Latent Information Extraction from Text

Keywords: information networks, natural language processing, information extraction, latent variable models

Introduction

In order to provide decision makers with the information necessary to make informed choices as well as predict the effects of those choices, we must have a structured representation of information to model the state of the world. *Information networks* (as described in Army BAA: II. A. c. iii. (3)) provide a graphical representation of information and how it propagates through a network. I focus on knowledge graphs, information networks where each node contains a set of facts about an entity and each edge describes how the facts in one node influence the facts in another. Knowledge graphs provide an intuitive and queryable representation of knowledge. A decision maker may query the relevant nodes to gain situational awareness, and when simulating a decision that alters the information in one node the graph can easily propagate those changes by virtue of its representation.

Information extraction is the task of producing structured representations given unstructured text. In the context of knowledge graphs, information extraction is used to populate the nodes of a knowledge graph conditioned on text. A typical approach to an information extraction system is the following pipeline: (i) segment the text into mentions and values, (ii) align the mentions to an entity, a node in a knowledge graph, (iii) identify the relationships between segments. The relationships between these segments entail facts. Given the large cost of obtaining annotations to train an extraction system, a model that can perform well with fewer annotations is appealing. My proposal focuses on scenarios where there is a surplus of text but a lack of annotations, making it difficult to train an extraction system.

Proposal

In order to circumvent the cost of obtaining annotations, I propose to explore the use of text, rather than the facts, as the main source of learning signal for extracting knowledge graphs from textual sources. This can be accomplished through the use of *latent variable models* (LVMs), which are defined by a generative process or recipe for how a dataset is generated. An LVM for information extraction specifies the probabilistic relationship between a knowledge graph and the text. Such a model permits probabilistic queries such as (i) how likely a sentence is given a set of facts, as well as (ii) what the distribution over facts is given a text. Training an LVM relies on queries of type (i), while using the LVM for information extraction requires queries of type (ii). **My proposal is to develop a latent variable model that describes the process of generating text, where there is an abundance of data, based on information from a knowledge graph.**

I will apply this model in the semi-supervised information extraction setting with the following goals:

1. Demonstrate that a conditional model of text can provide signal for training an information extraction model.
2. Show that we can achieve competitive performance on knowledge graph extraction with fewer annotations.
3. Incorporate more linguistic signals and latent structure into the LVM to improve modeling performance.

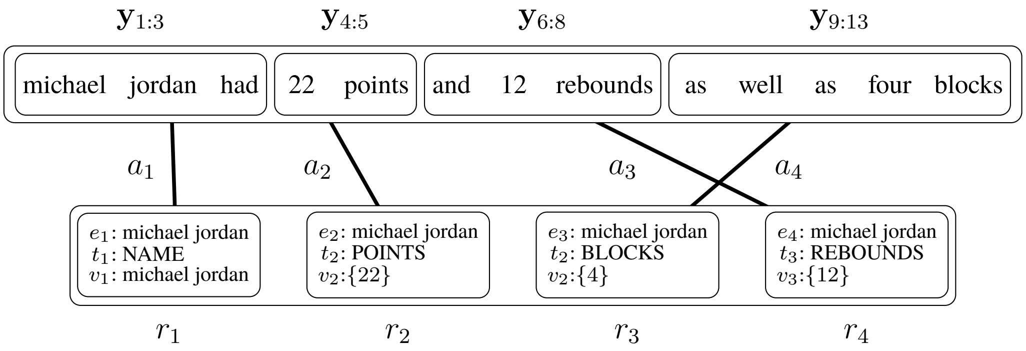

Figure 1: An example of information extraction. The generative process constructs the text from the given facts. The inference process aims to infer missing facts in curly braces.

The model will perform sequence-wise generation of text using facts from the nodes of a knowledge graph. This approach is inspired by the work of Liang et al. [4], who learn to align facts to text without supervision. Their work relies on the insight that a model of text provides signal for learning the alignments: if the likelihood of a segment of text improves when moving from one alignment choice to a another, then the new alignment is more likely to be correct given a suitably strong likelihood model. In my proposed LVM, I will extend this approach to a deep learning based model that incorporates state-of-the-art extraction and generation techniques.

Detailed Approach

Problem Setup To model the problem of information extraction, consider a dataset consisting of data and text pairs $\left\{(\mathbf{r}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{r}^{(2)}, \mathbf{y}^{(2)}) \dots\right\}$. Each datum $\mathbf{r} = \{r_1, \dots, r_N\}$ is a set of N records or facts, where each record $r = (e, t, v)$ is a tuple containing an entity, type, and value. We denote the aggregated collection of any elements as a bold variable. The datum $\mathbf{r}$ is a simple knowledge graph or information network. Each text $\mathbf{y} = y_1, y_2, \dots$ is a sequence of tokens each from a vocabulary V. As an example consider a dataset consisting of summaries of basketball games $\mathbf{y}$ aligned with the respective statistics $\mathbf{r}$ of those games in Figure 1. Here there are three records and the statement $\mathbf{y} =$ “michael jordan had 22 points and 12 rebounds as well as four blocks”. For this example, the process of information extraction is to infer the values $\mathbf{v}$ of the records given the entities $\mathbf{e}$, types $\mathbf{t}$, and the text $\mathbf{y}$.

Method My proposed model is a conditional LVM that specifies the relationship between the different components of the data, specifically how the entities and types generate the text. Assume that the entities and relation types $\{\mathbf{e}, \mathbf{t}\}$ of the knowledge graph are known. The model requires a generative process for $\{\mathbf{y}, \mathbf{a}, \mathbf{v}\}$, i.e. the values, the text, and the alignment between the text and the facts. This requires specifying:

$$p(\mathbf{y}, \mathbf{a}, \mathbf{v} \mid \mathbf{e}, \mathbf{t}) = p(\mathbf{v} \mid \mathbf{e}, \mathbf{t}) \times p(\mathbf{a} \mid \mathbf{v}, \mathbf{e}, \mathbf{t}) \times p(\mathbf{y} \mid \mathbf{a}, \mathbf{v}, \mathbf{e}, \mathbf{t}).$$

The full generative process is the following:

1. Generate Prior Values: $p(\mathbf{v} \mid \mathbf{e}, \mathbf{t})$. For each entity and type pair in our datum of records, predict a prior value. For example, given ‘michael jordan’ and POINTS we predict 19. Each value is predicted independently using a neural network over the entity e and type t.
2. Generate Record Alignments: $p(\mathbf{a} \mid \mathbf{v}, \mathbf{e}, \mathbf{t})$. Conditioned on the knowledge graph, choose an ordered subset of records $\mathbf{a} = a_1, \dots, a_T$ for the text to describe. This sequence is predicted using a Markov chain parameterized with a neural network transition function that also takes into account a global encoding of the knowledge graph.

3. Generate the Document: $p(\mathbf{y} \mid \mathbf{a}, \mathbf{v}, \mathbf{e}, \mathbf{t})$. For each record selected as a_i, choose a sequence of words $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iJ}\}$ to describe the record indicated by the alignment. This sequence is predicted using a recurrent neural network (RNN) that also takes into account an embedding of the current selected record.

This model is an instance of a hidden semi-Markov model (HSMM), a classical LVM that has been used for generation and extraction tasks. Our version extends the HSMM to incorporate methods from deep learning that have been shown to perform well on supervised extraction tasks.

To use this model for semi-supervised information extraction system we invert this generative process to obtain the posterior distribution over alignments and values:

$$p(\mathbf{a}, \mathbf{v} \mid \mathbf{y}, \mathbf{e}, \mathbf{t}) = \frac{p(\mathbf{y}, \mathbf{a}, \mathbf{v} \mid \mathbf{e}, \mathbf{t})}{p(\mathbf{y} \mid \mathbf{e}, \mathbf{t})} = \frac{p(\mathbf{y}, \mathbf{a}, \mathbf{v} \mid \mathbf{e}, \mathbf{t})}{\sum_{\mathbf{a}, \mathbf{v}} p(\mathbf{y}, \mathbf{a}, \mathbf{v} \mid \mathbf{e}, \mathbf{t})}.$$

Although the HSMM formulation allows the summation over alignments to be carried out efficiently, the sum over value assignments is intractable. I propose to instead apply variational inference, where an approximation of the posterior distribution is learned with a separate model. The LVM and the approximate posterior are trained jointly by maximizing a lower bound on the log marginal likelihood of $\mathbf{y}$ with gradient-based methods. The resulting approximate posterior can be used independently of the original model as an information extraction system that gives a distribution over values in a table of records and alignments from text to records.

Plan

I plan to develop a system based on these ideas for information extraction with distant supervision. I will evaluate the approach on reduced annotation versions of standard information extraction benchmarks including the TAC KBP 2015 slot filling task and the TACRED dataset [7]. The goal will be to demonstrate competitive performance on extraction metrics such as precision, recall, and F1, while using as little supervision as possible by ignoring subsets of the given data. For example, we can allow the model to only learn from subsets of the given values (or none at all). Given success in that goal, the next step would be to extend the model to capture more of the joint distribution with the aim of boosting sample and label efficiency. Possible extensions in this direction include coreference resolution [2], learning the types of relations in a semi-supervised manner, leveraging discourse structure to improve the model [6], explicitly modeling nuisance variables such as author style [3], and incorporating multi-hop reasoning in order to leverage relationships between entities [1, 5].

Given admittance to the NDSEG Fellowship Program, I will evaluate the application of LVMs to the problem of information extraction. As a result of the digital age, the ubiquity of information networks as well as their enormous growth makes it clear that a method for training information extraction systems with minimal supervision is a necessity. I will push for scalable information extraction systems that require minimal supervision by recasting information extraction as a text modeling problem.

References

[1] Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. Variational knowledge graph reasoning. *CoRR*, abs/1803.06581, 2018. URL `http://arxiv.org/abs/1803.06581`.

[2] Aria Haghighi and Dan Klein. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 385–393, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. ISBN 1-932432-65-5. URL `http://dl.acm.org/citation.cfm?id=1857999.1858060`.

[3] Wei-Ning Hsu, Yu Zhang, and James R. Glass. Learning latent representations for speech generation and transformation. *CoRR*, abs/1704.04222, 2017. URL `http://arxiv.org/abs/1704.04222`.

[4] Percy Liang, Michael I. Jordan, and Dan Klein. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 91–99, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9. URL `http://dl.acm.org/citation.cfm?id=1687878.1687893`.

[5] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. *CoRR*, abs/1705.11040, 2017. URL `http://arxiv.org/abs/1705.11040`.

[6] Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 208–216, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9. URL `http://dl.acm.org/citation.cfm?id=1687878.1687909`.

[7] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1004. URL `http://aclweb.org/anthology/D17-1004`.