Appendix: Synthetic QA Corpora Generation with Roundtrip Consistency

Chris Alberti **Daniel Andor** **Emily Pitler** **Jacob Devlin** **Michael Collins**
Google Research
{chrisalberti, danielandor, epitler, jacobdevlin, mjcollins}@google.com

A Supplementary Material: a Sketch of a Formal Justification for the Approach

This section sketches a potential approach to giving a formal justification for the roundtrip method, inspired by the method of Balcan and Blum (2005) for learning with labeled and unlabeled data. This section is intentionally rather speculative but is intended to develop intuition about the methods, and to propose possible directions for future work in developing a more formal grounding.

Assume that we have parameter estimates $\hat{\theta}_A$ and $\hat{\theta}_Q$ derived from labeled examples. The log-likelihood function for the remaining parameters is then

$$L(\theta_{A'}) = \sum_i \log p(a^{(i)}|q^{(i)}, c^{(i)}; \theta_{A'})$$

Estimation from labeled examples alone would involve the following optimization problem:

$$\hat{\theta}_{A'} = \text{argmax}_{\theta_{A'} \in \mathcal{H}} L(\theta_{A'}) \qquad (1)$$

where $\mathcal{H}$ is a set of possible parameter values—typically $\mathcal{H}$ would be unconstrained, or would impose some regularization on $\theta_{A'}$.

Now assume we have some auxiliary function $\beta(\theta_{A'})$ that measures the roundtrip consistency of parameter values $(\theta_{A'}, \hat{\theta}_A, \hat{\theta}_Q)$ on a set of unlabeled examples. We will give concrete proposals for $\beta(\theta_{A'})$ below. A natural alternative to Eq. 1 is then to define

$$\mathcal{H}' = \{\theta_{A'} \in \mathcal{H} : \beta(\theta_{A'}) \geq \gamma\}$$

for some value of γ, and to derive new parameter estimates

$$\hat{\theta}_{A'} = \text{argmax}_{\theta_{A'} \in \mathcal{H}'} L(\theta_{A'}) \qquad (2)$$

The value for γ can be estimated using cross-validation of accuracy on tuning data.

Intuitively a good choice of auxiliary function $\beta(\theta_{A'})$ would have the property that there is some value of γ such that: (1) $\mathcal{H}'$ is much "smaller" or less complex than $\mathcal{H}$, and hence many fewer labeled examples are required for estimation (Balcan and Blum (2005) give precise guarantees of this type); (2) $\mathcal{H}'$ nevertheless contains 'good' parameter values that perform well on the labeled data.

A first suggested auxiliary function is the following, which makes use of unlabeled examples $c^{(j)}$ for $j = 1 \ldots m$:

$$\beta(\theta_{A'}) = \frac{1}{m} \sum_j \sum_{a,q} p(a, q|c^{(j)}; \hat{\theta}_A, \hat{\theta}_Q) \log p(a|q, c^{(j)}; \theta_{A'})$$

where $p(a, q|c^{(j)}; \hat{\theta}_A, \hat{\theta}_Q) = p(a|c^{(j)}; \hat{\theta}_A) \times p(q|a, c^{(j)}; \hat{\theta}_Q)$.

This auxiliary function encourages roundtrip consistency under parameters $\theta_{A'}$. It is reasonable to assume that the optimal parameters achieve a high value for $\beta(\theta_{A'})$, and hence that this will be a useful auxiliary function.

A second auxiliary function, which may be more closely related to the approach in the current paper, is derived as follows. Assume we have some method of deriving triples $(c^{(j)}, q^{(j)}, a^{(j)})$ from unlabeled data, where a significant proportion of these examples are 'correct' question-answer pairs. Define the following auxiliary function:

$$\beta(\theta_{A'}) = \frac{1}{m} \sum_j f(p(a^{(j)}|q^{(j)}, c^{(j)}; \theta_{A'}))$$

Here f is some function that encourages high values for $p(a^{(j)}|q^{(j)}, c^{(j)}; \theta_{A'})$. One choice would be $f(z) = \log z$; another choice would be $f(z) = 1$ if $z \geq \mu$, 0 otherwise, where μ is a target 'margin'.

Thus under this auxiliary function the constraint

$$\beta(\theta_{A'}) \geq \gamma$$

would force the parameters $\theta_{A'}$ to fit the triples $(c^{(j)}, q^{(j)}, a^{(j)})$ derived from unlabeled data.

A remaining question is how to solve the optimization problem in Eq. 2. One obvious approach would be to perform gradient ascent on the objective

$$L(\theta_{A'}) + \lambda\beta(\theta_{A'})$$

where $\lambda > 0$ dictates the relative weight of the two terms, and can be estimated using cross-validation on tuning data (each value for λ implies a different value for γ).

A second approach may be to first pre-train the parameters on the auxiliary function $\beta(\theta_{A'})$, then fine-tune on the function $L(\theta_{A'})$. In practice this may lead to final parameter values with relatively high values for both objective functions. This latter approach appears to be related to the algorithms described in the current paper; future work should investigate this more closely.

References

Maria-Florina Balcan and Avrim Blum. 2005. A pac-style model for learning from labeled and unlabeled data. In *Proceedings of the 18th Annual Conference on Learning Theory*, COLT'05, pages 111–126, Berlin, Heidelberg. Springer-Verlag.