Prof. Ing. Luděk Müller, PhD.,
Fakulta aplikovaných věd ZČU, Katedra kybernetiky
Technická 8 306 14 Plzeň
Tel.: 377 632 508, 377 632 523
Email: muller@kky.zcu.cz

Oponentský posudek disertační práce Mgr. Jana Oldřich Krůzy:

Iterativní zdokonalování přepisu zvukových nahrávek s využitím zpětné vazby posluchačů´

Posuzovaná práce pana Mgr. Jana Oldřicha Krůzy se zabývá problematikou přepisu řeči do textu, speciálně se pak věnuje možnosti postupného zdokonalování výsledného přepisu do textu s využitím systému automatického přepisu a postupného rozšiřování trénovací množiny, to vše se zaměřením na úzkou doménu, kterou je soubor zvukových nahrávek obsahujících hovory českého filozofa ing. Karla Makoně.

Předložená písemná práce o délce 88-ti stran (včetně odkazů na literaturu) je psaná v češtině a má dobrou jazykovou i grafickou úroveň s minimálním počtem překlepů. Práce je rozdělena do logických částí (uvedených v obsahu práce) a má zajímavé členění, kdy poměrně velký prostor je věnován popisu života a nauky ing. Karla Makoně a jen velmi málo stran je věnováno výkladu použitých metod strojového učení. Uvítal bych, kdyby obsah práce čítal i samostatnou kapitolu vytyčující cíle práce, kapitolu popisující stav současného vědeckého poznání a kapitolu shrnující přínos disertační práce k posunu tohoto poznání. Práce je ale čtivá a dobře srozumitelná, její výklad je vesměs pochopitelný, i když na některých místech užívá ne zcela přesnou terminologii.

Praktickým cílem a pravděpodobně i hlavním hybatelem práce bylo dosáhnout co nejpřesnějšího přepisu audionahrávek ing. Karla Makoně zaznamenaných na kazetách a magnetofonových kotoučích a zpřístupnit je široké komunitě zájemců využitím metod vyhledávání v (automaticky) přepsaném audiu. Pro dosažení tohoto cíle bylo třeba vyřešit řadu podcílů, a to od přípravy a analýzy dat (přípravy korpusů zvukových nahrávek), postavení systému automatického rozpoznávání řeči, jeho adaptaci na cílové prostředí a doménu, testování a vývoj metod pro postupné interaktivní rozšiřování množiny trénovacích dat s využitím kontroly (se vzdáleným přístupem) prováděné posluchači a čtenáři přepsaného textu.

K práci mám řadu připomínek, z nichž nejpodstatnější je k celkovému spíše aplikačnímu vyznění práce s relativně menším přínosem inovativních metod z oblasti zpracování přirozeného jazyka, jimž by se práce z oboru Matematická lingvistika měla především věnovat. Dále je práce zatížena množstvím nepřesných tvrzení, která by též kromě naplnění své pravdivostní hodnoty měla mít (s ohledem na obor disertace) exaktnější formu než prosté vágní vyjádření v přirozeném jazyce.

V následujícím seznamu uvádím jednotlivé dílčí připomínky:

1. V části 2.2.1 na str. 17 autor uvádí, že automatický přepis slova „mithraismus“ vždy selhával, tj. toto slovo nebylo nikdy rozpoznáno, a uvádí, že cituji: „by bylo užitečné mít k dispozici pro vyhledávání i přepis získaný bez použití jazykového modelu. Toto je předmětem budoucí práce.“ Zde za prvé nevím, co se míní přepisem bez jazykového modelu (pravděpodobně prakticky ve vztahu k dekodéru CTC u SpeechDeep), a za druhé možnou příčinu selhání vidím spíše ve špatném jazykovém modelu. Při této příležitosti bych se rád dotázal, zda autor zkoumal OOV (Out-Of-Vocabulary Word)?

2. V tabulkách 2.1 a 2.2 by bylo vhodné uvádět i recall, počty „výsledků v automat. přepisu“ tabulek totiž nejspíše zahrnují i false positive a uváděné precision je jen velmi hrubým odhadem.
3. V části 3.4.1 „Spektrální odečet šumu“ autor konstatuje, že mu „Chybovost na slovech vzrostla skoro na sto procent.“ Je otázka, zda má smysl tuto část uvádět, když se již autor nezabývá příčinou s odůvodněním, že od této metody nemá žádná očekávání.
4. V části 3.4.2 „Neurální doménový transfer“ (autor by se asi neměl psát dle ruského přístupu Žu, ale spíše Zhu) je síť CycleGAN použita pro odstranění přebuzení nebo nízké kvality „nízkootáčkových“ nahrávek. Výsledky uvedené v části 3.4.3 „Vyhodnocení“ nejsou dobré. Doporučoval bych použít augmentovaná trénovací data a to transferem dobrých nahrávek na přebuzené či „nízkootáčkové“.
5. Nerozumím na straně 38 tomu, proč je pro konstrukci jazykového modelu bráno jako nedostatek, že se z přepisu jednání PS ČR vynechávají odkazy na jiné schůze. Odkaz („link“) přece není relevantním tokenem. Nebo se tím myslí něco jiného?
6. Na straně 38 je též zmíněna Levenshteinova vzdálenost počítající s editačními operacemi. Není však řečeno, s jakými jednotkami pracují editační operace (předpokládám, že pracují s písmeny porovnávaných slov.
7. V části 4.1.3 „Výběr trénovacích vzorků“ autor uvádí, že musel vyřadit minimálně 40 % vzorků z důvodu, že „z každého čtrnáctiminutového souboru je jen deset minut pokryto odpovídajícím přepisem", což výpočetně nesedí.
8. Na str. 40 bych oponoval tvrzení „je přípustné, aby některá slova měla i nulovou spolehlivost, tedy aby v nich všechna písmena byla špatně“ – zkoušel autor identifikovat taková slova? Využívá se zde nějak zpětná vazba pro trénink systému ASR? Zdokonaluje se trénovací korpus i aplikací nových modelů na trénovací data?
9. V části 4.2. „Číslovky a zkratky” opravdu nelze očekávat, že pouhé “zahrnutí číslic do abecedy a tedy přepis číselných výrazů přímo na číslice” bude při známé velikosti trénovacích dat (cca 500 tis. čísel) fungovat. Autor správně navrhuje nejprve před vlastním trénováním aplikovat (automatický) rozpis číslicového výrazu (asi nikoliv “číslic”, jak je psáno na straně 40), ale ze všech pak vybírá jen “nejpravděpodobnější variantu”. V jakém smyslu je zde chápána nejpravděpodobnější varianta (bere algoritmus např. v úvahu akustiku)?
10. V kapitole 5. “Automatický přepis” je uváděn rozdíl mezi HMM a modelem založeném na neuronových sítích. To není přesné. HMM mohou používat neuronové sítě a tedy rozdělení modelů uváděné autorem je nesprávné (předpokládám, že autor zná rozdíly mezi pojmy jako DNN, RNN, CTC, Attention NN, RNN-Transducer, atd.). S tím pak samozřejmě souvisí i tvrzení „Dosud neznám žádný nástroj založený na hlubokých neuronových sítích, který by toto [zarovnání – poznámka oponenta] poskytoval“.
11. Tvrzení “Všechny fonémy se inicializují jako shodné” na str. 43 u popisu trénování akustického modelu HMM v systému HTK je opravdu velmi vágní.
12. Podobně vágní jsou výrazy typu: “Každá [složka GMM – poznámka oponenta] má svůj střed, svoji varianci a svoji váhu”.
13. Co znamená pojem “virtuální trifoném” (str. 45)?
14. Neobvyklou psanou formu má též výraz “gaußovská distribuce” (str. 45).
15. Na straně 45 se pravděpodobně zcela mění význam pojmu “mixture” a myslí se jím jedna složka modelu směsi. Též výraz „mixture" by mohl být nahrazen výrazem “směs”, když autor používá české výrazy pro jiné pojmy, jako např. “foném”.

16. I další termíny jako např. “markovovský model” (např. str. 43), “trifonémy” (např. str. 44), popř. i „dizambiguace“ (nikoliv „disambiguace“) (na str. 55) jsou neobvyklé a v literatuře nezavedené.
17. Obrat „Z každé dvacáté jsem snížil na polovic nejen abych neplýtval trénovacími daty, nýbrž také protože vyhodnocování mixtur zabírá při trénovásní zdaleka nejvíce času, a ten je přímo úměrný velikosti sady heldout.“ je nejasný (str. 48).
18. Kepstrální normalizace MFCC se standardně již dlouhá léta provádí na částech neobsahujících neřečový úsek (ticho). Metoda uváděná v části 5.6 „Experiment s kepstrální normalizací“ na straně 48 tedy není nijak nová, navíc často obsahuje nejen CMN ale i CVN.
19. Co znamená obrat “ručně přepsaná slova a automaticky přepsaná slova mají tendenci se shlukovat”? (str. 61)
20. V části 6.3.2 “Metody hledání bodů předělu” na straně 65 autor píše: “Tam, kde pořízení přepisu nebo jeho automatické zarovnání selhalo, lze použít detekci ticha prostou akustickou analýzou. Tato metoda je velice náchylná k chybám v případě nahrávek s malým poměrem signálu k šumu, a těch je v korpusu Karla Makoně mnoho”. Existují ale metody pro konstrukci VAD, který je použitelný v hlučném prostředí. Zkoušel autor tuto úlohu řešit nebo na základě úsudku prezentované ve výše citované větě se tímto problémem nezabýval?
21. Nerozumím tomu, proč by úloha nalezení bodu předělu (tj. místa, kde lze zvukovou nahrávku rozdělit pro další zpracování, aniž bychom ji dělili uprostřed slova) neměla mít řešení pro případy, že délka ticha trvá déle než 120 sec. Lze přeci libovolně dlouhý úsek ticha vystřihnout. (část 6.3.3 Výběr bodů předělu, str. 66)
22. Algoritmus výběru bodů předělu na téže straně 66 popsaný slovy: “Začneme s množinou všech tich a iterujeme přes ně od nejkratšího po nejdelší. Ticho z množiny odebereme, pokud sloučením sousedních segmentů nevznikne segment delší než 60 sekund. Přes vybraná ticha znova iterujeme a ticho odebereme, jestliže jeden z jeho sousedů má méně než 30 sekund.” by zasluhoval vysvětlení.
23. Datová reprezentace slova (v části 6.4.2 „Pořízení fonetického přepisu“ na str. 68-69) obsahuje v bodě 3. výslovnost v podobě seznamu fonémů. Opět se jedná o nesprávný termín – pravděpodobně autor myslí posloupnost fonémů, nikoliv seznam.
24. Ohledně fonetického zápisu mi není jasné, zda autor uvažuje více výslovnostních tvarů slova, minimálně u HTK systému se toto dá s úspěchem využít. CTC systém pak nemusí používat výslovnostní slovník.
25. V části 6.4.4 „Vyhodnocení kvality přepisů“ pro vyhodnocení kvality přepisů bylo náhodně vybráno 20 odmítnutých přepisů “zarovnávačem” a provedena jejich analýza. Znamená to, že neodmítnuté přepisy nebyly pak vůbec hodnoceny z hlediska kvality?
26. Na straně 72 pak věta “Čtvrtá kombinace fonetického zápisu a špatné výslovnosti se pochopitelně nevyskytuje.” není psána příliš exaktně a pravděpodobně znamená “Čtvrtá kombinace správného fonetického zápisu a špatné výslovnosti se pochopitelně nevyskytuje.”.
27. Tvrzení na str. 74: “Bohužel se zdá, že nucené zarovnání funguje v HTK pouze s monofonémovým modelem, takže přesnost v rozlišování přesných a chybných příspěvků není optimální.” není správné. Samozřejmě lze použít příkaz HVite i k zarovnání s kontextovým akustickým modelem.
28. V kapitole 7 „Vyhledávání“ na straně 76 autor plánuje používat systém přepisu bez jazykového modelu. Co je tím přesně myšleno? A proč je to cílem? (souvisí s připomínkou č. 1 výše)
29. Konstatování “Jakékoli velké soubory audio nahrávek s komunitou příznivců by tak mohly být přepsány a dále podrobněji zpracovávány metodami, které jsem v této práci rozvinul.” je diskutabilní. Jaké konkrétní metody a jejich rozvinutí má autor na mysli?

Přes uvedené nedostatky musím kladně hodnotit opravdovou snahu autora o dosažení co nejlepšího výsledku v úloze automatického přepisu korpusu audionahrávek ing. Karla Makoně zaznamenaných na kazetách a magnetofonových kotoučích. Jedná se o aktuální úlohu zcela jistě i s pozitivním praktickým dopadem. Přínosem práce je především návrh, realizace a vyhodnocení metody pořízení kvalitního zarovnaného přepisu velkého množství dat se zapojením pouze malého počtu laických přispěvatelů, metoda získávání specifických trénovacích dat pro aktivní učení od dobrovolných anotátorů, nový tisícihodinový korpus pro úlohu S2T, který autor dává svobodně k dispozici, i dostupnost všeho autorem použitého kódu pro řešení úlohy. Domnívám se, že z tohoto důvodu lze dovozovat, že výsledky provedených experimentů jsou reprodukovatelné. Kladně lze hodnotit i snahu autora o aplikování aktuálních state-of-the-art metod, které se v průběhu řešení disertační práce poměrně výrazně vyvíjely. Práci proto doporučuji k obhajobě s tím, že při ní od autora očekávám zdůraznění jejího významu pro rozvoj studovaného vědního oboru a přínosu pro vědeckou komunitu.

V Plzni 7. 9. 2020

.....
Luděk Müller