

I. IDENTIFIKAČNÍ ÚDAJE

Název práce:	Named Entity Recognition Using Recurrent Neural Networks
Jméno autora:	Hoang Long Nguyen
Typ práce:	Bakalářská práce
Fakulta/ústav:	Fakulta elektrotechnická
Katedra/ústav:	Katedra cybernetics
Oponent práce:	Mgr. Jan Hajič
Pracoviště oponenta práce:	Ústav formální a aplikované lingvistiky, MFF UK

II. HODNOCENÍ JEDNOTLIVÝCH KRITÉRIÍ

Zadání	A (výborně)
<i>Hodnocení náročnosti zadání závěrečné práce.</i>	
<p>Student se rozhodl vypracovat výzkumnou práci spíše než implementační, což samo o sobě práci řadí mezi obtížnější. Náročnost zadání dále hodnotím ze čtyř hledisek: obtížnost řešené úlohy, náročnost použití zvolených metod, složitost implementace, a jelikož se v práci užívá strojové učení, také náročnost práce s použitými daty.</p> <p>Řešitel zpracovává jednu z podstatných otevřených úloh zpracování přirozeného jazyka (NLP). Rozpoznávání pojmenovaných entit (vlastních jmen) je kritické pro mnoho praktických aplikací NLP (např. dialogových systémů), neboť tyto „kotví“ konverzaci do skutečného světa a umožňují tak systémům interakci s objekty, které uživatelé pojmenovávají (např. automatická rezervace místa v restauraci, vyhledání spojení v MHD přes hlasové rozhraní, či sledování zpráv o jednotlivých osobnostech). Obtížnost úlohy je daná tím, že (a) vlastních jmen je velmi mnoho, avšak (b) až na nejčastější výjimky (názy hlavních měst, prezidentů, velkých firem...) se vyskytují velmi zřídka, (c) používají často slova obvyklá mimo pojmenování (např. názvy ulic: „Dlouhá“), (d) víceslovná pojmenování entit nutí jít nad rámec příznaků na úrovni jednotlivých tokenů (např. používání velkých písmen). Navíc je v češtině nutné se vypořádat s bohatou morfologií. Z hlediska zadání úlohy tedy práci hodnotím jako velmi náročnou.</p> <p>Zároveň řešitel používá metody tzv. hlubokého učení. Tyto se v nedávné době dostaly do centra pozornosti a s tím i vzrostl počet open-source nástrojů, které jejich využívání zjednodušují, avšak je stále relativně náročné se s těmito metodami seznámit a umět je úspěšně použít na daný problém, neboť se jedná o vysoce nelineární modely s často neintuitivním chováním. (Zvláštní zmínku si však zaslouhuje práce s literaturou: většina článků, které v oblasti hlubokého učení vychází, je psána poměrně „kondenzovaným“ způsobem, jednotliví autoři často odkazují na své předchozí výsledky, a přesné čtení tak vyžaduje určitou zběhlost v oboru.) Z hlediska použitých metod práci tedy hodnotím jako náročnou.</p> <p>Díky využití open-source nástrojů implementujících metody hlubokého učení (což je samozřejmě správná volba) je implementační náročnost neuronových sítí relativně nízká. Stojí však za zmínku, že extrakce příznaků z lingvistických dat – a obzvláště z češtiny – je často problematictější, než se z jejího popisu (sekce 4.2.5, 5.1, 5.2) může zdát. Z hlediska implementace hodnotím práci jako netriviální, avšak stále relativně jednoduchou.</p> <p>Používá se standardní dataset CNEC, čímž se autor nevystavuje problematice sběru dat a definice „pravdy“, tj. kategorizace pojmenovaných entit. To je moudré rozhodnutí, neboť tvorba datasetu pro tuto úlohu v potřebném rozsahu by vydala na samostatnou práci a není to dobrá praxe kvůli replikovatelnosti. Z hlediska práce s daty je práce tedy jednoduchá.</p> <p>Celkově hodnotím práci jako nadstandardně náročnou.</p>	

Splnění zadání

B (velmi dobře)

Posud'te, zda předložená závěrečná práce splňuje zadání. V komentáři případně uveďte body zadání, které nebyly zcela splněny, nebo zda je práce oproti zadání rozšířena. Nebylo-li zadání zcela splněno, pokuste se posoudit závažnost, dopady a případně i příčiny jednotlivých nedostatků.

Práce splňuje zadání beze zbytku, na druhou stranu nepřekračuje jeho rámec. Vyhodnocení použitých metod by se dalo doplnit o detailnější rozbor úspěšnosti na jednotlivých kategoriích pojmenovaných entit a analýzu chyb, což možná mohlo pomoci autorovi v průběhu řešení zlepšovat rekurentní modely nejen standardními postupy jako úspěšně použitá batch-normalization, nýbrž i pomocí lingvistických postřehů. Rozhodnutí pro implementaci neuronových sítí zvolit knihovnu Keras místo přímého použití frameworku TensorFlow považuji za správné rozhodnutí, neboť řešitel nenavrhuje nové modely, jež by knihovna Keras neposkytovala.

Mírně pochybnosti mám pouze o splnění bodu 2, "Propose a suitable neural network architecture based on previous research" (zvýraznění dodáno). Vztah k dosavadnímu výzkumu mohl být v kapitole 5 popsán explicitněji.

Zvolený postup řešení

B (velmi dobře)

Posud'te, zda student zvolil správný postup nebo metody řešení.

Zvolený postup je adekvátní, plně v souladu s metodologií oboru. Volba neuronových sítí pro řešení úlohy rozpoznávání pojmenovaných entit je logická, neboť tyto metody dnes běžně v různých úlohách zpracování přirozeného jazyka dosahují nejlepších (a stále lepších) výsledků, a stejně tak i konkrétně ve zvolené úloze. Je také třeba vyzdvihnout úsilí věnované zajištění solidního výchozímu modelu CRF, proti kterému se měří případné zlepšení dosažené pomocí neuronových sítí. Pouze ve vyhodnocování výsledků je poněkud škoda chybějící analýzy jednotlivých typů chyb z lingvistického hlediska, která mohla pomoci výsledky zlepšit.

Odborná úroveň

A (výborně)

Posud'te úroveň odbornosti závěrečné práce, využití znalostí získaných studiem a z odborné literatury, využití podkladů a dat získaných z praxe.

Po odborné stránce je práce jednoznačně na velmi dobré úrovni. Řešitel pronikl do problematiky hlubokého učení a orientuje se v ní, a prokázal schopnost provést replikovatelný pokus a dostatečně jej popsat. V tomto směru je mojí jedinou připomínkou, že v práci není popsáno rozdělení dat na trénovací, validační a testovací množinu. Použité datasey sice poskytují data rovnou rozdělená do těchto sad, takže chybějící údaj neohrožuje replikovatelnost pokusů, bylo by však vhodné tuto skutečnost v práci uvést, spolu s velikostmi jednotlivých množin.

Formální a jazyková úroveň, rozsah práce

A (výborně)

Posud'te správnost používání formálních zápisů obsažených v práci. Posud'te typografickou a jazykovou stránku.

Po formální stránce práce splňuje všechny náležitosti. Jazyková úroveň je velmi dobrá a stylisticky je práce také velmi kvalitní. (Jedinou podstatnější stylistickou připomínkou mám k abstraktu: je nevhodné v abstraktu používat budoucí čas.) Jak je u bakalářských prací poměrně obvyklé, v textu se nachází drobné formální nedostatky, které se u běžných recenzovaných článků řeší camera-ready verzí (např. popis obrázku 3.1 má neúplnou poslední větu, chybí diakritika u bibliografického údaje 36, neodmazané „We believe—“ na s.30). Žádné z těchto přehlédnutí však nebrání srozumitelnosti textu.

Rozsah práce je dostatečný a zároveň práce není zbytečně dlouhá.

Výběr zdrojů, korektnost citací

A (výborně)

Vyjádřete se k aktivitě studenta při získávání a využívání studijních materiálů k řešení závěrečné práce. Charakterizujte výběr pramenů. Posuďte, zda student využil všechny relevantní zdroje. Ověřte, zda jsou všechny převzaté prvky řádně odlišeny od vlastních výsledků a úvah, zda nedošlo k porušení citační etiky a zda jsou bibliografické citace úplné a v souladu s citačními zvyklostmi a normami.

Autor využívá širokou paletu zdrojů z odborné literatury: jak starší články či knihy, které představují základní literaturu oboru, a state-of-the-art literaturu. Vzhledem k tomu, že práce využívá metody hlubokého učení vyvinuté během několika posledních let, podstatná část použité literatury jsou pak konferenční příspěvky na relevantních fórech, neboť ještě nejsou napsané přehledové práce ani učebnice. Rozsah bibliografie považují za nadstandardní, především co se literatury o nových postupech u neuronových sítí týče.

Student korektně používá citační a bibliografický aparát. Zacházení se „známými fakty“, tj. široce přijímanými skutečnostmi, u kterých není nutné uvádět citaci, také odpovídá zvyklostem v oboru. Výjimečně je citace ve formě [XY] nevhodně zasazena do textu (s.3: „...using a maximum entropy classifier by [22]“ → „...using a maximum entropy classifier [22]“, případně uvést v textu jména autorů: s.6 „There are also special tags defined in [36]...“ → „Ševčíková et al. [36] also define special tags...“). V bibliografii je pouze jediný formální nedostatek: v bibliografickém údaji [34] je uveden pouze autor a název práce. Všechny ostatní bibliografické údaje jsou úplné a v souladu se zvyklostmi v oboru.

Další komentáře a hodnocení

Vyjádřete se k úrovni dosažených hlavních výsledků závěrečné práce, např. k úrovni teoretických výsledků, nebo k úrovni a funkčnosti technického nebo programového vytvořeného řešení, publikačním výstupům, experimentální zručnosti apod.

Neúspěch neuronových sítí trénovaných autorem ve srovnání se state-of-the-art článkem Straky a kol. souvisí především se zkušeností, a na úrovni bakalářské práce tím pádem nelze očekávat, že student bude konkurovat nejlepším výsledkům na příslušné problému, obzvláště pokud se jedná o metody hlubokého učení. Volby architektury a hyperparametrů těchto modelů jsou často intuitivní, neboť v oboru je stále mnoho teoretických mezer a state-of-the-art iterativní optimalizační algoritmy z rodiny Stochastic Gradient Descent (používané kvůli absenci analytických řešení či aproximací) trvají příliš dlouho na to, aby se reálně daly použít hyperoptimalizační postupy založené např. na evolučních algoritmech. (Autor práce správně poznamenává v sekci 4.3.4.1, že výzkum hyperoptimalizace architektury neuronových sítí probíhá – je však třeba dodat, že to je výzkum extrémně obtížný a citovaný článek je v době odevzdání práce pouze šest měsíců starý.)

Proto ale myslím, že by práci prospěla analýza chyb – které lingvistické jevy jsou zodpovědné za problémy modelů RNN? Spadají chyby do jasně vymezených skupin, nebo jsou rozloženy vůči kategoriím pojmenovaných entit náhodně? Hodilo by se i srovnání těchto aspektů chyb mezi CRF a RNN, neboť se jedná o modely operující odlišným způsobem. Mají příznaky jako *is_capitalized* podobný vliv na distribuci chyb v CRF jako v RNN? Podobné analýzy nejsou vždy v článcích o hlubokém učení obvyklé, avšak pro úlohy zpracování přirozeného jazyka je přirozené takové rozborů provádět. Navíc na rozdíl od spekulací o nejasném vlivu jednotlivých rozhodnutí v návrhu neuronových sítí jako velikost vnitřních vrstev je taková analýza relativně jasně interpretovatelná a přináší tak silnější, konkrétnější závěry. (Toto je podnět jak pro případný navazující výzkum autora práce, tak pro jejího vedoucího.)

III. CELKOVÉ HODNOCENÍ, OTÁZKY K OBHAJOBĚ, NÁVRH KLASIFIKACE

Shrňte aspekty závěrečné práce, které nejvíce ovlivnily Vaše celkové hodnocení. Uveďte případné otázky, které by měl student zodpovědět při obhajobě závěrečné práce před komisí.

Student se rozhodl svou bakalářskou práci směřovat k výzkumu v oboru zpracování přirozeného jazyka. Zvolil poměrně náročné otevřené téma rozpoznávání pojmenovaných entit a pro řešení vybral na základě prostudování situace v oboru pokročilé metody hlubokého učení, které dnes dosahují nejlepších výsledků.

Nejsilnější stránkou práce je skutečnost, že řešitel prokázal schopnost vykonávat vědeckou práci v souladu s metodologickými zvyklostmi oboru zpracování přirozeného jazyka. Baseline dosažená pomocí metody CRF je solidní a autor provedl dále relativně velké množství pokusů s rekurentními neuronovými sítěmi – neúspěch rozhodně není důsledek nedostatečné snahy, spíše naopak ilustruje, jak je nečekaně obtížné obecné modely RNN přizpůsobit pro řešení konkrétní netriviální úlohy.

Slabší stránky práce jsou dvě. Pro analýzu souvisejících prací byla zvolena velmi hutná forma textu, která podstatu dosavadních přístupů shrnuje jen velmi stručně a v podstatě očekává od čtenáře, že si příslušné články přečte. Toto vnímám jako určitý „průsak“ stylu konferenčních článků, které mají výrazně omezený rozsah – v bakalářské práci, která rozsah omezený nemá, by však bylo vhodnější související práce rozebrat podrobněji, tedy například uvést, jakým způsobem state-of-the-art článek od Straky a kol. využívá rekurentní sítě na úrovni jednotlivých znaků. Ocenil bych také podrobnější srovnání, co řešitel očekává od konkrétních zvolených architektur – jaké mohou mít oproti existujícím řešením výhody? (Správnost, rychlost, nižší paměťové nároky?)

Druhá slabší stránka je již zmiňované opomenutí podrobnější analýzy chyb, a s tím souvisí otázka, kterou řešiteli k obhajobě pokládám:

Jak vypadá rozložení chyb na jednotlivých kategoriích pojmenovaných entit při použití nejlepšího modelu CRF vs. nejlepší RNN? (Případně: jsou nějaká výrazně frekventovanější sousloví, která dělají potíže?) Lze v chybách vysledovat nějaké notoricky problematické vzorce a navrhnout způsoby zlepšení, ať už pomocí nových lingvisticky motivovaných featur, nebo pomocí změn architektury, případně dat? Je kupříkladu chybovost v dlouhých větách vyšší než v krátkých?

Předloženou závěrečnou práci hodnotím klasifikačním stupněm **A (výborně)**

Datum: 4. 6. 2017

Podpis: