„Multilingvální systémy rozpoznávání řeči a jejich efektivní učení“

Oponentský posudek disertační práce Ing. Radka Šafaříka

Předložená disertační práce se zabývá adaptací existujícího systému automatického rozpoznávání řeči (ASR) pro další jazyky. Zaměřuje se zejména na slovanské jazyky, které jsou si dostatečně podobné, přičemž výchozím jazykem je čeština. Při řešení se autor nejdříve zaměřil na tvorbu jazykového korpusu, výslovnostního slovníku a jazykového modelu. Dalším cílem práce byl návrh a implementace metod pro tvorbu akustických modelů. Funkčnost navržených metod byla úspěšně ověřena na jedenácti slovanských a třech dalších jazycích. Chybovost slov (WER) byla ve většině případů nižší než 20 %.

1. **Význam disertační práce pro obor**

   Vícejazyčnost a vývoj systémů, které dokáží zpracovávat řeč / přirozený jazyk ve více jazycích je dnes **velmi perspektivní téma**, kterým se zabývá řada výzkumných institucí i firem. Do této oblasti patří i návrh a implementace komponent pro vývoj multilinválního ASR systému, čímž se zabývá předkládaná práce. Hlavním přínosem práce je vývoj jazykově závislých modulů (slovníky, jazykové modely, akustické modely) pro další jazyky přidávané do systému tak, aby byla minimalizována manuální činnost. Navržené postupy a metody představují, podle mého názoru, **podstatný přínos** v dané oblasti. Jakýkoliv posun je zde žádoucí z pohledu výzkumu i praxe.

2. **Použité metody, postup řešení a splnění stanovených cílů**

   Autor použil při řešení práce tradiční postupy spolu s metodami založenými na hlubokém učení. Výběr metod pro řešení problémů je logický.

   Student také volil vhodné postupy pro dosažení všech šesti dílčích cílů, které je možno shrnout následujícím způsobem:

   i) návrh efektivního přístupu pro tvorbu jazykově závislých modulů s minimální supervizí;
   ii) návrh a implementace sady nástrojů pro podporu automatizace potřebných úloh;
   iii) průzkum možností využití strojového učení pro sběr a anotaci dat pro tvorbu akustických modelů;
   iv) aplikace navržených postupů pro všechny slovanské jazyky;
   v) praktické využití metod v doméně automatického přepisu a monitoringu médií ve všech slovanských jazycích;
   vi) ověření navržených metod a postupů na dalších vybraných neslovanských jazycích.

   Definované cíle jsou jako celek velmi ambiciózní a vyžadují na straně jedné dostatečné pochopení problematiky a na straně druhé velké časové nároky. Po prostudování práce musím s potěšením konstatovat, že byly všechny **uvedené cíle splněny**.

3. **Využití výsledků práce v praxi**

   Za **největší přínos** disertační práce považuji její praktické využití. Práce je z větší části inženýrského charakteru, nenavrhuje nové metody a algoritmy, ale spíše využívá / adaptuje / spojuje existující osvědčené přístupy pro další jazyky. Nicméně výsledky této práce byly reálně nasazeny do praxe partnerskou firmou Newton technlogies v rámci projektů *Multilinmedia* a *DeepSpot*.

4. **Formální stránka práce**

   Předložená práce je napsaná v češtině a skládá se z devíti kapitol + přílohy. Celkem se jedná o 117 stran textu. Práce je dobře členěna. Mám jen výtku k umístění cílů práce, které jsou uvedeny v kap. 3 (str. 37). Bylo by vhodnější cíle práce umístit na začátek práce.

   Po jazykové stránce je disertace na velmi dobré úrovni, neobsahuje překlepy ani pravopisné chyby. Jen některé použité formulace nejsou vhodné pro odborný text (např. „ .. vzešlého z akustického a jazykového modelu “ na str. 23 nebo „ .. jsou řešeny moduly a nástroje .. “ na str. 54).

   Dále obsahuje několik drobných formálních nedostatků: nepřeložené anglické termíny (např. sekce 5.1.6 Textový *preprocessing*), některá tvrzení by bylo vhodné doplnit referencemi (např. str. 39, sekce 4.1) a první číslovaná stránka by měla být kap. 1 Úvod.

5. **Publikační aktivita studenta**

   Autor publikoval výsledky své práce celkem ve **dvanácti článcích** na mezinárodních konferencích, z nichž dvě jsou prestižní konference Interspeech, která je hodnocena jako „A“ podle portálu CORE. Tento počet publikací převyšuje požadavky kladené na studenta doktorského studia. Kvalita publikací je potvrzena **výborným citačním ohlasem** (13 x DB Scopus).

   Na druhou stranu v uvedeném seznamu postrádám alespoň jeden časopisecký článek. Dále mě lehce zaráží, že poslední publikace je z roku 2018. Proč autor nepublikoval nic později?

6. **Připomínky / Dotazy**

   - V Prohlášení autor mylně uvádí „bakalářskou práci“, ačkoli se jedná o práci disertační.
   - Řada termínů a pojmů by si zasloužila přesnější popis vč. souvisejícího matematického aparátu (např. jazykový model - str. 21 - matematický popis je pouze pro bigramový model a to na str. 53, nebo Skrytý Markovův model).
   - Na str. 23 autor popisuje modulární ASR systém. V popisu obrázku uvádí, že se systém skládá z „několika modulů“. Práce je odborný text, proto je třeba počet uvést přesně.
   - V sekci 1.6 (str. 25 a 26) jsou definovány metriky WER, OOV a OOL. Další evaluační metriky jsou uvedeny v sekci 6.4 (str. 70 a 71). Proč nejsou uvedeny všechny dohromady na jednom místě práce? Navíc u metrik OOV a OOL chybí matematický popis. Jako další důležitá metrika je F-míra, která v práci použita není. Proč jste tuto metriku nepoužil?

- V práci mi obecně chybí srovnání se souvisejícími pracemi. Připravte si toto srovnání pro jednotlivé úlohy, které jste řešil zejména pak pro:
  - identifikaci jazyka (str. 49).
  - výsledky ASR ve více jazycích (tab. 7.3 a 8.4).
- Text v některých tabulkách (např. 5.6, 5.7 nebo 6.1) by si zasloužil český překlad.
- V dnešní době považuji za zbytečné provádět srovnání rozpoznávání řeči s GMM a DNN, když je zřejmé, že DNN bude mít lepší výsledky.
- Na str. 81 uvádíte, že testovací sady jsou veřejně dostupné. Máte představu, jak jsou tyto sady v ASR komunitě využívány (např. počet stažení nebo relevantní citace)?

7. **Shrnutí**

   Na závěr konstatuji, že disertační práce pana Ing. Radka Šafaříka je samostatná vědecká práce, která obsahuje řadu nových výsledků. Práci proto **DOPORUČUJI** k obhajobě.

V Plzni dne 7. února 2021

doc. Ing. Pavel Král, Ph.D.
Katedra informatiky a výpočetní techniky
Západočeská univerzita v Plzni

POSUDEK NA DISERTAČNÍ PRÁCI

Téma práce: Multilingvální systémy rozpoznávání řeči a jejich efektivní učení.

Doktorand: Ing. Radek ŠAFAŘÍK

Posudek vypracoval: Doc. Ing. Petr POLLÁK, CSc.
ČVUT FEL K13131, Technická 2, 166 27 Praha 6

Předložená práce ing. Šafaříka se věnuje problematice multilingválního rozpoznávání řeči, která nachází v současné době mnohé aplikace v systémech hlasových technologií. Uvedená problematika je velmi aktuální a je intenzivně řešena v celosvětovém měřítku. Přestože mnohé systémy dosahují již velmi nízké chybovosti, je nepochybné, že je v dané problematice stále velký prostor pro zlepšení a disertabilní výzkum na různých úrovních, zejména pro řešené slovanské jazyky.

Po úvodu a přehledném shrnutí aktuálního stavu řešené problematiky definuje autor hlavní cíle práce, tj. studium a vývoj efektivních metod učení jazykově závislých modulů s užším zaměřením na slovanské jazyky. Stanovené cíle jsou jistě disertabilní a byly v předložené práci bezpochyby naplněny. Nejvýznamnější přínosy lze shrnout v následujících bodech.

- V první řadě autor podává ucelený pohled na současný stav řešené problematiky. Zmiňuje nejvýznamnější práce realizované v dané oblasti a souhrn prezentovaný v 2. kapitole je bez nadbytečných detailů velmi zdařilým a srozumitelným přehledem standardně používaných technik rozpoznávání spojité řeči s užším zaměřením na multilingvální systémy. Forma zpracování této části potvrzuje autorův nadhled nad řešenou problematikou.
- Ve 4. kapitole pak autor podává přehled vlastností a charakteristik slovanských jazyků, které jsou důležité pro účely rozpoznávání řeči. Tento stručný a přehledný popis je dalším z přínosů předložené práce, neboť na několika stranách jsou stručně popsány všechny slovanské jazyky z hlediska ortografie, fonetiky i morfologie. V přílohách jsou pak uvedeny systematicky úplné tabulky používaných ortografických znaků i přehledy fonetického obsahu všech slovanských jazyků. Toto považuji za velmi účelné a použitelné v širší komunitě řešící tuto problematiku na vědecké i aplikační úrovni. Čistě formálně bych pro přímější použití v implementacích řečových rozpoznávačů osobně doporučil ve fonetických tabulkách reprezentaci pomocí počítačově čitelné abecedy X-SAMPA (místo použité IPA). Autor později pro tyto účely používá interní fonetickou abecedu definovanou na školícím pracovišti, ta je ovšem optimalizovaná pro vývojáře s rodilým jazykem češtinou a není tedy úplně vhodná pro použití na mezinárodní úrovni.
- Hlavním přínosem práce je jednoznačně vyvinutá metodika a implementace tvorby jazykově závislých modelů rozpoznávácích systémů, popisovaná v kapitolách 5 a 6. Autor nejprve prezentuje obecnou metodiku tvorby textového korpusu pro nový jazyk, kde byl nucen řešit širokou škálu problémů počínaje vlastním stahováním dat z dostupných veřejných zdrojů (především WEBových stránek), přes filtraci a normalizaci textů, až po detekci jazyků či jejich vnitřní kódování. Z textových korpusů pak generuje autor potřebné slovníky a jazykové modely. Samostatnou problematikou je generování výslovnosti pro jednotlivá slova ve slovníku pomocí G2P převodníku na bázi produkčních pravidel.

  Klíčovou částí je pak především tvorba databáze nahrávek pro trénování akustického modelu (AM). Autor opět prezentuje systematickou metodiku zahrnující řešení tématicky různých problémů, od hledání zdrojů, přes vlastní získávání dat, jejich těžení, využití různých AM modelů, až po identifikaci cizího jazyka ve zpracovávaných nahrávkách.

  Uvedená obecně použitelná metodika je demonstrovaná na 3 východoslovanských jazycích (RU, UK, BE), přičemž jazykově závislé moduly řečového rozpoznávače pro UK a BE jsou

výhradním dílem autora. Shromáždění dostatečného množství dat textových i akustických, včetně důležité definice pravidel pro G2P, je určitě významným přínosem předložené práce.

K této části bych měl několik otázek. V textu zmiňujete v dílčích krocích určitý potřebný podíl manuální práce resp. korekce, dále je pak pochopitelná určitá znalost daného jazyka včetně jeho fonetiky. Zajímalo by mě jaké bylo množství nutných manuálních (expertních) zásahů při tvorbě jazykově závislých modulů pro RU, UK a BE? Jaká je Vaše znalost těchto tří jazyků?

Dále při vytváření databáze akustických dat zmiňujete dostupnost dat v různých formátech, mimo jiné v MP3/MP4. Jaké bylo procento stažených dat v tomto formátu, jehož ztrátová komprese obecně nepříznivě ovlivňuje přesnost rozpoznávání? Máte také přibližný odhad, do jaké míry toto mohlo ovlivnit výsledky prezentované v tabulce 7.3? Podobnou otázku bych měl i k množství a vlivu více zarušených promluv (hudba na pozadí, větší šum, apod.), které se vyskytují ve vytvořených testovacích sadách a předpokládám, že stejným způsobem jsou zastoupeny i v sadách trénovacích.

- Ve svém výzkumu se autor nespokojil jen s vlastní tvorbou požadovaných jazykových modulů resp. kompletací potřebných dat. Oceňuji také analýzu vlivu možných nepřesností ve vytvořených modulech/datech popisovanou v kapitole 6.7. Uvedené experimenty simulující možné chyby ve zkompletovaných datech demonstrativně ukazují na větší či menší vliv některých chyb na výslednou přesnost rozpoznávače vytvořeného z daných dat.
- Autor dále popisuje souhrnné výsledky rozpoznávání celkem pro 11 slovanských jazyků, kromě češtiny a slovenštiny se autor podílel na tvorbě všech ostatních jazykově závislých modulů. Dozažené výsledky odpovídají očekávání. Výsledky s nově vytvořenými moduly pro UK a BE jsou sice horší, to je však pochopitelné pro první vývojové verze. Jako nezanedbatelný přínos je nutno zmínit i vytvořené standardizované testovací sady pro studované slovanské jazyky, které jsou veřejně dostupné na CESNET cloudu.

  V 8. kapitole autor prezentuje i výsledky aplikace popisovaných metod na další 3 neslovanské jazyky, a to španělštinu, lotyšštinu a albánštinu. Dosažené výsledky potvrzují obecnou použitelnost navržených metod pro tvorbu jazykových modulů rozpoznávačů řeči.
- Zvláště ocenitelná je nakonec i skutečnost, že vytvořené moduly byly použity nejen v reálném provozu vývojových systémů na školícím pracovišti autora, ale jsou také postupně nasazovány do praxe v komerčních aplikacích monitorování médií, automatického titulkování či v obecných diktovacích systémech partnerské firmy Newton technologies, as.

Po formální stránce je předložená práce na velmi dobré úrovni, je přehledně členěna, text je psaný velmi pěknou češtinou, grafická úprava je na dobré úrovni. Práce obsahuje jen minimum typografických chyb, které jistě nemá smysl konkrétně zmiňovat.

Originální přínos autora v dané oblasti potvrzuje i kvalitní publikační činnost autora zmiňující 12 publikací (v 6 případech jako první autor). 2 příspěvky byly publikovány na prestižních konferencích řady Interspeech, avšak i ostatní publikace jsou z kvalitních mezinárodních konferencí a workshopů, jejichž sborníky jsou zaindexovány v prestižní databázi WoS.

Na základě výše uvedených skutečností lze tedy konstatovat, že předložená práce popisuje originální výsledky vědecké práce, které jsou využitelné na mezinárodní úrovni. Autorem zvolený "technicko pragmatický přístup" (slovy autora) považuji za velmi zdařilý, jistě pak pro dosažení výsledků v prezentované šíři. Práci **jednoznačně doporučuji** k obhajobě za účelem získání vědecké hodnosti doktora na Technické univerzitě v Liberci.

V Praze dne 22. února 2021