Definition Response Scoring with Probabilistic Ordinal Regression

Kevyn COLLINS-THOMPSON[a*], Gwen FRISHKOFF[b] & Scott CROSSLEY[b]
[a] *Microsoft Research, Redmond, U.S.A.* [b]*Georgia State University, Atlanta, U.S.A.*
*Contact author: kevynct@microsoft.com

Abstract: Word knowledge is often partial, rather than all-or-none. In this paper, we describe a method for estimating partial word knowledge on a trial-by-trial basis. Users generate a free-form synonym for a newly learned word. We then apply a probabilistic regression model that combines features based on Latent Semantic Analysis (LSA) with features derived from a large-scale, multi-relation word graph model to estimate the similarity of the user response to the actual meaning. This method allows us to predict multiple levels of accuracy, i.e., responses that precisely capture a word's meaning versus those that are partially correct or incorrect. We train and evaluate our approach using a new gold-standard corpus of expert responses, and find consistently superior performance compared to a state-of-the-art multi-class logistic regression baseline. These findings are a promising step toward a new kind of adaptive tutoring system that provides fine-grained, continuous feedback as learners acquire richer, more complete knowledge of words.

Keywords: Intelligent tutoring systems, definition scoring, ordinal regression.

1. Introduction

Because knowledge of a word's meaning is often acquired gradually, by exposure to the word over time and in different contexts, learners may have partial or incomplete knowledge of many words (Frishkoff, Collins-Thompson, Perfetti & Callan, 2008; Frishkoff, Perfetti, & Collins-Thompson, 2011). Further, they may benefit from instruction that is tuned to support different kinds of interactions with words that are partially known versus ones that are unknown (Frishkoff, Perfetti, & Collins-Thompson, 2010). In previous work we have described a method called MESA (Markov Estimation of Semantic Association), for estimating degrees of word knowledge by applying a random walk model to compute the distance between a user-generated synonym for a newly learned word and the actual (target) meaning. We used MESA to examine average trajectories across words within different instructional categories, e.g., when encounters with a new word are massed *vs*. spaced or when contexts that provide more *vs*. fewer clues to a word's meaning. These average measures have provided insights into how word knowledge develops through time and as a function of different learning and instructional variables, an important step towards an interactive and dynamic approach to vocabulary training.

In this study, we describe an extension of MESA for estimating degrees of word knowledge on a trial-by-trial basis. As in previous work, each learner is presented with a sentence that contains a rare target word, such as *aleatoric*, and is asked to provide a synonym for this word. The inputs to our model predictor are a *response word* from the student, along with a *target word* that the student is aiming to learn. The prediction output is a number on a four-point ordinal scale that captures how closely the student's response word matches the meaning of the target word. Our definition response scoring approach is based on probabilistic ordinal regression. It exploits rich semantic features on multiple types of

word relations, provides probabilistic scores and confidence estimates, and can achieve satisfactory performance on relatively small sets of human-labeled examples for training. We provide an initial evaluation of this approach by evaluating its prediction accuracy and the ability of the input features to discriminate between four levels of response accuracy.

2. Method

2.1 Dataset

Our target word list was a set of 60 English adjectives, verbs, and nouns selected by trained psychology experts. These words ranged from 'rare' to 'very rare' according to their frequency in the Kučera-Francis corpus (1979) with the requirement that very rare words appeared no more than 1 time out of 1 million tokens. We also provided, for each target word, a list of 1 to 3 reference words, which were higher-frequency synonyms that summarized the meaning of the target. For example, the reference words for the rare word 'limpid' were 'clear' and 'transparent'. The scoring algorithm uses these reference words as secondary targets when the target word itself is extremely rare and is found in few resources.

To create the labeled examples, each target word was paired with a short definition and two instances of the word in context: a high-constraint context, which provided rich cues to meaning (e.g., "I could not see a thing in the *X* room until I found the light switch."), and a low-constraint context that provided few if any cues to meaning (e.g., "Sharon did not expect to find that it would be this *X*."). For each target word, coders were asked to provide four kinds of responses:

(1) **Best Fit**: A word that matches the target and can be used in both contexts;
(2) **Strongly Related**: A word related to the target that can be used in both contexts;
(3) **Weakly Related**: A word that is weakly related to the target definition and can only be used in the low-constraint contexts;
(4) **Unrelated**: A word unrelated to the target, which cannot be used in either context.

We also evaluated scoring of *antonyms* of the 60 target words. Antonyms are challenging for many word similarity algorithms to score correctly because their semantic qualities are easily confused with those of synonyms. To label the responses, we employed three expert coders, resulting in three response files, each with 240 responses (60 targets x 4 response words, corresponding to the 4 ordinal levels as above)[1].

2.2 MESA Features

Prior work has used multiple resources to compute semantic distance, such as co-occurrence information, WordNet 2.0 (Harabagiu, Miller & Moldovan, 1999) dictionaries, and other resources (Mihalcea, Corley, & Strapparava, 2006). Each of these resources covers only a fraction of the potential relations between word pairs. By combining multiple resources using probabilistic chains of inference, it may be possible to bridge key gaps in a semantic network model. The MESA model adopts this approach (Collins-Thompson & Callan, 2007), and assigns a likelihood score to each target word on each learning trial. The target word's likelihood is derived from the stationary distribution of the Markov chain, which is approximated using a random walk. Details on the multiple word relations used by MESA are described in (Collins-Thompson & Callan, 2007).

[1] Researchers interested in using this dataset (non-commercial research purposes only) should contact the authors.

We used MESA synonymy, association, and morphology relations since these are the most effective combination for scoring synonyms. We added the ability to dynamically create new edges between terms that are not in the current word graph, and all current graph nodes. This is required us to handle (spell-corrected) free-form response words, such as those that would be likely to occur within an intelligent tutor. The edges are given uniform probability here, but more semantically focused schemes are possible and might give further prediction gains. In practice, we use a small number of walk steps (five) on a sparse representation of the word graph to perform the random walk. The random-walk based features we derived were the minimum, maximum, and average MESA log-likelihood scores of a response word, computed over all reference words for the target.

2.3 Features based on LSA

To add features that exploit word co-occurrence as a source of semantic information, we computed the LSA similarity score between each coder's response and the target concept, averaged over all reference words for the target, using the LSA Pairwise Comparison term-term comparison (http://lsa.colorado.edu/) with the default parameter settings. The resulting score between 0 and 1 was our LSA feature for that (target, response) pair.

2.4 Ordinal regression baseline

We used Gaussian Process Ordinal Regression, a state-of-the-art method recently introduced by Chu and Ghahramani (2005). GPOR outperforms previous approaches such as SVM-based ordinal regression or metric regression (Chu and Ghahramani, 2005). The GPOR method provides probabilistic prediction with confidence estimates for prediction and incorporates feature weighting as part of its model learning. GPOR also explicitly models the ordinal nature of the ratings.

2.5 Multi-class logistic regression baseline

We compare the effectiveness of GPOR for definition response scoring with a multi-class classification baseline using regularized logistic regression (Andrew and Gao, 2007). For each of the four response levels, a one-vs-all log-linear model was learned using the same training data as used for GPOR, also using 3-fold cross-validation. The L1 and L2 regularization weights were both set to default values of 1.0.

Prediction method	Precision (Micro-averaged)	Precision (Macro-averaged)
Random	0.250	0.250
Multi-class logistic regression	0.473	0.376
GPOR (all features)	**0.500**	**0.461**

Table 2. GPOR achieves higher precision than multi-class logistic regression baseline

3. Evaluation

We evaluated prediction effectiveness of GPOR using MESA and LSA features with the standard quadratic weighted kappa measure (Fleiss, 1971) and micro/macro-averaged precision (correct vs. incorrect label prediction). We used GPOR settings of a Gaussian kernel with noise variance set to $S = 0.40$, and default settings for other parameters. We

pooled the labeled datasets from the three raters, and used 3-fold cross-validation to produce train/test splits. Examples for the case where two raters' data are used to predict the ratings for the third rater are shown in Table 1. The results for all rater combinations (where R1+R2 → R3 means raters 1 & 2 are used for training and rater 3 for testing) are given, showing the full confusion matrix: each row corresponds to a different 'true' label, and columns show the predicted labels in each column. Weighted kappa varied from a minimum of 0.416 to a maximum of 0.519 (with kappa being on a [0,1] scale). Precision for individual labels was best for labels 1 and 4 in a range of 70 to 80%. Predicting the intermediate labels 2 and 3 was more difficult and had lower precision of 12 to 35%. Space does not permit showing a full learning curve analysis, but we found that as the amount of training data was varied: a) GPOR had consistently higher precision than LR across all training set sizes and b) this difference increased for small training sets (less than 20% of the original size).

3.1 Baseline comparison

The micro- and macro-average precision comparisons between GPOR and baselines based on random labeling and multi-class logistic regression are shown in Table 2. The random baseline results in a precision of 0.25 per category, since we have the same number of training examples for all four categories, and randomly picking a category is correct 25% of the time. GPOR prediction attains superior prediction accuracy over logistic regression for both micro- and macro-averaging (with each test slice having 240 instances).

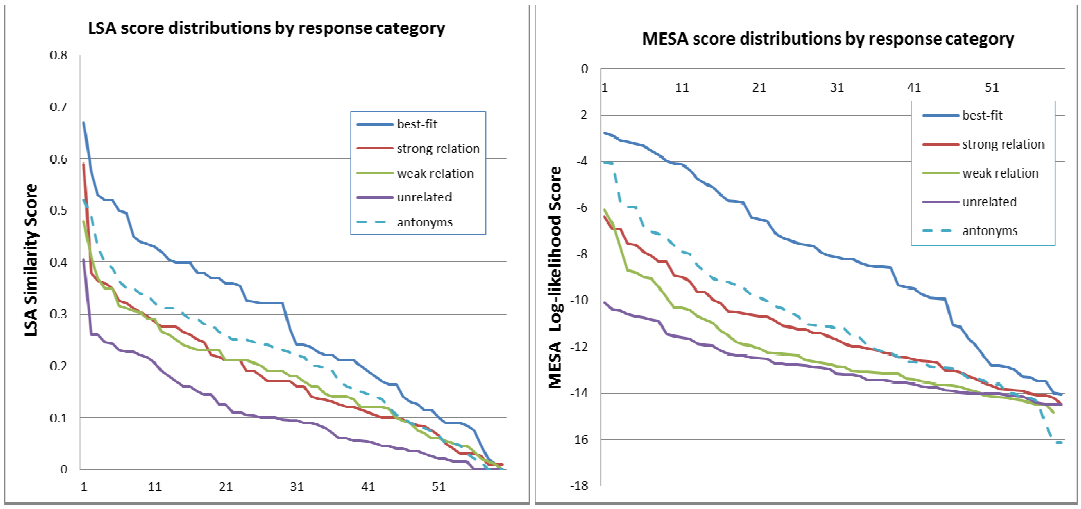

Figure 1. Distribution of LSA (left) and MESA (right) scores by 'true' response category, showing the relative ability of each method to discriminate between the response types. The x-axis gives the rank of response instance when sorted by score. The y-axis gives the feature value of the LSA score (left) or log-likelihood under the MESA probability model (right).

3.2 Feature comparison

To compare LSA and MESA (random-walk) score properties across response categories, we plotted their scores for responses categorized by their 'true' label (here, from rater 3). The results are shown in Figure 1. Both methods discriminate among the best-fit and unrelated categories: LSA was less effective at discriminating strongly- from weakly-related words. MESA scores were less effective at distinguishing weakly- from unrelated words. Antonyms were scored by both methods as intermediate between 'synonym' and 'strongly related', which seems appropriate for this task.

4. Conclusion

To summarize, we have shown that a supervised approach based on Gaussian Process ordinal regression can exploit relatively small amounts of expert-labeled training data for competitive performance on a difficult prediction problem: scoring definition responses on an ordinal scale. Our approach uses classification features that combine MESA and LSA scores to capture complementary aspects of word relationships. With these features GP regression achieves consistently higher precision than a multi-class logistic regression baseline, over a range of training set sizes. Further performance gains are likely with additional feature sets or by refining the prediction model.

Our long-range goal is to develop a robust method for online scoring that can be embedded in an intelligent tutoring system (ITS). By tracking changes in word-specific knowledge on a single trial basis, the ITS will be able to provide feedback to students in real time, enabling them to adjust their focus and strategies on subsequent trials. Further, the ITS will be able to adapt the presentation of stimuli based on student performance, combined with cognitive models of robust word learning. In this context, we view our results as a promising step towards an adaptive tutoring system that provides fine-grained, continuous feedback as learners gain richer and more complete knowledge of words.

References

G. Andrew and J. Gao. (2007) Scalable Training of L_1-Regularized Log-Linear Models. *Proc. of ICML-2007*.

W. Chu and Z. Ghahramani. (2005) Gaussian Processes for Ordinal Regression. *J. Machine Learning Research.* 6 (July 2005). pp. 1019–1041.

K. Collins-Thompson and J. Callan. (2007) Automatic and Human Scoring of Word Definition Responses. HLT-NAACL 2007. pp. 476–483.

S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. (1990) Indexing by latent semantic analysis. Journal of American Society for Information Science. 41:391–407, 1990.

J.L. Fleiss. (1971) Measuring Nominal Scale Agreement Among Many Raters. Psychological Bulletin, Vol. 76, No. 5. pp. 378–382.

W.N. Francis and H. Kučera. (1979) Standard Corpus of Present-Day American English. Dept. of Linguistics, Brown Univ., Providence, 1979.

G. Frishkoff, K. Collins-Thompson, C. Perfetti, J. Callan. (2008) Measuring incremental changes in word knowledge: Experimental validation and implications for learning and assessment. *Behavior Research Methods*, Vol. 40, No. 4. pp. 907–925.

G. Frishkoff, C. Perfetti, K. Collins-Thompson. (2010) Lexical quality in the brain: ERP evidence for robust word learning from context. *Developmental Neuropsychology*, Vol. 35, No. 4. pp. 376–403.

G. Frishkoff, C. Perfetti, K. Collins-Thompson. (2011) Predicting Robust Vocabulary Growth from Measures of Incremental Learning. *Scientific Studies of Reading*, Vol. 15, No. 1. pp. 71–91.

S. Harabagiu, G. Miller, D. Moldovan. (1999) WordNet2 - a morphologically and semantically enhanced resource. In Proceedings of SIGLEX-99, pp. 1–8.

R. Mihalcea, C. Corley, and C. Strapparava. (2006) Corpus-based and Knowledge-based Measures of Text Semantic Similarity. AAAI 2006.