Information Extraction with Distant Supervision

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

Relevance to Army BAA: II. A. c. iii. (3) Information Networks 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 an efficient representation of relevant information and a predictive model for the resultant state of the world after a choice is made. Information networks 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 an 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.

Given that a knowledge graph must represent an extremely large number of facts and relationships, it is infeasible to specify these completely by hand. I propose to fill in the nodes of a knowledge graph by leveraging a generative model of corpora. Many recent works have taken an orthogonal direction known as link prediction, which fills in the missing *edges* of a knowledge graph [1]. Through link prediction one uses the relationships specified by the edges to reason about the facts contained in a node conditioned on the facts of its neighbours. However, we generally also have access to large amounts of unlabeled and unstructured data in the form of natural language corpora. I aim to leverage natural language corpora by learning an information extraction system that is able to extract facts from text to fill in the nodes of a knowledge graph given only the assumption that the text was generated from the information represented in the knowledge graph. Although this is a strong assumption, we believe it is a reasonable starting point.

In this proposal we present a method towards automating the training of information extraction systems with unlabeled corpora by recasting the information extraction problem as a conditional generative modeling problem. Specifically, we plan to use the performance of a deep generative model of text as signal for learning an extraction system. Once the generative model is trained, the extractive system is obtained through posterior inference.

Introduction The goal of information extraction is to produce structured representations of information given unstructured text. A typical approach to an information extraction system is the following pipeline, where each stage has a model that is trained independently from the other stages: first segment the text into entities and values, then extract named entities and possibly perform coreference resolution by predicting whether segments refer to the same entity, and finally relation extraction where we identify the relationships between the extracted entities and values. Even if each stage is modeled end-to-end with a neural network [8], an approach without a generative model still fails to propoagate beliefs in a principled probabilistic manner. Incorporating a generative model of text will allow the model to better capture the relationship between the text and underlying knowledge graph, resulting in a stronger information extraction system.

A generative model specifies the relationship between all variables of interest in the form of a joint distribution; we also consider conditional variants which condition on other variables. Latent variable models (LVMs) are a class of (conditional) generative models that include latent or unobserved variables. LVMs provide a framework for formulating an information extraction pipeline in a probabilistically principled fashion. More specifically, a LVM allows us to utilize the relationship between variables in the joint distribution to propagate our beliefs through inference. For instance, if we are uncertain about whether a segment of text corresponds to an entity, that uncertainty will influence our beliefs about the values

we can extract through the relationships defined by the LVM. LVMs have the additional benefit of being amenable to semi-supervised training, as shown in Kingma et al. [6] who demonstrated that LVMs with neural network parameterizations achieve strong performance on semi-supervised classification on MNIST with few labels.

In this proposal, I focus on the class of LVMs known as hidden semi-Markov models (HSMMs), used in Liang et al. [7] as a conditional generative model for the task of aligning segments of text to nodes in a knowledge graph without supervision. As in Liang et al. [7], I am interested in using a conditional generative model of text so that the amount of supervision necessary for training an information extraction system can be minimized. The performance of their conditional generative model of text provides signal for learning the alignments: if the likelihood of a segment of text improves when moving from a past alignment choice to a new one, then the new alignment is more likely to be correct given a suitably strong likelihood model. The same intuition can be used to formulate a related LVM for a semi-supervised information extraction which aims to model not just the alignments from segments of text to nodes in a knowledge graph but also the values in the nodes themselves. We build on the formulation in Liang et al. [7] by parameterizing our LVM with neural networks, drastically increasing the generative model's capacity. By parameterizing the HSMM with neural networks as in Wiseman [12], where a neural network-based HSMM was used to induce templates in a data-to-text generation task, I hope to incorporate recent progress in parameterizing LVMs with neural networks so as to learn a more accurate information extraction system by using a more powerful generative model.

Background We consider datasets consisting of aligned data and text $\left\{(\mathbf{r}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{r}^{(2)}, \mathbf{y}^{(2)}) \dots\right\}$. For brevity, I refer to a single datum and text as $\mathbf{r}, \mathbf{y}$, omitting the superscript. Each datum $\mathbf{r} = \{r_1, \dots, r_N\}$ is a set of N records, where each record $r_i = (e_i, t_i, v_i)$ is a tuple containing an entity, type, and value. The datum $\mathbf{r}$ is a knowledge base, or equivalently an information network without a representation of the causal effects between records. We refer collectively to all the entities, types, and values in a given datum $\mathbf{r}$ as $\mathbf{e}, \mathbf{t}, \mathbf{v}$ respectively. Each text $\mathbf{y} = \{y_1, \dots, y_T\}$ is a sequence of tokens each from a vocabulary V.

The Rotowire dataset [13] is a concrete example of such a dataset. Rotowire contains summaries of basketball games $\mathbf{y}$ aligned with the respective box scores $\mathbf{r}$ of those games. Consider the datum in Figure 1 that consists of 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 $\mathbf{y}$. We want to infer the value v_2 of the second record r_2 given the entity "michael jordan", the type POINTS, as well as the text $\mathbf{y}$.

Proposal We propose to verify the efficacy of the LVM framework in the semi-supervised information extraction setting, with the following goals:

1. By formulating a LVM for generating text conditioned on data, obtain an information extraction system through posterior inference.

2. Demonstrate strong extractive performance with minimal labels.

3. Move towards a model for knowledge graph completion that captures the full joint distribution.

We present one instance of a LVM and outline how it can be used to obtain an information extraction model without direct supervision, then argue that the same approach can be applied in even more ambitious settings. Our proposed LVM is a conditional generative model that specifies the relationship between data, specifically the entities and types, and text. We denote this model `Values`. Similar to the models defined in Wiseman [12] and Liang et al. [7], our model takes the form of a hidden semi-Markov model (HSMM).

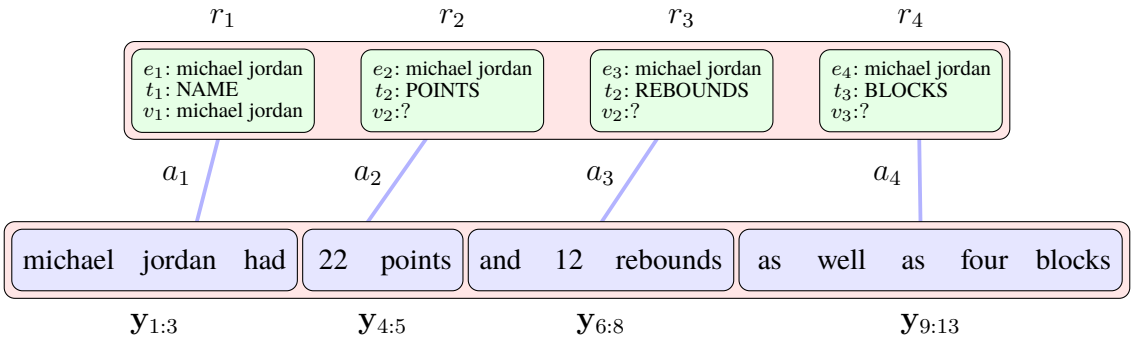

Figure 1: An example of our proposed information extraction procedure with an inferred segmentation. In addition to the segmentation, we aim to infer the missing values in records. The missing values v_j are predicted independently. Records r_i are then chosen, and their ordering is given by $\mathbf{a}_{1:4}$. Each a_i then generates a sequence of words to describe its record r_i.

The primary difference is that the other models simply assumed the records were complete and conditioned on them, whereas ours learns to generate the values.

Our model assumes observed $\{\mathbf{e}, \mathbf{t}\}$ and latent $\{\mathbf{v}, \mathbf{a}\}$, where $\mathbf{e}$ are the entities, $\mathbf{t}$ are the types, $\mathbf{v}$ are the values, and $\mathbf{a}$ are record indices. `Values` is given by the following generative process:

1. Prior Value Choice: $p(\mathbf{v} \mid \mathbf{e}, \mathbf{t})$. For each entity and type pair in our datum of records, predict a value. For example, given 'michael jordan' and POINTS, we predict 19. Each value is predicted independently.

2. Prior Record Choice: $p(\mathbf{a} \mid \mathbf{v}, \mathbf{e}, \mathbf{t}) = \prod_i p(a_i \mid a_{i-1}, \mathbf{v}, \mathbf{e}, \mathbf{t})$. Conditioned on our choices of values as well as the given entities and records, in other words a completed data with no missing values, choose a sequence of records $\mathbf{a} = \{a_1, \ldots, a_T\}$ to describe with a Markov model.

3. Word Choice: $p(\mathbf{y} \mid \mathbf{a}, \mathbf{v}, \mathbf{e}, \mathbf{t})$. For each alignment a_i, choose a sequence of words $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iJ}\}$ to describe the record indicated by the alignment. With the HSMM formulation, we have that within each segment aligned to a record, the words are modeled autoregressively.

An information extraction system is obtained by using 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 (marginalization) over alignments to be carried out efficiently, one cannot marginalize over value assignments. Instead one must resort to variational inference as in Deng et al. [3], where an approximation of the posterior distribution $q(\mathbf{a}, \mathbf{v} \mid \mathbf{y}, \mathbf{e}, \mathbf{t})$ is learned with a separate model. The conditional generative model and the approximate posterior is trained jointly by maximizing a lower bound on the log marginal likelihood of $\mathbf{y}$ with gradient-based methods. The resulting approximate posterior $q(\mathbf{a}, \mathbf{v} \mid \mathbf{y}, \mathbf{e}, \mathbf{t})$ can be used independently of the generative model as an information extraction system that gives a distribution over values in a table of records and alignments from text to records.

I will evaluate the approach on the TAC KBP 2015 slot filling, TACRED [14], and ROTOWIRE [13] datasets in order to compare to previous work. 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 information. For example, with the model `Values`, we can allow the

model to only learn from subsets of the given values (or none at all) in ROTOWIRE, such as only the home team's statistics, at training time. 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 [4], learning the types of relations in a semi-supervised manner, leveraging discourse structure to improve the conditional generative model [11], explicitly modeling nuisance variables such as author style [5], and incorporating multi-hop reasoning in order to leverage relationships between entities [1, 10].

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 generative modeling problem.

Outline

1. Introduction
   (a) Relevance to BAA
      i. Intro to information networks and KG
         A. Information networks and decision making
         B. Specify knowledge graphs as the information networks we are interested in
         C. Brief outline of knowledge graphs, ie nodes are entities, contain sets of facts, edges specify relationships between facts (can include self-loops)
      ii. Argument that KGs cannot be populated by hand. (Brief outline of methods for populating, no in-depth descriptions provided in this proposal)
         A. Link prediction
         B. Multi-hop link prediction
         C. Corpora-based
      iii. Proposal: train information extraction systems by recasting as a generative modeling problem.
   (b) Real Introduction
      i. We focus on information extraction, which is the task of producing structured representations of text.
      ii. We believe a generative model of text will allow the model to better capture the relationship between the text and underlying knowledge graph, resulting in a stronger information extraction system.
      iii. A latent variable model is simply a generative model that incorporates latent or unobserved variables alongside the observed variables in the joint distribution.
      iv. By formalizing the generative process of text given data as a LVM, we are afforded a principled approach to information extraction through posterior inference.
   (c) Information Extraction at a high level
      i. Define information extraction
         A. The goal is to produce or fill in structured representations of information from a given unstructured text.
         B. A typical pipeline for information extraction includes text segmentation, named entity recognition, coreference resolution, relation extraction, and finally producing structured representations of the unlabeled text.
      ii. Previous work utilize phrases as additional relations but do not model the underlying generative process [2, 9].
      iii. Argue for LVM approach to unify all parts of the pipeline in a single joint distribution in order to utilize the interaction between parts during inference.
      iv. We aim to correct this through recently developed techniques for training LVMs parameterized with neural networks.
   (d) LVMs for joint modeling
      i. What is a LVM?
      ii. What is the benefit? Propagate uncertainty through belief propagation.

iii. Draw parallels to progress in translation, mainly in terms of modeling the joint
   A. Deterministic structure via attention Bahdanau
   B. Latent attention Deng et al. [3]

(e) Recent advances in neural LVMs
   i. Semi-supervised LVMs Kingma et al. [6]?
   ii. Demonstrate that parameterization with a neural network does not affect computational complexity of inference.
   iii. Then the same technique can be applied to model with more structure, as long as the graphical model itself permits tractable inference.
   iv. In this proposal, we focus on the hidden semi-Markov model (HSMM), used in Liang et al. [7] for the task of aligning segments of text to records in a knowledge base without supervision.
   v. As in Liang et al. [7], we are interested in learning a generative model of text so that we can minimize the amount of supervision necessary for training an information extraction system.
   vi. Also that although worse sample complexity, using an approximate posterior with monte carlo sampling achieves comparable performance.

2. Background and Notation

(a) Formal notation for elements of the dataset

(b) Define the distribution we would like to learn: $p(z \mid y, x)$. z and x are placeholders and will change, but y is always the text.

(c) Link to rotowire example (argument is that ACE is made up of ontonotes-like sentences, so all short-form)

(d) Clarify that the scope of posterior inference is very general.

3. Proposal

(a) Outline approach
   i. Choose a subset of available data as conditioning, and thus it is not modelled.
   ii. The joint distribution of the remaining variables, both observed and unobserved, will be modelled.

(b) Link back to motivation. We want to scale information extraction by requiring less supervision.

(c) We present one model as an example, which we will serve as a starting point for the proposed research.

(d) Define generative model: HSMM as in [7], and [12]. The generative story (a picture would be helpful):
   i. Fill in values
   ii. Choose alignments
   iii. Choose words

(e) Define IE as the distribution we would like to learn
   i. Values: $p(c, \mathbf{v} \mid \mathbf{y}, \mathbf{e}, \mathbf{t})$ (Just this one)

(f) We either use the posterior distribution of the conditional model or learn an approximation of it.

(g) Training and Inference

i. As we are dealing with large state spaces, we train with an approximate posterior in order to satisfy memory constraints.

ii. Highlight that the approx posterior is a SEPARATE model that can be used completely independently from generative model, i.e. we throw away generative model after training.

iii. We maximize a lower bound on the log marginal likelihood, called the evidence lower bound.

(h) Experiments, evaluation, and expectation

i. Evaluate on Rotowire.

ii. We evaluate `Values` using the precision, recall, and F1 score on the task of predicting the values associated with entities, otherwise known as slot-filling.

iii. What would success look like?

A. Competitive to supervised methods when supervision is available

B. But able to be applied when supervision is not available

C. Able to leverage lots of unlabeled data during training, and success would see a marked improvement over purely supervised methods as well as the purely supervised version of this model.

D. Would provide explanations for the answers (ie segmentations).

iv. Also on ACE?

(i) Future work

i. Incorporate more structure into the generative model, for example entity tracking or coreference resolution Haghighi and Klein [4].

ii. Model more structure in the data, for example the edges between nodes in the knowledge graph Chen et al. [1].

iii. 'Multi-hop' reasoning, where we try to compose relationships to infer new ones, i.e. through unification Chen et al. [1], Rocktäschel and Riedel [10].

(j) Conclusion

i. Please accept!

ii. Recap: Minimal supervision IE systems so that they can scale to extracting information for large information networks from large bodies of text.

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] Lei Cui, Furu Wei, and Ming Zhou. Neural open information extraction. *CoRR*, abs/1805.04270, 2018. URL http://arxiv.org/abs/1805.04270.

[3] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander M. Rush. Latent alignment and variational attention. *CoRR*, abs/1807.03756, 2018. URL http://arxiv.org/abs/1807.03756.

[4] 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.

[5] 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.

[6] Diederik P. Kingma, Danilo Jimenez Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *CoRR*, abs/1406.5298, 2014. URL http://arxiv.org/abs/1406.5298.

[7] 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.

[8] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using lstms on sequences and tree structures. *CoRR*, abs/1601.00770, 2016. URL http://arxiv.org/abs/1601.00770.

[9] Meng Qu, Xiang Ren, Yu Zhang, and Jiawei Han. Overcoming limited supervision in relation extraction: A pattern-enhanced distributional representation approach. *CoRR*, abs/1711.03226, 2017. URL http://arxiv.org/abs/1711.03226.

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

[11] 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.

[12] Sam Wiseman. Learning neural templates for text generation. *CoRR*, abs/lol, 2018. URL http://arxiv.org/abs/lol.

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

[14] 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.