Posudek oponenta diplomové práce

Autor práce: **Lukáš Witz**

Název práce: **Rozpoznávání názvů značek v sociálních médiích**

Práce se zabývala rozpoznávání pojmenovaných entit, konkrétně detekcí názvů značek v diskuzích na Webu.

Po velmi obecném úvodu a kapitole o strojovém učení se práce dostává k detailně popsanému rozpoznávání pojmenovaných entit (NER). Následuje popis dat a realizační část. V některých místech byla práce šitá horkou jehlou. I když je realizační část rozsahem kratší, ukazuje zajímavé porovnání NER na různých typech textu.

Komentáře k textu:

- Z popisu SVM není zřejmé, jak by se klasifikátor dal použít na úlohu rozpoznávání entit.
- Vzorce: popis CRF (str. 16): index t ve vzorci =p(y|x), P místo TP ve vzorcích na straně 17.
- Některé anglické pojmy by bylo asi raději lepší nepřekládat („pokutová funkce", „podpůrné vektory", „Markovův model"-> „Markovský model").
- Odstavce na sebe ne vždy ideálně navazují (např. začátek odstavce na straně 11).
- 3.2.6 – zasahuje příliš daleko, až do detekce referencí na entity, což je samostatná úloha.
- Některým úlohám je věnována až přílišná pozornost (např stemming).
- Strukturování: od strany 28 popisovány algoritmy pro stemming a lemmatizaci, následuje pak krátká kapitola o lematizaci.
- Formulace: úvod kapitoly 4 nebo poslední věta prvního odstavce na straně 35, „data 1 oproti 2" (str 35), „příslušník entity" (str. 36).
- Část 4.1.2 popisuje anotační nástroj bez reference.
- Nenašel jsem popis formátu značek, strana 33.
- Singletony v UML (str.64).

Formální stránka:

- Občas příliš dlouhé věty (např. poslední v odstavci 3.1.),
- Velká/malá písmena na začátku vět,
- P(.) na straně 11,
- představení zkratek (např. ME) nebo
- odkazy (např. na obrázek 3.1).

Otázky:

1. Jak velké jsou vytvořené korpusy (např. počet výskytů entit, počet jedinečných entit)?
2. Dokážete odhadnout úspěšnost v případě použití celé hierarchie tříd (příloha B)? Stačila by trénovací data?
3. Jak byly vyřešeny zkratky při tokentizaci (5.3.1)?
4. Rozdíl oproti práci [21]: která nastavení nejvíce ovlivnila úspěšnost?

Programové vybavení je funkční, kvalitní a dobře okomentované. Experimenty jsou v práci dobře popsány. Zadání bylo splněno.

Navrhuji hodnocení známkou **velmi dobře** a práci doporučuji k obhajobě.

V Plzni 15. 8. 2016

doc. Ing. Josef Steinberger, Ph.D.

SOUHLASÍ S ORIGINÁLEM

Západočeská univerzita v Plzni
Fakulta aplikovaných věd
katedra informatiky a výpočetní techniky