Rozpoznání pojmenovaných entit v textu

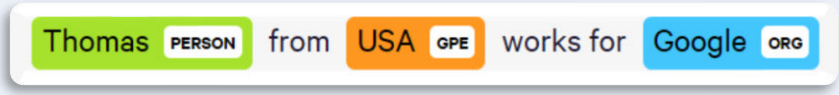
Autor: Martin Süss
Vedoucí práce: doc. Ing. František Dařena, Ph.D.
Ústav informatiky, Provozně ekonomická fakulta, Mendelova univerzita v Brně

Mendelova univerzita v Brně

MENDELU Provozně ekonomická fakulta

Rozpoznání pojmenovaných entit

- Jedna ze **základních úloh** v oblasti **zpracování přirozeného jazyka.**
- Hledání **význačných** slov či slovních spojení v textu.
- Osoby, organizace, lokality, produkty, události, …
- Lze řešit několika způsoby – seznamy entit, pravidla, strojové učení.



Vektorová reprezentace slov

- Množina technik přiřazující všem unikátním slovům v textu mnohodimenzionální číselný vektor.
- Vektory mohou nést zakódované **informace o vztazích mezi slovy** (sémantická či syntaktická podobnost) při dostatečně rozsáhlém textu.
- **Známé techniky**: Word2Vec, GloVe, FastText.
- Pomocí algebraických operací s vektory lze získat užitečné informace:

vektor(*„král“*) – vektor(*„muž“*) + vektor (*„žena“*) ≈ vektor(*„královna“*)

Motivace a cíl práce

- Moderní systémy pro rozpoznání pojmenovaných entit využívají **neuronové sítě** společně s **vektorovou reprezentací slov.**
- **Vektory** slov **různě kvalitní** (dle použitého algoritmu a konkrétního nastavení dimenze, velikosti kontextového okna, textového korpusu).
- Autoři systémů detailněji **nezkoumají kvalitu vektorů**, nýbrž pouze až konečnou kvalitu rozpoznávání.
- Cílem práce je prozkoumat **dopady kvality vektorů** na **výslednou kvalitu rozpoznávání** a přinést řadu **poznatků** a **doporučení.**

Návrh systému pro rozpoznávání

- **Rekurentní** neuronová síť s **obousměrnými LSTM** buňkami.
- Hyper-parametry sítě získány empiricky z provedeného experimentu.
- Vstup = **anotovaná trénovací data** (vstup → výstup) a **vektory slov.**
- Realizace v **Python 3** (balíčky Gensim, TensorFlow, MorphoDita, …).
- **Postup**: čištění a předzpracování textových dat ⇒ tvorba vektorů slov z korpusu ⇒ trénování neuronové sítě ⇒ testování systému (metriky precision, recall, F1) ⇒ uložení modelu pro budoucí využití.

Výsledky

- Provedena řada experimentů s **faktory ovlivňující kvalitu vektorů.**
- Zkoumána kvalita a kvantita korpusu, dimenze vektorů, techniky předzpracování textu, specifické algoritmy a jejich nastavení.

Systémy pro rozpoznávání	F1-míra [%]
Bez modelu vnoření slov	49,83
Nejlepší zjištěné nastavení	68,16
Demir a Ozgür (2014)	64,72
Straková a kol. (2016)	63,91

Výstup práce a její využití

- Sada poznatků a doporučení aplikovatelná při tvorbě vektorů slov:
  - Potřebná **kvalita a kvantita** textových datových množin.
  - Správné nastavení **dimenzí** vektorů.
  - Vhodné **techniky předzpracování** textu v **různých korpusech**.
  - Výběr vhodného **algoritmu** včetně jeho **nastavení.**
- Možnost vylepšení existujících rozpoznávačů pojmenovaných entit za pomocí zvýšení kvality vektorů, nikoliv změnou architektury algoritmu.