

Latent Data to Document

August 17, 2018

1 Introduction

2 Problem

We would like to learn a conditional model over sentences $\mathbf{y} = \{y_0, y_1, \dots\}$ and latent structure \mathbf{z} given a table \mathbf{x} . We are primarily interested in the respective conditional distributions: both the posterior distribution over structure given a sentence and table $p(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$, which is an information extraction model, as well as the conditional distribution over summaries $p(\mathbf{y} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z} \mid \mathbf{x})} [p(\mathbf{y} \mid \mathbf{z}, \mathbf{x})]$.

3 Data to Document

The conditional copy model in Wiseman et al. (2017)

3.1 The Model

Let y_t be the current token, $\mathbf{y}_{0:t-1}$ all previous tokens,

4 Latent Table Model

4.1 Generative Model

We decompose structure into content selection and ordering respectively: $\mathbf{z} = \{\mathbf{c}, \pi\}$. Actually, content selection might be bad and we may want to perform content selection and ordering jointly.

Content Selection $p(\mathbf{c} \mid \mathbf{x})$ where $\mathbf{c} \in \{0, 1\}^n$ is a distribution over binary masks over relations. If a mask value is 1 then that specific relation is used to produce a summary.

Content Ordering $p(\pi \mid \mathbf{c}, \mathbf{x})$, where π is a permutation matrix. We may model this implicitly with a language model over relations, i.e. debugging. Error: we may have repeated records. It may be possible that certain records are only referred to a single time while we should be available for use multiple times.

Relation Realization $p(\mathbf{y} \mid \pi(\mathbf{c}), \mathbf{x}) = \prod_t p(y_t \mid \mathbf{y}_{<t} \pi(\mathbf{c})[t], \mathbf{x})$.

4.2 Information Extraction Model

Recall the information extraction model from Wiseman et al. (2017). q

4.3 Learning Conjunctions

Introduce latent variable \mathbf{h} and

5 Training and Inference

6 Related Work

References

Sam Wiseman, Stuart M. Shieber, and Alexander M. Rush. Challenges in data-to-document generation. *CoRR*, abs/1707.08052, 2017.