

Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Vladan Glončák

Název práce Using Syntactic Features for Opinion Target Identification

Rok odevzdání 2019

Studijní program Informatika **Studijní obor** Matematická lingvistika

Autor posudku Jindřich Helcl **Role** oponent

Pracoviště Ústav formální a aplikované lingvistiky

Text posudku:

Obsah práce

Diplomová práce Vladana Glončáka se zabývá zkoumáním vlivu explicitní syntaktické informace na kvalitu systémů pro identifikaci hodnocených entit (OTE). Text práce je rozčleněn do sedmi kapitol, úvodu a závěru. V úvodu práce autor seznamuje se základními pojmy a s úlohou OTE a stanovuje cíl práce, jímž je zjistit, jak přidání syntaktických rysů ovlivňuje kvalitu modelů.

První kapitola je pokračováním úvodu, stručně popisuje úlohu analýzy sentimentu, třídy metod jimiž se řeší a reprezentaci slov pomocí embeddingů. Ve druhé kapitole jsou rozebrány jednotlivé syntaktické a morfologické rysy použité v experimentech. Textová data jsou zpracována nástrojem UDPipe. Třetí kapitola popisuje použitá data. Experimentuje se s daty v angličtině a češtině; v angličtině jde o hodnocení restaurací, v češtině o recenze elektronických zařízení. Čtvrtá kapitola pojedává o výsledcích souvisejícího výzkumu, který byl dosud nad použitými daty proveden. V páté kapitole autor popisuje použité metody, jmenovitě conditional random fields (CRF) a rekurentní neuronové sítě; dále je zde uvedena metodika vyhodnocování výsledků. Šestá kapitola stručně uvádí konkrétní varianty použitých modelů a jejich parametry. V sedmé kapitole jsou prezentovány podrobné výsledky navrhovaných experimentů.

Práce je psána anglicky, má 65 stran a je k němu přiložen archiv se skripty a konfiguračními soubory pro zpracování dat, trénování a evaluaci modelů.

Hodnocení

Úvod do tematiky. Přehled tematiky působí úplně s dostatkem odkazů na související literaturu. Všiml jsem si nicméně v textu několika následujících drobných nesrovnalostí:

V úvodu práce autor míní, že přidání syntaktických anotací do modelu nepřináší do dat žádnou novou informaci. Tento výrok je jen zčásti pravdivý. Anotace vygenerované automaticky modelem mohou přinést například chyby nebo jistou (statistickou) úroveň disambiguace.

Systémy analýzy sentimentu jsou v textu rozčleněny na pravidlové, automatické a hybridní. Zde mi slovo *automatické* nepřijde vhodně zvolené – automatické metody jsou všechny tři. Snad by se více hodilo označení *statistické*.

V kapitole 5.4 se objevuje tvrzení, že GRU se od LSTM liší tím, že mají jen forget gate. Ve skutečnosti mají GRU dva gating mechanismy (reset a update). Dále v téže kapitole není předkládána rovnice pro softmax, jak autor tvrdí, nýbrž pro normovaný vektor. Softmax normalizuje exponenciálu posloupnosti čísel tak, aby formovala pravděpodobnostní rozdělení (tj. sčítala do jedničky). Prvky jednotkového vektoru však do jedničky sčítat nemusí.

V kapitole 5.5 je rovnice pro precision asociována s výpočtem permissive precision. Přestože je u rovnice textový dodatek, jasnější by bylo napsat rovnice pro základní precision a recall a pak dodefinovat permissivní varianty.

Experimentální část. Experimenty jsou navrženy v souladu se zásadami strojového učení a parametry modelů jsou přizpůsobeny hardwarovým omezením. Níže předkládám několik otázek či komentářů k samotným experimentům.

- Subjectivity lexicon: Nejsem si jist, je-li přiložen popis formátu lexikonu, ale předpokládám, že jde o kombinaci sekvencí znaků (slov) a jejich polarity. Co kdyby k formě byl ještě v lexikonu napsaný slovní druh, který by nám říkal, že sloveso “like” je kladné, zatímco předložka/spojka “like” neutrální? Existuje takový lexikon? Mohl by takový lexikon pomoci?
- Data: V popisu dat by bylo vhodné uvést OOV rate, zvlášť mezi target expressions. Autor uvádí, že v datech je každý hodnocený výraz pouze jednou. Znamená to, že všechny hodnocené výrazy v testovacích datech jsou OOV? Jaký mělo vliv nahrazování unikátních slov OOV tokenem na zastoupení OOV tokenů mezi hodnocenými výrazy?
- Learning rate schedule: Není jasné, proč se autor nerozhodl pro (standardní) exponenciální snižování LR. Pokud je to rozhodnutí založeno na předchozím výzkumu, chybí zde patřičná citace.
- Autor zmiňuje, že pro experimenty s češtinou neexistovaly žádné předtrénované embeddingy slov s požadovanou velikostí. Tyto by se ale daly jednoduše (a výpočetně nenáročně i za použití CPU) natrénovat například pomocí nástroje word2vec na libovolných českých datech.

Forma a styl textu Samotný text práce působí poněkud nedokončeně. Velmi časté jsou překlapy, opakující se slova, nedokončené věty, chybějící členy. Autor očividně nepoužil kontrolu pravopisu, což by nepochybně zvládat měl.

Víceciferná čísla jsou zapsaná nejednotným stylem, buď česky nebo bez mezer, desetinná čísla v tabulkách anglicky.

Tabulek s výsledky je příliš mnoho a text tím ztrácí na přehlednosti. Tabulky se tak dostávají až o několik stránek před text, který tím postrádá potřebný kontext. Například v kapitole 7.3 na stránce 36 je zkoumán nejlepší model (*LSTM-emb*) z tabulky 4.16 ze stránky 40. Na stránce 37 pokračují odkazy na tabulku výsledků 7.19 zpět na stránku 40. Popisek tabulky 7.19 je navíc nedopsaná věta. Místo tří zvláštních tabulek pro “I”, “B” a “permissive” pro každý typ experimentu by byla přijatelnější jediná tabulka s devíti sloupci. To by počet tabulek v kapitole 7 zredukovalo z dvaceti pěti na devět.

Příloha k práci by pro snazší replikovatelnost měla obsahovat soubory s natrénovanými modely.

Abstrakt práce je jen zadání přeložené do angličtiny s jednou přidanou větou o výsledcích. V ideálním případě je to snad možné, ale bylo by lepší, kdyby abstrakt více odrážel strukturu a obsah textu.

Rozsah práce a splnění zadání Zadání práce spočívá ve zkoumání, jak ovlivňuje přidaná lingvistická informace daný model pro OTE. Ačkoliv není cílem práce najít co nejlepší model, kvalita zkoumaného modelu a vliv přidané syntaktické informace spolu úzce souvisí. V triviálním případě, kdy nebude model umět nic, mu zřejmě syntaktická informace pomůže. Bude-li to naopak model dokonalý, žádné další vstupy nikdy potřebovat nebude. Místo otázky vlivu přidání syntaktické informace do daného modelu by nám dle mého názoru více informací přineslo zkoumat, *kterým modelům a za jakých podmínek syntax ještě pomáhá a proč*. Tento směr výzkumu už by ale velice přesahoval rámec diplomové práce.

Autor v diplomové práci prokázal schopnost navrhovat a provádět experimenty se zpracováním přirozeného jazyka a třebaže nezjistil pozitivní vliv přidání syntaktické informace na kvalitu modelů OTE, na svou výzkumnou otázku podal odpověď a diplomovou práci splnil zcela v rozsahu zadání.

Práci doporučuji k obhajobě.

Práci nenavrhují na zvláštní ocenění.

V Praze dne 1. 9. 2019

Podpis: