Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Karolína Burešová
Název práce	Text Simplification in Czech / Zjednodušování textu v češtině
Rok odevzdání	2017
Studijní program	Informatika **Studijní obor** Matematická lingvistika

Autor posudku	Eduard Bejček **Role** oponent
Pracoviště	Ústav formální a aplikované lingvistiky, MFF, UK

Text posudku:

Popis práce

Práce popisuje poměrně rozsáhlou úlohu zjednodušování (simplifikace) textu. Autorka úlohu popisuje do značné šíře, probírá zjednodušování na různých úrovních (lexikální, syntaktické, pragmatické), pro cílové skupiny s různými potřebami (např. studenti jazyka, dyslektici, děti, laici) apod. Upozorňuje na velké množství drobných úskalí i zásadních překážek, které nám stojí v cestě, i na možná řešení. Často popisované dokládá množstvím příkladů.

Pro dobré porozumění tématu provedla autorka tři experimenty s lidmi, kteří hodnotili „čtivost“ vět, porovnávali složitost dvojic vět a ručně zjednodušovali předložené věty. Autorka implementovala a v práci popsala systém na zjednodušení českých vět pomocí identifikace složitých slov, vygenerování kandidátu, kteří by je mohli nahradit, a určení nejvhodnějšího z nich.

První kapitola práce je úvod s popisem úlohy, terminologie, současný stav poznání. Protože je celý text anglicky, ale pojednává o češtině, následuje kapitola shrnující některé její vlastnosti pro čtenáře, které s ní nejsou obeznámeni. Třetí kapitola se podrobně a obecně věnuje úloze zjednodušování textu na třech úrovních, rozebírá také pojmy „čtivost“ a „srozumitelnost“ a možnosti jejich měření, popisuje možné cílové skupiny; pro některé z nich je potřeba zadání úlohy mírně pozměnit. Čtvrtá kapitola velmi detailně rozebírá tři úvodní experimenty s lidmi. Pátá kapitola se věnuje systému na lexikální zjednodušování, který je součástí práce, a v šesté kapitole dochází k jeho vyhodnocení. Práce je zakončena implementačními poznámkami a závěrem, kde autorka prochází řadu možností na zlepšení systému, kterým se chce dále věnovat.

Hodnocení

Celá práce je velmi přehledná, napsaná dobrou angličtinou a řečeno jejími pojmy: má vysokou „čtivost“ i „srozumitelnost“. Drobné formulační nedostatky jsem autorce poslal e-mailem. Poněkud však trpí náročností a rozsáhlostí zvoleného tématu. Zatímco problematika zjednodušování

textu je zde popsaná jako ucelená mnohovrstevnatá úloha, systém, který je představen v druhé polovině, musel být od samého začátku omezen. Jak autorka v úvodu upozorňuje, implementace se omezuje na úroveň pouze lexikální a v kontextu jediné věty, tedy na nahrazování složitých slov jejich synonymy v češtině. V důsledku toho je většina dílčích úloh, které do zjednodušování textu spadají či s ním souvisí, pouze představena v první části. Tam je také doplněna řadou příkladů a problémů, jež by provázely pokusy o vyřešení úloh, i návrhů na řešení, ovšem bez implementace – ta není obsahem této práce. V rámci značně zúženého problému autorka i tak naráží na místa, kde by raději použila jiný přístup, ale snad kvůli nedostatku času k tomu již nedošlo. Svědomitě na taková místa poukazuje a uvádí jiná možná řešení, která nebyla vyzkoušena, ale očekává, že by mohly přinášet lepší výsledky, poukazuje na opomenutí a nedostatky (například jsou nahrazovány pouze jednoslovné výrazy v nejužším smyslu – a zanedbány tak jsou i reflexivní slovesa; při hledání synonym s použitím „word embeddings“ je vždy použito přesně pět nejbližších slov; generování správného tvaru výsledného slova je zřejmě příliš jednoduché). Na příkladu s výběrem pěti podobných slov je patrné, že autorka si dobře uvědomuje celou řadu jiných a lepších možností, navrhuje je – ale nevyzkoušela je a zvolila jednoduše konstantu 5.

Co se týče tzv. experimentů (tedy lidského hodnocení složitosti vět), vedou k velmi zajímavé diskusi, autorka přináší podnětné nápady a pohledy. Je však bohužel potřeba brát takové závěry poněkud s rezervou, neboť anotátorů bylo poměrně málo (s výjimkou druhého experimentu se účastnili jen dva a tři anotátoři, vět nebylo mnoho).

Na závěr práce s implementační částí bych si rád prohlédl výstupy, které představovaný systém nabízí. Nejlépe ve formě ukázkových dvojic uvedených přímo v diplomové práci: původní a zjednodušená věta. Takové ukázky v práci nejsou, jen na přiloženém CD jsou k disposici použité skripty. (Autorka uvádí: „…which allows to try any of the tested strategies on one's own corpus“.) To bohužel není pravda. Menší vada je, že chybí nějaký připravený skript, který by spustil systém na ukázkový text. Je potřeba podle nepřesných instrukcí v práci sestavit dlouhý příkaz s mnoha parametry, předzpracovat vstupní větu atp. Pak je ovšem také potřeba z mnoha důvodů číst zdrojové kódy a zasahovat do nich (ať už kvůli opravě cesty ke slovníku, nebo kvůli opravě chyb, kvůli kterým skript práci vůbec nedokončí). Vzhledem k absenci slovníků a frekvenční databáze není možné dobře reprodukovat experimenty – nejúspěšnější strategie právě tyto dva zdroje používají. Pokud je vše znemožněno licenčními podmínkami, doporučuji to v práci zmínit spolu s návodem, jak je možné některé ze zdrojů získat. Výsledek, který jsem nakonec získal, se tedy pravděpodobně odlišuje od správného chování systému. Jsou z něj patrné dvě věci: zaměňuje zejména koncovky za hovorové, což skutečně může srozumitelnost pro mnohé lidi zvýšit („starověcí“→„starověký“, „nezávisí“→„nezávisej“, „o nějakém“→„o nějakym“), nicméně taková změna nemění lemma, a drobné opomenutí: nezachovává velké písmeno na začátku věty.

Otázky

- Co bylo zdrojem korpusu nazvaného „complex“? Jiný uvedený korpus, beletrie, novinové články, …? Na str. 57 popisujete jeho velikost a rozdělení do souborů (na co bylo rozdělení použito?) a předvýběr vhodných vět – neuvádíte však, odkud věty pocházejí. Která data

jsou v tomto případě trénovací a která testovací (dělení uváděné v tabulkách 6.1 až 6.3)?

- Při ručním označování komplexních slov dostali anotátoři již anotované věty a anotaci jen opravovali, což – jak uvádíte – mohlo ovlivnit jejich rozhodování. Jaký byl tedy pro tento zvláštní postup důvod? Podobně několikrát narážíte na to, že způsob, jakým byly věty do korpusu „complex“ vybrány, nahrává vyhledání komplexních slov podle frekvence. Opět se chci zeptat: Co Vás tedy k takovému předvýběru vedlo?
- Na straně 58 připomínáte, že zatímco pro řešenou úlohu je zřejmě Recall důležitější než Precision, F-measure je hodnotí se stejnou důležitostí. Zvažovala jste použití F-measure s jiným koeficientem? Například pro F_3-measure již vychází výsledek pro „content“ (0,80) lépe než pro „frequency“ (0,68).
- Sekce o Joint evaluation (str. 65) v části o generování substitucí je velmi krátká. Myslí se tím testování obou přístupů (pomocí slovníku i pomocí „embeddings“) na stejných datech? Dostali tedy anotátoři třikrát stejné texty: jednou schvalovali vhodná synonyma ze slovníkové nabídky, podruhé z nabídky získané z „embeddings“ a potřetí z obojího naráz? (Zmiňujete nekonsistentní anotaci, která způsobila, že podruhé je úspěšnost pro slovník vyšší.) Proč bylo potřeba data anotovat vícekrát? Výsledky Precision a Recall zvlášť pro „embeddings“ uvedené nejsou – je to proto, že ty se na rozdíl od výsledků pro slovník neliší? Je při Joint evaluation počítán nově definovaný Recall (viz níže) proti synonymům získaným všemi prostředky?
- Popis systému končí ohodnocením možných synonym a výběrem nejvhodnějšího kandidáta. V poznámkách je kratičká zmínka o morfologické značce takového kandidáta. Jakým způsobem je tedy nakonec vygenerována morfologická značka zmíněného neutra nahrazujícího femininum? Bylo provedeno také testování (alespoň namátkou na několika větách) kvality konečného výstupu celého systému, tedy plynulosti a gramatičnosti výsledné věty, srozumitelnosti a zachování původního významu? Jaký z toho je výsledný dojem?

Jednotlivé poznámky

- Tvrzení, že subjekt v pasivní konstrukci odpovídá vždy akuzativu v aktivní, je platné nanejvýš v případě, že nám bude stačit zpracovat pouze ideální věty. V praxi se však i genitiv stává subjektem po pasivizaci často – někdy to vypadá lépe, někdy hůře:
  „Ministr práce říká, že sociální dávky nejsou a nebudou krizí dotčeny.“ (dotknout se *čeho*)
  „Vrchol byl dosažen v pozdních večerních hodinách.“ (dosáhnout *čeho*)
  „Nemohla být nepovšimnuta totální nespolupráce.“ (povšimnout si *čeho*)
  (poznámka pod čarou 12 na str. 31; protipříklady z Vallexu 3.0).
- Počty tokenů a vět v PDT uváděné na str. 45 jsou ve skutečnosti vyšší i pro tektogramatickou rovinu; v celém PDT je pak více než čtyřikrát tolik tokenů (1 957 247) a dvakrát tolik vět (115 844).
- Na str. 69 je uvedeno, že zatímco tři korpusy byly morfologicky označkovány s použitím MorphoDiTy, čtvrtý korpus (PDT) byl použit s původními značkami. To jsou bohužel značky podle starého morfologického slovníku a od nových se mohou v některých jednotlivých slovech

výrazně lišit (jak morfologickou značkou, tak počtem možných lemmat apod.) a bylo by bývalo lepší označkovat nově i PDT, aby všechna data byla v souladu.

Závěr

Práce přináší přehledný a ucelený pohled do problematiky zjednodušování textu. Autorka dobře popisuje různé důvody, které ke zjednodušování vedou a které – což úkol dále ztěžuje – kladou často rozporné požadavky. Dala si záležet na popisu tří základních úrovní řešení (náhrady slov, změny syntaxe a pragmatickém zjednodušení), čímž jistě zcela naplnila cíl práce („prozkoumat problematiku zjednodušování textu v češtině“). Navrhla také a uskutečnila tři experimenty, při kterých lidé poskytli podklady k tomu, co je to složitá a co jednoduchá věta.

Naplnění úkolu „navrhnout a implementovat experimentální metodu pro tuto úlohu“ bych si však na informatické škole dovedl představit trochu rozsáhlejší a v detailech pečlivější. Ať už jde o kvalitu software na přiloženém CD nebo o nevyzkoušení některých v práci navržených postupů. Implementační část je nutno chápat skutečně jen jako pilotní experimenty a ověření evaluace. Z budoucích kroků navržených v závěru a z vyjádření školitele je však vidět, že autorka chce na tématu dále pracovat.

Práci doporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 2. 6. 2017

Podpis: