Semi-supervised Open Domain Information Extraction with Conditional VAE

Zhengbao Jiang (zhengbaj) [1] **Songwei Ge (songweig)** [1]
Ruohong Zhang (ruohongz) [1] **Donghan Yu (dyu2)** [1]

Abstract

Open information extraction (OpenIE) is the task of extracting open-domain assertions from natural language sentences. However, the lack of annotated data hurts the performance of current model and made it barely satisfactory. In this paper, we aim to improve the OpenIE model with the help of the semantic role labeling (SRL) data, which has a very similar goal of identifying predicate-argument structure from natural language sentences, but with more labeled instances available. We propose a semi-supervised OpenIE model, which jointly optimizes supervised loss and unsupervised loss by treating OpenIE labels as hidden variables to reconstruct observed SRL labels. Conditional variantional autoencoder (CVAE) is used to optimize the lower bound of the data log-likelihood. Different from traditional multitask or transfer learning, we apply a more direct way to exploit the correlation between OpenIE and SRL. We compare our model with transfer and multi-task learning, and the results corroborate that our framework is able to better utilize such correlation information.

1. Introduction

Open information extraction (OpenIE) (Banko et al., 2007a; Fader et al., 2011; Mausam et al., 2012) aims to extract structured information from unstructured natural language. The target is usually in the form of a n-tuples, consisting of a predicate, and several arguments. OpenIE is beneficial to many downstream tasks, such as question answering, text summarization, and knowledge base construction. Unlike traditional IE where a small set of target relations are provided in advance, Open information extraction aims at extracting as many potential relations as possible in a text based on the semantic information. In that way, it facilitates the domain-independent discovery of relations extracted from text and scales to large, heterogeneous corpora.

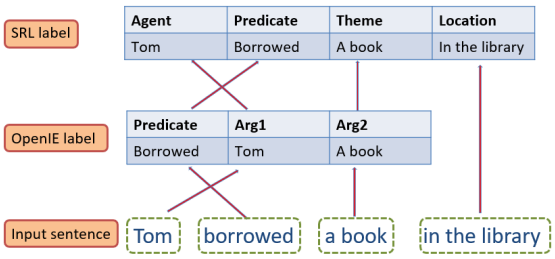

Figure 1. Illustration of correlation between OpenIE and SRL.

However, manually creating training data for this task is very expansive and time-consuming, because the same relation could be expressed in many ways in the text. Therefore, distant supervision, hand-crafted rules and bootstrapping (Mintz et al., 2009a; Pantel & Pennacchiotti, 2006) are heavily used in this area due to their advantage of only requiring no or a small amount of annotation data. However, these methods usually make strong assumptions which yield low data quality. Furthermore, some of manually defined rules and patterns generalize poorly to the different datasets.

In this paper, we show another effective way to improve OpenIE with non-annotated datasets by jointly learning with semantic role labeling (SRL). Both OpenIE and SRL can be formulated as a sequence tagging problem. In addition, SRL contains very similar output to OpenIE as shown in Figure 1. More importantly, the labels of SRL are relatively easy to obtain. To better exploit such correlation information between the outputs of two tasks, we design a semi-supervised learning framework based on the conditional variational autoencoder. We corroborate that our model can take better advantage of the SRL information than traditional multitask learning and transfer learning on this task. Our contributions are listed as follows:

- We propose a semi-supervised OpenIE model, which jointly optimizes supervised loss and unsupervised loss by treating OpenIE labels as hidden variables to reconstruct observed SRL labels.
- We propose using conditional variational autoencoder

[1]Carnegie Mellon University, Pittsburgh, PA 15213, USA.

10708 Class Project, Spring, 2019.

to optimize the lower bound of the data log-likelihood and serveral parameter sharing techniques to enable better representation learning and stablize the training.

- Experiments on benchmark dataset OIE2016 show that our model performs the best, comparing to other transfer learning and multitask learning models.

2. Background

2.1. Relation Extraction

With the explosion of data on the internet and the need to extract useful and sophisticated information from the data, the technology of Information Extraction (IE) and Information Retrieval (IR) has become more and more popular in NLP researches. In particular, relation extraction is an important task in IE, where raw texts are taken as input data, and the relations between entities are identified from those sentences in an automatic manner.

Since the labeled data is expensive to produce and thus limited in quantity, the supervised relation extraction suffers from a lot of problems. Various unsupervised and semi-supervised solutions are proposed.

In (Shinyama & Sekine, 2006), a pure unsupervised learning algorithm was used to extract relations from text documents. The algorithm employed a clustering algorithm for documents in which similar entities names appear, and then use "basic patterns" to group entities that play the same role together. In this way, the group entities entail the same relation and all the relations are extracted automatically. (Banko et al., 2007b) utilizes minimal data to extract relations from large corpus. In their *TextRunner* architecture, a self-supervised learner was trained a small corpus of samples to distinguish whether a relation tuple is trustworthy or not. Then, a single-pass extractor uses the learned classifies to extract potential relations from large corpus. Relations were obtained based on a redundancy-based assessor which assigns confidence level to each potential relation.

Alternatively, (Mintz et al., 2009b) uses the relation information in Freebase to provide distant supervision for relation extraction to avoid the lack of labeled data. The distant supervision method assumes that if two entities participate in a relation, any sentences which contains both of the entities are likely to express that relation. In the system by (Mintz et al., 2009b), the relation entities were extracted from Freebase and then matched to Wikipedia sentences. If a sentence contains a pair of entities in a relation, the system will extract features from that sentence. For example, *(Virginia, Richmond)* are both present in *Richmond, the capital of Virginia*, then features from the sentence are extract as a positive example for location-contains relation. Since any sentence can give incorrect relation, a negative sampling technique is used to train a multi-class logistic regression.

2.2. Semantic Role Labeling

In Natural Language Processing, Semantic Role Labeling is the process that assigns labels to words or phrases, in order to discover the predicate-argument structure of a sentence, such as "Who did what to whom", "when" and "where".

The early approaches of SRL utilizes the full syntactic tree, and the task has been usually divided into two phase procedures consisting of *recognition* and *labeling* of arguments. Various models had been applied to this two procedure SRL task, such as probabilistic models(Gildea & Jurafsky, 2002), Max Entroy(Fleischman et al., 2003), generative models(Thompson et al., 2003), etc.

Later approaches of the SRL systems(Carreras & Màrquez, 2005) try to reduce the dependencies on syntactic parsing and use only the paritial syntactic information. This avoids the use of full parser and external lexico-semantic knowlege basis. Most of the systems are based on a SVM tagging system, using IOB decisions on the chunk of sentences, and exploring a various choices of partial syntactic features, such as local information on contexts of words, internal structure of candidate argument, properties of target verb predicates, or the relation between the verb predicate and the constituent under consideration.

Recently, (He et al., 2017b) proposed a method using deep learning models to tag sentences with SRL labels without any syntactic parsing, which is considered as a prerequisite for all the previous works. Their model used 8 layers of Bi-LSTM with highway connection, orthogonal initialization and locked-dropout. They also used BIO, SRL and syntatic constrain decoding to improve the quality of the final tagging. The deep learning model improved the F1 accuracy and is found to be excel at long-range depencies compared with previous syntactic labeling methods. In our project, we will use a very similar LSTM based encoder-decoder architecture for tagging the SRL data, but we will extend the base model to fit in a semi-supervised learning scheme.

2.3. Semi-supervised learning

Semi-supervised learning can be effective when labeled data is limited or hard to obtain while the number of unlabeled data is much richer. With the recent advance of deep learning, modeling the distribution of unlabeled data at scale using neural based generative model is getting essential. Variational Auto-Encoder (VAE) (Kingma & Welling, 2013) is very successful in modeling data distribution and data generation. However, vanilla VAE can not generate data based on given context. Thus, Conditional VAE (Sohn et al., 2015) was proposed to solve this problem, where the input observations modulate the prior on Gaussian latent variables that generate the outputs. That is, for given observation x, z drawn from the prior distribution $p_\theta(z|x)$, and the output y

My	dog	also	likes	eating	sausage
B-ARG0	I-ARG0	O	B-V	B-ARG1	I-ARG1

Figure 2. OpenIE as a sequence tagging problem.

Seat	currently	are	quoted	at	$	361,000	bid	.
B-ARG1	B-ARGM-TMP	O	B-V	B-ARG2	I-ARG2	I-ARG2	I-ARG2	O

Figure 3. SRL as a sequence tagging problem.

is generated from the distribution $p_\theta(y|x,z)$.

In our setting, the unlabeled data can also have labels of another task different from OpenIE, such as SRL. Then we have two kinds of dataset which are target dataset $\langle \mathcal{X}_t, \mathcal{Y}_t \rangle$ and auxiliary dataset $\langle \mathcal{X}_a, \mathcal{Y}_a \rangle$ respectively. Transfer Learning (TL) (Pan & Yang, 2010) or Multi-Task Learning (MTL) (Caruana, 1997) can be used to solve this kind of problem. In TL, we usually train a model using auxiliary dataset and replace certain layers with new layers adopted to the target task, then only utilize target dataset to train the new layers while keep the parameters of other layers fixed. MTL jointly train different models for different tasks, and share some parameters or latent feature to constrain the model. MTL is popular in NLP (Collobert et al., 2011; Zhang & Weiss, 2016; Swayamdipta et al., 2017; Strubell et al., 2018; Yang et al., 2018). For example, LISA (Strubell et al., 2018) combines multi-head self-attention with multi-task learning across dependency parsing, part-of-speech tagging, predicate detection and SRL. However, the different tasks share the same $\mathcal{X}$ in their setting. while in our paper, there's much less overlapping between SRL dataset and OpenIE dataset. What's more, our model is more expressive due to the probabilistic latent variable while LISA is totally deterministic.

3. Methods

In this section, we explain our methods to train a semi-supervised OpenIE model with the help of additional SRL data. From the previous section, OpenIE and SRL tasks share a lot things in common, i.e they share a similar tag space, and they have correspondences among the different tags. To improve the performance of OpenIE model by utilizing large SRL datasets, we treat the OpenIE tag sequences as hidden variables, and decode the SRL labels based on that hidden representations. Specifically, we use a conditional variational autoencoder (CVAE) as the OpenIE model. In the following parts, we first formulate the semi-supervised problem; then introduce our proposed models; in the end, we discuss the model implementation in practice.

3.1. Problem formulation

Two data sources are available in our task: a small dataset with OpenIE annotations $\langle \mathcal{X}_{\text{oie}}, \mathcal{Y}_{\text{oie}} \rangle$, and a large dataset with SRL annotations $\langle \mathcal{X}_{\text{srl}}, \mathcal{Y}_{\text{srl}} \rangle$. $\mathcal{X}_{\text{oie}}$ and $\mathcal{X}_{\text{srl}}$ are two sets of sentences with minimal or no overlap. In our case, among these two datasets, there is a small amount of parallel data, i.e. sentences with both SRL annotations and OpenIE annotations. Each sentence $\mathbf{X}$ contains a sequence of words $\{w_1, w_2, \ldots, w_n\}$. For notation brevity, we omit the index and just use $\mathbf{X}$ to denote a sentence either from the OpenIE dataset or the SRL dataset. $\mathcal{Y}_{\text{oie}}$ contains the corresponding OIE labels for each sentence in $\mathcal{X}_{\text{oie}}$, and $\mathcal{Y}_{\text{srl}}$ contains the corresponding SRL labels for each sentence in $\mathcal{X}_{\text{srl}}$. Although the ultimate goal of OpenIE and SRL is to extract predicate-argument structure, we can formulate both problems as a sequence tagging problem (Stanovsky et al., 2018; Jia et al., 2018; He et al., 2017a).

Specifically, given a sentence $\mathbf{X} = (w_1, w_2, ..., w_n)$, the goal of OpenIE and SRL is to extract n-tuples $\mathbf{r} = (\mathbf{p}, \mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_m)$, composed of a single predicate $\mathbf{p}$ and several "arguments". We assume all components in $\mathbf{r}$ are contiguous spans of words and there is no overlap between them. The major different between OpenIE and RL is in the definition of "arguments". In OpenIE, arguments are just components in the sentences that are related to the predicate. For example, in Figure 2, we have two arguments: `ARG0` that specifies the subject of the predicate *likes* and `ARG1` that specifies the object of the predicate. In SRL, the case becomes a little complex. A predefined set of roles are used to explicitly represent the relation between each argument and the predicate. A SRL example is shown in Figure 3, where `ARGM-TMP` is a role indicating the temporal information of the predicate. As a result, we can interpret SRL as a more fine-grained predicate-argument structure identification task. However, it's worth to mention that there is no trivial mapping between two tag spaces. Instead, the correspondence usually also depends on the other factors such as semantic information and the context of the sequence.

Within this framework, a tuple $\mathbf{r}$ can be mapped to a

unique BIO (Stanovsky et al., 2018; He et al., 2017a) label sequence $\mathbf{Y} = (y_1, y_2, ..., y_n)$ by assigning `O` to the words not contained in $\mathbf{r}$, `B-V/I-V` to the words in $\mathbf{p}$, and `B-ARGi/I-ARGi` or other roles to the words in $\mathbf{a}_i$ respectively, depending on whether the word is at the beginning or inside of the span. We use $\mathbf{Y}^{\text{oie}}$ to denote a OIE label sequence and $\mathbf{Y}^{\text{srl}}$ to denote a SRL label sequence. Note that the OpenIE and SRL datasets have different tag spaces, i.e., $\{y_i | y_i \in \mathbf{Y}^{\text{oie}}, \mathbf{Y}^{\text{oie}} \in \mathcal{Y}_{\text{oie}}\} \neq \{y_i | y_i \in \mathbf{Y}^{\text{srl}}, \mathbf{Y}^{\text{srl}} \in \mathcal{Y}_{\text{srl}}\}$.

Given a sentence $\mathbf{X}$, the ultimate goal is to improve the OpenIE model $p(\mathbf{Y}^{\text{oie}}|\mathbf{X})$ using both OIE dataset $\langle \mathcal{X}_{\text{oie}}, \mathcal{Y}_{\text{oie}} \rangle$ and SRL dataset $\langle \mathcal{X}_{\text{srl}}, \mathcal{Y}_{\text{srl}} \rangle$.

3.2. Semi-supervised learning with conditional VAE

Given a sentence, we want to predict OpenIE tag sequence using $p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X})$, where θ represents the parameters of the model. Under supervised learning setting, one can directly optimize this model on the OpenIE dataset $\langle \mathcal{X}_{\text{oie}}, \mathcal{Y}_{\text{oie}} \rangle$. This can be achieved by minimizing the negative log-likelihood using the ground truth OpenIE tags:

$$\mathcal{L}_{\text{sup}} = -\log p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X}).$$

However, this dataset is very limited. As a result, the model can easily overfit this dataset and has poor generalization ability to the other practical datasets. Therefore, we propose to combine this supervised learning objective function with another unsupervised learning objective function from the SRL dataset. Considering the fact that the SRL task is very similar to the OpenIE task with respect to the resulting tag sequence, we can explicitly leverage SRL annotations to provide supervisions for OpenIE task, which can be achieved by a conditional variantional antuencoder (CVAE) model.

Generative Story In unsupervised learning, given a input sentence $\mathbf{X}$, we treat the OpenIE tag sequences $\mathbf{Y}^{\text{oie}}$ as hidden variables, which are then used to reconstruct the SRL labels $\mathbf{Y}^{\text{srl}}$. The basic rationale behind this is that only the proper OpenIE tag sequences are useful to reconstruct the SRL tag sequences due to the correspondence between them. The plate notation of our graphical model is shown in Figure 4. The generative model is:

$$p(\mathbf{Y}^{\text{srl}}|\mathbf{X}) = \sum_{\mathbf{Y}^{\text{oie}}} p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X}) p_\omega(\mathbf{Y}^{\text{srl}}|\mathbf{X}, \mathbf{Y}^{\text{oie}}), \quad (1)$$

where θ is the parameter of the OpenIE model and ω is the parameter of the reconstruction model (i.e., decoder), which predicts SRL tags conditioned on both sentence $\mathbf{X}$ and OpenIE tags $\mathbf{Y}^{\text{oie}}$.

Learning with conditional VAE Due to the large space of the hidden variables $\mathbf{Y}^{\text{oie}}$, it is intractable to exactly compute the marginal distribution in Equation 3.2. To mitigate this problem, we introduce a variational posterior distribution, i.e., the encoder $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{Y}^{\text{srl}}, \mathbf{X})$, to approximate the true posterior distribution.

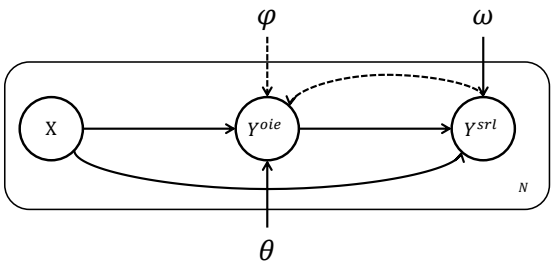

Figure 4. The plate notation of our conditional VAE. The solid lines represent prior model $p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X})$ and reconstruction model (decoder) $p_\omega(\mathbf{Y}^{\text{srl}}|\mathbf{Y}^{\text{oie}}, \mathbf{X})$ respectively. And the dashed line represents the variational approximation (encoder) $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{Y}^{\text{srl}}, \mathbf{X})$ to the intractable posterior distribution.

Instead of directly maximizing the marginal distribution which is intractable, we maximize the evidence lower bound objective (ELBO). After sampling some OpenIE tag sequences from the distribution implied by the encoder, the decoder aims to reconstruct the SRL tag seqeuces based on both the sentence and these OIE tag samples. In fact, only using the OpenIE tags may not be sufficient to reconstruct SRL tags because SRL contains more information than OIE.

The unsupervised loss defined as the negative ELBO is:

$$\text{ELBO} = \mathbb{E}_{\mathbf{Y}^{\text{oie}} \sim q_\phi}[\log p_\omega(\mathbf{Y}^{\text{srl}}|\mathbf{Y}^{\text{oie}}, \mathbf{X})] \quad (2)$$

$$- \text{KL}[q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{Y}^{\text{srl}}, \mathbf{X})||p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X})], \quad (3)$$

which includes three components:

- **encoder** (posterior model): $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{Y}^{\text{srl}}, \mathbf{X})$, which approximate the real posterior distribution.
- **decoder** (reconstruction model): $p_\omega(\mathbf{Y}^{\text{srl}}|\mathbf{Y}^{\text{oie}}, \mathbf{X})$, which reconstructs SRL tags conditioned on both the sentence and the OpenIE tags.
- **prior** (OpenIE model): $p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X})$, which is our targe model that we are eventually interested in.

Based on our assumption that only the correctly predicted OpenIE tag sequences are useful to reconstruct the SRL tag sequences due to the correspondence, maximizing the reconstruction loss in Equation 3.2 allows the model to learn a better posterior distribution of the OpenIE tag sequences. Consequently, the posterior model is expected to be more powerful than the prior model due to the extra guidance provided by the SRL labels. Simultaneously, by minimizing the KL distance between the posterior and prior distribution, the prior model is optimized to follow the steps led by the

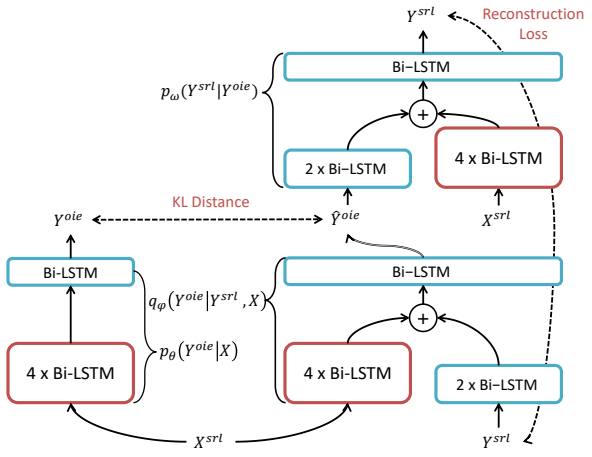

Figure 5. Illustration of our conditional VAE model. The solid lines represent the forward direction and the double line represents the sampling operation. The dashed lines represent the loss computation. The parameters in the red blocks are shared across different modules while the parameters in the blue block are model-specific.

posterior distribution. In addition, we don't want the prior distribution to move too far from the original prediction, so we also minimize the supervised loss in the meanwhile. The KL distance also constrains the solution space searched by the posterior model to be valid OpenIE tag sequence space.

Semi-supervised Learning The overall semi-supervised learning loss function is:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \cdot \mathcal{L}_{\text{unsup}} = \mathcal{L}_{\text{sup}} - \lambda \cdot \text{ELBO},$$

where we use λ to control the tradeoffs between supervised loss and unsupervised loss. During the training, the model parameters θ, ϕ, and ω are optimized jointly.

3.3. Model implementation

In this section, we will elaborate more about the implementation of our model in the setting of using neural networks. The framework of our semi-CVAE model for semi-supervised learning is illustrated in the Figure 5. As stated above, there are three components: encoder $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{Y}^{\text{srl}}, \mathbf{X})$, decoder $p_\omega(\mathbf{Y}^{\text{srl}}|\mathbf{Y}^{\text{oie}}, \mathbf{X})$, and prior model $p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X})$. Since all of these three components are conditioned on $\mathbf{X}$, they can share the parameters used in modeling $\mathbf{X}$ to reduce the training difficulty and risk of overfitting. As a result, each component is implemented with two modules: the base module (red blocks in Figure 5) and the specific module (blue blocks in Figure 5). Note that the base module is shared across three components. Since we are dealing with sequence labeling tasks, we use BiLSTM (Graves et al., 2013) as our building block. We will introduce the detail of each components in the the following sections.

Algorithm 1 Batch Gradient Descent for Conditional Variantional Auto-Encoder

input : SRL pairs $\langle \mathcal{X}^{\text{srl}}, \mathcal{Y}^{\text{srl}} \rangle$ and OpenIE pairs $\langle \mathcal{X}^{\text{oie}}, \mathcal{Y}^{\text{oie}} \rangle$ as minibatch with size B
$\theta, \phi, \omega \leftarrow$ Initialize parameters
repeat
  compute $p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X}^{\text{oie}})$, $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{X}^{\text{srl}}, \mathbf{Y}^{\text{srl}})$
  $\hat{\mathbf{Y}}^{\text{oie},(i)} \leftarrow$ Sample N fake OpenIE tag sequences from posterior distribution $q_\phi(\mathbf{Y}^{\text{oie}}|\mathbf{X}^{\text{srl}}, \mathbf{Y}^{\text{srl}})$
  $\mathcal{L}_r \leftarrow -\sum_B \sum_{i=1}^{N} \log\Big(p_\omega(\mathbf{Y}^{\text{srl}}|\hat{\mathbf{Y}}^{\text{oie},(i)})\Big)$ (Reconstruction loss)
  $g_{\phi,\omega,\theta} \leftarrow \nabla_{\phi,\omega}(\mathcal{L}_r + \text{KL}\left[p_\theta||q_\phi\right])$ (Gradients of minibatch negative ELBO)
  $\phi, \omega, \theta \leftarrow$ Update encoder and decoder using Gumbel Softmax or REINFORCE with $g_{\phi,\omega,\theta}$
  $\mathcal{L}_s \leftarrow -\sum_B \log\big(p_\theta(\mathbf{Y}^{\text{oie}}|\mathbf{X}^{\text{oie}})\big)$ (Supervised loss)
  $g_\theta \leftarrow \nabla_\theta \mathcal{L}_s$
  $\theta \leftarrow$ Update prior using gradient g_θ
until *convergence of parameters* (θ, ϕ, ω);

BiLSTM Building Block We use stacked BiLSTM with highway connections (Zhang et al., 2016; Srivastava et al., 2015) and recurrent dropout (Gal & Ghahramani, 2016) as our building blocks. All of the base module and specific modules are implemented using this architecture. Depending on the input of the module, we have three concrete instantiations:

- If the module takes $\mathbf{X}$ as inputs, the input embedding is the concatenation of word embedding and another embedding indicating whether this word is predicate:

  $$\boldsymbol{x}_t = [\boldsymbol{W}_{\text{emb}}(w_t), \boldsymbol{W}_{\text{mask}}(w_t = v)],$$

  and the resulted module is denoted as BiLSTM^x

- If the module takes $\mathbf{Y}$ (either $\mathbf{Y}^{\text{oie}}$ or $\mathbf{Y}^{\text{srl}}$) as inputs, the input embedding is the of tag embedding:

  $$\boldsymbol{x}_t = \boldsymbol{W}_{\text{tag}}(y_t),$$

  and the resulted module is denoted as BiLSTM^y

- If the module builds upon other modules, the input is the hidden state of previous layer:

  $$\boldsymbol{x}_t = \boldsymbol{h}_t,$$

  and the resulted module is denoted as BiLSTM^h

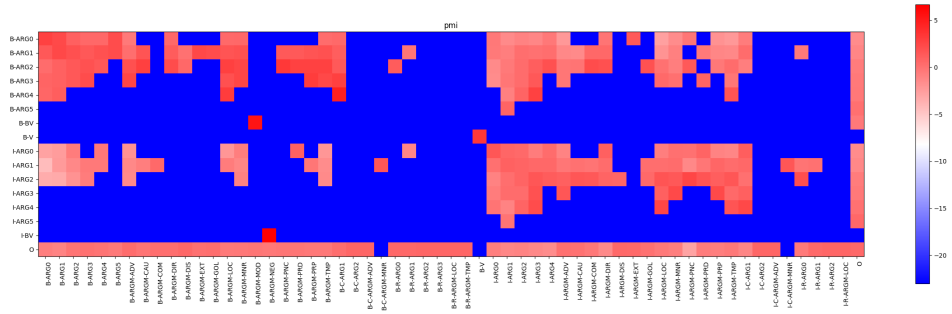

Figure 6. Correlation matrix between OpenIE and SRL labels. The red regions represent a larger normalized frequency of the co-occurrence between the corresponding OpenIE labels and SRL labels. The blue regions represent no co-occurrence in the parallel data.

Three Components Implementation Using the building blocks introduce above, we can build encoder, decoder, and prior by stacking based module and specific modules:

- encoder: $\text{BiLSTM}^{x} + \text{BiLSTM}^{y_{\text{srl}}} \rightarrow \text{BiLSTM}^{h_{\text{oie}}}$.

  $\mathbf{X}$ and $\mathbf{Y}^{\text{srl}}$ is modeled by BiLSTM^{x} and $\text{BiLSTM}^{y_{\text{srl}}}$ respectively, and their hidden state on last layer is added and fed to $\text{BiLSTM}^{y_{\text{oie}}}$ to make OpenIE tag predictions.

- decoder: $\text{BiLSTM}^{x} + \text{BiLSTM}^{y_{\text{oie}}} \rightarrow \text{BiLSTM}^{h_{\text{srl}}}$.

  $\mathbf{X}$ and $\mathbf{Y}^{\text{oie}}$ is modeled by BiLSTM^{x} and $\text{BiLSTM}^{y_{\text{oie}}}$ respectively, and their hidden state on last layer is added and fed to $\text{BiLSTM}^{y_{\text{srl}}}$ to make SRL tag predictions.

- prior: $\text{BiLSTM}^{x} \rightarrow \text{BiLSTM}^{h_{\text{oie}}}$.

  $\mathbf{X}$ is modeled by BiLSTM^{x}, and its hidden state on last layer is fed to $\text{BiLSTM}^{y_{\text{oie}}}$ to make OIE tag predictions.

The probability of the label at each position is calculated independently using a softmax function. At decoding time, we use the Viterbi algorithm to reject invalid label transitions (He et al., 2017a), such as `B-ARG0` followed by `I-ARG1`.

Parameter Sharing in Semi-CVAE The base module BiLSTM^{x} takes sentence $\mathbf{X}$ as inputs and is a 4-layer stacked BiLSTM, whose parameters are shared in three components. $\text{BiLSTM}^{y_{\text{oie}}}$ and $\text{BiLSTM}^{y_{\text{srl}}}$ takes tag sequence as inputs and is a 2-layer stacked BiLSTM. $\text{BiLSTM}^{h_{\text{oie}}}$ and $\text{BiLSTM}^{h_{\text{srl}}}$ takes previous hidden state as inputs and is a 1-layer BiLSTM. Parameters for the specific modules are not shared.

To avoid collapsing in decoder, which means that the decoder only use $\mathbf{X}$ to reconstruct SRL, instead of direct aggregation, we add a parameter μ to control the proportion between BiLSTM^{x} and $\text{BiLSTM}^{y_{\text{oie}}}$ in the merged information:

$$\mu \cdot \text{BiLSTM}^{x} + (1 - \mu) \cdot \text{BiLSTM}^{y_{\text{oie}}}$$

In practice, this is equivalent to adding dropout on the sentence information.

	Train	Dev.	Test
# sentence	1 688	560	641
# extraction	3 040	971	1 729

Table 1. Statistics of the OpenIE Dataset.

Optimization with Discrete Variables A basic approach to compute the stochastic gradients is shown in algorithm 1. The training process of our model is very similar to VAE (Kingma & Welling, 2013) but has several differences:

- Since the prior distribution is also a conditional distribution formulated by neural nets, we also need to calculate the gradients to update the prior model.

- Since the posterior distribution is discrete in our setting, the gradients are not able to be backpropogated properly. Either Gumbel Softmax (Jang et al., 2016) or REINFORCE (Miao & Blunsom, 2016) can be used to circumvent this issue. To be specific, we gradually decrease the temperature in Gumbel Softmax to stablize the training. For REINFORCE, we use average reward as baseline to reduce the gradient variance.

- The KL distance between the prior and posterior distribution can be calculated directly instead of sampling since both distributions are sequences of independent categorical distribution.

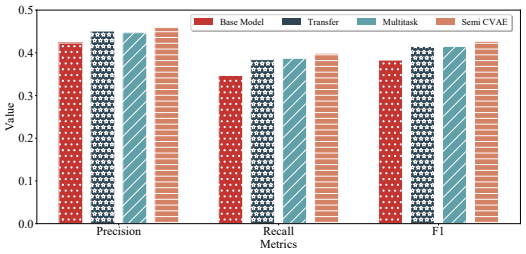

Figure 7. Overall performance on span-based metrics.

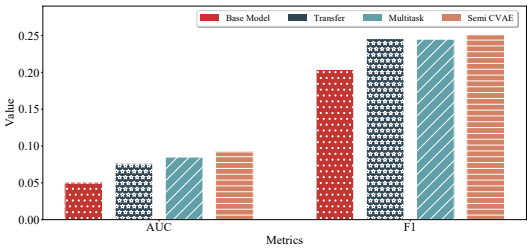

Figure 8. Overall performance on metrics proposed by (Stanovsky & Dagan, 2016).

4. Experiments

4.1. Experimental Settings

Dataset We use the OIE2016 dataset (Stanovsky & Dagan, 2016) to evaluate our method, which only contains verbal predicates. OIE2016 is automatically generated from the QA-SRL dataset (He et al., 2015), and to remove noise, we remove extractions without predicates, with less than two arguments, and with multiple instances of an argument. The statistics of the resulting dataset are summarized in Table 1. The PropBank-style, span-based SRL datasets: CoNLL-2012 (Pradhan et al., 2013) is used for semi-supervised learning. It provides gold predicates and their index in the sentence as part of the input. We follow the train-development-test split used in official evaluation. A quick look at the correlation between OpenIE and SRL labels are shown in Figure 6.

Evaluation Metrics We follow the evaluation metrics described by (Stanovsky & Dagan, 2016): area under the precision-recall curve (AUC) and F1 score. Note that this evaluation metric is very challenging. An extraction is judged as correct only if the predicate and all the arguments include the syntactic head of the gold standard counterparts, which is hard considering that lots of extractions in OIE2016 contain multiple arguments. For example, if an extraction contains three arguments, two of them are correctly identified and one is missing, the system still gets zero credit. Considering the fact that OpenIE is similar to SRL and to better measure incremental progress, we also report the standard evaluation metrics used for SRL systems: span-based precision, recall, and F1 measure.

Implementation Details The build block of our semi-CVAE model is a stacked BiLSTM. The network consists of 4+1 BiLSTM layers with 300-dimensional hidden units. ELMo (Peters et al., 2018) is used to map words into contextualized embeddings, which are concatenated with a 100-dimensional predicate indicator embedding. The recurrent dropout probability is set to 0.1. Adadelta (Zeiler, 2012) with $\epsilon = 10^{-6}$ and $\rho = 0.95$ and mini-batches of size 80 are used to optimize the parameters. We sample 5 OIE tag sequences in ELBO estimation. The weight of unsupervised loss is set to 0.3. Our implementation is based on AllenNLP (Gardner et al., 2018).[1]

overgenerated predicate	wrong argument	missing argument
41%	38%	21%

Table 2. Proportions of three errors.

Baselines We compare our method with three baselines:

- **Base Model** The standard BiLSTM sequence tagging model trained on OpenIE dataset from scratch.
- **Transfer Model** We pretrain all of the parameters of the BiLSTM sequence tagging model except for the tag prediction layer using SRL datasets. Then we use this pretrained weights to initialize the training on OpenIE datasets.
- **Multitask Model** Both the SRL model and the OpenIE model are trained simultaneously by sharing the lower layers, which can learn a general representation by hard parameter sharing.

4.2. Experimental Results

The span-based evaluation metrics are displayed in Figure 7, and AUC and F1 measure are reported in Figure 8.

(1) Overall, semi-CVAE achieves best performance across all of the span-based metrics, demonstrating the effectiveness of semi-supervised learning. Unsupervised learning on SRL datasets combined with supervised learning on OpenIE

[1] https://allennlp.org/models#open-information-extraction

A CEN	forms	an important but small part of a Local Strategic Partnership .		
A Democrat , he	became	the youngest mayor in Pittsburgh's history	in September 2006 at the age of 26 .	
An animal	that cares for its young but shows no other sociality traits is said to	be	" subsocial" .	

Table 3. Case study of extractions. Green for arguments and red for predicates.

datasets can successfully explore the correlation between SRL tag sequence and OpenIE tag sequence, leading to better representation learning.

(2) Transfer model, multitask model, and semi-CVAE model all significantly improve the performance over base model, which indicates that using SRL dataset to enhance OpenIE model is beneficial. Both SRL and OpenIE aims to identify predicate-argument structure from natural language sentences, and this similarity can be explicitly leveraged to do transfer learning and multitask learning.

(3) The improvement of semi-CVAE model over transfer model and multitask model can be explained from the graphical model in Figure 4. Given $\mathbf{X}$, $\mathbf{Y}^{\text{oie}}$ and $\mathbf{Y}^{\text{srl}}$ are not independent, which means that SRL ($\mathbf{Y}^{\text{srl}}$) prediction accuracy could be further improved by conditioning on OpenIE tags ($\mathbf{Y}^{\text{oie}}$). This is the core assumption of our model: better OpenIE tags lead to better SRL tags reconstruction. Furthermore, if we take a closer at our processed model in Figure 5, we can see a clear connection between semi-CVAE model and multitask model: if we set the trade-off parameter μ in decoder as 1, or unfortunately the decoder fail to use any information from OpenIE labels, the encoder will only be updated to mimic the prior model and they become almost identical. As a consequence, our model degenerates to multitask learning where the decoder is fully conditional on the sentence and thus becomes a standard SRL model. It shows that multitask learning is a special case of our proposed semi-CVAE model, and provides an explanation on why our model is superior to transfer and multitask learning from another perspective.

5. Case Study and Error Analysis

Table 3 showcase some extractions generated by our system. We can see that SRL are very similar to OpenIE. Arguments in OpenIE usually corresponds to arguments in SRL. Note that there are approximately 1000 sentences annotated with both SRL and OpenIE tag sequence. We analyze this parallel dataset by visualizing the correspondence between OpenIE labels and SRL labels in Figure 6, which is a pointwise mutual information matrix. We can see a clear pattern of the correspondences among different labels indicated by the red regions. To be specific, given SRL labels of the input sentence, the distribution of OpenIE labels are restricted in some certain spaces. Furthermore, there is not a straightforward mapping among the labels, which means complex model such as neural nets are necessary to solve the ambiguity.

To better understand the relatively low performance in this task, we randomly sample 50 extractions generated by our model and conduct an error analysis to answer this question. To count as a correct extraction, the number and order of the arguments should be exactly the same as the ground truth and syntactic heads must be included, which is challenging considering that the open IE datasets have complex syntactic structures and multiple arguments per predicate. We classify the errors into three categories, as shown in Table 2: (1) "Overgenerated predicate" is where predicates not included in ground truth are overgenerated, because all the verbs are used as candidate predicates. An effective mechanism should be designed to reject useless candidates. (2) "Wrong argument" is where extracted arguments do not coincide with ground truth, which is mainly caused by merging multiple arguments in ground truth into one. (3) "Missing argument" is where the model fails to recognize arguments. These two errors usually happen when the structure of the sentence is complicated and coreference is involved.

6. Conclusion and Future Work

Open domain information extraction is increasingly important as a result of the growing demand of extracting structured data from tremendous unconstrained data. However, the poor quality due to the limited number of labeled data has prevented current model from being used in the real settings. We proposed to improve the performance of OpenIE model with the data of SRL. Considering the connection between the property of the two tasks, we propose a semi-supervised learning framework based on the conditional VAE. The performance of our model compared with the baseline models attests that our assumption that the proposed semi-supervised setting can take advantage of the additional correlation information. Furthermore, our model provides a new idea on dealing with multitask learning which could be potentially extended to other problems with similar settings.

References

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. Open information extraction from the web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 2670–2676, 2007a. URL http://ijcai.org/Proceedings/07/Papers/429.pdf.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. Open information extraction from the web. In *IJCAI*, volume 7, pp. 2670–2676, 2007b.

Carreras, X. and Màrquez, L. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL-2005) at HLT-NAACL 2005*, 2005.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(Aug):2493–2537, 2011.

Fader, A., Soderland, S., and Etzioni, O. Identifying relations for open information extraction. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 1535–1545, 2011. URL http://www.aclweb.org/anthology/D11-1142.

Fleischman, M., Kwon, N., and Hovy, E. Maximum entropy models for framenet classification. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 49–56. Association for Computational Linguistics, 2003.

Gal, Y. and Ghahramani, Z. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016*, pp. 1019–1027, 2016. URL http://papers.nips.cc/paper/6241-a-theoretically-grounded-\application-of-dropout-in-recurrent-\neural-networks.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M. E., Schmitz, M., and Zettlemoyer, L. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018. URL http://arxiv.org/abs/1803.07640.

Gildea, D. and Jurafsky, D. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding*, pp. 273–278. IEEE, 2013.

He, L., Lewis, M., and Zettlemoyer, L. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 643–653, 2015. URL http://aclweb.org/anthology/D/D15/D15-1076.pdf.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 473–483, 2017a. doi: 10.18653/v1/P17-1044. URL https://doi.org/10.18653/v1/P17-1044.

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 473–483, 2017b.

Jang, E., Gu, S., and Poole, B. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Jia, S., Xiang, Y., and Chen, X. Supervised neural models revitalize the open relation extraction. *CoRR*, abs/1809.09408, 2018. URL http://arxiv.org/abs/1809.09408.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Mausam, Schmitz, M., Soderland, S., Bart, R., and Etzioni, O. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 523–534, 2012. URL http://www.aclweb.org/anthology/D12-1048.

Miao, Y. and Blunsom, P. Language as a latent variable: Discrete generative models for sentence compression. *arXiv preprint arXiv:1609.07317*, 2016.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pp. 1003–1011, 2009a. URL http://www.aclweb.org/anthology/P09-1113.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. Distant supervision for relation extraction without labeled data. 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 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics, 2009b.

Pan, S. J. and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359, 2010.

Pantel, P. and Pennacchiotti, M. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006*, 2006. URL http://aclweb.org/anthology/P06-1015.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2227–2237, 2018. URL https://aclanthology.info/papers/N18-1202/n18-1202.

Pradhan, S., Moschitti, A., Xue, N., Ng, H. T., Björkelund, A., Uryupina, O., Zhang, Y., and Zhong, Z. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pp. 143–152, 2013. URL http://aclweb.org/anthology/W/W13/W13-3516.pdf.

Shinyama, Y. and Sekine, S. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pp. 304–311. Association for Computational Linguistics, 2006.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. In *Advances in neural information processing systems*, pp. 3483–3491, 2015.

Srivastava, R. K., Greff, K., and Schmidhuber, J. Training very deep networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*, pp. 2377–2385, 2015. URL http://papers.nips.cc/paper/5850-training-very-deep-networks.

Stanovsky, G. and Dagan, I. Creating a large benchmark for open information extraction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2300–2305, 2016. URL http://aclweb.org/anthology/D/D16/D16-1252.pdf.

Stanovsky, G., Michael, J., Zettlemoyer, L., and Dagan, I. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 885–895, 2018. URL https://aclanthology.info/papers/N18-1081/n18-1081.

Strubell, E., Verga, P., Andor, D., Weiss, D., and McCallum, A. Linguistically-informed self-attention for semantic role labeling. *arXiv preprint arXiv:1804.08199*, 2018.

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. Frame-semantic parsing with softmax-margin segmental rnns and a syntactic scaffold. *arXiv preprint arXiv:1706.09528*, 2017.

Thompson, C. A., Levy, R., and Manning, C. D. A generative model for semantic role labeling. In *European Conference on Machine Learning*, pp. 397–408. Springer, 2003.

Yang, Z., Dhingra, B., He, K., Cohen, W. W., Salakhutdinov, R., LeCun, Y., et al. Glomo: Unsupervisedly learned relational graphs as transferable representations. *arXiv preprint arXiv:1806.05662*, 2018.

Zeiler, M. D. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701, 2012. URL http://arxiv.org/abs/1212.5701.

Zhang, Y. and Weiss, D. Stack-propagation: Improved representation learning for syntax. *arXiv preprint arXiv:1603.06598*, 2016.

Zhang, Y., Chen, G., Yu, D., Yao, K., Khudanpur, S., and Glass, J. R. Highway long short-term memory RNNS for distant speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5755–5759, 2016. doi: 10.1109/ICASSP.2016.7472780. URL https://doi.org/10.1109/ICASSP.2016.7472780.