Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Richard Ejem
Název práce	Relation extraction in police records
Rok odevzdání	2017
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	David Mareček **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky

Obsah práce

Předložená práce se zabývá extrakcí relací mezi pojmenovanými entitami v policejních spisech. Po úvodní kapitole následuje kapiola shrnující sočasné nejlepší metody pro extrakci relací. Rozděluje je na řízené, částečně řízené a neřízené a popisuje principy několika existujících nástrojů. Třetí kapitola se zabývá daty, které měl diplomant k dispozici. Jsou to pro češtinu data poskytnutá policií ČR a pro angličtinu pak dataset z LDC. Popisuje jejich formát, nedostatky a nutnou anonymizaci pro účely prezentace příkladů. Čtvrtá kapitola pak uvádí výslednou aplikaci a její rozhraní. V pátá kapitola popisuje připravu dat, konverzi do jednotného formátu, rozdělení na části a jejich lingvistickou analýzu (tagging, parsing). Součástí je i hledání skutečných pozic pojmenovaných entit v policejních textech, neboť anotace v nich je prováděna pouze výčtem za každým dokumentem bez referencí do textu. Samotné experimenty se strojovým učením těchto relací jsou popsány v šesté kapitole. Diplomant zkoušel nejprve jednoduché rysy založené na závislostních stromech a trénování pomocí SVM, následně pak přechází ke kernelovým metodám pro závislostní stromy, vyhodnocuje je pomocí SVM a KNN, zkouší vyvažování pozitivních a negativních instancí v datech a další kernelové metody založené na bag-of-words na závislostních podstromech a jejich kombinaci. Bohužel nedosahuje žádných rozumných výsledků. Navrhuje ještě jednoduchý pravidlový přístup založený pouze na třech triviálních pravidlech a ukazuje, že hlavním problémem je samotný dataset, který zřejmě obsahuje velké množství neoznačených relací, které zabraňují správnému naučení se daného problému. V sedmé kapitole se tedy diplomant zabývá analýzou dat a v osmé pak navrhuje pravidla která více odpovídají potřebám a anotacím policie. Poslední kapitola pak práci uzavírá. K práci je přiloženo CD obsahující všechny skripty potřebné pro popsané experimenty.

Hodnocení práce

Práce má 66 stránek čistého textu a je psaná dobře srozumitelnou angličtinou. Našel jsem pouze pár jazykových chyb, jako "semantical" místo "semantic", "junctions" místo "conjunctions" nebo

"collaborants" místo "collaborators", jednu chybějící poznámku pod čarou a chybějící odkaz na tabulku 6.4. Jinak je práce přehledně strukturovaná a všechny související články jsou správně citovány. Kladně hodnotím především popis a analýzu dat a závěrečnou podrobnou analýzu vlastností policejních anotací a návrh pravidel. Naopak výhrady mám k začátku kapitoly 6.5, kde by měly být definované kernelové funkce na závislostních stromech. Ke vzorcům, které jsou opsané z jiného citovaného článku bohužel chybí vysvětlení dost velké části proměnných (r_1, r_2, a, b, c). Vzorce navíc obsahují chyby (např. na straně 32 špatné horní indexy nebo nerozlišení vektrorů od skalárů) což současně s velmi stručným popisem dělá text naprosto nepochopitelným.

Diplomant provedl velké množství experimentů a jejich výsledky řádně prodiskutoval. Přestože se ukázalo, že výsledky použitých metod strojového učení mají velmi nízkou kvalitu, dokázal, že na vině jsou především nedostatečně anotovaná data a v práci pak pokračuje jejich analýzou. Trochu mi v práci chybí příklady některých výstupů, chápu ale, že vzhledem k výsledkům i kvalitě dat by na nich možná nic zajímavého vidět nebylo.

Dotazy a připomínky

1. Pracoval jste se skutečnými, nebo již anonymizovanými daty? V práci jsem to nikde přímo nenašel. Předpokládám, že anonymizace proběhla jenom pro prezentační účely a uváděné výsledky jsou na skutečných datech. Mohlo by to totiž leccos ovlivnit.

2. Uvádíte, že jednoduchý způsob, jakým hledáte pozice anotovaných entit ve skutečném textu (tedy hledáním pouze prvního tokenu entity v textu) je dostatečný. Zkoušel jste vyhodnotit chybovost tohoto přístupu? Z příkladu jsem zjistil, že v textu se vyskytující entitu "5kg OPL - kokain" pak policisti anotují jako "kokain, 5kg". Je to tedy pravidlem, že to nejdůležitější slovo, se kterým se pak následně i v závislostním stromě pracuje je v anotované entitě vždy na první pozici? Pokud by bylo v anotaci "5 kg kokainu", mohlo by zůstat pouze číslo "5", což by možná nebylo úplně žádoucí.

3. Jak se závislostní parser trénovaný na PDT vypořádává se specifickým textem plným zkratek a nevyskloňovaných jmen? Máte představu o chybách parseru? A zkoušel jste jednodušší metody, které by pracovaly například jenom se slovy v okolí a mezi danými entitami?

4. Píšete o rozdělení dat na trénovací, held-out a testovací. K čemu jste použil ta held-out data a na kterých datech děláte zmíněnou 5-fold cross-validaci?

5. Proč jste nevyhodnotil precision recall a f-measure pro vámi navrhnutá pravidla v kapitole 8? Chápu že data jsou špatná, ale pro srovnání s metodami strojového učení by to bylo zajímavé porovnání. Je totiž otázka, zda cílem má být správná anotace dat, nebo přiblížit se tomu, co policisti skutečně anotují.

Závěr

Předložená diplomová práce splnila vytyčený úkol automaticky extrahovat relace mezi pojmenovanými entitami. Přestože data, která byla k dispozici, nebyla kvalitní, student prokázal, že se umí s tímto vážným nedostatkem vypořádat a po kvalitní analýze navrhl nejlepší možná řešení. Mezi slabší stránky práce pak patří některé nedodělky, které na několika místech ubírají na srozumitelnosti.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 5. 9. 2017

Podpis: