Vyjádření vedoucího disertační práce

Onřej Dušek: Novel Methods for Natural Language Generation in Spoken Dialogue Systems

Ondřej Dušek se ve své práci zaměřil na generování přirozeného jazyka (NLG) v dialogových systémech, což je jedna ze základních aktivit nutných pro tvorbu konverzačních agentů. V této práci p. Dušek navrhnul několik originálních metod s cílem zlepšit adaptační schopnosti NLG. Mezi hlavní výsledky patří využití perceptronového rerankeru, který se dá natrénovat bez zarovnaných trénovacích data (promluv a jejich sémantické reprezentace) a užití neuronového generátoru, který se adaptuje na předchozí vstup uživatele.

Popis práce:

Práce má 147 stran a je členěna na úvod, sedm obsahových kapitol, závěr, poděkování, literaturu, seznam zkratek, a obsah. V úvodu autor popisuje základní model dialogového systému, možnosti reprezentace významu promluvy a kam v dialogovém systému zasadit generování přirozeného jazyka. Dále se v úvodu věnuje základní motivaci pro předkládanou práci, cílům práce, a dosažených výsledkům. Ve druhé kapitole autor popisuje základní přístupy k NLG, adaptivní NLG, hlavní zdroje dat pro NLG. Ve třetí kapitole ve větším detailu popisuje jednotlivé části generování mluveného jazyka, jako jsou reprezentace významu, použití nezarovnaných dat, delexikalizace, hluboká syntaktická reprezentace, a evaluační metriky. V čtvrté kapitole autor již popisuje první experimenty provedené na téma povrchové realizace, kdy generování probíhá z tzv. t-stromů. Tento povrchový generátor se používá v dalších kapitolách. V rámci tohoto převáženě pravidlového generátoru se používá originální trénovatelný morfologický generátor založený na logistické regresi. V páté kapitole se autor věnuje experimentům generátoru promluv založeným na perceptronu, který se trénuje z nezarovnaných dat. Vstupem je sémantická reprezentace promluvy a výstupem je t-strom, který se následně syntetizuje pomocí povrchové ho generátoru (popsaného v předchozí kapitole) do textové podoby. V šesté kapitole autor popisuje přístup generování pomocí neuronových sítí založených na sekvenčních modelech. Autor experimentuje jak s generováním do t-stromů tak přímo do textu. V sedmé kapitole autor popisuje generátor promluv, který se adaptuje na předchozí promluvy uživatele. Pro tyto účely byl vytvořen nový zdroj trénovacích data. V osmé kapitole je detailně probráno generování českého jazyka pomocí předchozích metod, srovnána obtížnost generování vzhledem k experimentům s angličtinou. Závěr shrnuje přínosy předložené práce, naplnění jednotlivých cílů a publikace kde jsou jednotlivé výsledky uveřejněny.

Hodnocení:

Práce je psána výbornou angličtinou a přístupnou formou. Oceňuji zejména časté shrnutí toho, co bude popsáno nebo co zrovna bylo diskutováno. Tento přístup ocení zejména čtenář, který nečte práci tzv. od začátku do konce, ale na přeskáčku. Autor v kapitolách 4 – 8 popisuje vlastní experimenty na přípravě dat nebo implementaci jednotlivých metod. Přínos práce je zjevný, metody dobře popsány, kdy jejich popis umožňuje re-implementaci. Vytvořený software je volně dostupný spolu se všemi pořízenými daty. Objem a kvalita prezentované práce je více než dostatečná. Je nutné poznamenat, že tato práce neobsahuje popis všech aktivit autora na UFALu, kde se navíc intenzivně věnoval strojovému překladu například v EU projektu QT Leap. Z formálního hlediska práce vyhovuje požadavkům a standardům ohledně disertační práce: formát práce je obvyklý, grafická úprava dobrá, seznam zkratek úplný, seznam literatury je relevantní.

Dotaz:

V práci srovnáváte pravidlové systémy, částečně pravidlové, statistické systémy včetně tzv. „end-to-end“ systémů založených na neuronových sítích na jazycích, jako je angličtina a čeština. Mohl byste při obhajobě práce zhodnotit vhodnost tzv. „end-to-end“ systémů pro morfologicky bohaté jazyky jako je čeština, a to s ohledem na kvalitu výstupu a množství trénovacích data? Dále by mě zajímal Váš názor na úvahu, že na český jazyk (morfologicky bohatý) není možné jednoduše aplikovat přístupy vyvinuté například pro angličtinu v kontextu vaší práce.

Závěr:

Disertační práci Ondřeje Duška považuji za mimořádně kvalitní a významně převyšující průměr. Toto hodnocení se týká jak objemu odvedené práce, její kvality, tak i počtu publikovaných prací na mezinárodních prestižních konferencích. Tato práce v mnohém posunula současný stav poznání a to lze doložit již existujícími citacemi na publikované práce autora, tak i zájmem o vyvinutý software nebo pořízená data. Na základě tohoto doporučuji práci k obhajobě a věřím, že práce splňuje všechna kritéria kladená na disertační práci v oboru Matematická lingvistika na MFF UK v Praze.

Praha, 29. 5. 2017, Filip Jurčíček