# Hloubková automatická analýza angličtiny

## Diplomová práce

Autor: Ondřej Dušek
Vedoucí práce: Prof. RNDr. Jan Hajič, Dr.

Ústav formální a aplikované lingvistiky
MFF UK

6. září 2010

# Úvod

## Automatická analýza významu

- Popis významu závislostními stromy vs. propozicemi
- Identifikace propozic:
  - Identifikace a disambiguace predikátů
  - Identifikace a klasifikace argumentů
- Použití strojového učení

## Cíl práce

Navhrnout a implementovat systém pro automatickou sémantickou analýzu v angličtině, otestovat na existujících datech (CoNLL 2009).

# Použitá data

## CoNLL 2009 Shared Task

- Data pro 7 jazyků, vč. angličtiny, v jednotném formátu
- Disambiguace predikátů, identifikace a klasifikace argumentů
- Jednotná evaluace
- Popisy soutěžních systémů a výsledná data jsou zveřejněná

## Anglický korpus

- Penn Treebank – morfologická a syntaktická anotace
- PropBank / NomBank – sémantické propozice pro slovesa a substantiva

# Sémantické značkování dat CoNLL 2009 – příklad

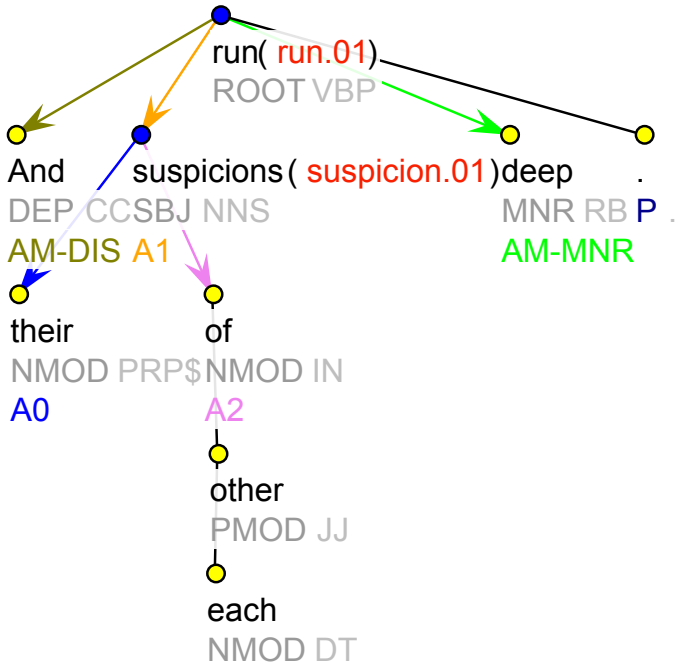
## Disambiguace predikátů

run.01 operate, procede *(John runs the company.)*

run.02 walk quickly
*(John ran the Boston Marathon.)*

run.03 cost *(This watch runs $30.)*

run.04 range, extend
*(Shelves ran from floor to ceiling.)*




And their suspicions of each other run deep.

## Identifikace a klasifikace argumentů

„run.01“ – valenční argumenty:

A0 operator

A1 machine, operation, procedure

A2 employer

A3 coworker

A4 instrumental

...

# Běhové prostředí pro experimenty

## Cíle

- Snadná konfigurovatelnost
- Dávkové zpracování úkolů, závislosti
- Paralelní provádění experimentů

## Implementace

- Integrace existujících nástrojů (WEKA, LP_Solve)
- Modularita (obecný interface a nezávislost jednotlivých úkolů)
- Dávkové zpracování – expanze úkolů
  - Wildcards, kombinace pouze odpovídajících si souborů

# Základní techniky (1)

## Klasifikátory

- Logistická regrese
- Support Vector Machine

## Konverze dat

- „Zploštení“ formátu – rozdělení lemmat / predikátů
  - Zvětšení objemu dat
- Rysy z anotace + generované rysy
- Filtrování
- Výběr rysů
  - Hodnocení rysů (6 kritérií) a lineární přidávání
  - Hladový algoritmus

# Základní techniky (2)

## Generované rysy

- Implementace dříve popsaných rysů
  - Syntaktické – sourozenci, děti, rodič
  - Morfologické, topologické
- Nové rysy
  - *Children Types* – děti v závislosti na morfologii
  - *Clusters* – sémantický clustering, implicitní informace (syntakticky / morfologicky vztažené k danému slovu)

## Evaluace

- Využití existující evaluace CoNLL 2009
- Experimenty: standardní metriky, bootstrapový test

# Implementace systému

## Disambiguace predikátů

- Rozdělení dat podle lemmat (nezávislé významy)
- Různé techniky vyběru rysů podle počtu významů lemmatu (kompromis rychlosti výpočtu a kvality výsledku)

## Klasifikace argumentů

- Spojení identifikace a klasifikace argumentů
- Oddělení klasifikace pro valenční argumenty / adverbiální doplnění
- „Post-Inference“ – unikátnost valenčních argumentů (dvě varianty)

# Výsledky systému na datech CoNLL 2009

## Celkové výsledky

| | | |
|---|---|---|
| Disambiguace predikátů | 95.06 % | accuracy |
| Klasifikace argumentů | 64.42 % | labeled $F_1$ |
| Celkem | 75.00 % | labeled $F_1$ |

Pro přímočarý algoritmus post-inference.

## Srovnání s ostatními systémy ze soutěže CoNLL 2009

- Celkový výsledek – 15./21
- Velmi dobrá disambiguace predikátů (4.)
- Průměrná kvalita klasifikace argumentů

# Analýza výsledků

## Podrobný pohled

- Disambiguace predikátů – *Children Types* i *Clusters* často vybírané
- Klasifikace argumentů – precision vs. recall
- Problém s adverbiálními doplněními u substantiv

## Možná zlepšení

- Další generované rysy
- Podrobnější filtrování / výběr rysů
- Oddělení klasifikace adverbiálních doplnění pro substantiva a slovesa
- Speciální řešení koreferencí

# Následný experiment: Aplikace na češtinu

- Data pro češtinu z PDT 2.0
  - Větší objem dat – další slovní druhy, více významů
- Úpravy pro jiné značení lemmat a významů predikátů
- Úprava pro složitější morfologické značky
- Nutné úpravy generovaných rysů – na základě morfologie
- Predicate disambiguation: 94.89 % accuracy