačové lingvistiky – příprava na zkoušku

Tento text vznikl jako pomocný text k přípravě na zkoušku z předmětu Úvod do počítačové lingvistiky v roce 2011. Může obsahovat mnoho chyb, dohadů a nepřesností. Tedy jeho autoři se vzdávají zodpovědnosti za důsledky z toho vyplývající... Za každou větší kapitolou následuje seznam možných otázek, jejichž zdrojem jsou wiki matfyz a matfyz fórum. Markéta Popelová a Jakub Tomek.

1 MORFOLOGIE

Předmětem morfologie je studium vztahů mezi jednotlivými částmi slov, studium vnitřní struktury slov. Zabývá se tvořením tvarů slov a jejich významem, dále i tvořením nových slov. Studuje způsoby skloňování (deklinace) a časování (konjugace).

S morfologií se pojí tyto pojmy:

- *lexikologie* – slova jsou studována jako jednotky slovní zásoby
- *lexikografie* – sestavování slovníků

Základní jednotkou je *morfém* = nejmenší jednotka nesoucí význam. Existují dva typy morfémů:

- *lexikální* morfém = kmen slova, který nese význam slova,
- *gramatický* morfém, který určuje gramatickou roli slovního tvaru.

Např. za-hrad-ou má předponu „za“, lexikální morfém „hrad“ a gramatický morfém „ou“, který určuje 3 elementární jednotky: pád, číslo a rod.

Další pojmy:

- *tvaroslovné dublety* – stejné slovní tvary odvozené od více slovních základů (žena, stát atd.), neboli slova víceznačná, která mají různé slovní druhy,
- *alternace* – změna hlásek uvnitř kmene (vůz → vozu, švec → ševce, prkno → prken) – problematické např. pro vyhledávání v textu
- *alomorfy* – varianty kmene odvozené od stejného slovního základu (matka-matce- matek-matčin).

Slova se dělí na *autosémantická* = plnovýznamová a *synsémantická* = pomocná.

Morfologická typologie jazyků dělí jazyky na:

- *analytické* (slovo = morfém) → izolační – každé slovo je morfém, bez předpon/přípon: vietnamština, částečně angličtina
- *syntetické* (slovo > morfém)
  - *flektivní* – mají předpony, přípony, koncovky (ale míra řetězení je nějak omezená), daný tvar morfému nese více významů (koncovka určuje pár, rod, ...): latina, stará řečtina, slovanské jazyky
  - *aglutinační* – také různé předpony apod., ale jeden morfém nese jeden význam (tedy např. koncovka přidá jeden význam): maďarština, japonština
- *polysyntetické* (slovo = věta): eskymácké a indiánské jazyky

Jak zpracovávat morfologii? Podle toho, na čem je morfologie založená:

- na *morfémech* – vidí slovo jako řetízek morfémů.
- na *lexémech* – vidí slovo jako výsledek aplikace pravidel, která slovo mění a tím vytváří nový tvar (např. tvorba množného čísla v angličtině)
- na *slovech* – centrální roli mají vzory slov. Známe-li základní tvar slova a jeho vzor, dokážeme vygenerovat všechny zbylé tvary. Vhodné i pokud jeden morfém reprezentuje více gramatických kategorií (např. 3.os, sg., r.ž.), kde předchozí přístupy selhávají.

Two-Level Morphology

Populární spolehlivý systém zpracování morfologie vyvinutý Lauri Karttunenem a Kimmo Koskeniemmin na začátku 80. let. Jednalo se o první obecný model zpracování morfologie přirozeného jazyka. Systém je založen na konečněstavových automatech a na nich definovaných oboustranných přechodech. Mechanismus morfologie byl pro všechny jazyky společný (to byl požadavek pro nezávislost na jazyku), ale pro každý jazyk se musel vytvořit slovník a pravidla (přechody mezi stavy – zachytí přechod mezi konečným stavem a lemmatem).
První úroveň zpracování je lexikální (hloubková), druhá povrchová. Pravidla se aplikují paralelně, nikoliv sekvenčně – oproti předchozím modelům nezáleží na pořadí. Podmínky se mohou vztahovat k oběma úrovním zároveň či k jedné z nich. Lexikální vyhledávání a morfologická analýza probíhají současně.

Česká morfologie

Vyvíjena od r. 1989 prof. Hajičem a kol. Využívá poziční značky, tedy každé slovo má 15-ti písmennou značku, která ji určuje. Z těchto 15 písmen se využívá jen 13 (2 jsou rezervní), každá kategorie má své pořadí ve výsledné značce. Některé značky se vzájemně vylučují (např. příslovce nemá osobu), pak se na dané pozici píše pomlčka. Kromě značky se každému slovu ještě přiřadí jednoznačné lemma, což je základní tvar slova. Značka a lemma dohromady jednoznačně určují slovo a jeho tvar se vším všudy.
K některým tvarům daného slova pasuje více značek, pak se tam napíšou obě. Např. slovo „funkci“:
`<f>funkci<MMl>funkce<MMt>NNFS3-----A----<MMt>NNFS4-----A---- <MMt>NNFS6-----A----`
Lemma je „funkce“ a značky jsou 3, neboť pád může být dativ, akuzativ i lokál(3., 4. i 6).

Činnosti využívající morfologii

- *Morfologická analýza* – správná analýza dá všechny možné tagy
- *Morfologické značkování* (tagging) – proces výběru jediné správné značky v daném kontextu na základě statistických metod nebo algoritmických pravidel. Úspěšnost až 96%.
- *Částečná morfologická desambiguace založená na pravidlech* – Oliva a Petkevič vytvořili “spolehlivá” pravidla, která platí bez výjimek. Pomocí těchto pravidel se redukuje počet značek, ponechává všechny, které nelze spolehlivě odstranit (asi 40%). Náhodou se takto ovšem přišlo na jinou zajímavou aplikaci – pokud se u některého slovního tvaru vyškrtá vše, znamená to, že ve větě musí být gramatická chyba. Tedy jednou z aplikací je kontrola gramatických chyb.
- *Lemmatizace* – proces výběru správného základního tvaru (lemmatu), ze kterého byl odvozen daný vstupní tvar. Klíčová operace pro vyhledávání v textech. U nás má úspěšnost 99,9%.
- *Stemming* – odříznutí koncovky, na rozdíl od lemmatizace je základním tvarem kmen slova. Populární je tzv. Porterův stemmer.
- *Generování* – proces výběru správného slovního tvaru ze zadaného lemmatu a množiny značek (gram. kategorií).

Kontrola překlepů jako aplikace morfologie

Obtížné požadavky: najít a opravit všechny překlepy, přezkoušet kontextové podmínky korigované verze, neznámá slova se nemají hlásit jako chybná (nesplněný požadavek u většiny současných systémů), nedávat falešná chybová hlášení, co nejvíce automatická korektura, krátký čas zpracování.
Používají se 2 základní metody:

1. *Porovnávání řetězců se slovy ve slovníku:* Buď se slovníkem všech možných slovních tvarů

daného jazyka (wordlist) nebo se slovníkem lemmat, provádí se morfologická analýza. Je to spolehlivé a jednoduché, ale pomalé, náročné na kvalitu slovníku, místo, nerozezná to chybná slova od neznámých a každé zlepšení musí zařídit autor či uživatel.
2. *Porovnávání skupin znaků (dvojic, trojic) a hledání nedovolených kombinací znaků:* Nezávislé na slovníku a rychlé. Ale také neúplné a neodhalí překlepy ve slovech, která se skládají jen z vhodných kombinací znaků.

Možná vylepšení: vzít v úvahu okolnosti chyb (např. blízké klávesy), zohlednit statistiku chyb a časté pravopisné chyby (mně x mě, jsem x jsme), různé heuristiky na oddělení chyb a neznámých slov, zapojení syntaxe a sémantiky, pracovat s kontextem (např. porovnávat korpusy).
Související pojmy: *precision* = počet nahlášených chyb / počet nahlášených slov, *recall* = počet nahlášených chyb / počet všech chyb. Lepší je odhalit méně chyb, ale jen to, co jsou opravdu chyby. To znamená více se snažit *maximalizovat precision* (aby co nejvíce nahlášených slov byly opravdu chyby) a tolik neřešit recall (aby co nejvíce z chyb bylo odhaleno).

Systém ASIMUT (Automatická Selekce Informací Metodou Úplného Textu)

ASIMUT (1990 Králíková, Panevová) měl dva moduly:

1. *Vyhledávací modul:* Sloužil pro automatické vyhledávání ohýbaných slov v textu na základě parametrů (uživatel zadá vyhledáváný výraz doplněný o značky). Na vstupu jsou výrazy složené z podstatných a přídavných jmen doplněné o sadu operátorů (**!** vyskloňovat, **-1-** obě slova musí být bezprostředně vedle sebe, **-2-** ob slovo, **-3-** ve stejné větě, **-4-** ve stejném odstavci).

   Příklad vyhledávání: `vzdálenost!, rodinný! -1- domek!` znamená všechny výrazy, které obsahují buď slovo „vzdálenost“ v nějakém tvaru, či sousloví „rodinný domek“, opět v libovolném tvaru (neřeší, jestli gramaticky správném). Předpokládá se členění textu na slova, věty a odstavce.
2. *Jazykový modul* pro automatické skloňování českých slov: Pro dané vstupní slovo vrátí všechny jeho možné tvary. Využívá retrográdní (seřazený odzadu) slovník dr. Slavíčkové (1975). Vychází z myšlenky, že slova se stejnou koncovkou se skloňují stejně. Výjimky z tohoto pravidla je možno uložit do zvláštního slovníku (jsou jich stovky, maximálně tisíce). Retrográdní slovník nevyužívá přímo, ale na jeho základě byl vytvořen klíč pro určování vzorů slov (seznam pravidel dle konců slov).

ASIMUT – algoritmus

Porovnává písmena vstupního slova (musí být v základním tvaru, tedy 1. pád sg.) odzadu, dokud nenajde jednozačný vzor pro skloňování. Pak slovo vyskloňuje dle vzoru (umí i základní alternace). Má různé problémy: ne vždy lze jednoznačně určit vzor (právník i trávník mají stejnou koncovku, ale liší se v životnosti), problém přegenerování (systém vygeneruje i neexistující tvary), malý rozsah retrográdního slovníku (40 000 slov) → je nutno přidávat výjimky, pro slovesa už nefunguje spolehlivě.

Další pojmy – ASIMUT:
Negativní slovník obsahuje ta slova, která nejsou při vyhledávání v textu důležitá (spojky, citoslovce), proto jsou odstraněna z textu ještě před vyhledáváním.
Konkordance – při tomto procesu je všem slovním tvarům mimo negativní slovník přiřazena adresa a frekvence výskytu v textu. Slova z negativního slovníku dostanou jen adresu (kvůli počítání vzdáleností). Samotné vyhledávání pak neprobíhá na původním textu, ale na této konkordanci.

Systém MOSAIC (Morphemic Oriented System of Automatic Indexing and Condensation)

MOSAIC (70. léta Kirschner a kol.) byl vytvořen pro indexaci dokumentů, tvoření souhrnů a seznamů klíčových slov. Podobně jako ASIMUT nepoužívá rozsáhlé slovníky, ale lingvistické poznatky: využívá toho, že řada přípon a koncovek nese význam (AJ: -er/-or konatel děje, -tion činnost, -ity/-ness vlastnosti; ČJ: -ič/-ač/-čka/-ér/-or/-dlo/-metr/-graf/-fon/-skop přístroje a nástroje, -ací/-ecí účel, atd.) .

MOSAIC – algoritmus

1. Na vstupu je nepředzpracovaný text se zachovaným typografickým členěním.
2. Lemmatizace a morfologická analýza → získáme lemata a morfologické značky.
3. Jsou odstraněna lemmata, jejichž kmen nemá vztah k dané tematické oblasti (k tomu se využívá malý negativní slovník), či jsou příliš krátká nebo obsahují nevhodné kombinace hlásek.
4. Syntaktická analýza jmenných skupin. Využívá jednoduchou gramatiku v jazyce Systémů Q. To pomůže odhalit tématicky významné několikaslovné termíny (operační zesilovač TESLA KC 415).
5. Vážené ohodnocení termínů podle důležitosti. Záleží na tom, v jak důležité části textu jsou (nadpis, první/poslední odstavec, první/poslední věta). Váhy jsou exponenciální.
6. Normalizace vah vzhledem k délce dokumentu. (Nejčastější termín získá 100 bodů, zbytek poměrně.) Umožňuje porovnávat různě dlouhé dokumenty.
7. Výstupem je 10 nejvýznamnějších termínů, seřazených podle četnosti výskytu.

Výhody: Není nutno vytvářet specializované slovníky odborných termínů. Stačí množiny relevantních koncovek a přípon, daný negativní slovník a několik pravidel. Lokální syntaktická analýza umožňuje větší flexibilitu při hledání termínů.
Problémy: Neřeší nevyjádřené podměty, odkazování v textu pomocí zájmen apod. Pracné vymýšlení negativního slovníku, pravidel a koncovek.

Otázky:

1. Alomorfy.
2. Co je to morfém a jak ho klasifikujeme?
3. Lemmatizace – co to je a kde se používá.
4. Two-level morphology.
5. Morfologie a kontrola pravopisu.
6. Kontrola překlepů: podrobně popište metody používané při automatické kontrole překlepů (i volby dialogu s uživatelem).
7. ASIMUT (Co to je, jak funguje, na čem je založen jeho jazykový modul.)
8. MOSAIC (Používá se u systémů MOSAIC syntaktická analýza? Proč ano/ne.)

2 SYNTAX

Syntax (skladba) se zabývá vztahy mezi slovy ve větě, tvořením větných konstrukcí a slovosledem. Existují dva typicky používané datové typy:

1. *Závislostní strom* se podobá větnému rozboru ze základní školy, kořen je jediný a obsahuje přísudek. Uzly stromu jsou slova věty. Každé slovo závisí na jiném, závislost je popsána orientovanou hranou mezi slovy. Závislostní strom tedy dobře a přehledně zachycuje vztahy mezi jednotlivými větnými členy, ale nedává návod, jak strom získat (strom nezachycuje postup výpočtu). Ne vždy je jednoduché určit, jak (a zda) jsou slova závislá. Proto mají závislostní stromy další možnosti pro zaznamenávání následujících jevů:
    1. *koordinace* – různé větné členy se stejnou sémantickou rolí (Jan a Marie; černý nebo bílý).
    2. *apozice* – různé větné členy se stejnou syntaktickou rolí a shodnou gramatickou kategorií (tzn. gramaticky kongruentní). Např. Matematicko-fyzikální fakulta (MFF); Ivo Truchlivý, učitel matematiky.
    3. *parenze* (vsuvka) – věta či větný člen, který syntakticky nesouvisí s okolím, ale upřesňuje, o čem se v okolí mluví (Mohl bych, prosím, zavřít okno?)

2. *Složkový (derivační) strom o*dpovídá derivačnímu stromu bezkontextové gramatiky. Věta se rozdělí do částí, které se zase rozdělí do částí (…) . Slova věty (tokeny) odpovídají uzlům stromu. Je méně přehledný, má větší množství uzlů a má jeden hlavní problém – přirozené jazyky nebývají bezkontextové. Problematické jsou tedy právě neprojektivní konstrukce. Derivační strom se dá znázornit pouhým uzávorkováním věty, kde uvnitř závorky jsou vždy právě dva prvky – jiný uzávorkovaný výraz nebo samotné slovo.
   `př. ((Malý chlapec) ((psal dopis) (na stole)))`

Neprojektivní konstrukce

Neprojektivní konstrukce je závislost mezi dvěma slovy ve větě oddělenými slovem třetím, které (ani nepřímo) nezávisí na žádném z nich. V češtině jsou běžné.

`př. Soubor se nepodařilo otevřít. / Tuto knihu jsem se mu rozhodl dát.`

Závislostní strom s tím nemá problém (jen hrany se v něm jakoby kříží), ale složkový ano (viz slidy).

Transformační gramatika

Navazuje na předválečnou americkou lingvistiku, která se snažila explicitně popsat jazyková pravidla. Mezi “předchůdce” transformační gramatiky patří:

- *Deskriptivismus* (*Language* 1933 Bloomfield). Jazyková fakta popisuje, klasifikuje a registruje, ale nevysvětluje. Zabývá se spíše povrchovou větnou strukturou.
- *Analytická syntax* (*Analytic Syntax* 1937 Jespersen).
- *Logický přístup* (Kazimierz Ajdukiewicz) – kategoriální gramatika. Obecně se zavádí koncept povrchové a hloubkové syntaktické struktury (surface & deep structure). Povrchová struktura se zabývá zápisem, hloubková významem. Pak je běžné, že jedné povrchové reprezentaci může odpovídat více hloubkových (jedna věta je významově víceznačná), stejně jako naopak (jeden význam se dá vyjádřit různě).

Dosud jazyk typicky nebyl popisován formální matematickou strukturou, často se míchala syntax a sémantika (syntaktické jevy se popisovaly pomocí sémantiky).

*A transformational grammar has a "natural tripartite arrangement": **phrase structure rules**, **transformational rules** and **morphophonemic rules**.*

- *The phrase structure rules are used for the expansion of grammatical categories and for substitutions. These yield a string of morphemes.*
- *A transformational rule "operates on a given string...with a given constituent structure and converts it into a new string with a new derived constituent structure." It "may rearrange strings or may add or delete morphemes."Transformational rules are of two kinds: obligatory or optional. Obligatory transformations applied on the "terminal strings" of the grammar produce the "kernel of the language", which are simple, active, declarative and affirmative sentences. To produce passive, negative, interrogative or complex sentences, one or more optional transformation rules must be applied in a particular order to the kernel sentences.*
- *At the final stage of the grammar, morphophonemic rules convert a string of words into a string of phonemes.*

Noam Chomsky v knize *Syntactic Structures* popsal 3 základní komponenty transformační gramatiky:

- *Báze*: soubor bezkontextových pravidel. Tato pravidla generují složkové stromy, tzv. frázové ukazatele (phrase makers). S → NP VP, kde NP je noun phrase, VP je verb phrase.
- *Transformační komponenta*: soubor transformačních pravidel nad frázovými ukazateli. Z nich vytváří povrchovou strukturu věty. Dva typy transformačních pravidel:
  - Obligatorní – transformace musí být provedena (pokud je to možné).
  - Fakultativní – transformace je volitelná.
- *Fonologická komponenta*: soubor regulárních přepisovacích pravidel. Řetězcům morfémů přidělují fonetickou interpretaci a význam. (Fonetika i fonologie zkoumají zvukovou stránku jazyka. Fonetika zkoumá, jak se hlásky v těle tvoří a vnímají, zatímco fonologie zkoumá funkci hlásek a zvukovérozdíly, které mají v jazyce nějakou funkci.)

Cílem je tvořit věty – jsou vytvářeny generativní procedurou, která používá různá přepisovací pravidla (někdy kontextová, jindy bezkontextová). Není ale schopná zachytit vztahy mezi variantami vět, např. mezi větou tázací a oznamovací. Transformace (v transformační komponentě) jsou definovány strukturním indexem řetězců (řez stromem, výraz se matchuje na množinu vrcholů) a strukturní změnou (co se má s namatchovanými vrcholy provést). Pravidla mohou být bezkontextová. Pak má tato složka sílu Turingova stroje, což je moc. V dalších verzích byla tato složka oslabena.

Vývoj transformační gramatiky

Obecně se vývojem táhne omezování informací.

- 1965 *Standard Theory*, popsaná v *Aspects of the Theory of Syntax* (Chomsky)
- 1968 *Extended Standard Theory*
- začátek 80. let *Government-binding Theory* – teorie založená na obecných principech univerzální gramatiky všech jazyků a parametrech platných pro jednotlivé jazyky.
- začátek 90. let *Teorie minimalismu* – obsahuje jen dvě roviny: rovinu *logické formy* (reprezentace jazyka a významu) a *fonetickou* rovinu (zvuková stránka jazyka).

Tree Adjoining Grammars (TAGs)

TAG je formalismus pro popis gramatik. Elementární strukturou TAG jsou stromy. V gramatickém stromu se uzly nahrazují jinými stromy (např. X miluje Y=Emil, tak za X se dá dosadit „Milan“, ale i strom „Milan, Ferda a Dežo“), ale jen tehdy, když se uzel a kořen stromu shodují. Uzly, které je možno substituovat, jsou označeny šipkou. Proces končí, když už nelze žádný uzel nahradit. Svou generativní silou mohou po drobných modifikacích TAG dosahovat až kontextových gramatik.

Typy základních stromů:

- základní (initial) strom udává valenční vztahy a strukturu věty
- pomocný (auxiliary) strom: pomocí těchto se tvoří rekurze ve stromu

Typy změn:

- *substituce* – list stromu je nahrazen pomocným stromem, jehož kořen je značený stejně, jako list původního stromu.
- *adjungace* – vnitřní uzel je nahrazen pomocným strom, kořen opět značen stejně jako list původního stromu.

Lexical-Functional Grammar (LFG)

Využívá dva typy struktur:

- *c-struktura* (constituent), která spojuje slova do frází, datovým typem je složkový strom
- *f-struktura* (functional), která reprezentuje funkční vztahy ve větě (např. vazby sloves). Používá datový typ matice atribut-hodnota (že by mapa).

Každá c-struktura se spojuje s jednou f-strukturou. Opačně jich může být i více.

Unifikační gramatiky

Každý objekt je reprezentován množinou vlastností, tzv. rysů. Stylem <název_vlastnosti>: <hodnota_vlastnosti>. Této množině říkáme sestava rysů. Vlastnosti mohou být např. grafémický zápis, slovní druh, rod, číslo, pád, atd. Jejich hodnotami mohou být i další sestavy rysů či proměnné.
Máme-li dvě sestavy rysů popisující objekt, můžeme je unifikovat, jestliže nejsou v konfliktu v žádné vlastnosti.
Problém je, že lze unifikovat i vlastnosti, které spolu nesouvisejí, třeba pád podmětu a způsob přísudku.
Typové sestavy rysů využívají toho, že některé typy objektů mají společné vlastnosti. Např. u sloves určujeme osobu, číslo, čas, způsob atd. Typy se pak většinou řadí hierarchicky. Slova se dělí na ohebné a neohebné druhy. Ohebné zase na časované a skloňované atd.
Příklady:
FUG (Funkční unifikační gramatika, Martin Kay)
HPSG (Head Driven Phrase Structure Grammar) – zahrnuje principy, gramatická pravidla a slovníkové položky (tříděné, dle různých kategorií). Slovo má dva základní rysy – phon (zvuk, fonetická forma) a synsem (syntaktické, sémantické informace) – tyto rysy dále děleny. Základním typem je znak, který se dělí na slova a fráze.
GPSG (Generalized Phrase Structure Grammar 1985)

Kategoriální gramatiky

Každému slovu je přiřazena kategorie – množina syntaktických vlastností daného slova. Zápis kategorií X/Y, či X\Y, podle toho, zda argument je vlevo, či vpravo. Existují dvě základní pravidla: X/Y Y → X; Y X\Y → X.

Nástroje na syntaktickou analýzu.

Augmented Transition Networks (Woods, 1970).

- Má tři typy hran: CAT (přechod do stavu, nalezne-li příslušnou kategorii), JUMP (přechod do stavu bez hledání kategorie), SEEK (přechod k podsíti).

Q-systémy (Colmerauer 1969).

- Formalismus pro transformaci grafů. Grafy jsou linearizovány, např. `S(NP,VP(V,NP))`. Používají se tři typy objektů – atomy (písmena A-J), stromy (L-N) a seznamy stromů (U-Z) – a operátory `DANS`, `HORS`, `ET`, `NON`, `OU`, `=`, `"`.
- Např. `S(NP,VP(V,NP))` může být popsáno jako `A*(U*)` nebo `S(NP,L*)`, kde '*' signalizuje, že se jedná o proměnnou.
- Q-systémy používal např. RUSLAN, ale jinak byly oblíbené téměř jen ve francouzštině. Jinde byly oblíbenější rozšířené přechodové systémy.

Funkční generativní popis (Sgall 1967, Panevová, Hajičová)

- Stratifikační teorie – řeší popis na více vrstvách:
  - fonetická, fonologická, morfématická, povrchová, tektogramatická.
- *Princip forem a funkcí*: jednotka na vyšší rovině reprezentuje funkci jenotky na nižší rovině. Na vyšších úrovních (povrchová a tektogramatická) se jazyk popisuje závislostní reprezentací, typicky závislostními stromy.
- Teorie valence, viz níže.

Valence

Teorie valence existuje od 60. let, její základy vytvořili Kurylowicz a Tesniere, rozpracoval ji Charles Fillmore ve své *Case Grammar*, kde studoval sémantické role jednotlivých slovesných aktantů. Valence je schopnost některých slov (především sloves) „vyžadovat" jiné větné členy a tvořit s nimi věty. V dané teorii se rozlišují na *tektogramatické rovině* dva typy základních členů:

- *aktanty* : Konatel (aktor, agens), Patient, Adresát, Origo, Efekt – každý může být ve větě zastoupen pouze jednou
- *volná doplnění* se můžou vyskytovat vícekrát

Dále se vazby v tektogramatické rovině dělí na *obligatorní* (povinné) a *fakultativní* (nepovinné) – obligatorní aktant nesmí ve větě chybět (může ale chybět na povrchové rovině, pokud ho známe z kontextu). *Valenční rámec* je seznam aktantů (i fakultativních) a obligatorních volných doplnění. 2007 sestavili Lopatková a Žabokrtský valenční slovník *Vallex* (viz http://ufal.mff.cuni.cz/vallex)

Kontrola gramatické správnosti:

Kontrolovat lze (např. typicky pro češtinu) shodu podmětu s přísudkem, interpunkci, neprojektivní konstrukce, tvar zájmen ...
Kontroluje se za pomocí těchto nástrojů:

- Chybové vzorky: Vhodné hlavně pro jazyky s pevným slovosledem, kde chybné konstrukce zůstávají v lokálním kontextu a nerozlézají se daleko po větě.
- Gramatika: Nelze ale rozeznat, zda je věta špatně, nebo zda je správně vzhledem k neúplné gramatice. To se snaží řešit RFODG (Robust Free-Order Dependency Grammar).
- RFODG: Výpočet probíhá ve fázích. Interpret gramatiky rozhoduje, jak se bude stejné gramatické pravidlo používat. Je snaha o co nejplynulejší fázování výpočtu, což 2001 zlepšil např. i Holan.

LanGR (Květoň 2003)

To, co používá MS Word. Pravidla jsou psána ručně na základě korpusu, pracují v cyklech. Snaha o vysokou precision (85%) oproti recall (30%), uživatele „otravuje“ tak jednou za 3-4 stránky. Každé pravidlo má 4 části: kontext, desambiguace, report a akce. Pravidla byla tvořena speciálně pro češtinu.
LanGR používá desambiguaci na to, že když se odstraní všechny tagy, tak víme, že je něco špatně. V tu chvíli se ale musí určit, co je špatně a jak to opravit. Neřeší to ten problém, že věta může být správně, ale až v dalekém kontextu přes jiné věty. (Tatínek šly do práce.)
Je problém, jak hodnotit kvalitu grammer-checkeru. Je nutno to dělat ručně. Obecně používá tuto přípravu (klasický postup při zpracování psaného textu): segmentace (rozseká na věty) → tokenizace (rozseká na slova) → morfologická analýza (každému tokenu dá seznam dvojic lemma – tag) → morfologická desambiguace (každému tokenu vybere ideálně jeden token) → syntaktická analýza (větný rozbor) → sémantická analýza (rozbor významu věty).

Otázky:

1. Syntaktická analýza.
2. Závislostní (D-tree) a složkový (C-tree) strom pro větu "Ve včerejším závodu startovali výborní skokani."
3. Převeďte složkový strom na závislostní.
4. Chomského transformační gramatika (v jiné otázce jako teorie popsaná v knize Syntactic structures).
5. Na co slouží strukturní index u Chomského gramatiky?
6. Sestavy rysů a jejich použití.
7. Co jsou to unifikační gramatiky, jejich výhody, nevýhody.
8. Hloubková a povrchní syntaxe, vztahy mezi nimi.
9. Valence.
10. Co je to HPSG, LFG, FGD.
11. Co znamená zkratka TAG, stručne vysvětlit princip.
12. Na čem je založena teorie funkčního generativního popisu.
13. Teorie minimalismu - autor a na které teorie navazuje.

3 STROJOVÝ PŘEKLAD

Problémy strojového překladu:

- Cílem překladu není převést slovo na slovo, ani větu na větu, ale převést sdělení z jednoho jazyka do druhého, aby se dalo pochopit, co chtělo říci. Nestačí tedy překládat doslovně, dokonce nestačí ani překládat se znalostí morfologie, ustálených slovních spojení, či syntaxe, ale je třeba znát i kontext, který v jazyce obsažen vůbec není.
- Existují různé počty výrazů pro určité slovo, některé jazyky jsou „jemnější“ (Inuité a sníh). Dále podobné významy pro danou oblast nemusejí jít přímo namapovat na obdobné výrazy v jiném jazyce (vaření v angličtině/japonštině).
- Význam slova závisí na kontextu – „otevřít“ může být různé pro program, okno, plechovku...

Historie ve světě:

Automatickým překladem se lidé zabývali již od 30. let, postupně zkoušeli překlady slovo od slova → slovníky pro předpony, kmeny, přípony a koncovky zvlášť → zaveden preediting (ručně se text oseká o různé zvláštnosti, zkrátí se věty, víceznačná slova se nahradí jednoznačnými, stroj pak tyto snadné věty už přeloží) a postediting (stroj přeloží, co umí a zbytek přeloží člověk) → zaveden *pivotní jazyk* (politický problém, který jazyk se má zvolit + kumulují se zbytečné chyby, např. pro překlad příbuzných jazyků např. Čs-Aj-Sk) → experiment překladu jednoduchých vět mezi Aj a Ru (úspěšné).
V roce 1966 ALPAC, zpráva, která v USA utlumila výzkum, tedy objevy po tomto roce jsou především odjinud:

- TAUM METEO (1976) – z Montrealu, překlad meteorologických zpráv z angličtiny do francouzštiny – byla vymezena a rozumně omezena podmnožina syntaxe a sémantiky. Díky vhodné implementaci (Q-systém) šlo rozpoznat, že se neví, jak text přeložit. Což se pak udělalo ručně. První úspěšný komerční systém.
- SYSTRAN (konec 60. let). Překlad dokumentů EU, přímo (každý pár jazyků odděleně, uspokojivě jen několik málo prvních (A-F-N). Data oddělena od programu.
- EUROTRA. Mělo nahradit systran, ale každý s každým se měl dohodnout (opět přímý překlad) na analýze, dohodnout rozhraní atd. – nezvládlo se. Příliš megalomanské (72 jazykových párů).
- VERBMOBIL – Německý nástupce Eurotry, v universitním prostředí, překlad mluvené řeči,domluva obchodníků na další schůzce. V současnosti asi nežije.

Historie v ČR:

- APAČ (80. léta): Z angličtiny do češtiny,překlad z oblasti vodních pump. Využíval transdukční slovník – překladač koncovek (-ation → -ace; -ic → -ický) + slovník s asi 1500 výrazy.
- RUSLAN (1985-1990): Překlad manuálů k OS sálových počítačů – pomalý (1 věta trvala asi 4 minuty). Slovník o velikosti cca 8500 slov + transdukční slovník. Využit transfer. Gramatika zapsána pomocí Q-systémů (TAUM), navíc záchranná pravidla pro případ problémů při analýze.
- Česílko (90. léta) – lokalizace velkých SW systémů. Snaha minimalizovat podíl na překladu. Místo lokalizace z jazyka typově odlišného, bylo myšlenkou překládat z jazyka blízkého.Typ překladu – FAHQ (plně automatický, vysoce kvalitní), blízké jazyky, plné morfologické slovníky, statistická analýza češtiny. Blízké jazyky typu Čj-Sk mívají shodnou syntaxi a slovosled, jiné slovníky, avšak ne úplně, odlišné tvarosloví. Např. tedy funguje doslovný překlad slovo od slova. Myšlenkou bylo pomocí lidí přeložit ze zdrojového jazyka test do češtiny, a z češtiny strojově do polštiny, slovenštiny, ruštiny...

Dnešní situace:
Neexistují obecně použitelné systémy, přitom překlad je potřebný, určitá automatizace se hodí; je třeba spojit síly člověka a počítače.
Překlad podpořený počítačem (Computer Assisted Translation) – není strojový překlad:

- Pracuje s překladovou pamětí – text je dělen na segmenty (věty, či polygramy), které lidský překladatel překládá a systém ukládá dvojici: text v původním jazyce a překlad. Pokud se v textu časem objeví podobná, či stejná fráze, bude nabídnut uložený překlad.
- Je vhodné pro techničtější texty, pro beletrii ne tolik.
- Taktéž využívá terminologickou databázi – opět plněna ručně, sestává z konkrétních termínů zdrojového i cílového jazyka – např. carbon dioxide – oxid uhličitý (ne dioxid uhlíku apod.).

▪ Dnes se kombinuje se statistickým překladem.
◦ Strojový překlad (v základu bez intervence člověka)
▪ Pravidlový – systémy se snaží překládat slova a pomocí pravidel je k sobě skládat.
▪ Statistický
• Především se využívají korpusová data (dvojjazyčná).
Základním přístupem je generování určitého počtu možných překladů, kterým je pak přisouzena pravděpodobnost, že se jedná o překlad správný.
• Pokud jsou použity specializované korpusy, dobré výsledky jsou především v dané oblasti, není to úplně zobecnitelné.
• Jednoduchý pravděpodobnostní model – uváží se frekvence slova v trénovacích datech, což je výsledná pravděpodobnost slova.

Model zašuměného kanálu (Noisy Channel Model)

Uvažoval se model P(e|f) – udává pravděpodobnost anglické věty e, za předpokladu francouzské věty f. Model zašuměného kanálu má dvě složky:
▪ P(e) – jazykový model – např. trigramový model (z libovolných dat, ne nutně paralelní korpus).
▪ P(f|e) – překladový model – trénovaný z paralelního korpusu francouzsko-anglického.
◦ Pak P(e|f)=P(e,f)/P(f)
◦ Na základě překladového modelu byli vytvořeni určití kandidáti, pomocí P(e) – jazykového modelu, vybráno mezi nimi.
K hodnocení používán tzv. *Bleu index*, který porovnává kvalitu automatického překladu vůči lidskému překladu. Bohužel nebere v úvahu morfologii – drobnosti typu chybné koncovky, které srozumitelnosti nebrání, jsou brány stejně jako zcela špatný překlad.

Jiné dělení překladů:

- Přímý překlad – z jazyka do jazyka. Problém je, že je pro n států třeba hodně párů překladačů.
- Přes mezijazyk:
  - přes *pivotní jazyk* – pokud je třeba překládat mezi více jazyky, místo stylu „každý z každým“ se každý naučí převést text ze svého jazyka do jazyka pivotního, typicky přirozeného jazyka.
  - přes *Interlingua* – hypotetický formální logický zápis sdělení. Konstrukce obecné interlingui zatím moc neexistuje, neboť význam se těžko zapisuje (viz kapitola sémanticka). Když se interlingua používá, bývá to umělý mezijazyk, volně založený na románských jazycích. S Interlinguou se pojí tzv. Vauquoisův trojúhelník – čím vyšší patro, tím obtížnější provedení

Typicky se používá následující schéma: Text v prvním jazyce projde morfologickou, syntaktickou a sémantickou analýzou. Dále proběhne transfer do interlingui. A z ní se generuje text v novém jazyce. Některé části dělají lidé ručně (nějaké části syntaktické a hl. asi sémantické analýzy), některé se dějí strojově (např. ten samotný transfer). Opět do souvislého pěkného textu to asi upravuje typicky zase člověk.

Otázky:

1. Strojový překlad (kategorie, principy u jednotlivých kategorií + uvést příklady, metody využívající člověka).
2. Noise Channel v překladu. (Podrobně popište automatický překlad metodou zašuměného kanálu.)
3. Překladová paměť.
4. Interlingua a k čemu se používá, rozdíl mezi interlinguou a pivotním jazykem.
5. Transfer v automatickém překladě.
6. České překladové systémy.
7. Česílko.
8. Ruslan.
9. Popište Vauquoisův trojúhelník. (trojúhelník s interlinguou na vrcholu).

4 KORPUSOVÁ LINGVISTIKA

Korpus je rozsáhlý soubor textů (v digitální podobě) v daném jazyku, většinou anotovaný (značkovaný) na základě předchozí morfologické a někdy i syntaktické analýzy. Je to cenný soubor dat, ale někdy se chybně považuje za reprezentativní vzorek či rovnou celý jazyk.

Brown Corpus of Standard American English 1961 Francis, Kučera

První moderní elektronický korpus, skládal se z textů, které toho roku v Americe vyšly (v novinách, krásné literatuře apod.) Obsahoval milion slov, 15 druhů textu, dohromady 500 textů, každý cca 2000 slov. Texty byly vybírány schválně náhodně. Celé to bylo pečlivé, ale milion slov není moc. Texty nebyly anotované.

Penn Treebank 1990 Pennsylvania

První a nejznámější syntakticky anotovaný korpus. Také milion slov, asi 2500 článku. Všechno byly ale články z Wall Street Journal za poslední 3 roky, což je dosti omezující. Navíc články byly různě dlouhé. Syntaktická analýza využívala složkové systémy, tedy anotace pomocí uzávorkování a různých značek. Byla snaha ho přeložit do češtiny (PCEDT), což se podařilo, ale s obtížemi – vyžadovalo to lidi, kteří by uměli dobře česky i anglicky a vyznali se v prostředí burzovních textů z 90. let. Motivace pro překlad včetně podrobného značkování byl takový aby se nějaký statistický program mohl učit rozdíly.

Český národní korpus (CNC)

Od 1994 společně UK, MU a Ústav pro jazyk český. Morfologicky označkovaný - ne ručně, ale automatickými nástroji pro morfologickou analýzu. Současně obsahuje 500 miliónů slov a je složený z převážné části z novinových článků, dále z literatury a odborných textů.
Morfologická analýza používá výše popsané 15-ti poziční značky. V korpusu je rozeznáno 700 000 lemat, 15 miliónů slovních forem a po stochastické desambiguaci zůstane u každého slova průměrně 4,29 tagů. Používá statistické metody. Na učení se využívá ručně označkovaný korpus s 1,2 milióny tokenů (slov). K automatickému učení používá kontextová pravidla (asi 11 000 pravidel). Automaticky určuje váhy. Dosahuje rychlosti 200 tokenů za sekundu a výsledná úspěšnost je přes 94%.

Pražský závislostní korpus (Prague Dependency Treebank) 1967 Sgall

Automaticky anotovaný korpus PDT obsahuje 100 000 vět a 1,25 miliónů slov. Je anotovaný na několika rovinách:

- *Slovní 'w' rovina*: Pouze surový text bez anotace, ovšem včetně členění.
- *Morfologická 'm' rovina* : Každému slovu ve větě přiřadí několik atributů (lemma, tag (15-ti poziční značka), jednoznačné id využité při propojování rovin, odkaz do slovní roviny, atd.). Anotace probíhala dvoufázově: nejdříve anotoval automatický morfologický analyzátor → a pak dva lidští anotátoři na sobě nezávisle vybírali správná lemmata a tagy z výsledků automatického → nakonec třetí lidský anotátor vybral nejlepší možnost z předchozích dvou.
- *Analytická 'a' rovina*: Každá věta je reprezentována stromem orientovaným do kořene s ohodnocenými hranami mezi uzly. Uzly jsou právě prvky morfologické roviny, hrany jsou ohodnoceny podle závislostních vztahů uzlů, či určují další jevy (koordinace – s předchozí větou, apozice, interpunkce). Každý uzel si i pamatuje své pořadí ve větě kvůli grafickému znázornění. Byl použit automatický parser na předzpracování textu a dále automatický nástroj, který na základě pravidel určoval ohodnocení hran, ale výstup byl často chybný či neúplný, tedy museli nastoupit ruční anotátoři. Následně byly provedeny automatické kontrolní testy (např.

slovenský jmenný predikát závisí na být) a porušení byla ručně opravena. Nakonec byla provedena společná revize morfologické a analytické roviny (např. shoda v pádě, rodu a čísle závislého a nadřízeného uzlu, atd.).

- *Tektogramatická 't' rovina*: Opět je každá věta reprezentována stromem. Nicméně jeho uzly už nemusí být právě prvky morfologické analýzy (některé prvky zde nejsou (např. předložky) a některé uzly tu jsou navíc (např. nevyjádřený podmět)). Zachycuje hloubkovou strukturu věty. K některým uzlům jsou připojeny gramatémy poskytující o uzlu informaci, kterou nelze jinak odvodit. K uzlům reprezentujícím sloveso či některé typy podstatným jmen je přiřazen valenční rámec (odkaz do Vallexu). Dále nějaké koreference.

Otázky:

1. Korpusy. Charakterizovat korpusy, které jsme probírali (zdroje textu, co je v nich značkovano atd.). K čemu jsou korpusy dobré v teoretickém i aplikovaném výzkumu?
2. Brownův korpus.
3. PennTreeBank.
4. Český národní korpus (složení, velikost, typy značek).
5. Co víte o Pražském závislostním korpuse? (Tady toho chtěl trochu víc - velikost, zdroj, použité značky,...)

5 PRAVDĚPODOBNOSTNÍ A STATISTICKÉ METODY V AUTOMATICKÉM PŘEKLADU

Motivace: víme, že existují 3 překlady pro dané slovo. Je těžké určit, který je pro danou situaci vhodný. Nicméně mohl by nám k tomu pomoci kontext okolních slov. Statistické překladové metody zkoumají, jakou mají různé kombinace slov v daném jazyce pravděpodobnost – a dle toho se rozhodují o překladu. Pravděpodobnost výskytu slova w v textu T je P(w) = počet výskytů slova S v textu T / počet slov textu T.

Modelování jazyka je technika, která se snaží předpovídat, co bude následující slovo na základě předchozího kontextu. Nechť jsme před slovem w. Označme h dosavadní historii (text před slovem w). Pak nás zajímá P(w|h). Což z Bayesovy věty spočítáme jako P(w|h) = P(h|w)*P(w)/P(h). Díky větě o úplné pravděpodobnosti pak můžeme počítat pravděpodobnost celé věty W jako:

$$p(W) = p(\langle w_i \rangle_{i=1..n})$$
$$= p(w_n|\langle w_i \rangle_{i=1..n-1}) * p(w_{n-1}|\langle w_i \rangle_{i=1..n-2}) * p(w_{n-2}|\langle w_i \rangle_{i=1..n-3}) * ... * p(w_2|w_1) * p(w_1)$$

Jelikož příliš dlouhá historie by byla výpočetně náročná a zároveň by mnohé pravděpodobnosti byly příliš malé (kombinace dlouhých sousloví nejsou příliš pravděpodobná), tak se historie typicky omezuje pouze na 3 slova, což se nazývá *trigramový model*: $p(W) = p(w_3|w_2w_1) * p(w_2|w_1) * p(w_1)$

Termín *n-gram* znamená n-tice slov za sebou (lépe by však bylo upřesnit „slovní n-gram“, jindy se n-gramem totiž myslí n-tice písmen).

Vyhlazovaní – smoothing. Ve velkém slovníku je příliš mnoho nulových pravděpodobností (kombinací trigramů je hodně, ale v daných textech se jich vyskytne jen malá část). To se řeší tak, že nulové pravděpodobnosti se nahradí nějakými malými hodnotami.

Noise Channel viz kapitola strojový překlad.

Jak měřit kvalitu překladu? Bleu.

Jedná se o obtížnou záležitost i ručně, natož automaticky. V roce 2002 vznikla míra Bleu.
Pro porovnání dvou překladů vyžaduje mít daný text ještě alespoň jednou kvalitně přeložený. Následně zkoumá, zda se dané slovní n-gramy vyskytují v některých z
referenčních překladů. Čím více je překladů, tím lépe nám to řeší problém synonym nebo pouze jinak správně uspořádaného slovosledu.
$BLEU = BP*(p_1*p_2*p_3*p_4)^{1/4}$.
Problémy jsou zřejmé – jiný slovosled může způsobit velmi špatné výsledky v této metrice. Nebere v úvahu morfologii, tedy pouze chybná koncovka (ale správný význam) pokazí skóre stejně jako úplně špatný překlad. Je to hodně náročné na velikost trénovacích dat. Proto se lépe překládá mezi „velkými jazyky“, kde se data lehko shání.

Otázky:

1. Podrobně popište statistické metody v automatickém překladě.
2. Co to jsou n-gramy? (Pozor na to, že zde se mluví o slovních n-gramech, ne písmenkových.)
3. Co je vyhlazování?
4. Bleu metoda.

6 SÉMANTIKA

Sémantika přirozeného jazyka

Pomocí syntaxe můžeme rozlišovat gramaticky správné a nesprávné věty. Nicméně nic to neříká o jejich pravdivosti. Zároveň je nutno rozlišovat mezi významem a pravdivostí věty. (Naopak sémantika *formálních* jazyků tyto pojmy často ztotožňuje.) Pravdivost je dána kontextem, není obsažena v jazyce. Jsou k ní potřeba různá pravidla a předpoklady světa, ze kterého vycházíme. Navíc i nepravdivé věty mohou mít svůj význam. U některých vět zase není možno ověřit pravdivost. Obdobně je těžké obecně rozlišit věty se stejným významem: `Pozorovali ho dobrovolně` x `Byl jimi pozorován dobrovolně`.
Vyplývání – z pravdivé věty často vyplývají různé další skutečnosti (na základě obecných pravidel a zákonitostí), nicméně tyto zákonitosti nejsou stoprocentní, mohou mít výjimky, které nás předem nenapadnou: `Tučnáci jsou ptáci → Tučnáci mají křídla a létají .`

Fregeho princip kompozicionality (1925 Gottlob Frege).

Význam složeného výrazu je jednoznačně určen významy jeho částí a způsobem jejich kombinace. Tedy např. význam textu je určen významy jednotlivých vět a jejich poskládáním. Obdobně význam vět je určen významem jejich slov, atd.

Lexikální sémantika

Pro popis významu slov bychom potřebovali opět nějaký (meta)jazyk – buď formální (např. vycházející z něčeho již vystavěného, např. predikátové logiky) nebo přirozený (ten stejný nebo jiný) + se dá využívat okolní svět (Toto je křída.). Přirozený jazyk používají například výkladové slovníky,

slovníky synonym, definice slov apod.

Problémy lexikální sématiky: Význam slova závisí na kontextu okolních slov a vět (např. Střílení poslanců ohrožuje naši demokracii). Význam slov není jednoznačný (oko, list, ...).

Jednou z možností popisu významu slov jsou významové třídy (rysy).

Ontologie je množina tříd objektů, která představuje klasifikaci objektů universa U na různé třídy (např. fyzické objekty, vlastnosti, vztahy, činnosti, živé bytosti apod.), které lze dále dělit. Dané slovo (objekt) pak popíšeme pomocí příznaků ke každé třídě: + (patří do ní), - (nepatří do ní), 0 (nezávisí na ní). Ontologie jsou buď doménové (někde jsem našla, že zpracovává jen jednu doménu – obor; jinde že to je množina názvů oborů) či vrcholové (Top Ontology – prý množina nejzákladnějších výrazů, nezávislých na jazyku).

Jinou možností popisu významu slov jsou *sémantické sítě*, které umožňují určit různé vztahy a směry vztahů mezi pojmy, tj. nejen hierarchii sémantických tříd, ale i vztahy napříč nimi. Zabývají se vztahy jako *hyperonimie* (slovo nadřazené) a *hyponimie* (slovo podřazené), *synonymie* (ekvivalentní význam, ale jiná forma) a *antonymie* (slova protikladná), *meronymie* (býti částí) a *holonymie* (obsahovat). Navíc se dobře zpracovávají počítačově.

WordNet 1993 G. A. Miller

Rozsáhlá lexikální databáze anglických slov (podstatná a přídavná jména, slovesa a příslovce) seskupených do množin synonym, tzv. synonymických řad neboli synsetů. Každý synset vyjadřuje určitý koncept. Mezi synsety jsou propojení v podobě sémantických a lexikálních relací. Síť lze procházet pomocí počítače, nicméně vznikala hlavně ručně.

EuroWordNet 1997 Vossen

Rozšíření WordNetu do více jazyků (nejdříve přidány holandština, italština a španělština, později francouzština, němčina, čeština a estonština). Navíc byla zavedena vrcholová ontologie, což byla množina 63 nejzákladnějších výrazů (konceptů), nezávislých na jazyku. Ke každému jazyku pak bylo vybráno 1000 základních jazykově závislých konceptů (Base Concepts), tvořících jádra slov. V Aj WordNetu získal každý synset jednoznačný identifikátor, díky kterému vznikl mezi-jazykový index (Inter-Lingual Index, ILI). Pak byly na sebe různojazyčné WordNety navázány a vznikly vztahy ekvivalence (EQ-relations).

Aplikace

- Překlad. Jednak může pomáhat v počítačem asisotvaném překladu (Computer Aided Translation). Překladatel si v něm může hledat významy slov, jejich synonyma, antonyma, příklady použití, slova odvozená apod. Druhak společně s morfologickou a syntaktickou analýzou může díky tomu, že ukládá valenční rámce pro slovesa, sloužit k automatickému strojovému překladu.
- Extrakce informací. Např. může sloužit při vícejazyčném vyhledávání a kdekoliv kde jsou potřeba sémantické vztahy jako synonymie apod.
- Určování významů slov (Word Sense Disambiguation).
- Vyhodnocování kvality překladu – zlepšení automatických metrik typu BLEU.

Problémy

• Především je problém, že cizojazyčné WordNety vznikaly především jako překlad toho anglického. Tedy nezachycují typické vlastnosti jazyků. Také tam vzniklo mnoho chyb a nekonzistencí. A navíc pak přestaly být projekty financovány a přestaly se rozvíjet. Jiný problém je, že podobných výsledků (a lepších, rychlejších, s méně usilím) lze dnes dosahovat pomocí statistických metod. A obecně v době Googlu apod. některé projekty jako je WordNet už nemají tak dobrý smysl.

Reprezentace významu vět

Pomocí predikátové logiky 1. řádu + se přidávají nové vlastnosti. Vychází z principu kompozicionality. Složkám věty odpovídají části sémantického zápisu. Problémy začíná tvořit modalita, čas, postoj, předpoklady (presupozice), neurčitost (fuzziness), apod. Jsou tedy vytvořeny nové operátory (např. possible(F), necessary(F)). Ale tento přístup nemá dostatečnou sílu. Nastávají problémy při nahrazování částí vět.
Extenze je souhrn věcí, které pod pojem spadají. *Intenze* je samotný popis (charakteristika, definice) pojmu. Např. intenzí pojmu „čtyřúhelník“ je rovinný mnohoúhelník se čtyřmi vrcholy a čtyřmi stranami. Extenzí stejného pojmu jsou asi pojmy různoběžník, rovnoběžník = kosodélník (tedy obdélník, kosočtverec, čtverec), lichoběžník, deltoid, atd.

Základní přístupy k sémantice

◦ Modelově-teoretická sémantika . Pracuje s pravdivostními podmínkami vztaženými k určitému modelu.
Syntaktické kategorie odpovídají sémantickým typům. Obsahuje základní lexikální výrazy a jejich interpretaci, syntaktická a sémantická pravidla.
Montagueovská gramatika.
Původně Universal Grammar 1971. Založena na formální logice, lambda kalkulu, teorii množin, používá pojmy intenzionální logiky a teorie typů. Vychází z předpokladu, že neexistuje zásadní rozdíl mezi sémantikou přirozených a formálních jazyků. Obsahuje syntaktické kategorie s množinami konkrétních slov a syntaktická pravidla pro slova z těchto kategorií.
◦ Kompozicionální sémantika . Vychází z principu kompozicionality. Používá různé reprezentace (sémantické rysy a jejich skládání, koncepty a převod ze syntaktické reprezentace, logickou reprezentaci a zjišťování pravdivosti).
◦ Další příklad Transparentní intezionální logika (TIL). Je založen na typovém lambda kalkulu. Nemá vlastní logické spojky, kvantifikátory apod. Nějak řeší možné světy. Univerzum je množina společná všem možným světům. Používá nějaké nálepky individuí.

Rozpoznávání vztahů v textu

Pochopení smyslu textu je ještě těžší než smyslu věty. Problém je, že věty v textu na sebe navazují a odkazují se (např. nevyjádřeným podmětem). Jsou tyto typy vztahů v textu (resp. spíše referencí):

- *Exofora* – odkazování mimo text – zájmeno poukazuje k mimotextové situaci či skutečnostem. `Slyšíš ji? Dej mi, prosím, tyhle tři.`
- *Endofora* – odkazování v rámci textu.
  - *Anafora* – odkazování zpětně v textu. Typy anaforických vztahů:
    1. Zájmena a nevyjádřený podmět či jiný větný člen (tzv. nulové výrazy). `Jakub šel udělat čaj. To neměl dělat. On se u toho vždy opaří.`
    2. Určité jmenné skupiny. Dvojí sousloví označující to samé. `Budeme první, prohlásil Pavel Nedvěd. To mělo mužstvo rádo, když jejich kapitán takto promluvil.`
    3. Elipsa. Vynechání části věty obsahující informaci, která je příjemci známa a bez níž větu dokáže pochopit. `Jak se máš? Dobře. Kolik je? Petr přinesl dva stoly. Dřevěný a kovový.`
    4. Textové spojovací výrazy: spojky, například, na jedné straně – na druhé straně, atd.

Nejdříve si dáme jídlo, pak siestu.

- *Katafora* – odkazování dopředu v textu. Věřte tomu nebo ne, máme schodkový rozpočet. Dej mu pohlavek, tomu uličníkovi.

Aplikace: získávání informací z textu, automatický překlad, dialogové systémy.

Řešení anafory: Můžeme využít morfologické značky (shoda v rodě a pádě), syntaktickou strukturu věty (valence pomáhá doplnění elipsy), statistické přístupy, aktuální členění (odkazujeme se jinak na něco, co bylo zmíněno na začátku a uprostřed (resp. v základu a ohnisku) a pomocné znalosti (ontologie, sémantická síť, atd.).

▪ Např. Stock of Shared Knowledge. Každému podstatnému jménu přidělí určitý index podle „důležitosti“, resp. pravděpodobnosti, že se na něj pak někdo bude odkazovat.

Otázky:

1. Rozdíly mezi významem a pravdivostí věty.
2. Fregova koncepce (Fregeho princip kompozicionality).
3. Ontologie – co to je a jak se používá.
4. EuroWordNet, WordNet.
5. Rozdíl extenze / intenze v sémantice.
6. Rozdíly mezi modelově teoretickou a kompozicionální sémantikou.
7. Uveďte 4 typy anaforických vztahů v textu + příklady.