Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce	Kateřina Macková		
Název práce	Question Answering		
Rok odevzdání	2020		
Studijní program	Informatika	**Studijní obor**	Matematická lingvistika

Autor posudku	Rudolf Rosa	**Role**	oponent
Pracoviště	Ústav formální a aplikované lingvistiky		

Text posudku:

Diplomová práce se zabývá úlohou Question Answeringu, tedy automatického odpovídání na otázky, pro češtinu. Konkrétně se zabývá jistým podtypem této úlohy, kdy vstupem je delší text a otázka, a požadovaným výstupem je úsek vstupního textu, který je odpovědí na otázku. Pro tuto úlohu existují různé hotové nástroje, z nichž autorka v práci používá a porovnává BiDAF a BERT.

Pro češtinu neexistuje Question Answering dataset, autorka proto přeložila anglický dataset SQuAD 1.1 do češtiny; k překladu využila automatický nástroj Lindat Translation. Zde autorka narazila na problém s nekonzistencí překladu, daný povahou úlohy a způsobem překladu. Úloha vyžaduje, aby odpověď byla úsekem vstupního textu, autorka ale překládá vstupní text a odpověď nezávisle, proto asi v polovině případů dochází k tomu, že nezávislý překlad odpovědi se v překladu vstupního textu nenachází, například z důvodu použití synonymního tvaru, změny slovosledu, či skloňování. Autorka proto navrhuje a implementuje jednoduchý algoritmus, který odhaduje pozici odpovědi ve vstupním textu na základě hledání nejdelšího společného podřetězce. Míru shody odpovědi se vstupním textem následně autorka používá pro filtrování datasetu, kdy v následných experimentech porovnává, zda k lepší úspěšnosti vede trénování nástrojů na větším množství zahrnujících i data s nižší shodou, anebo na menším množství dat s vysokou shodou. Ukazuje se, že pro BiDAF je lepší trénovat na větších datech obsahujících i nepřesné shody, zatímco pro BERT je lepší využít pouze data s přesnými shodami; jako development data je pak pro oba nástroje vhodnější použít pouze data s přesnými shodami. Odhadování pozice odpovědi a využití míry shody odpovědi s textem pro filtrování trénovacích dat je tedy **prvním přínosem práce**, byť z mnoha důvodů poněkud problematickým.

Moje první výhrada směřuje k tomu, že řešený problém si autorka do značné míry způsobila sama tím, že překládá odpověď nezávisle na překladu vstupního textu. Pokud by místo toho překládala pouze vstupní text, a odpověď na otázku následně hledala přímo v tomto překladu, přímo tento problém by nevznikl. Vzhledem k tomu, že anglický dataset obsahuje informaci o poloze odpovědi v anglickém textu, bylo by možné odhadnout polohu odpovědi v přeloženém textu například pomocí technik slovního zarovnání (word alignment). Neumím odhadnout, jak úspěšná by tato metoda byla, nicméně autorka kořen problému a možnost alternativního řešení vůbec nezvažuje, metodu nezávislého překladu vstupního textu a odpovědi bere jako jaksi danou.

Následně pro nápravu problému používá poměrně "tupý" string matching, který mnohdy selhává. Přestože autorka správně identifikuje některé vlastnosti překladu a českého jazyka, které problém způsobují (synonymie, skloňování, slovosled, překlad názvů, slovní přepis čísel), nijak se nepokouší tyto problémy řešit. Kromě toho nesprávně tyto problémy uvádí jako chyby automatického překladu, tedy svádí je na užitý systém, přestože přinejmenším ve všech uvedených příkladech je překlad zcela korektní, pozorované neshody jsou inherentní vlastností jazyka, a následné chybné namapování odpovědi na text je tedy nedostatkem autorčina algoritmu, nikoliv překladového nástroje. Přitom synonymii lze odhadovat například pomocí slovních embedinků (například pomocí nástroje word2vec, který sama autorka jinde v práci užívá), skloňování lze řešit pomocí taggerů, lematizátorů a morfogenerátorů (například pomocí nástroje MorphoDiTa, který autorka taktéž v práci užívá, avšak pouze pro lematizaci v evaluaci; úplně stejně by šla přece dělat

lematizace zde), změna slovosledu je snadno řešitelná i bez zapojení jazykových nástrojů (autorka jako příklad problému ukazuje případ, kdy text obsahuje frázi "pes domácí", zatímco odpověď je přeložena jako "domácí pes"), slovní přepis čísel lze pak řešit pravidlově (ovšem v příkladu, který k tomu autorka uvádí, se vůbec nejedná o tento problém).

Obecně platí, že autorka v zásadě nijak nevyužívá to, že pracuje s češtinou, které rozumí a pro kterou lze mnohé problémy snadno řešit; to by trochu dávalo smysl, pokud by se autorka snažila vytvořit jazykově nezávislé řešení, ale takto práce nebyla zadána, a ani autorka ji takto neprezentuje a vyhodnocuje ji pouze na češtině, přestože v principu asi nic nebrání vyhodnocení i na dalších jazycích (mBERT podporuje 100 jazyků, Lindat Translation podporuje překlad mezi angličtinou a 5 jazyky, tj. pro další 4 jazyky by bylo provedení experimentů absolutně triviální). Následně je tedy otázkou, do jaké míry je skutečně přínosné filtrování dat na základě neshody odpovědi s textem; ostatně ať už by byla odpověď na text namapována jakkoliv, bylo by následně možné přeloženou neshodující se odpověď nahradit odpovídajícícm úsekem textu, a tím neshody odstranit. Prvním přínosem práce je tedy ve skutečnosti triviální automatický překlad existujícího datasetu pomocí existujícího nástroje s následným poněkud pochybným řešením problému, který možná ani nemusel vzniknout; následné experimenty ohledně vlivu filtrování dat pak mají platnost omezenou na dataset vzniklý zvoleným postupem, a není zřejmé, zda z nich lze vyvozovat obecnější závěry. Ostatně diskuze výsledků je také pochybná, neboť vůbec nezmiňuje dle mého názoru klíčový rozdíl mezi nástroji BiDAF a BERT, totiž že BiDAF predikuje pozici odpovědi v textu (a dle popisu v práci tedy vůbec nepracuje s překladem odpovědi), zatímco BERT predikuje, který z úseků vstupního textu je nejlepší odpovědí na otázku (a tedy naopak nepracuje s pozicí odpovědi v textu); to má pravděpodobně zásadní vliv na chování těchto nástrojů v situacích, kdy odpověď neodpovídá textu.

Druhým přínosem práce je pak porovnání 5 různých nastavení experimentu. Pro BiDAF i pro BERT autorka zkouší dva postupy. Prvním postupem je trénování nástroje na datasetu přeloženém do češtiny. Druhým postupem je trénování nástroje na původním anglickém datasetu, kdy pro použití nástroje na češtině je třeba nejprve přeložit vstup do angličtiny (nástrojem Lindat Translation), a následně výstup nástroje opět automaticky přeložit do češtiny. Třetím postupem je pak využití multilingválního nástroje mBERT, kdy je tento natrénován na anglických datech, a následně aplikován přímo na česká data bez jakéhokoliv překladu; díky své multilingvalitě pak nástroj generuje výstupy v češtině. Autorka ukazuje, že BERT je obecně úspěšnější než BiDAF, přičemž nejlepších výsledků dosahuje BERT natrénovaný na anglických datech, využívající automatický překlad vstupů a výstupů; kompetitivních výsledků ale dosahuje i mBERT, který je přitom snadnější na použití díky absenci nutnosti využití automatického překladu.

Žádný z těchto postupů není nový, tj. autorka pouze přímočaře využívá již známé a ověřené nástroje a postupy, vesměs doporučené již v zadání práce. Jejich výběr je nicméně velmi vhodný, podložený jejich vysokou úspěšností v současném výzkumu. Druhým přínosem práce je tedy zejména aplikace existujících nástrojů a postupů a jejich porovnání na úloze Question Answeringu pro češtinu, což pravděpodobně nebylo dříve učiněno, a tedy nebylo známo, který z postupů v této situaci dosáhne nejvyšší úspěšnosti. Tento přínos nezpochyňuji, je ovšem otázkou, zda v tomto lze spatřovat dostatečné množství "samostatné tvůrčí odborné práce" autorky, očekávané od diplomové práce; dle mého názoru je zde vlastní tvůrčí vklad autorky zanedbatelný.

Zřejmý prostor pro vlastní tvůrčí činnost u práce tohoto typu spatřuji zejména v důkladné analýze a interpretaci výstupů a výsledků. Evaluace v práci je ale velmi plochá, zaměřující se pouze na prezentaci a porovnání dvou automaticky počítaných skóre (která jsou zmatečně zavedena v jedné z podpodsekcí kapitoly o datasetu a popsána nepřesně či přímo chybně, takže ani není zřejmé, co za skóre se vlastně počítá). Porovnání různých nastavení experimentu pak obvykle sestává z pouhého konstatování, pro které nastavení vyšly vyšší hodnoty těchto skóre; pouze příležitostně se autorka pouští do "odvážnějších" hypotéz, zejména u vyšší úspěšnosti systémů trénovaných na angličtině oproti stejným systémům trénovaným na češtině, kde usuzuje, že důvodem je nejspíše větší menší dat a menší morfologická bohatost v případě češtiny. Tyto hypotézy nicméně již dále nezkoumá, přestože by bylo poměrně snadné je ověřit – jednak natrénováním systémů na podmnožině anglických dat, tak aby odpovídala velikostí českým datům (která jsou menší z důvodu filtrování), jednak lematizací dat (čímž by se v zásadě eliminovala morfologická bohatost). Hlouběji autorka

již nejde, tj. zejména v práci nenalézám žádný doklad toho, že by autorka přímo vlastníma očima zkoumala výstupy systémů či se na ně vůbec kdy dívala; práce neobsahuje ani jeden příklad výstupu (nějaká ruční analýza výstupů je provedena pouze u výstupů strojového překladu datasetu). Je přitom dobře známo, že automatická evaluace má pouze omezenou vypovídací hodnotu a je užitečná zejména pro trénování systému a porovnání většího množství experimentů; finální systém by měl vždy být vyhodnocen i ručním zkoumáním jeho výstupů. Vůbec se tedy nedozvíme, jaké výstupy systémy vlastně generují, jakých chyb se dopouštějí, zda mají výstupy nějaké zvláštní vlastnosti, zda se výstupy některých systémů nějak konzistentně liší, které vstupy jsou pro ně snadné a které obtížné, apod. Taková analýza je přitom užitečná nejen pro lepší vyhodnocení kvality vytvořeného systému, ale zejména pro lepší porozumění vlivu různého nastavení experimentů a pro potenciální odhalení problémů, které by bylo možné odstranit například jiným nastavením systému či nějakým dalším zpracováním či filtrováním dat.

Další příležitostí je pak dostatečně široká rešerše a porovnání možných nástrojů a postupů. Autorka přímo experimentálně porovnává dva nástroje pro Question Answering a tři postupy pro mezijazykový přenos, avšak bez odpovídající rešerše, která by použití právě těchto nástrojů motivovala. Teoretická část práce zaměřená na Question Answering je zcela odtržená od vlastní práce, neboť popisuje zejména systémy z minulého tisíciletí, zaměřující se navíc na jiný podtyp úlohy než autorka řeší (vstupem je pouze otázka, odpověď systém nehledá v textu, ale v zabudované bázi znalostí); užití systémů BiDAF a BERT pak jaksi "spadne z nebe", přitom na internetu lze mimo jiné snadno nalézt například přehled téměř 200 Question Answering systémů z posledních let a jejich úspěšnosti na autorkou užitém datasetu SQuAD 1.1 (*https://paperswithcode.com/sota/question-answering-on-squad11*); proč tedy právě BiDAF a BERT, a proč nejsou zmíněné žádné jiné systémy? Pro podúlohu mezijazykového přenosu pak není v práci obsažena rešerše či motivace vůbec žádná, dokonce zde pro žádnou ze tří užitých metod není uvedena žádná reference na literaturu zavádějící či popisující dané metody, což u práce výslovně zaměřené na mezijazykový přenos znalostí považuji za hrubý nedostatek (byť zde proti samotné volbě užitých metod nemám námitek).

Samotný text práce je velmi slabý. Zhruba polovinu ze 42 stran textu tvoří pouze "převyprávění" několika zdrojových článků. Text nelze považovat za rešerši v pravém slova smyslu, neboť ta by měla ze zdrojových článků vycházet, avšak s informacemi z nich nějakým způsobem dále pracovat, interpretovat je, dávat je do souvislostí či kontrastů, vzájemně je porovnávat, spojovat, apod. To se ale neděje. Navíc se autorka mnohdy při převypravování článků dopouští různých omylů, nepřesností či zmatení, často uvádí informace vytržené z kontextu, které samy o sobě nedávají mnoho smyslu, takže celková hodnota některých částí textu je spíše záporná – aneb čtenář udělá lépe, pokud si rovnou přečte původní zdroje těchto pasáží.

Kapitola 2, která je úvodem do problematiky zpracování přirozeného jazyka a question answeringu (5 stran), postupně shrnuje tři články z Wikipedie (!), aniž by citovala jakékoliv primární zdroje (i v případech, kdy se v textu mluví například o konkrétních systémech, tak k těmto není uvedena žádná citace). Mimoto je tato kapitola dost odtržená od toho, co se pak v práci reálně dělá, a je spíše jakýmsi téměř nesouvisejícícm historickým exkurzem; o postupech užívaných v práci (hluboké učení, slovní embedinky, jazykové modely, attention, atd.) se zde čtenář nedozví prakticky nic. U samotného Question Answeringu pak autorka popisuje převážně jiný podtyp úlohy než který následně v práci řeší, jak již jsem uvedl výše; toto nikde není výslovně řečeno či rozlišeno, navíc je zde terminologické zmatení, kdy autorka vstupní text, ve kterém hledá odpovědi, střídavě označuje jako "text", "content", či "context" (tedy já se pouze domnívám, že pokaždé myslí totéž).

Kapitola 3 (7 stran) se tváří jako rešerše existujících datasetů pro Question Answering, motivující volbu datasetu SQuAD, a následný podrobný popis tohoto datasetu. Ve skutečnosti je ale celá tato kapitola pouze převyprávěním článku Pranav Rajpurkar and Liang [2016], který je popisem datasetu SQuAD, a jako takový obsahuje i rešerši dalších datasetů. Zdánlivě autorkou provedená rešerše je tedy ve skutečnosti pouze adaptována z tohoto článku (a reference na datasety jsou také pouze přejaty z tohoto článku, neobjevil jsem v práci doklad toho, že by autorka datasety skutečně zkoumala i sama). V samotném popisu datasetu SQuAD pak autorka ze zdrojového článku adaptuje i zcela irelevantní pasáže, popisující například metody, které autoři použili

pro analýzu vzniklého datasetu (ale již neuvádí výsledky této analýzy, které by snad mohly být zajímavé), či dokonce experimenty, kde autoři článku využili vytvořený dataset pro trénování Question Answering systému; tyto metody autorka mylně popisuje jako metody pro analýzu datasetu (a výsledky experimentů neuvádí).

Kapitoly 4 a 5 (celkem 7 stran) popisují použité nástroje pro Question Answering, tedy BidAF a BERT, a jsou opět autorčiným převyprávěním dvou článků: Seo et al. [2016] a Devlin et al. [2018]. Popisy považuji za poněkud zmatené, pro pochopení toho, jak nástroje fungují, jsem si musel přečíst původní zdrojové články. V kapitole 5 pak za kriticky nedostatečnou považuji sekci 5.2, která popisuje vlastní aplikaci nástroje BERT na úlohu Question Answeringu, tedy nastavení, které se následně v experimentech ukáže být nejúspěšnějším. Tato sekce, čítajících pouhých 11 řádek textu, je zcela nesrozumitelná, vůbec jsem z ní nepochopil, jak se ten Question Answering dělá. Například se zde hovoří o nějakých kandidátech na odpovědi, o kterých ale nikde jinde není zmínka, a není vůbec zřejmé, kde se vezmou. Očekával bych, že popisu hlavní metody bude věnováno výrazně více prostoru a bude zpracována výrazně kvalitněji.

Kapitoly 6 a 7 (celkem 18 stran) konečně popisují vlastní práci autorky, tedy tvorbu českého datasetu a experimenty. Kapitola 6 zbytečně podrobně popisuje JSON formát datasetu, což by navíc patřilo spíše do kapitoly 3. Samotný algoritmus je pak popsán poněkud nešikovně, ne zcela jsem pochopil jeho princip; zejména čtvrtý řádek od konce je poměrně kryptický, ale i další obraty a značení použité v algoritmu nejsou dostatečně dobře vysvětleny. Kapitola 7 je, zejména ve srovnání se zbytkem práce, poměrně dobrá a srozumitelná. Mísí se zde popis metody, experimentů a evaluace, ale strukturování je i tak relativně přehledné. Samozřejmě je i zde co vytknout, například spojnicový graf 7.1 spojující související i nesouvisející body, ale k tomu, co kapitola obsahuje, v zásadě nemám závažnějších námitek. Problemtické je zejména to, co kapitola neobsahuje, tj. chybějící podrobnější analýza a vyhodnocení výsledků, jak jsem již uvedl.

I po formální stránce je text práce slabý. Práce je psaná anglicky (přestože v SISu se uvádí jako jazyk práce čeština) s velkým množstvím chyb a překlepů, přičemž textu by velmi pomohla jazyková korektura; mnoho překlepů by ale jistě zvládla opravit i sama autorka, pokud by si byla práci po sobě pozorně přečetla. V některých případech chyby v textu ztěžují či znemožňují porozumění významu textu.

Jak již bylo naznačeno, za velký nedostatek považuji citování, kdy velké množství relevantních citací není uvedeno, a to i na zřejmých místech (například pokud se v textu hovoří o konkrétním nástroji či metodě). Mnohdy jsem až po určité době rozklíčoval, co je (asi) vlastní práce autorky a co je pouze převyprávěno z jiné práce. Práce celkově cituje asi 20 článků, včetně 3 článků z Wikipedie; z několika málo článků vychází v podstatě polovina obsahu práce, ostatní články jsou většinou naopak pouze zmíněny bez dalšího popisu. Formátování citací také není optimální (např. "Therefore, we have lemmatized by MorphoDiTa Straková et al. [2014] all answers in the file").

Význam práce nelze spatřovat ani v její implementační části – přiložené zdrojové kódy mají v součtu pouze asi 500 řádků kódu, přičemž s výjimkou algoritmu pro mapování přeložených odpovědí na přeložené vstupní texty (asi 100 řádek kódu) se jedná zejména o volání existujících nástrojů a předávání dat mezi nimi. Jde spíše o pomocné skripty, tj. k práci není přiložen výsledný natrénovaný Question Answering systém, který by šel spustit (přiloženy jsou pouze vstupy a výstupy).

Název práce "Question Answering" neodpovídá názvu v SISu "Crosslingual Transfer in Question Answering", špatné je též vročení (2019 místo 2020).

Samotné zadání práce je poměrně stručné, zde mohu konstatovat, že práce jej splňuje.

Celkově práci neupírám jistý přínos, vzhledem ke všem uvedeným problémům ji ale nepovažuji za dostatečně kvalitní pro udělení magisterského titulu.

Práci nedoporučuji k obhajobě.

Práci nenavrhuji na zvláštní ocenění.

V Praze dne 27. 1. 2020

Podpis: