Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Bc. Richard Ejem
Název práce	Relation Extraction in Police Records
Rok odevzdání	2017
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	doc. Ing. Zdeněk Žabokrtský, Ph.D. **Role** Vedoucí
Pracoviště	ÚFAL

Text posudku:

Obsah práce:

Cílem posuzované práce bylo navrhnout a implementovat systém, který bude schopen v doméně textových záznamů Policie ČR rozpoznávat relace mezi entitami v textu.

Práce je členěna následovně. Po úvodní kapitole následuje kapitola popisující vybrané existující metody pro extrakci relací z textu. Třetí kapitola představuje datové zdroje, které by bylo možné k řešení úlohy využít. Čtvrtá kapitola stručně popisuje softwarový rámec, ve kterém budou experimenty realizovány. Pátá kapitola dokumentuje technické aspekty nutného předzpracování policejních dat. V šesté kapitole jsou představeny evaluační metriky a několik různých přístupů k rozpoznávání relací. Sedmá kapitola kriticky rozebírá problém stávajících anotovaných dat. Jelikož nekonzistence přítomné v trénovacích datech jsou značné a zásadně komplikují využitelnost řízených metod strojového učení, osmá kapitola překládá alternativní řešení založené na pravidlech. Následuje závěr, seznam literatury a přílohy. Práce je psána anglicky, včetně příloh má 80 stran.

Hodnocení:

Práce má přehlednou strukturu, formální náležitosti jako seznam literatury a další rejstříky jsou rovněž v pořádku. Práce je napsaná dobře srozumitelnou angličtinou.

Realizace experimentů byla zásadně ztížena tím, že student ze zjevných bezpečnostních důvodů nemohl s kolekcí policejních textů pracovat na svém počítači ani na fakultní síti, práce s daty se mohla odehrávat pouze během předjednaných návštěv na jednom z útvarů PČR. Jedním z důsledků toho, že nebylo možné s daty pracovat flexibilně, byl i značně netypický průběh experimentů. Relativně sofistikované metody nasazené na začátku (klasifikátory se stromovými kernely), které byly úspěšně použité v literatuře, vedly k velmi skromným výsledkům. Nicméně ani klasifikátory bez kernelů nevedly k úspěšnosti použitelné v praxi. V navazující kritické analýze dat se ukázalo, že stávající anotace jsou natolik nejednotné, že metody strojového učení optimalizované na úspěšnost na stávajících trénovacích datech nevedou k použitelným výsledkům (a stejně tak tato data nejsou přímočaře aplikovatelná ani pro evaluaci). Na základě další analýzy dat byla tedy vytvořena sada kontextových pravidel, která rozpoznává relace s přijatelnou přesností (jak bylo studentem ověřeno při testování na daných textech), nicméně celkové pokrytí (recall) této sady pravidel změřit nelze.

Na rozdíl od počátečního záměru tedy není hlavním výstupem práce nějaký model standardně natrénovaný a vyhodnocený na ručních anotacích, ale spíše kritický rozbor stávajících anotací a zmíněná sada relativně spolehlivých pravidel, které pokrývají alespoň některé typy relací. To, že se student dovedl zorientovat i v poměrně nečekané situaci, hodnotím velmi kladně, stejně jako to, že jeho první experimenty s touto datovou kolekcí umožnily pracovišti ÚFAL zúčastnit se veřejné soutěže o zakázku na další zpracování těchto dat.

Závěr:

Autor práce prokázal, že je schopen navrhnout, implementovat a vyhodnotit komplexní výzkumný experiment. Práce podle mého názoru práce splňuje požadavky kladené na diplomovou práci.

Práci doporučuji **k obhajobě.**

Práci nenavrhuji **na zvláštní ocenění.**

Datum 5. 9. 2017 **Podpis**