Korpusová Lingvistika

Úvod

Studie jazyka

• Studie struktury
identifikuje strukturní jednotky a třídy jazyka (např. morfémy, slova, fráze, gramatické třídy) a popisuje, jak se mohou malé jednotky kombinovat a tvořit velké

• Studie užití
jak mluvčí a pisatelé využívají prostředků svého jazyka; skutečné používání jazyka v přirozeně se vyskytujících textech – pro tyto přístupy potřebujeme velké objemy dat, nejlépe s doplněnými informacemi, které dovolí objevovat hlubší spojení mezi jednotlivými jednotkami

Co je jazykový korpus

CORPUS (13c: z latiny *corpus* tělo.)

(1) Sbírka textů, v ideálním případě kompletní a samostatná: *the corpus of Anglo-Saxon verse*.

(2) V lingvistice a lexikografii soubor textů, promluv nebo jiných vzorků považovaný za více méně reprezentativní vzorek jazyka, obvykle uložený v elektronické databázi. V současné době korpusy obsahují mnoho milionů slov běžného textu, jehož vlastnosti mohou být analyzovány ponocí *značek (tagů)* (doplněných informací identifikujících a klasifikujících slova nebo jiné útvary) a konkordačních programů. **Korpusová lingvistika** studuje data v takových korpusech.
(viz. McArthur, Tom "Corpus" , in: McArthur, Tom (ed.) 1992. *The Oxford Companion to the English Language*. Oxford, 265-266)

Komentáře ke korpusům

Jakýkoli přirozený korpus bude zkreslený. Některé věty se neobjeví, protože jsou příliš zřejmé, jiné proto, že jsou nepravdivé, ještě jiné proto, že jsou nezdvořilé. Přirozený korpus bude tak divoce zkreslený, že popis [jazyka založený na korpusu] nebude ničím jiným než prostým seznamem. (**Chomsky, Noam**. **1957**. Syntactic structures, strana 159)

Musím zmínit dvě pozorování. První je to, že si nemyslím, že by mohl existovat korpus, ať již jakkoliv veliký, který by obsahoval informace o všech oblastech anglické slovní zásoby a gramatiky, které bych chtěl zkoumat; všechny, které jsem dosud viděl, nejsou adekvátní. Druhé pozorování říká, že jakýkoliv korpus, který jsem měl příležitost zkoumat, byť sebemenší, mě naučil fakta, u nichž si nedovedu představit, že bych se o nich mohl dozvědět jakýmkoli jiným způsobem. (**Fillmore, Charles J. 1992**. "Corpus linguistics" nebo "Computer-aided armchair linguistics", v publikaci: Svartvik, Jan. (ed.) Directions in Corpus Linguistics. Berlin/New York, 35)

Charakteristika moderních korpusů

• Výběr vzorků a reprezentativnost
jazyk je nekonečný – jakýkoli korpus je konečný. To neznamená, že bychom se korpusovou lingvistikou neměli zabývat, spíš bychom měli hledat způsoby, pomocí kterých budeme konstruovat korpusy mnohem méně jednostranné a více vyvážené.

• Konečná velikost
s výjimkou tzv. **monitorovacích korpusů** (kde jsou data stále přidávána) mají korpusy pevnou velikost, která umožňuje kvantitativní výzkum

• Strojem čitelná forma
Strojově čitelné korpusy mají v porovnání s mluvenými nebo psanými výhodu ve snadném prohledávání a rychlé manipulaci a mohou být snadno doplněny o dodatečné informace.

• Standardní reference
Aby korpus mohl sloužit širšímu publiku, musí dodržovat určité standardy

Brown korpus

Brown Corpus of Standard American English
- první moderní elektronický korpus, sestaven W.N. Francisem a H. Kučerou na Brown University v Providence

• 1 milion slov textů v americké angličtině vytištěných v roce 1961

• 15 druhů textu, 500 textů, každý cca 2000 slov, v různých kategoriích různý počet textů, např.:

novinové reportáže 44 textů,
humor 9 textů,
krásná literatura 75 textů

Brown korpus – příklad textů

G01. Edward P. Lawton, "Northern Liberals and Southern Bourbons"
G02. Arthur S. Miller, "Toward a Concept of National Responsibility"
G03. Peter Wyden, "The Chances of Accidental War"
G04. Eugene Burdick, "The Invisible Aborigine"
G05. Terence O'Donnell, "Evenings at the Bridge"
G06. The American-German Review, October-November, 1961
G07. Richard B. Morris, "Seven Who Set Our Destiny"
G08. Frank Murphy, "New Southern Fiction: Urban or Agrarian?"
G09. Selma Jeanne Cohen, "Avant-Garde Choreography"
G10. Clarence Streit, "How the Civil War Kept You Sovereign"
G11. Frank Oppenheimer, "Science and Fear-- A Discussion of Some Fruits of Scientific
G12. Tom F. Driver, "Beckett by the Madeleine,"
G13. Charles Glicksberg, "Sex in Contemporary Literature"
G14. Helen Hooven Santmyer, "There Were Fences"
G15. Howard Nemerov, "Themes and Methods: The Early Stories of Thomas

PennTreebank

Penn Treebank (PTB) – první a nejznámější syntakticky anotovaný korpus

- University of Pennsylvania

- cca 1 milion slov

- obsahuje 2 499 článků ze souboru 98 732 článků Wall Street Journal (WSJ) nasbíraných v průběhu 3 let (přelom 80. a 90.let)

Autoři:

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz a Ann Taylor

PennTreebank – příklad textů

((S (NP-SBJ (NP (NNP Pierre) (NNP Vinken)) (, ,) (ADJP (NP (CD 61) (NNS years)) (JJ old)) (, ,)) (VP (MD will) (VP (VB join) (NP (DT the) (NN board)) (PP-CLR (IN as) (NP (DT a) (JJ nonexecutive) (NN director))) (NP-TMP (NNP Nov.) (CD 29))))

Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.

(. .)))((S (NP-SBJ (NNP Mr.) (NNP Vinken)) (VP (VBZ is) (NP-PRD (NP (NN chairman)) (PP (IN of) (NP (NP (NNP Elsevier) (NNP N.V.)) (, ,) (NP (DT the) (NNP Dutch) (VBG publishing) (NN group)))))) (. .)))

Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.

PennTreebank – příklad textů

HEALTHDYNE Inc., Atlanta, said its subsidiary, Home Nutritional Services Inc., registered with the Securities and Exchange Commission an initial public offering of four million shares of common. The in-home health care services provider said it will sell 1.8 million of the new shares, while Home Nutritional Services will sell the remaining 2.2 million. The company estimates the offering price at between $14 and $16 a share. The company said it expects to use the proceeds to repay certain bank debt and for general corporate purposes, including establishing new operating centers and possible acquisitions Home Nutritional currently has 10 million shares outstanding. It will have 11.8 million shares outstanding after the offering, with Healthdyne owning about 65% of the total.

Další důležité korpusy angličtiny

British National Corpus – 100 milionů slov, morfologická anotace, psaný a mluvený jazyk, http://www.natcorp.ox.ac.uk/

American National Corpus – 22 milionů slov současné americké angličtiny, morfologická anotace, http://www.anc.org/

Corpus of Contemporary American English – 410 milionů slov, 20 milionů za rok mezi lety 1990 and 2010, morfologická anotace, http://www.americancorpus.org/

Oxford English Corpus – 2 biliony slov, morfologická anotace, gramatické vztahy získané pomocí nástroje Sketch Engine, http://www.oxforddictionaries.com/page/oec
http://www.sketchengine.co.uk/

Další známé korpusy

- Negr@ Corpus (German) – syntakticky anotovaných 20000 vět
  www.coli.uni-saarland.de/projects/sfb378/negra-corpus/
- Tiger Treebank (German) – 50000 vět z novinových textů
  www.coli.uni-saarland.de/projects/tiger
- Canadian Hansard – a paralelní anglicko-francouzský korpus parlamentních debat
  www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20
- EUROPARL - European Parliament Proceedings Parallel Corpus 96-06
  Obsahuje 11 evropských jazyků: románské (Fr., It., Šp., Port.), germánské (Angl., Hol., Něm., Dán., Šv.), řečtinu a finštinu, až 44 milionů slov pro jeden jazyk.

  www.statmt.org/europarl/index.html

Český národní korpus

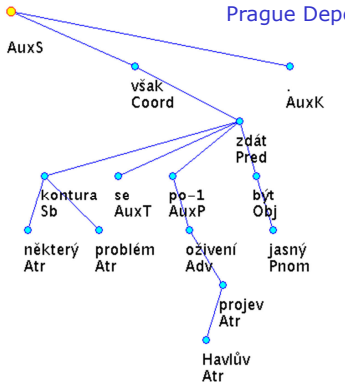
- Výsledek společného úsilí Karlovy univerzity, Masarykovy univerzity a Ústavu pro jazyk český České akademie věd
- Ústav českého národního korpusu byl založen na FF UK v roce 1994.
- ČNK je značkován na morfologické úrovni
- V současné době obsahuje asi 600 milionů slov, 100 milionů bylo uvolněno pro veřejnost jako SYN2000
- SYN2000 se skládá z:
  - 15 % literatury (11% krásná literatura)
  - 60% novinové texty
  - 25% Technické a odborné texty

Pražský závislostní korpus (PDT)

Propracované anotační schéma aplikovatelné na jazyky různých typů. Data jsou podmnožinou ČNK.

- Teoretický základ - založené na teorii Funkčního generativního popisu prof. P.Sgalla
- 100 000 vět, 1.25 milionu běžných slov
- úrovně anotace:
  - morfologie
  - analytická rovina (povrchově syntaktická)
  - tektogramatická rovina (4 podroviny):
    - závislostní struktura, (detailní) funktory
    - jádro/ohnisko (topic/focus) a hloubkový pořádek slov
    - koreference (většinou pouze gramatická)
    - vše ostatní (gramatémy):
      - detailní funktory
      - hloubkový rod, číslo, ...

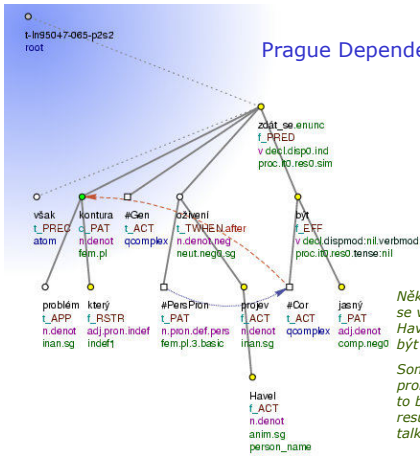
Prague Dependency Treebank



Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.

Some contours of the problem neverthelles seem to be more clear after the resurgence by the Havel's talk.

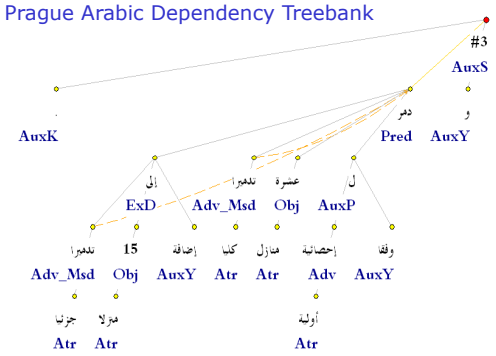
Prague Dependency Treebank



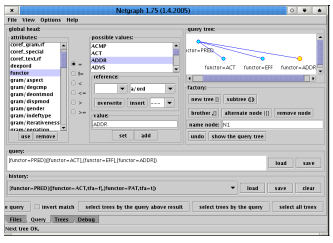
Některé kontury problému se však po oživení Havlovým projevem zdají být jasnější.

Some contours of the problem neverthelles seem to be more clear after the resurgence by the Havel's talk.

Prague Arabic Dependency Treebank



Anotační a vyhledávací nástroje



Netgraph

- prohledávání stromů
- víceuživatelská internetová aplikace client-server
- Netgraph server je napsán v C a C++ a pracuje pod Linuxem, jinými unixovými systémy a Apple Mac OS. (Experimentální verze existovala i pro MS Windows.)
- Vyžaduje kódování UTF-8.
- Klient Netgraph je napsán v Javě a je nezávislý na platformě.

Anotační a vyhledávací nástroje

TrEd – tree editor

1. Hlavní okno
2. Holá textová forma věty.
3. Stavový řádek.
4. Aktuální kontext.
5. Aktuální styl.
6. Editace stylu.
7. Tlačítko odhalující seznam vět v aktuálním souboru. Index aktuálního stromu v aktuálním souboru se zobrazuje nad tímto tlačítkem.
8. Tlačítka na otvírání, ukládání a opětovné nahrávání souboru. Ikony znamenají *Undo*, *Redo*, *Previous* a *Next File*, *Print*, *Find*, *Find Next*, *Find Previous*.
9. Tlačítka na posouvání na předchozí/následující strom v aktuálním souboru a pro správu rámců

Pravděpodobnostní a statistické metody

Pravděpodobnost vs. Relativní četnost

Příklad:
překlad předložky „in" do francouzštiny:
3 možnosti – dans, à, de
Jak zjistit pravděpodobnost překladu?

Těžko.

Potřebujeme k tomu totiž obrovské množství událostí, neboť při dostatečně dlouhé sérii pokusů se relativní četnost jednotlivých výsledků začne blížit jejich pravděpodobnosti.

Relativní četnost:

$f(E) = c(E)/N$

Odhad pravděpodobnosti

Jednoduchá pravděpodobnost výskytu konkrétního slova v textu:
- máme 2 miliony slov v textu, slovo „read" se v něm vyskytuje 720 krát
=> $p(read) \approx 720/2000000 = 0{,}00036$

Překlad „in":
Posbíráme 500 vět obsahujících překlad „in" do francouzštiny
- 250 krát dans, 150 krát à, 100 krát de
- p(dans)=250/500=0,5
- p(à) = 0,3
- p(de) = 0,2

Základní vzorce

Dva nezávislé jevy A a B:

$p(A,B)=p(A)p(B)$

Podmíněná pravděpodobnost (Bayesův vzorec) :

$p(A|B)=p(B|A)p(A)/p(B)$

Další důležité vzorce:

$p(A,B)=p(A|B)\ p(B)=p(B,A)$

$p(A,B,C)=p(A|B,C)p(B|C)p(C)$ etc.

Pokud A a B jsou nezávislé, pak $p(A|B)=p(A)$ a tedy

$p(A,B)=p(A)p(B)$

Modelování jazyka

Hlavní úkol: Předpovědět následující slovo v běžném textu nebo promluvě.

Jak?

Pomocí podmíněné pravděpodobnosti na základě kontextu (historie) předpovídáme následující slovní tvar

$p(w|h)$ w – předpovídané slovo, h – historie, vše, co bylo dosud řečeno (napsáno)

Cíl: spočítat pravděpodobnost celé věty:

$p(W)=p(<w_i>_{i=1..n})$

N-gramy

$p(W)=p(<w_i>_{i=1..n})=p(w_n|<w_i>_{i=1..n-1})* p(w_{n-1}|<w_i>_{i=1..n-2})*$
$p(w_{n-2}|<w_i>_{i=1..n-3})*...* p(w_2|w_1)*p(w_1)$

Problém:

příliš dlouhá historie znamená nedostatek (řídkost) dat a obrovské nároky na výpočetní kapacitu => je nutné historii včas „useknout"

n=3 trigramový model

$p(W)= p(w_3|w_2w_1)*p(w_2|w_1)* p(w_1)$

Kratší jsou bigramy (n=2), unigramy (n=1)

Vyhlazování

Problémem je velikost dat

Máme-li slovník (V) o 40000 slovech =>

-$|V|=40k$, velikost modelu = $|V|^3 = 6{,}4 \times 10^{13}$

- typická velikost trénovacích dat – stamiliony (10^8) slov

- příliš mnoho nulových pravděpodobností (nenulová pouze jedna ze 100000 !) – řídká data

- některé z nich ale zastupují existující kombinace

- pokus o řešení – nahradit nulovou pravděpodobnost nějakou velmi malou hodnotou

Stručný úvod do statistického překladu

- Základní myšlenka – použít paralelní korpus jako trénovací množinu příkladů dobrého překladu
- Paralelní korpusy existují jak pro dvojice jazyků, tak i pro větší množiny (korpus dokumentů EU)
- Klasický příklad – kanadský Hansard (A-F, trénovací data pro pokusy IBM v 90.letech, má 1.7 milionu vět)
- Inspiraci můžeme vysledovat až k Warenu Weawerovi, který v roce 1949 navrhoval použití kryptoanalytických metod na překlad
- Díky soustavnému růstu výpočetní síly a dostupnosti paralelních dat se statistický překlad stal v posledním desetiletí převládající metodou

Metoda zašuměného kanálu

- Příklad – chceme překládat z F do A
- Hledáme pravděpodobnostní model P(A|F), který vyjádří podmíněnou pravděpodobnost libovolné anglické věty *a*, máme-li francouzskou větu *f* . Tréninkový korpus nám pomůže nastavit parametry
- Bayesův vzorec: $P(a|f) = P(f|a)P(a)/P(f)$
- Tím jsme „obrátili" směr překladu, budeme hledat dva modely: překladový model P(f|a)
  jazykový model cílového jazyka P(a)
- Proč zašuměný kanál? Protože předstíráme, že jsme obdrželi větu „pokaženou" přenosem přes nespolehlivý kanál a hledáme její správný originál.

Metoda zašuměného kanálu

- Jazykový model P(A) může být trigramový model založený na mnohem rozsáhlejším korpusu cílového jazyka, řádově stamiliony slov
- Překladový model P(F|A) je založen na mnohem menším paralelním korpusu (miliony slov)
- Důležité:
  - překlad probíhá obráceně
  - jazykový model odfiltruje nepodařené překlady, vyrovná chyby překladového modelu
  - Jazykový model vybírá pouze „hezké věty," nemá vztah k originálu
  - Hledání překladových hypotéz (dekódování) je obtížným problémem samo o sobě

Příklad (Koehn and Knight)

Překlad ze španělštiny do angličtiny, možné překlady jsou založeny pouze na P(Š | E) :

Que hambre tengo yo!

What hunger have P(Š | E) = 0.000014

Hungry I am so P(Š | E) = 0.000001

I am so hungry P(Š | E) = 0.0000015

Have i that hunger P(Š | E) = 0.000020

: : :

Přidáme jazykový model P(E):

Que hambre tengo yo!

What hunger have P(S | E)P(E) = 0.000014 x 0.000001
Hungry I am so P(S | E)P(E) = 0.000001 x 0.0000014
I am so hungry P(S | E)P(E) = 0.0000015 x 0.0001
Have i that hunger P(S | E)P(E) = 0.000020 x 0.00000098
: : :

Evaluace systémů automatického překladu

metrika BLEU (Papineni, Roukos,Ward and Zhu, 2002):

Kandidát 1: It is a guide to action which ensures that the military always **obeys** the commands of the party.
Kandidát 2: It is to **insure the troops** forever **hearing** the **activity guidebook** that party **direct**.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.
Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Reference 3: It is the practical guide for the army always to heed the directions of the party.

Přesnost překladu v n-gramech

Unigramová přesnost a kandidátů na správný překlad: **C/N**
kde N je počet slov v daném kandidátu, C je počet slov z tohoto kandidáta, které se vyskytují alespoň v jednom referenčním překladu.
V našem příkladu tedy:
Kandidát 1: 17/18
(pouze *obeys* chybí ve všech referenčních překladech)
Kandidát 2: 7/18

Modifikovaná N-gramová přesnost
Unigramovou přesnost je možné zobecnit na další n-gramy, postupně bereme dvojice, trojice a čtveřice.
Například pro kandidáty 1 a 2 platí bigramová přesnost:
Přesnost$_1$(bigram) =10/17
Přesnost$_2$(bigram) =1/13

BLEU - Celkové skóre

Penalizace za stručnost (Brevity penalty)
Míra založená pouze na počtu správných n-gramů by měla tendenci favorizovat krátké věty, jejichž všechny n-gramy by existovaly v referenčních překladech, přestože by referenční překlady byly podstatně delší – např. při překladu pouze části originálu. Proto je nutné výsledek upravit o číslo, které zachycuje rozdíl v délkách a penalizuje jej

Celkové skóre je tedy:

$$BLEU = BP * (p_1 p_2 p_3 p_4)^{1/4}$$

Tedy: Penalizace za stručnost vynásobená geometrickým průměrem n-gramové přesnosti pro n=1..4
Výsledkem je vždy číslo mezi 0 a 1