# Information Extraction with Weak Supervision

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

**Relevance to BAA** As information networks get larger and more complex, acquiring explicit supervision for the training of information extraction systems becomes extremely expensive. Both the representation of the information in the 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.

## Background

1. Train NLU and use the NLG system as signal
2. Train an NLU system by extracting information that makes it easy to reconstruct text (NLG)
3. TODO: maybe redundant with prev section, can merge later
4. TODO: Cite recent IE work to show all of them have strong supervision (hopefully)

Natural language processing (NLP) can largely be decomposed into two separate but closely related tasks: natural language understanding (NLU) and natural language generation (NLG). In this proposal we argue that the two tasks are highly complementary, and we capitalize on their duality by proposing a method to train a NLU system without direct supervision. More specifically, we elaborate on a framework that allows an information extraction system to use the performance of a generative model of text as learning signal. We focus on the task of summarizing a table of data, which we refer to as Data-to-Text (D2T).

## Why is D2T important?

1. Convince readers that this is a good information extraction task, similar to semantic parsing / instruction following?
2. Highlight: Closed domain
3. TODO: probably need to rewrite this, rethink framing

Unlike tasks such as translation or summarization where one would need to first understand a source passage then generate a target passage, D2T isolates the task of NLU from NLG by introducing an intermediate representation in the form of structured data. This makes the task similar to semantic parsing, program induction, and instruction following, except for the lack of compositionality in D2T's intermediate representation. Whereas the other tasks assume an executable intermediate representation, D2T has a latent representation that takes the form of alignments from words in the summary to their respective referent data. This allows D2T to serve as a useful benchmark for conditional text generation. In addition to establishing a convenient representation, D2T makes the assumption that the text is generated conditioned only on functions of the data table. This assumption is central to the task as it allows practitioners to define heuristics both for providing weak supervision to an information extraction system as well as evaluating a summary's fidelity to the underlying data.

### Rotowire

1. Motivating extraction example
2. Transition towards more formal definitions so that we can define the tasks
3. TODO: Can maybe replace this whole section with a picture.

We provide a concrete example from a D2T dataset: In this proposal we focus on the recently proposed the Rotowire dataset [4]. Rotowire contains summaries of basketball games aligned with the respective box scores of those games. We consider a box score to be a collection or set of facts, where a fact is a tuple of an entity, relation type, and value. For example, consider the collection: a collection that consists of a single fact, (entity = Jeremy Lin, type = POINTS, value = 19), and a simple statement "Jeremy Lin scored 19 points". The relation type is the label in the box score, for example POINTS, REBOUNDS, etc. In order to automatically evaluate this summary, thanks to our assumption the summary was generated from only the data, we are able to use an information extraction system (IE) to extract entity-value pairs from text, predict the type of the relation between the entity and value, and compare resulting tuple (entity, type, value) to the given collection of facts.

**Our proposal** Our goal is to maximize the coverage of the information extraction system in terms of the number of words contained within segments that are aligned to data, while minimizing the amount of supervision needed. We propose to learn an information extraction (IE) system as the approximate posterior distribution over alignments from segments of text to their generating data. We also propose to view the given data as incomplete and learn boolean-valued functions of the data.

### Problem Definition

1. Define notation
2. Define generation, mention sam and puduppully work.
3. Define extraction and its subtasks, mention regina and percy's work.
4. TODO: maybe combine align and values? Although I prefer to keep them separate since values will probably turn into fill in the away team's values given home team's

Rotowire consists of aligned box score data and basketball game summaries $\left\{(\mathbf{r}^{(1)}, \mathbf{y}^{(1)}), (\mathbf{r}^{(2)}, \mathbf{y}^{(2)}) \ldots\right\}$. For brevity, we refer to a single data and summary as $\mathbf{r}, \mathbf{y}$, omitting the superscript. Each data $\mathbf{r} = \{r_1, \ldots, r_N\}$ is a set of $N$ records, each of which has an entity, type, and value $r_i = (e_i, t_i, v_i)$. We refer collectively to all the entities, types, and values in a given data $\mathbf{r}$ as $\mathbf{e}, \mathbf{t}, \mathbf{v}$ respectively. Each summary $\mathbf{y} = \{y_1, \ldots, y_T\}$ is a sequence of tokens that makes up the text of the game description.

For generation, the goal is to learn a conditional model $p(\mathbf{y} \mid \mathbf{r})$ of the text given the data. This is simple to evaluate, as we can use the log-likelihood of a given summary under our model as a measure of performance. In Wiseman et al. [4] model takes the form of a conditional language model that can copy values directly from records in $\mathbf{r}$. Subsequent work in Puduppully et al. [2] decomposed the distribution $p(\mathbf{y} \mid \mathbf{r}) = \sum_{\mathbf{c}} p(\mathbf{y} \mid \mathbf{c}) p(\mathbf{c} \mid \mathbf{r})$ by introducing a content plan $\mathbf{c}$, which is a sequence of records drawn from $\mathbf{r}$. This was also previously implemented in prior work [1], which modelled the text generation process through a hierarchical hidden semi-Markov model.

We divide the information extraction task into three subtasks. Before outlining the tasks, we propose two measures of performance through which to evaluate an unsupervised information extraction system. The first is how well the information extracted from a summary allows a generative model to reconstruct the summary measured by the likelihood of the summary given the extracted information. We refer to this as reconstruction. The second is coverage, which we define as the number of words contained in segments that are aligned to a record in $\mathbf{r}$. (Do I need to argue why these are useful? And also how the tasks aim at increasing them by weakening assumptions or constraining model flexibility compared to previous work)

The first task (ALIGN) is to align segments of text to the records that generated them. This is similar to learning a content plan $\mathbf{c} \mid \mathbf{r}$ as in Puduppully et al. [2], however we are interested in the **posterior** distribution $p(\mathbf{c} \mid \mathbf{r}, \mathbf{y})$ of the content plan $\mathbf{c}$ after observing the text $\mathbf{y}$. Liang et al. [1] utilize the fact that they define a model in which posterior inference is tractable, however tractability does not hold once the latent distribution becomes autoregressive. In Wiseman et al. [4] and subsequently in Puduppully et al. [2] this was accomplished by separately training a classifier to predict the type $t$ of an entity $e$ and value $v$ pair within a sentence. The entity and value are extracted heuristically by checking exact string matches within $\mathbf{e}$ and $\mathbf{v}$, and the supervision over $t$ is obtained through the following function [4]: $\text{findType}(\hat{e}, \hat{v}) = \{r.t : x \in \mathbf{r}, r.e = \hat{e}, x.r = \hat{v}\}$. However, this limits alignments exclusively to entities and values explicitly in $\mathbf{r}$. We would like to align whole segments of text in order to increase the coverage of our information extraction system.

The second task (VALUES) is to reconstruct values $v$ in the table $\mathbf{r}$. This is implemented on top of task (ALIGN). In particular, we want to learn $p(\mathbf{v} \mid \mathbf{y}, \mathbf{c}, \mathbf{e}, \mathbf{t})$, the distribution over all values given the summary, the content plan, all entities, and all types.

The third task (FUNCTIONS) is the most ambitious. In order to demonstrate the flexibility of the framework, we propose a method to further learn functions of $\mathbf{r}$ in an unsupervised manner. (TODO)

## Model

1. TODO: flesh this out later.

We begin by defining a model for (ALIGN) and proceed to (VALUES) and (FUNCTIONS).

Our generative model factors into the likelihood and prior: $p(\mathbf{y}, \mathbf{c} \mid \mathbf{r}) = p(\mathbf{y} \mid \mathbf{c})p(\mathbf{c} \mid \mathbf{r})$. Our likelihood is given by a conditional language model with a copying mechanism as in Wiseman et al. [4] (cite Caglar later) in addition to monotonic attention [3, 5]. The prior is simply an autoregressive model over records. (This is the simplest baseline, can make things much more complicated) (Posterior is actually tractable for this, since it's a HSMM, will change this later)

Our initial IE model for (ALIGN) is given by $p(\mathbf{c} \mid \mathbf{y}, \mathbf{r})$, which includes both a segmentation of the summary as well as the alignments. Initially we parameterize $p(\mathbf{c} \mid \mathbf{y}, \mathbf{r}) = \prod_t p(c_i \mid \mathbf{y}, \mathbf{r})$ as a fully factored distribution over alignments. (Structured attention for pairwise potentials later)

TODO: values, functions

## Learning and Inference

1. TODO: flesh this out later
2. Question: Do I need to motivate approximate inference? I could also argue that rather than computing the posterior exactly, at test time using the approximation directly can be more efficient, especially if it is fully factored

(Is this the right place for this?) A conditional latent variable model $p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})$ models an observed $\mathbf{y}$ as well as an unobserved $\mathbf{z}$ conditioned on $\mathbf{x}$. When fitting such a model, we would like to maximize the evidence $p(\mathbf{y} \mid \mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})$ which marginalizes over the latent $\mathbf{z}$. Depending on the structure of $p(\mathbf{y}, \mathbf{z} \mid \mathbf{x})$, the marginalization procedure may be intractable to perform exactly. For example, this is the case with an autoregressive model for the latent variable $p(z_t \mid \mathbf{z}_{<t})$, where variable elimination's runtime would be exponential in the length $|\mathbf{z}|$. This also precludes tractable posterior inference, i.e. $p(\mathbf{z} \mid \mathbf{y}, \mathbf{x})$, since by Bayes' Rule we have $p(\mathbf{z} \mid \mathbf{y}, \mathbf{x}) = p(\mathbf{y}, \mathbf{z} \mid \mathbf{x}) / p(\mathbf{y} \mid \mathbf{x})$ which requires evaluating the intractable sum in the evidence. Therefore we resort to learning an approximation of the posterior through the variational principle: the best approximation within a family of distributions is the one with minimal KL-divergence to the model's posterior.

For task (ALIGN), we have the fully observed summary $\mathbf{y} = \mathbf{y}$, the unobserved content plan $\mathbf{z} = \mathbf{c}$, and all records as conditioning $\mathbf{x} = \mathbf{r}$. For task (VALUES), we again have the observed summary $\mathbf{y} = \mathbf{y}$, but we pretend the values are unobserved $\mathbf{z} = \{\mathbf{c}, \mathbf{v}\}$, and use the rest of the records as conditioning $\mathbf{x} = \{\mathbf{e}, \mathbf{t}\}$.

TODO: functions, and then cover KLpost = -ELBO + evidence

# References

[1] 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`.

[2] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. *CoRR*, abs/1809.00582, 2018. URL `http://arxiv.org/abs/1809.00582`.

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

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

[5] Lei Yu, Jan Buys, and Phil Blunsom. Online segment to segment neural transduction. *CoRR*, abs/1609.08194, 2016. URL `http://arxiv.org/abs/1609.08194`.