A Supplemental Material

A.1 Deep Probabilistic Logic

Since human-labeled evidence sentences are seldom available in existing machine reading comprehension datasets, we use distant supervision to generate weakly labeled evidence sentences: we know the correct answer options, then we can select the sentences in the reference document that have the highest information overlapping with the question and the correct answer option (Seciton 2.1). However, weakly labeled data generated by distant supervision is inevitably noisy (Bing et al., 2015), and therefore we need a denoising strategy that can leverage various sources of indirect supervision.

In this paper, we use Deep Probabilistic Logic (DPL) (Wang and Poon, 2018), a unifying denoise framework that can efficiently model various indirect supervision by integrating probabilistic logic with deep learning. It consists of two modules: 1) a supervision module that represents indirect supervision using probabilistic logic; 2) a prediction module that uses deep neural networks to perform the downstream task. The label decisions derived from indirect supervision are modeled as latent variables and serve as the interface between the two modules. DPL combines three sources of indirect supervision: distant supervision, data programming, and joint inference. We introduce a set of labeling functions that are specified by simple rules, and each function assigns a label to an instance if the input satisfies certain conditions for data programming, and we introduce a set of high-order factors for joint inference. We will detail these sources of indirect supervision under our task setting in Section A.3.

Formally, let $K = (\Phi_1, \cdots, \Phi_V)$ be a set of indirect supervision signals, which has been used to incorporate label preference and derived from prior knowledge. DPL comprises of a supervision module Φ over K and a prediction module Ψ over (X, Y), where Y is latent in DPL:

$$P(K, Y \mid X) \propto \prod_v \Phi_v(X, Y) \cdot \prod_i \Psi(X_i, Y_i) \quad (5)$$

Without loss of generality, we assume all indirect supervision are log-linear factors, which can be compactly represented by weighted first-order logical formulas (Richardson and Domingos, 2006). Namely, $\Phi_v(X, Y) = \exp(w_v \cdot f_v(X, Y))$, where $f_v(X, Y)$ is a feature represented by a first-order logical formula, w_v is a weight parameter for $f_v(X, Y)$ and is initialized according to our prior belief about how strong this feature is[3]. The optimization of DPL amounts to maximizing $\sum_Y P(K, Y|X)$ (e.g., variational EM formulation), and we can use EM-like learning approach to decompose the optimization over the supervision module and prediction module. See Wang and Poon (2018) for more details about optimization.

A.2 Denoising with DPL

Besides distant supervision, DPL also includes data programming (i.e., $f_v(X, Y)$ in Section 2.3) and joint inference. As a preliminary attempt, we manually design a small number of sentence-level labeling functions for data programming and high-order factors for joint inference.

For sentence-level functions, we consider lexical features (i.e., the sentence length, the entity types in a sentence, and sentence positions in a document), semantic features based on word and paraphrase embeddings and ConceptNet (Speer et al., 2017) triples, and rewards for each sentence from an existing neural reader, language inference model, and sentiment classifier, respectively.

For high-order factors, we consider factors including if whether adjacent sentences prefer the same label, the maximum distance between two evidence sentences that support the same question, and the token overlap between two evidence sentences that support different questions.

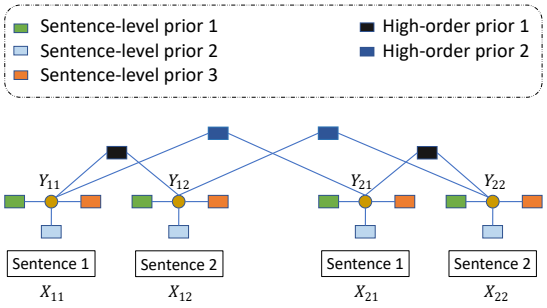

Figure 3: A simple factor graph for denoising.

We show the factor graph for a toy example in Figure 3, where the document contains two sentences and two questions. X_{ij} denotes an instance consisting of sentence i, question j and its associated options, Y_{ij} is a latent variable indicating the

[3]Once initial weights can reasonably reflect our prior belief, the learning is stable.

probability that sentence i is an evidence sentence for question j. We build a factor graph for the document and all its associated questions jointly. By introducing the logic rules jointly over X_{ij} and Y_{ij}, we can model the joint probability for Y.

A.3 Indirect Supervision Strategies

Besides distant supervision, DPL also includes data programming and joint inference. For data programming, we design the following sentence-level labeling functions:

A.3.1 Sentence-Level Labeling Functions

- Sentences contain the information asked in a question or not: for "when"-questions, a sentence must contain at least one time expression; for "who"-questions, a sentence must contain at least one person entity.
- Whether a sentence and the correct answer option have a similar length: $0.5 \leq \frac{\text{len(sentence)}}{\text{len(answer)}} \leq 3$.
- A sentence that is neither too short nor too long since those sentences tend to be less informative or contain irrelevant information: $5 \leq \#$ of tokens in sentence ≤ 40.
- Reward for each sentence from a neural reader. We sample different sentences and use their probabilities of leading to the correct answer option as rewards. See Section 3.2 for details about reward calculation.
- Paraphrase embedding similarity between a question and each sentence in a document: $\cos(e^{para}_q, e^{para}_{sent}) \geq 0.4$.
- Word embedding similarity between a question and each sentence in a document: $\cos(e^{w}_q, e^{w}_{sent}) \geq 0.3$.
- Whether question and sentence contain words that have the same entity type.
- Whether a sentence and the question have the same sentiment classification result.
- Language inference result between sentence and question: entail, contradiction, neutral.
- # of matched tokens between the concatenated question and candidate sentence with the triples in ConceptNet (Speer et al., 2017): $\frac{\#\text{ of matching}}{\#\text{ of tokens in sentence}} \leq 0.2$.
- If a question requires the document-level understanding, we prefer the first or the last three sentences in the reference document.

A.3.2 High-Order Factors

For joint inference, we consider the following high-order factors $f_v(X, Y)$.

- Adjacent sentences tend to have the same label.
- Evidence sentences for the same question should be within window size 8. For example, we assume S_1 and S_{12} in Figure 1 are less likely to serve as evidence sentences for the same question.
- Overlap ratio between evidence sentences for different questions is smaller than 0.5. We assume the same set of evidence sentences are less likely to support multiple questions.