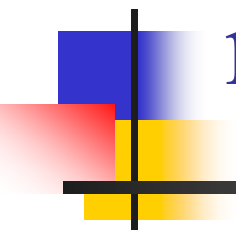
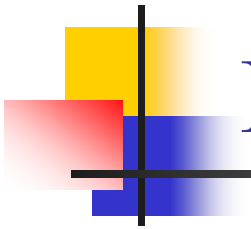


Využití neuronových sítí pro morfologické značkování češtiny

Petr Němec, ACM 2004

Co je morfologické značkování ?

- slovo (českého) jazyka může být morfologicky víceznačné
- **cíl**: určit správné hodnoty morfologických kategorií pro slova vstupního textu



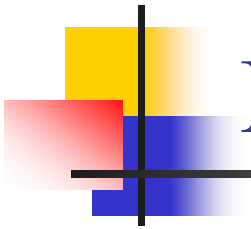
K čemu je to dobré ?

- redukce počtu nejednoznačností (ambiguit), které výrazně snižují úspěšnost syntaktické analýzy
- velmi důležité pro další úlohy zpracování přirozeného jazyka (rozpoznávání mluvené řeči, automatický překlad, kontrola gramatiky)

Dosavadní výsledky

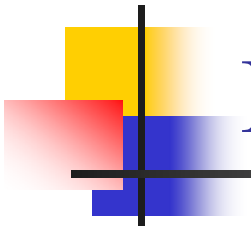
- statistické přístupy
  - na části trénovacích dat Pražského závislostního korpusu **93,47%**
  - výsledky trénování na celém korpusu nejsou k dispozici

- neuronové sítě použity v ojedinělých experimentech pro angličtinu, úspěšnost srovnatelná se statistikou



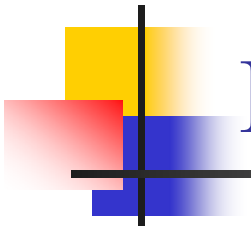
Má smysl ještě experimentovat ?

- česká věta obsahuje průměrně 16 slov
- i při značkovací úspěšnosti 95% je pravděpodobnost, že bude celá označkována správně < 50%
- nevíme, kde chyba je
- navíc je úspěšnost měřená pouze u nejednoznačných slov < 80%



Podoba vstupních dat

- konkrétní konfigurace morfologických kategorií daného slova je vyjádřena poziční značkou (např. **NNFP1-----A----**)
- morfologická analýza určí předem pro každý slovní tvar seznam možných značek – značkovač (tagger) pak vybírá správnou značku pouze z tohoto seznamu

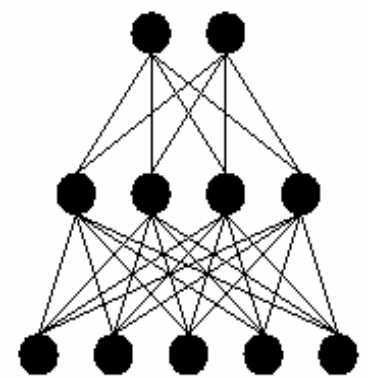


Datové zdroje

- trénovací data Pražského závislostního korpusu (PDT) čítající cca 1,5 mil. slov a obsahující
  - správnou značku
  - výsledek morfologické analýzy
  - výsledky statistických značkovačů
- testovací data (2 x cca 100 000 slov)

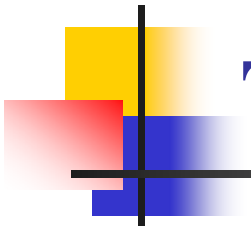
Neuronová síť zpětného šíření(1)

- množina propojených vrstev matematických neuronů (vstupní, skryté, výstupní)
- neurony vstupní/výstupní vrstvy ~ složky vektorů



Neuronová síť zpětného šíření(2)

- trénovací množina dvojic vektorů (**I**,**O**)
- síť při učení minimalizuje celkovou odchylku svých výstupů od požadovaných výstupů **O** pro všechny **I**
- **cíl**: generalizace (síť vrací správný výstup i pro netrénované instance)



Testovací prostředí

- testovací prostředí implementováno v C++ (MVS 6.0)
- automatizace experimentů
- vlastní optimalizovaná implementace učících algoritmů sítě zpětného šíření (práce s momentem, super-sab)

Předběžné experimenty: „spolehlivý“ kontext

- správnou značku určujeme na základě *n* předcházejících **správných** značek a sufixu délky *m*
- vstupní vektor = kód těchto komponent
- výstupní vektor = kód správné značky

„Spolehlivý“ kontext - příklad

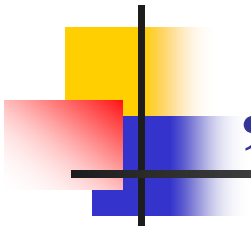
Prezident (`NNMS1-----A----`) *dnes* (`Db-------------`)
rezignoval (`VpYS---XR-AA---`) . (`Z:-------------`)

$n = 1$, $m = 2$, určujeme kategorii *číslo*:

C_t(`NNMS1-----A----`), C_t(`Db-------------`), C_t(`VpYS---XR-AA---`), C_t(`Z:-------------`) → $C_l(S)$

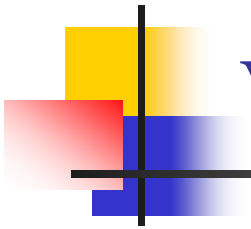
„Spolehlivý“ kontext - výsledky

- úspěšnost **89,22%**
  - rod: 94,51%
  - číslo: 97,12%
  - pád: 94,46%
- nalezení optimálního kódování
- určení optimálních parametrů
  - kontextu
  - sítě



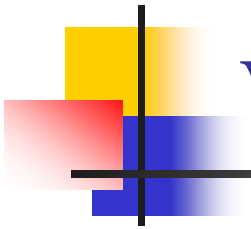
„Statistický“ kontext

- reprezentace stejná jako v předchozím experimentu, kontext je však tvořen značkami určenými předem statistikou
- přímo aplikovatelné
- úspěšnost **88,71%**



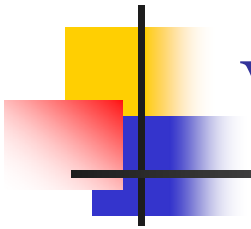
Voting – o co jde ?

- rozhodování mezi výstupy dvou statistických metod:
  - Feature-based tagger (**92,74**%)
  - Markov model tagger (**92,58**%)



Voting – vektorová reprezentace

- vstup
  - levý kontext + sufix (jako dříve)
  - 2 kandidátské značky – výstupy taggerů
  - pravý kontext
- výstup
  - dvousložkový vektor - složky určují, zda odpovídající tagger určil správnou značku

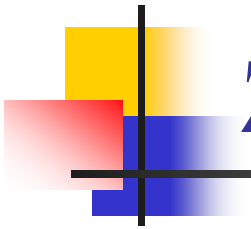


Voting – výsledky(1)

- optimální dosažená hodnota **93,56%**, podstatně lepší než vstupní statistika
- **více než nejlepší statistický výsledek** na PDT (93,47%)

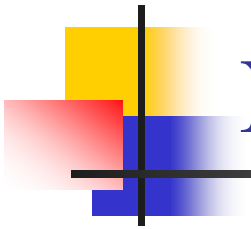
Voting – výsledky (2)

- cca 25% narůst úspěšnosti oproti náhodné „základní čáře“ (92,69%) v poměru k ideálnímu rozhodovacímu stroji (95,78%)
- distribuce chyb v jednotlivých kategoriích odpovídá distribuci chyb vstupních značkovačů



Závěr

- síť zpětného šíření dosáhla absolutně nejvyšší úspěšnosti na dané trénovací množině (použila přitom však statistický výstup)
- spojení metod se jeví jako velmi slibné



Náměty k další činnosti

- využití statistických značkovačů natrénovaných na celém trénovacím korpusu PDT
- použití rekurentních neuronových sítí pro morfologické značkování
- použití neuronové sítě pro určování optimálních parametrů kontextu