Abstract

- Language models often cannot predict entity names in text well due to their diverse vocabularies.
- We propose a language model for texts with entity names by leveraging the deterministic entity types.
- **Type Model**: predicts the type of the entity name
- **Entity Composite Model**: generates the conditional actual entity name given the entity type as prior
- We conduct experiments on two applications: Recipe generation and Code generation

Motivation

- Generating entity names in text is very challenging (e.g., olive oil, canola oil, grape oil, etc.—all are different varieties of oil).
- Existing language models:
  - fail to address this problem.
  - focus on copying/matching the entity names from the reference corpus
  - are based on generative models of low performance.

Methodology

Our model learns the probability distribution over the candidate words by decomposing it into two sub-problems:
- **Type Model** ($\theta_t$):
  - considers all of entities with the same type equal and represent them by their type information.
  - reduces the vocab size to a great extent.
  - predicts the type $s(w)$ of each entity $w$ more accurately.
- **Entity Composite Model** ($\theta_e$):
  - uses the entity type generated by the type model as a prior.
  - calculates the conditional probability distribution of the actual entity names.

### Figure 1. An illustration of our model Given a context, the type model (in bottom red block) generates the type of the words. Further, for that given context & type of each candidate generated by the type model, the entity composite model (in upper green block) generates the conditional actual entity names.

Experimental Results

- We create two benchmark datasets that contain many named entities.
- **Recipe Generation Corpus**:
  - 95,786 cooking recipes
  - Manually categorized 8 super-ingredients (i.e., types); e.g., proteins, vegetables, fruits etc.

References


Acknowledgement

This work was supported in part by National Science Foundation Grants IIS1760523, CCF-16-19123, CNS-16-18771 and an NVIDIA hardware grant.