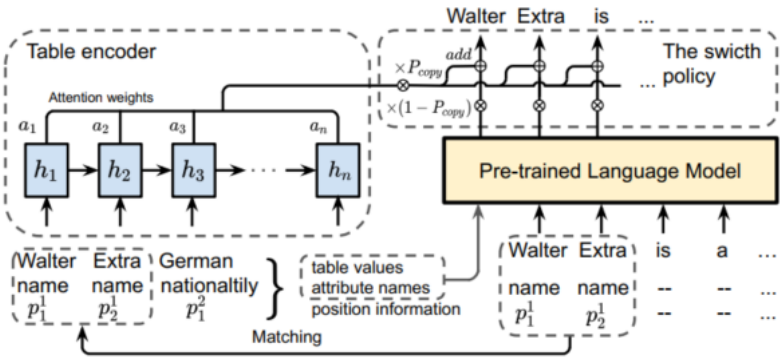
اعضای گروه: سید محمدعرفان موسوی‌منزه 401722199– غزل زمانی‌نژاد 401722244– محمد مصطفی رستم خانی 401722235

موضوع: *تولید متن زبان طبیعی با استفاده از مدل پیش آموخته به روش یادگیری با نمونه‌های معدود*

تولید متن توسط یک مدل مبتنی بر شبکه عصبی به صورت end-to-end (انتها به انتها) از یک داده ساختاریافته مانند جدول نیازمند دادگان زیادی است تا بتواند با حوزه جدید مطابقت یابد و این موضوع باعث می‌شود تا این کار با دادگان محدود در دنیای واقعی، امری دشوار باشد. با توجه به تمایل انسان به خلاصه‌سازی داده‌های جدولی، تسک جدید few shot natural language generation معرفی شده است. در این مقاله روش ساده ولی در عین حال تاثیرگذاری را پیشنهاد می‌دهیم که نه تنها کارایی را افزایش داده، بلکه تعمیم‌پذیری خوبی بین حوزه‌های مختلف فراهم می‌کند. طراحی معماری مدل بر پایه دو جنبه است: انتخاب محتوا از داده‌های ورودی و مدل‌سازی زبانی برای ترکیب و تولید جملات با معنی و منسجم. تنها با استفاده از ۲۰۰داده آموزشی در چند حوزه، نشان داده شد که روش مطرح شده در مقاله، به کارایی معقولی می‌رسد و می‌تواند بهترین مدل‌های پایه را در امتیاز BLEU با اختلاف 8 بهبود دهد. کدها و داده‌های مقاله در آدرس https://github.com/czyssrs/Few-Shot-NLG موجود است.

Input Table

Attribute (R)	Value (V)
Name	Walter Extra
Nationality	German
Occupation	Aircraft designer and manufacturer
...	...



مدل معرفی شده دارای دو قسمت است: یکی table encoder که داده‌های موجود در جدول را encode می‌کند و دیگری یک مدل زبانی پیش آموخته (pre-trained language model) است که برای تولید متن از آن استفاده شده است. بر اساس نوآوری مطرح شده در مقاله، مدل زبانی کلمه‌ای را که قرار است تولید کند، با احتمالی از جدول کپی و با احتمال مکمل، آن را تولید می‌کند. هدف از انجام این کار آن است که مدل زبانی اطلاعات نادرستی را تولید نکند.

مجموعه داده مورد استفاده در این مقاله WIKIBIO نام دارد که شامل ۷۰۰هزار مقاله از ویکی پدیای انگلیسی از اشخاص مشهور است. جعبه اطلاعات ویکی (Wiki infobox) به عنوان داده ورودی ساختاریافته و اولین جمله از مقاله به عنوان متن هدف در نظر گرفته می‌شود. علاوه بر آن برای نشان دادن تعمیم‌پذیری، یک مجموعه داده با دو حوزه جدید کتاب‌ها و آهنگ‌ها از خزش صفحات ویکی پدیا جمع‌آوری شده است. پس از فیلتر و پاکسازی، ۲۳۶۵۱ نمونه در حوزه کتاب و ۳۹۴۵۰ نمونه در حوزه آهنگ به دست آمد. از مجموع این دو حوزه همراه با بخش «افراد» دیتاست WIKIBIO، آزمایش‌هایی با تغییر اندازه داده آموزشی (سایز ۵۰، ۱۰۰، ۲۰۰ و ۵۰۰) انجام شد. سایر داده‌ها به عنوان داده اعتبارسنجی و آزمون مورد استفاده قرار گرفت.

مقاله دوم که مورد بررسی قرار می‌گیرد KGPT: Knowledge-Grounded Pre-Training for Data-to-Text Generation نام دارد. این مدل در دو بخش مورد بررسی قرار می‌گیرد: ۱) یک مدل generation مبتنی بر دانش برای تولید متن غنی‌شده با دانش ۲) یک پارادایم pre-train بر روی حجم زیادی از متن غنی شده با دانش که از وب خزش شده.

مقاله سوم که مورد بررسی قرار می‌گیرد، Few-shot Natural Language Generation for Task-Oriented Dialog است. در این مقاله یک معیار جدید به نام FewShotWoz معرفی کرده‌اند تا تنظیمات یادگیری few shot در سیستم‌های گفتگوی وظیفه محور را شبیه‌سازی کنند.