# Answering Reading Comprehension Tests using Multi-View Regression

Dean Foster, Daniel Kim, Andrew Pak and Lyle Ungar*
foster@wharton.upenn.edu, (seunghyu,apak2)@seas.upenn.edu, ungar@cis.upenn.edu
University of Pennsylvania
Philadelphia, PA 19104

* presenter

Computer "reading" of natural languages is a growing research topic, and raises the question: How do we tell if a program is actually capturing the "meaning" of a text? One answer is to define "meaning" as is done in humans, namely to ask: Can the computer answer the same reading comprehension questions that are given early learners of language? Looking at even easy reading assessment questions (see figure 1) suggests that many simple approaches will do poorly. We present an approach that represents "meaning" as a real-valued state vector associated with each word occurrence and use Multi-View Regression (MVR, a highly efficient way of estimating linear dynamical systems described below) to learn a dynamical belief net which, like a Kalman filter or HMM, estimates the state based on the preceding sequence of words and predicts the next word to be "emitted" based on that state.

We assess our MVR method using standard reading comprehension tests used for elementary school children. In these tests, students read a short passage and answer multiple choice questions about it. (See Figure 1.) We have a set of such questions developed by the Lexile corporation for assessing reading ability of elementary school students, and answers to these questions given by a large group of students. We compare the errors made by our software to those made by the students, and to errors on predictions from n-gram language models built using the Google n-gram collection. As one would expect, N-gram models do poorly on these tests, as they fail to capture the necessary longer range dependencies.

*It fell – splash! – into the well. The princess watched her golden ball sink deep into the water of the well, and she began to cry. She cried harder and harder.*

*Question: The princess* _____ *her ball*
*1. cleaned*
*2. threw*
*3. sold*
*4. lost*

Figure 1: Example reading comprehension test

Many problems in Natural Language Processing can be modeled by dynamic Bayesian networks such as HMM's, which take as input a sequence of words and estimate a state vector for

each word instance. These state vectors can then be mapped to labels such as part of speech tags or entity type or, as is done in this paper, used in the style of language models to predict deleted words. Dynamic belief networks are attractive, capturing local context information not included in Latent Semantic Analysis (LSA) or in Latent Dirichlet Allocation (LDA)-based topic models, which find latent vectors for entire documents. However, dynamic belief nets suffer a number of problems, including being computationally difficult to estimate for large corpora. In this paper, we present a new multi-viewed learning method based on Canonical Correlation Analysis (CCA) that offers a highly efficient method of estimating state for sequence problems. Our method yields optimal models for certain linear dynamical systems, and is potentially extremely useful for many other dynamic Bayesian networks.

Our MVR model, in the style of a Kalman filter or HMM, assumes that each word can be characterised by two views, the words preceding it in the text, and the words following it in the text, and that these two views are conditionally independent given some hidden state. Under this assumption, Canonical Correlation Analysis, (CCA, a generalization of PCA to a pair of matrices) between the views learns a mapping from the previous words to a latent state space. The model learned is linear, and only requires doing an eigenvalue calculation similar in cost to that of PCA. By using exponential smooths of the word counts, we are able to scale to reasonable vocabularies and state space sizes. We then use the hidden state estimated by CCA as features for supervised learning (e.g. logistic regression) to predict what word will be seen as a function of state.

MVR comes with a strong set of theoretical guarantees, such as the following

**Theorem 1** *Assume there exists some $k$-dimensional hidden state vector $S$ such that for the two views $X^{(1)}$ and $X^{(2)}$ (the words before and after the target word) and for the emission $Y$ to be predicted are all conditionally independent:*

$$\mathit{Cov}(X^{(1)}, X^{(2)}|S) = \mathit{Cov}(Y, X^{(1)}|S) = \mathit{Cov}(Y, X^{(2)}|S) = 0 \tag{1}$$

*Further suppose that the $\mathit{Cov}(X^{(1)}, S)$ and $\mathit{Cov}(X^{(2)}, S)$ both have rank $k$. Then,*

- *The unconditional cross correlation $\Sigma_{12}$ matrix between $X^{(1)}$ and $X^{(2)}$ is at least rank $k$.*
- *The best linear estimator of $Y$ based on $X^{(1)}$ is a vector lying in the subspace spanned by the first $k$ correlations between $X^{(1)}$ and $X^{(2)}$.*
- *The best linear estimator of $Y$ based on $X^{(1)}$ and $X^{(2)}$ is a vector in the $2k$ union of both sides.*

We trained MVR on thousands of novels taken from Project Gutenberg, and used the resulting model to predict multiple choice answers on our reading assessment stories. MVR combines both the short range dependencies typical of n-gram language models and the longer range dependencies more typically captured at the document level using LSA. It performs well above chance, but is not competitive with elementary school students. Error analysis suggests a number of enhancements to improve performance.

**topic: estimation, prediction, and sequence modeling**
**preference: oral/poster**