Automatická extrakce lexikálně-syntaktických údajů z korpusu

Diplomová práce

Ondřej Bojar

Motivace

- Vzdálené cíle počítačové lingvistiky:
  - Strojový překlad, kontrola gramatiky, souhrny textů. . .
- Tradiční kroky při zpracování vět přirozeného jazyka:
  - Morfologická analýza ($\sim$ hotova).
  - Syntaktická analýza (obtížná).

    Hlubší analýzy (obtížnější, přesahuje rámec této prezentace).
- Syntakt. analýza vyžaduje množství slovníkových informací.
  - Ruční výroba slovníků nákladná, extrahujme údaje z textů.
  - Slovesa “organizují” větnou strukturu nejvýznamněji, začněme s nimi.
  - Charakterizujme slovesa seznamem typických doplnění.

Dostupné zdroje dat

- Pražský závislostní korpus (PDT)
  - Syntaktická anotace (ruční).
  - ∼1,5 milionu slov, ∼98 tisíc vět.
  - Doplnění slovesa zřejmá ze struktury věty.
  - 2,8 tisíce různých sloves

    z toho jen 54 více než 50x a 235 více než 20x
- Český národní korpus (ČNK)
  - Morfologická anotace (automaticky zjednoznačněná).
  - ∼100 milionů slov, ∼1,8 milionu vět.
  - Doplnění slovesa lze extrahovat jen s obtížemi.
  - ∼22 tisíc různých sloves

    z toho jen ∼9 500 více než 50x a ∼12 000 více než 20x

Jak získat více anotovaných dat

- Jednoduchý postup:
  - (Získej více textů; např. z Internetu.)
  - Morfologicky analyzuj (a zjednoznačni) věty.
  - Použij jeden z dostupných parserů češtiny pro získání syntaktické struktury.
  - Extrahuj požadované lexikálně-syntaktické údaje.
- Přesnost současných parserů není dostatečná:
  - Nejvýše $\sim$83 % správně zapojených slov.
  - Nejvýše $\sim$55 % sloves se správnými rozvitími.

Jak získat lépe anotovaná data

- Parsery nejsou obecně dost přesné, proto:
  - Morfologicky anotuj (a zjednoznačni) věty.
  - Vyber “jednodušší” věty, snadněji analyzovatelné.
  - Použij jeden z dostupných parserů. . .
- Filtrace musí být *lingvisticky motivovaná*.
  - Ponech věty obsahující sledovaný jev.
  - Zamítni všechny věty, kde je jev “skrytý” nebo interferuje s dalšími složitými jevy.
  - (Částečná syntaktická analýza nutná pro některá rozhodnutí.)
- Filtrace závisí na cíli $\Rightarrow$ šikovný by byl “skriptovací” jazyk pro přípravu různých filtrů $\Rightarrow$ systém AX.

Systém AX

- Cíle navrženého jazyka:
  - Pohodlná formulace filtrů pro zamítání vět.
  - Pohodlná formulace částečné syntaktické analýzy vět.
- Celkové schéma:
  - Vstupní věta = posloupnost sestav rysů (feature structures).

    1 vstupní slovo = 1 sestava s (víceznačnou) morfologickou informací.
  - Skript je řada střídavých *filtrací* a *přepisovacích pravidel*.
  - Skupiny pravidel upravují vstupní posloupnosti sestav rysů, generují nová “čtení” věty.

    Není omezeno na CFG či CSG. Povolena nedeterministická pravidla.
  - Filtry zamítnou nepřijatelná čtení věty.

Ukázkové filtry a pravidla přepisů

```
filter "Zamítni věty obsahující dvě slovesa nebo spojku:"
  ^ .* [cat-verb] .* [cat-verb] .* $
  |
  ^ .* [cat-conj] .* $
end

     # Najdi celou jmennou skupinu a nahrad' ji jediným symbolem
rule nounph ---> (adv{0,2} adj)* noun ::
  adj = [cat-adj], noun = [cat-noun], adv = [cat-adv],
     # Příd. jm. a podst. jm. se ovšem musí shodovat:
  adj.case = noun.case,  # v pádě,
  adj.num = noun.num,    # v čísle,
  adj.gend = noun.gend,  # ve jmenném rodě.
  nounph = noun,         # Naplň výstupní sestavy rysů.
end
```

Extrakce doplnění sloves

- Fáze filtrace:
  - Ponechme jen jednoduchá souvětí, bez typicky problematických jevů.
  - Navíc v jedné větě nesmějí být dvě a více plnovýznamových sloves.
  - $\Rightarrow$ “velmi jednoduché věty” ($\sim$15–20 % vět z korpusů)
- Na vybraných větách se parsery zlepšily proti původní úspěšnosti:
  - o 5–10 % podle počtu správně zapojených slov.
  - o 10–15 % podle počtu sloves se správnými rozvítími.
  - o 15–20 % ($\sim$dvakrát) podle počtu vět zcela bez chyby.

Shrnutí

- Potřebujeme syntaktické slovníky.
- Závislostní korpusy jsou vynikající zdroj, ale:
  - jejich příprava je nákladná.
  - neposkytují dostatek dat (řada slov málo četná).
- Použijme “nestromečkové” korpusy či libovolný text:

  + dostupné parsery pro získání syntaktické struktury.
  - ale napřed vyberme věty, kde je sledovaný jev dobře pozorovatelný a které je snazší syntakticky rozebrat.
- AX je systém umožňující potřebnou filtraci.

  Ve skutečnosti silnější; umožňuje částečnou i úplnou syntaktickou analýzu.
- Byl použit konkrétně pro extrakci doplnění sloves.