Week 8 Homework

Samuel Cuthbertson SAMUEL.CUTHBERTSON@COLORADO.EDU

1. Answers about the reading

1.1. What are the actions used by the reinforcement learning system in this paper?

Either Wait, Commit, Next Word, or Verb. Wait simply enables the translator to read more input, Commit returns the current translation based on all the current input as the best translation, Next Word is similar to Wait, but uses the predicted next word instead of reading new input. Verb predicts the final verb of the input, placing the prediction at the end of the translated sentence.

1.2. What does Figure 5 tell you about how reliable the verb predictions are?

Figure 5 compares Batch, Monotone, and the learned SEARN predictors with the optimal predictor. Visually, it shows us how similar SEARN is to the optimal and how Monotone/Batch differ wildly. Since the optimal is the most reliable, it shows us how reliable SEARN is and how unreliable Batch/Monotone are by comparison.

1.3. Where is a classifier used in this system?

A classifier is used to predict the verb at the end of the sentence, and uses the model in Equation 2.

1.4. What are the features and labels of the classifier used in this system?

The features are described as being of three categories: Input, prediction, and translation. Input features are made up of both a bag of words from the input sentence, as well as the most recent word and bi-gram, and finally the length of the source sentence. Prediction features include the identity of the predicted verb as well as the predicted next word, along with the respective probabilities of both the above. Translation features are composed of a bag of words from the current translation, the score of the current translation, the score of the consensus translation, and the difference of the current and potential translation scores.

1.5. What is the reward in this system?

The reward is determined by `LBLEU` (Figure 3). `LBLEU` is a way of balancing reward for an accurate translation with reward for a "fast" translation. That is, a translation returned before all source sentence is read.

1.6. Explain each term in formula 3 and what the entire formula means.

Formula 3 describes how the next policy is created as the interpolation of the current policy and the ideal policy. π_{k+1} is the next policy. $\epsilon\pi_k$ is the portion of π_{k+1} from π_k, where π_k is the current policy. $(\epsilon - 1)\pi^*$ is the portion of π_{k+1} from π^*, where π^* is the ideal policy. ϵ is some mixer variable where $0 \leq \epsilon \leq 1$.

1.7. Explain each term in formula 4 and what the entire formula means.

Formula 4 describes the cost ($C(a_t, x)$) of an action a_t. Here, x_t is the current state and $Q(x, \pi^*(x_t))$ is the reward from taking the optimal action ($\pi^*(x_t)$) at state x_t. $Q(x, a_t(x_t))$ is the reward from taking the action current under consideration ($a_t(x_t)$). This leads to cost being defined as the regret from not taking the optimal action:

$$C(a_t, x) = Q(x, \pi^*(x_t)) - Q(x, a_t(x_t))$$

1.8. What are the baselines in this system?

Batch and monotone policies, as mentioned in Section 7.3.

1.9. How is the optimal policy calculated?

Section 5.1 is all that mentions calculating an optimal policy, and it is very vague on the specifics of how the oracle (optimal) policy is determined. To quote: "Using dynamic programming, we can determine such actions (constitute a oracle policy) for a fixed translation model." Using this passage to answer the question, it would appear that the answer is simply dynamic programming.

1.10. What is the policy here, and what does it learn to do?

The policy is a translator, and it learns how to translate quickly and accurately through predicting final verbs. More specifically, it learns what actions to take in order to accurately predict a translation in as few actions as possible (Figure 6).

1.11. What is the input to the policy?

An effective batch translation system and verb predictions.