Latent Data to Document

August 15, 2018

1 Introduction

2 Problem

We would like to learn a generative model over sentences $\mathbf{x} = \{x_0, x_1, \ldots\}$ as well as distribution over latent tables $\mathbf{z}$. We are primarily interested in the respective conditional distributions: both the posterior distribution over tables given a sentence $p(\mathbf{z} \mid \mathbf{x})$, which is an information extraction model, as well as the conditional distribution over summaries $p(\mathbf{x} \mid \mathbf{x})$.

3 Data to Document

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

3.1 The Model

Let x_t be the current token, $\mathbf{x}_{0:t-1}$ all previous tokens, z_t the prediction for the current score, and $\tilde{\mathbf{z}}_{0:t-1} \in \mathbb{R}_+^{t-1}$ be all previous scores. The model takes the form a language model, where the distribution over the next token

$$p(x_t \mid \mathbf{x}_{0:t-1}, \tilde{\mathbf{z}}_{0:t-1}, z_t) = f([h_{l_t:t-1}, h_t])$$

4 Latent Table Model

4.1 Generative Model

- $p(z)$ Is the prior over table completions, which we will not focus on. Or should we? Can we set the prior to simply be the nearest neighbour plus some noise?
- $p(x|z)$ is the likelihood of the summary given the table. We use the conditional copy model as in Wiseman et al. (2017).
- $q(z|x)$ is the posterior distribution over table completions given a summary.

4.2 Learning Conjunctions

Introduce latent variable $\mathbf{y}$ and

4.3 Extending the supervision

Content Selection $p(\mathbf{c} \mid \mathbf{x})$ where $\mathbf{c} \in \{0, 1\}^n$.

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

Relation Realization $p(\mathbf{y} \mid \pi(\mathbf{c}), \mathbf{x})$.

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.