Information Extraction with Weak Supervision

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

Relevance to BAA   As information networks grow larger and more complex, acquiring explicit supervision for the training of information extraction systems becomes extremely expensive. Information networks are also dynamic; over time one method for representing information may become inadequate. Currently, both the representation of the information in a knowledge base as well as the extraction process itself must be hand-designed. In this proposal we present a framework towards automating the training of information extraction systems with minimal supervision.

Introduction   Natural language processing contains two separate but closely related subfields: natural language understanding (NLU) and natural language generation (NLG). Recent approaches to information extraction use only the NLU perspective and frame extraction as a classification problem. We argue that the two perspectives, NLU and NLG, are complementary and capitalize on their duality by proposing a method to train a NLU system without direct supervision. More specifically, we train an information extraction system by using the performance of a deep generative model as signal. By training in this fashion we obtain an information extraction system that extracts facts that best explain the given text.

Recent work using neural network-based systems cast information extraction as a supervised problem [2, 8]. Although approaches do incorporate structure (?) into internal representations, for example to integrate multiple sources of information [6], the final output still uses strong supervision during training. Finding strong supervision is a difficult task, and as datasets get larger we must approach information extraction from a different perspective. Latent variable models (LVMs) are one method for alleviating the need for supervision. LVMs do not require manual labels, as they instead treat quantities of interest as latent random variables and deal with them in a probabilistically principled fashion. We will leverage recent advances in deep generative modeling in order to train without explicit supervision.

Neural models for language generation have seen much progress in recent years, with state of the art performance in both language modeling and translation [5, 9]. By integrating the powerful language modeling capabilities of neural networks with the inductive biases of graphical models through LVMs, we can leverage the LVM framework to derive a principled method for training an information extraction system with less supervision. Namely, we turn to an efficient technique for training hard attention that relies on variational inference [1]. This technique is applicable to sequential LVMs such as hidden Markov models and hidden semi-Markov models (HSMMs). HSMMs have been applied to small scale datasets for text generation [7] as well as without integrating neural networks [4]. We will scale sequential LVMs to large datasets through variational inference and use the signal from our generative model in order to train an information extraction system.

Background   We consider datasets consisting of aligned data and text $\left\{ (\mathbf{r}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{r}^{(2)}, \mathbf{y}^{(2)}) \ldots \right\}$. For brevity, we refer to a single datum and text as $\mathbf{r}, \mathbf{y}$, omitting the superscript. Each datum $\mathbf{r} = \{r_1, \ldots, r_N\}$ is a set of N records, each of which has an entity, type, and value $r_i = (e_i, t_i, v_i)$. The datum $\mathbf{r}$ is a flattened representation of an information network. Whereas an information network may also be represented as a hypergraph, $\mathbf{r}$ is a list of relations or 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 $\mathbf{y} = \{y_1, \ldots, y_T\}$ is a sequence of tokens.

⁴⁴ Let variables $\mathbf{z}$ be unobserved or latent, $\mathbf{y}$ observed, and $\mathbf{x}$ taken as conditioning and thus not
⁴⁵ modelled. For information extraction we are interested in distributions that can be specified as
⁴⁶ $p(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$, where $\mathbf{z}$ and $\mathbf{x}$ may correspond to various quantities depending on the task but $\mathbf{y}$ is
⁴⁷ always the text.
⁴⁸ As a concrete example, we use the Rotowire dataset [8]. Rotowire contains summaries of
⁴⁹ basketball games $\mathbf{y}$ aligned with the respective box scores $\mathbf{r}$ of those games. Consider a datum that
⁵⁰ consists of a single record, $\mathbf{r} = \{(e_1 = \text{Jeremy Lin}, t_1 = \text{POINTS}, v_1 = 19)\}$, and a simple statement
⁵¹ $\mathbf{y} =$ "Jeremy Lin scored 19 points". In its simplest incarnation, the process of information extraction
⁵² may be to infer any subset of $\mathbf{r}$, which in this case will be our latent $\mathbf{z}$, given the remaining elements
⁵³ in $\mathbf{r}$ which corresponds to $\mathbf{x}$, as well as the text $\mathbf{y}$. For example, we may want to infer the type
⁵⁴ of the relation t given the entity Jeremy Lin, the value 19, as well as the text $\mathbf{y}$. In this case, we
⁵⁵ would have $\mathbf{z} = \{t_1\}$ and $\mathbf{x} = \{e_1, v_1\}$. In an alternative task, we may want to infer the value v_1
⁵⁶ as well as the type t_1 given $\mathbf{y}$ and e_1, therefore $\mathbf{z} = \{t_1, v_1\}$ and $\mathbf{x} = \{e_1\}$. (Will switch to ACE
⁵⁷ dataset after I figure out what's going on with the data.)
⁵⁸ Note that we are not constrained to setting $\mathbf{z}$ to subsets of $\mathbf{r}$. We also consider the case
⁵⁹ where $\mathbf{z}$ includes alignments from individual words y_t to records r_i. We denote the alignments
⁶⁰ $\mathbf{a} = \{a_1, \ldots, a_T\}$, where each a_t is associated with y_t and selects a record r_i such that $a_t = i$.

⁶¹ **Proposal** We propose to verify the efficacy of the LVM framework in the weakly supervised
⁶² information extraction setting, with the goal of demonstrating strong extractive performance with
⁶³ minimal labels. We present one instance of a LVM and outline how it can be used to obtain an
⁶⁴ information extraction model without direct supervision, then argue that the same approach can
⁶⁵ be applied in even more ambitious settings. Our first LVM is a conditional model that specifies
⁶⁶ the relationship between data, specifically the entities and types, and text. We denote this model
⁶⁷ `Values`. Similar to the models defined in Wiseman [7] and Liang et al. [4], our model takes the form
⁶⁸ of a hidden semi-Markov model (HSMM). The primary difference is that the other models simply
⁶⁹ assumed the records were complete and conditioned on them, whereas ours learns to generate the
⁷⁰ values. `Values` is given by the following generative process:

⁷¹ 1. Value Choice: For each pair of entities and types in our datum of records, we predict a value.
⁷²   We assume that each record type constrains the values to be members of a finite set. Thus
⁷³   each record type is assigned a categorical distribution over its respective values, and the values
⁷⁴   are drawn independently from that respective distribution. Each $v_i \sim \text{Cat}(f_{t_i}(e_i))$ is drawn
⁷⁵   from a Categorical distribution, whose parameters f_{t_i} are output by a neural network that is
⁷⁶   shared across record types and takes as input the entity.

⁷⁷ 2. Record Choice: Conditioned on our choices of values as well as the given entities and records,
⁷⁸   we choose a sequence of records $\mathbf{a} = \{a_1, \ldots, a_I\}$ to describe. Each record alignment a_i points
⁷⁹   to the index of its corresponding record. The records are parameterized as a Markov model
⁸⁰   where each $a_t \sim \text{Cat}(f_\theta(a_{t-1}, \mathbf{v}, \mathbf{e}, \mathbf{t}))$, where f_θ is a neural network.

⁸¹ 3. Word Choice: For each record alignment a_i, we choose a sequence of words $\mathbf{y}_i = \{y_{i1}, \ldots, y_{iJ}\}$
⁸²   to describe the record. The words are modelled by an autoregressive emission model within
⁸³   each segment that is aligned to the same record: $y_{ij} \sim \text{Cat}(f_\theta(\mathbf{y}_{i<j}, v_{a_i}, e_{a_i}, t_{a_i}))$, where f_θ is
⁸⁴   another neural network.

⁸⁵ The value and record choices correspond to prior distributions over values $p(\mathbf{v} \mid \mathbf{e}, \mathbf{t})$ and alignments
⁸⁶ $p(\mathbf{a} \mid \mathbf{v}, \mathbf{e}, \mathbf{t})$ respectively, while the word choice model gives us the likelihood of some text given
⁸⁷ our value and alignment choices $p(\mathbf{y} \mid \mathbf{a}, \mathbf{v}, \mathbf{e}, \mathbf{t})$. In this case, we have latent $\mathbf{z} = \{\mathbf{v}, \mathbf{a}\}$ and

⁸⁸ observed $\mathbf{x} = \{\mathbf{e}, \mathbf{t}\}$. We obtain an information extraction by using the **posterior** distribution over
⁸⁹ alignments and values:

$$p(\mathbf{z} \mid \mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})}{p(\mathbf{y} \mid \mathbf{x})} = \frac{p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})}{\sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})}.$$

⁹⁰ Althought the HSMM formulation allows the summation (marginalization) over alignments to be
⁹¹ carried out efficiently, we cannot marginalize over value assignments. We instead resort to varia-
⁹² tional inference as in Deng et al. [1], where we learn an approximation of the posterior distribution
⁹³ $q(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$ with a separate model. Given that we are primarily interested in the values rather than
⁹⁴ the alignments, we can obtain an information extraction system over only values by marginalizing
⁹⁵ over alignments. By marginalizing over the alignment distribution, the model propagates uncer-
⁹⁶ tainty over alignments to uncertainty over values. We can train this model by maximizing a lower
⁹⁷ bound on the log marginal likelihood or evidence of $\mathbf{y}$, called the evidence lower bound (ELBO)
⁹⁸ which is given formally by the expression $\mathrm{ELBO}_q \triangleq \mathbb{E}_{q(\mathbf{z}|\mathbf{y},\mathbf{z})} \left[\log p(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) \right] \leq \log \mathbb{E}_{p(\mathbf{z}|\mathbf{x})} \left[p(\mathbf{y} \mid \mathbf{z}, \mathbf{x}) \right]$.
⁹⁹ This objective can be maximized with gradient-based methods. The resulting approximate pos-
¹⁰⁰ terior $q(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$ can be used independently of the generative model as an information extraction
¹⁰¹ system.

¹⁰²   As we are interested in applying the LVM framework to information extraction rather than a
¹⁰³ single model, we proceed to motivate and outline an extension to `Values`. In addition to inferring
¹⁰⁴ values, ideally an information extraction system would also be able to infer new entities and relation
¹⁰⁵ types. We propose a possible route towards defining model that can learn to parameterize new
¹⁰⁶ relation types in an unsupervised manner using the same LVM framework as `Values`, which we
¹⁰⁷ denote `Types`. The motivation behind `Types` is that a text may refer to multiple records at the same
¹⁰⁸ time. For example, in basketball games a 'triple-double' refers to more a player achieving a value
¹⁰⁹ of more than 10 of any three of the five categories: points, steals, rebounds, blocks, or assists. The
¹¹⁰ goal of `Types` is to attempt to capture the latent relationship behind a 'triple-double' appearing in
¹¹¹ the text and the underlying records in an unsupervised manner. We plan to introduce a new step
¹¹² to the generative process of `Values` that allows the model to learn new records as boolean-valued
¹¹³ functions of relations already defined in the data. By approximating these boolean functions with
¹¹⁴ neural networks, we hope to find a model parameterization that admits a low variance monte carlo
¹¹⁵ gradient estimator.

¹¹⁶   We will evaluate our initial approach on the Rotowire dataset, and extensions to our model
¹¹⁷ that will include entity tracking and event resolution on the automatic content extraction (ACE)
¹¹⁸ [3] and the Text Analysis Conference's Streaming Multimedia Knowledge Base Population (SM-
¹¹⁹ KBP) datasets. We expect the variance of the gradient estimator to be an issue, in particular
¹²⁰ its effect on sample complexity. In previous work, we observed that gradient estimators based on
¹²¹ exact inference resulted in better sample complexity than monte carlo gradient estimators [1], and
¹²² we expect that controlling the variance of the gradient estimator with the inductive bias of neural
¹²³ architectures will be of paramount importance for the success of our proposed method.

¹²⁴   Given admittance to the NDSEG Fellowship Program, I will evaluate the application of un-
¹²⁵ supervised LVMs to the problem of information extraction. As a result of the digital age, the
¹²⁶ ubiquity of information networks as well as their enormous growth makes it clear that a method
¹²⁷ for training information extraction systems with minimal supervision is a necessity. I will push for
¹²⁸ scalable information enformation extraction systems that require minimal supervision by recasting
¹²⁹ information extraction as a generative modeling problem.

130 Outline

131 1. Introduction

132   (a) Relevance to BAA

133   i. What are information networks? Graphical representation of objects, their charac-
134   teristics of interest, and their relationship to objects. Characterized by large graphs
135   and voluminous data.

136   ii. Information networks must also represent knowledge wihich is both uncertain as
137   well as dynamic; over time one method for representing information may become
138   inadequate.

139   iii. Obtaining supervision for large networks is expensive, especially if the underlying
140   schema is subject to change.

141   iv. TODO: Make expensive argument more concrete by hammering in the dynamic
142   schema aspect. This argument might be weak since the model would also have to
143   be retrained, but that's much cheaper than additional labeling.

144   v. TODO: possibly establish a more military-focused or at least information-gathering
145   running example, as opposed to basketball.

146   vi. In this proposal we present a framework towards automating the training of infor-
147   mation extraction systems with minimal supervision.

148   (b) Decomposition of NLP into NLU and NLG (Keep this paragraph high-level, just an
149   introduction to the thesis and proposal.)

150   i. NLP consists of two separate but closely related subfields: natural language under-
151   standing (NLU) and natural language generation (NLG).

152   ii. Although both subfields have been dominated by neural network-base models, there
153   is a striking difference.

154   iii. NLU models capitalize primarily on the representational power of neural networks.
155   Rather than requiring features as input, neural networks learn representations of
156   the data automatically. (Maybe move the whole NLU paragraph here, but probably
157   not since I feel like I have a strong opening paragraph.)

158   iv. However, NLG utilizes neural networks not only for representation learning but also
159   for generative modeling. Success on tasks such as language modeling and machine
160   translation demonstrate that generative and conditional models parameterized by
161   neural networks are able to approximate not only simple distributions over labels,
162   such as entity or not, but also complex joint distributions. (Maybe move the whole
163   generative paragraph here, X)

164   v. Recent approaches to information extraction use only the NLU perspective and
165   frame extraction as a classification problem, and thus do not take advantage of the
166   generative capability of neural networks.

167   vi. TODO: Possibly give examples QA, information extraction (dos santos + sam).

168   vii. We argue that the two subfields' approaches are complementary, and capitalize on
169   their duality by proposing a method to train a NLU system without direct supervi-
170   sion.

171   (c) Supervision in NLU (Need to hammer in the framing of supervision)

172   i. Recent work using neural network-based systems cast information extraction as a
173   supervised problem [2, 8].

ii. Even with structured representations [6], the final output is still supervised.

iii. Again, supervision does not scale.

iv. Latent variable models (LVMs) are one method for alleviating the need for supervision.

v. LVMs do not require manual labels, as they instead treat quantities of interest as latent random variables and deal with them in a probabilistically principled fashion.

(d) Recent advances in NLG (Need to argue the benefits of generative modeling, is the recent success enough?)

i. Deep generative models, namely LVMs with neural network components, have integrated the flexibility of neural networks with the inductive biases of graphical models.

ii. Most importantly, efficient techniques for training hard attention that rely on variational inference [1]. This technique is applicable to sequential LVMs such as hidden Markov models and hidden semi-Markov models (HSMMs).

iii. Use HSMM as signal for training IE.

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) We present one model as an example, then later demonstrate how the framework can handle extensions of the model.

(c) Define generative model: HSMM as in [4], and [7]. The generative story (a picture would be helpful):

i. Fill in values

ii. Choose alignments

iii. Choose words

(d) 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)

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

(f) Argue that the segmental model encourages more coverage by adding pairwise dependencies between labelings [4]. We will check how much structure in the generative model aids information extraction.

(g) Training and Inference

    i. As we are dealing with documents of significant length, we train with an approximate posterior in order to satisfy memory constraints.

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

(h) Extension, `Types`

    i. Introduce new step in generative process

    ii. Learn a boolean function that composes predicates applied to existing records

    iii. The search space is very large, so we must either constrain our model in a very clever way or obtain large amounts of data as any stochastic gradient estimator will have very high variance

(i) Extensions

    i. learn new types as functions of existing ones

    ii. learn a randomly initialized embedding of the type and a neural network directly to predict the value

    iii. let the input distribution be

(j) Experiments, evaluation, and expectation

    i. Evaluate on Rotowire?

    ii. Also on ACE

    iii. As corpora may be too large, we might need more hierarchy in the generative model

    iv. and also since memory is linear in the length of the sequence, we may have to resort to approximate inference. We can optimize a lower bound on the marginal likelihood with an approximation of the posterior distribution [1].

(k) 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] 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.

[2] Cícero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. *CoRR*, abs/1504.06580, 2015. URL http://arxiv.org/abs/1504.06580.

[3] T. George Doddington@NIS, C. Alexis Mitchell@LD, T. Mark Przybocki@NIS, N. Lance Ramshaw@BB, C. Stephanie Strassel@LD, and N. Ralph Weischedel@BB. The automatic content extraction (ACE) program–tasks, data, and evaluation. 2004. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.78.8442&rep=rep1&type=pdf.

[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] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[6] Dirk Weissenborn, Tomas Kocisky, and Chris Dyer. Dynamic integration of background knowledge in neural nlu systems. *CoRR*, abs/1706.02596, 2017. URL `http://arxiv.org/abs/1706.02596`.

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

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

[9] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. *CoRR*, abs/1711.03953, 2017.