Posudek návrhu disertační práce

Posudek vypracoval: Doc. RNDr. Ondřej Bojar, Ph.D.; bojar@ufal.mff.cuni.cz
ÚFAL MFF UK, Malostranské náměstí 25, Praha 1, 181 00

Dne: 15. 8. 2021

Název práce: Domain Adaptation for Natural Language Generation
Řešitel: Ing. Zdeněk Kasner
Vedoucí: Mgr. et Mgr. Dušek Ondřej, Ph.D.
ÚFAL MFF UK

Návrh disertační práce Zdeňka Kasnera se zabývá generováním přirozeného textu ze strukturovaných dat. Speciálně je pak zaměřen na moderní techniky hlubokého strojového učení, modely typu sequence-to-sequence, včetně velkých předtrénovaných modelů. Nová řešení autor již v prvních letech doktorského studia navrhl pro generování textu plynulého a přesného (tj. vyjadřujícího právě zadaná fakta) a pro vyhodnocování přesnosti systémů generování. Plán dalšího výzkumu si klade za cíl sjednotit způsob zápisu strukturovaných dat napříč doménami, zahrnout do procesu generování symbolické operace a explicitně značit, které úseky textů vycházejí ze kterých zadaných fakt, viz body (iv) až (vi) v návrhu.

Návrh je psán bezchybnou čtivou angličtinou, je dobře strukturován a ilustrace procesů jsou velice užitečné pro pochopení textu. Snaha o stručnost bohužel v některých popisovaných starších experimentech bohužel vede k textu až příliš hutnému a srozumitelnému jen při velmi důkladném a pomalém čtení. Pro účel těchto „tezí“ to není velký problém, v textu disertace bych doporučil věnovat metodám větší prostor (a nebo pečlivěji vážit, co je hlavní novinka, kterou chce autor představit, a zajistit, že právě popis této novinky bude úplný, plně vysvětlený a zcela plynulý, přičemž ostatní okolnosti se ještě více zestruční nebo zcela vynechají).

Postup i plán dalších pokusů shledávám velmi zajímavým a dobře navrženým. Mám několik otázek, nebo snad spíše námětů do diskuse:

- Váhám, zda cíl (v) „vykonávat v průběhu generování i symbolické operace“ opravdu patří do oblasti NLG. Vnímám úlohu NLG úžeji a uvedu-li to v extrémní analogii, překvapilo by mne, kdyby „moje ruka kontrolovala, zda píše správně vypočtený součet čísel“. Jinými slovy, myslel bych si, že symbolické operace jako vyvozování tvrzení, která nebyla v zadání, faktická kontrola proti externím zdrojům dat a logické nebo aritmetické výpočty patří do dřívějších fází přípravy strojové promluvy.
  Bude takto rozšířený model schopen lhát, tj. generovat fakticky chybné texty? Není to žádoucí, ale pro různá vyhodnocování ap. se to může hodit. A naopak, co až takto rozšířený model zalže, na čí straně bude „zodpovědnost“?
  Možná ale právě tento posun, že část sémantické kontroly by mělo převzít NLG, je součástí přirozeného vývoje a třeba význaným způsobem pomůže bojovat proti předpojatosti (biasu) velkých předtrénovaných modelů.
- Nápad iterativního vylepšování textu se mi velmi líbí. Konečná realizace používá specializované modely, které navrhují editační operace, ale také by mohlo být zajímavé použít prostě sequence-to-sequence to přímé generování nových formulací. A mohlo by být zajímavé postupovat ještě dál tímto směrem, kdy by i další vstupy do modelu byly zadávány rovnou textově, tj. sjednotit různé typy původně strukturované informace do sdělení zapsaných prostě přirozeným jazykem.
- Multijazykovost v sekci 4.2 není úplně jasně vysvětlena, u popisu WebNLG se o více jazycích nemluví. Chápu správně, že fakta, tj. trojice, zůstaly naprosto identické z anglické verze WebNLG a ruští mluvčí k nim doplnili ruské formulace, které tato fakta vyjadřují? Takový dataset je zajímavou výzvou z více důvodu, např. vlastní jména musí systém přeložit nebo alespoň transliterovat.
  Dále pozor, o více doménách ve WebNLG nebylo nic řečeno, jen sekce 4.2 říká, že model pracoval dobře i na neviděných doménách.
- V souvislosti se zmínkou o nejasném popisku „populationMetro“ mne napadlo, do jaké míry byly tyto popisky vysvětleny při crowdsourcingu. Jak moc dělali mluvčí chyby v generování z důvodu nejasného zadání, neznalosti významu popisovaných trojic?

- V plánech budoucích je popis přirozeně poněkud vágnější, přesto i zde by velmi pomohly ilustrace nebo náčrtky postupů, podobně jako tomu bylo u provedených experimentů. Týká se to všech částí sekce 6.
- V sekci 6.1 popisu rozumím tak, že pravidlové (resp. triviální šablonové) výstupy budou následně vylepšeny silným jazykovým modelem z hlediska plynulosti. V čem spočívá zásadní rozdíl oproti sekci 4.1? V tom, že nyní půjde o víc datových sad s jinou formální strukturou?
- Nápad využít duality NLG a NLU je velmi zajímavý. Mám však obavu, že NLU bude generovat často velmi odlišné reprezentace od té, která byla výstupem NLG. I lidé totiž často z jednoho textu vnímají mnoho různých informací.

Drobnosti k textu:

- Introduction: „generate texts with fluency comparable to human-written text“. Doporučuji tvrzení zmírnit minimálně přidáním „of individual sentences“.
- Introduction: „...require a considerable amount of human expertise“? Neměl autor na mysli „supervision“? Z textu totiž není moc jasné, jak se ta lidská specificky oborová znalost do systémů vkládá.
- V titulku bych psal velkými písmeny všechny součásti pomlčkových slov, tedy „Pipeline-Based Approaches“.
- Obrázek 2: Velmi by pomohlo ilustrovat i textové části datasetů, zejm. sebrané popisy ve WebNLG.
- Předposlední odstavec sekce 4.1: Není jasné, co se myslí výrazem „a single extra triple“.
- Obrázek 4: Překlep „direcly“.
- Sekce 5: Myslím, že správnější je „on *the* sentence/token level“.
- Popis vlastního postupu v sekci 5.1 je příliš stručný, těžko srozumitelný.
- Obrázek 6: Význam zkratek „C“, „N“, „E“ je sice jasný, ale přesto by měly být uvedeny aspoň v sekci 5.1 u příslušných termínů.
- U výsledků 77 % a 91 % není příliš jasné, k čemu se vztahují. Bylo by zajímavé vidět i výsledky zvlášť pro vynechání a halucinaci.

Návrh disertační práce Zdeňka Kasnera pokládám za velmi přehledný popis studované oblasti. Zdeněk již nyní představil zajímavé realizované nápady a výsledky a i navržené další směry postupu se mi jeví jako velmi relevantní a atraktivní. Jak jsem naznačil, téměř očekávám, že vedlejším výsledkem Zdeňkovy práce budou velmi cenné obecné poznatky o chování velkých jazykových modelů, včetně těch předtrénovaných. Jednoznačně doporučuji návrh přijmout a ve výzkumu pokračovat.

V Praze dne 15. 8. 2021,

Ondřej Bojar