NLP & strojové učení

Miloslav Konopík

2. dubna 2013

Co je to NLP?

NLP = Natural Language Processing (zpracování přirozeného jazyka)
Computational linguistic (komputační lingvistika)

Aplikační oblasti

- Vyhledávání textů (Google).
- Strojový překlad (IBM word model)
- Podpora marketingu (analýza sentimentu).
- Podpora PR (třídění e-mailů).
- Podpora rozpoznávání (řeči, skenovaných textů).
- Oprava pravopisu.
- Odpovídání otázek (SIRI, IBM Watson).
- další...

Z čeho s NLP skládá?

- Tokenizace (rozdělení textu na slova, věty, dokumenty).
- Normalizace (velikosti písmen, čísel, dat, ...).
- Jazykové modelování (určování posloupnosti tokenů).
- Morfologie slov (stemming, lematizace, morfologické značkování).
- Parsování vět (syntaktické / sémantické).
- Rozpoznávání pojmenovaných entit.
- Hledání kolokací.
- Klasifikace dokumentů (do jedné/více tříd).
- Analýza sentimentu (aspektová, sociálních médií).

Z čeho s NLP skládá?

- Sumarizace textů.
- Oprava pravopisu.
- Odpovídání otázek.
- Strojový překlad.
- Vyhledávání a získávání informací.
- (Stahování webů (získání HTML, parsování čištění)).

Tokenizace

Rozdělení textu na:

- **slova**,
- **věty, resp. souvětí**,
- **dokumenty**,
- odstavce,
- věty souvětí,
- slabiky,
- další jednotky.

Tokenizace

Složitější případy.

在北京，如果迷失方向，
完全不必着急。
北京是个大城市，
北京人对外国人都很热情。

- zkratky (s.r.o.),
- data, hodiny (12. března 2013, 12:30, 3.3), ...

Normalizace

Převod tokenů (slov) na standardní tvar. Snížení počtu jednotek.

Příklady

- **Lower casing**: první systém pro strojový překlad byl představen 7. ledna 1954 v ústředí firmy imb .
- **True casing**: první systém pro strojový překlad byl představen 7. ledna 1954 v ústředí firmy IBM .
- DATE(DAY=12, MONTH=3, YEAR=2013)

Jazykové modelování

Spadá do NLP, dle 5. přednášky.

Rozšíření

- Metody určování tříd – využití sémantiky slov.
- Topic modely.

Morfologie slov

Lexém: Jeden záznam ve slovníku.
Lemma: Normovaný tvar lexému.
Stem: Kořen slova. Většinou však bez lingvistického významu.

Lemmatizace: Hledání lemmat. OOV slova.
Stemming: Hledání "kořenů".

Morfologická analýza: Nalezení (všech) morfologických příznaků ke slovům.
Morfologické značkování: Nalezení kontextově správně posloupnosti morfologických příznaků slov.

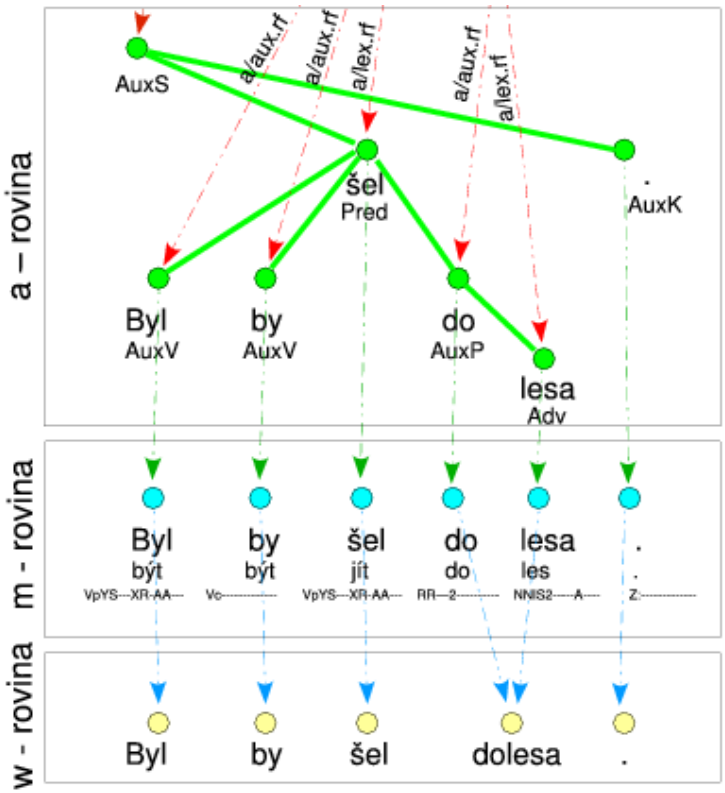
Morfologie slov - ukázka

Věta: “Září ve vědě.”
Zdroj: PDT 2.0.

```
<m id="m-vesm9211-031-p1s1w1">
      <src.rf>manual</src.rf>
      <w.rf>w#w-vesm9211-031-p1s1w1</w.rf>
      <form>Září</form>
      <lemma>září</lemma>
      <tag>NNNS1-----A----</tag>
</m>
```

Stem: “zář”.

Parsování vět



Zdroj: Manuál PDT 2.0.

Rozpoznávání pojmenovaných entit

NER - Určení významu slov a slovních spojení, která mají určitý předem definovaný význam. Například:

- Jména osobností.
- Názvy měst.
- Názvy států.
- Názvy společností.
- Data.
- Čísla.
- ...

Datum Společnost

První systém pro strojový překlad byl představen 7. ledna 1954 v ústředí firmy IBM .

Hledání kolokací

- Kolokace.
- Idiomy.

Například: “Strojové účení”

Klasifikace dokumentů

Do:

- jedné,
- více

tříd.
Například: určení tématu, detekce spamu, třídění e-mailů.

Přístupy

- Pravidlový přístup.
- Strojové učení.
- Kombinace

Strojové učení

Trénovací data - učení vztahů, závislostí, pravidel. Zobecňování.

- S učitelem (supervised).
- Bez učitele (unsupervised).
- Částečné učení s učitelem (semi-supervised).