Automatic Essay Scoring

Kenton W. Murray
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
kwmurray@cs.cmu.edu

Naoki Orii
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
norii@cs.cmu.edu

Abstract

Standardized tests are hampered by the manual effort required to score student-written essays. In this paper, we show how linear regression can be used to automatically grade essays on standardized tests. We combine simple, shallow features of the essays, such as character length and word length, with part-of-speech patterns. Our combined model gives significant reduction in prediction error. We discuss which features were effective in predicting scores.

1 Introduction

Efforts to extend standardized tests beyond multiple choice questions are limited by the ability to grade responses. Potential applications of machine learning and natural language processing could allow for these methods to scale to free text. A consortium of 44 US States and the Hewlett Foundation are looking into systems to automatically grade standardized test essays using small amounts of manually labeled training examples.

Analyzing natural language, or free-form text used in everyday human-to-human communications, is a vast and complex problem for computers regardless of the medium chosen, be it verbal communications, writing, or reading. Ambiguities in language and the lack of one "correct" solution to any given communication task make grading, evaluating or scoring a challenging undertaking. In general, this is a perfect domain for the application of machine learning techniques with large feature spaces, and huge amounts of data containing interesting patterns.

In this project, we explore the use of linear regression from text features to directly predict the score of a given essay. Using l_1 regularization, we take a large feature space consisting of a variety of linguistic features and determine the most predictive ones. We are able to significantly reduce prediction error and obtain state-of-the-art results, comparable to human annotators.

2 Data

The data was made publicly available to users of Kaggle[1], a platform for machine learning competitions. Users can freely sign up and download the labeled training set of the data during a specific time window of which the competition is held under. The data consists of 8 different essay sets of varying length and prompts: an overview of the data is shown in Table 1. Essays had been graded manually on various scales, depending on the prompt, with at least two scores given for each essay. The tests had been administered to US students from 7th to 10th grade students and were written in English.

The essays can mainly be divided into two types: Source Dependent Responses and Persuasive/Narrative/Expository responses. Source Dependent Responses are prompts based upon a passage that the students first has to read. Persuasive/Narrative/Expository responses asks students for

[1] http://www.kaggle.com

Table 1: Dataset statistics

Essay Set	Essay Type	Domain	Score Range	Average Length	train	dev.	test	total
1	Persuasive/Narrative/Expository	-	2-12	350 words	1,284	321	178	1,783
2	Persuasive/Narrative/Expository	Writing Applications	1-6	350 words	1,296	324	180	1,800
		Language Conventions	1-4					
3	Source Dependent Responses	-	0-3	150 words	1,244	310	172	1,726
4	Source Dependent Responses	-	0-4	150 words	1,276	319	177	1,772
5	Source Dependent Responses	-	0-4	150 words	1,300	325	180	1,805
6	Source Dependent Responses	-	0-4	150 words	1,296	324	180	1,800
7	Persuasive/Narrative/Expository	-	0-30	250 words	1,131	282	156	1,569
8	Persuasive/Narrative/Expository	-	0-60	650 words	521	130	72	723

stories, anecdotes, or formal arguments to persuade the reader in agreeing with the student's opinion on a particular topic. Essay Set 2 has two different grades associated for two different domains of writing: writing applications and language conventions. The other seven essay sets were holistically scored. Overall, the dataset provides a wide breadth of standardized essay prompts and domains on which one must learn robust algorithms in order to give high-precision predictions across the different essay sets.

Essays had been anonymized before being released to the public using the Named Entity Recognizer (NER) developed by the Stanford Natural Language Processing group (Finkel et al., 2005). Replacement IDs of the @ sign followed by words in all capitals were used instead. Name Entities of People, Organizations, Locations, Times/Dates, Numbers, Percents, E-mail Addresses, and Money were replaced.

3 Methodology

3.1 Existing Methods

Automated essay scoring is a highly commercialized market, and accordingly, not much is known about existing methodology in the public domain. We briefly compare our the performance of our model against those of existing methods in Section 4.5.

3.2 Generalized Linear Models

We use linear regression to predict the score, denoted y, based on features $\boldsymbol{x}$ extracted from a given essay. More precisely, given an input feature vector $\boldsymbol{x} \in \mathbb{R}^m$, we predict an output $\hat{y} \in \mathbb{R}$ using a linear model with a weight of $\boldsymbol{\beta}$:

$$\hat{y} = \beta_0 + \boldsymbol{x}^\intercal \boldsymbol{\beta}$$

To learn values for the parameters $\boldsymbol{\theta} = \langle \beta_0, \boldsymbol{\beta} \rangle$, we minimize the sum of squared errors for a training set containing n pairs of essays and scores, $\langle \boldsymbol{x}_i, y_i \rangle$, where $\boldsymbol{x_i} \in \mathbb{R}^m$ and $y_i \in \mathbb{R}$ for $1 \leq i \leq n$:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}=(\beta_0,\boldsymbol{\beta})}{\arg\min} \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - (\beta_0 + \boldsymbol{x_i}^\intercal \boldsymbol{\beta})\right)^2 + \alpha P(\boldsymbol{\beta})$$

where $P(\boldsymbol{\beta})$ is the penalty term for the weights. In particular, we use l_1 regularization, where $P(\boldsymbol{\beta})$ takes the form:

$$P(\boldsymbol{\beta}) = \sum_{j=1}^{m} |\beta_j|$$

For each essay set, we tune α on our development set using a 5-fold cross-validation.

We chose the linear regression model for its ease of interpretability. That is, we can directly infer the model's prediction "trend" by observing its trained weights: the sign of a weight indicates the direction of the prediction, and its magnitude indicates the feature's importance in the prediction. In

particular, the l_1 norm, otherwise known as the Lasso, drives most of the weights to zero, allowing for the model to be compact and easily interpretable. However, it should also be noted that l_1 regularization does not necessarily identify features that are truly predictive of the scores. For example, if two features are highly correlated, one of the corresponding weights will be driven to zero, even if both features are predictive. A weight of zero in this case does not imply that that the corresponding feature is not predictive.

3.3 Features

We extract two types of text features: (1) simple, shallow features of the essays, such as character length and word length, and (2) part-of-speech n-grams. As the former type of features mainly has non-zero values, while the latter contain mostly values of zero and few non-zero values, we denote the former as *dense features* and the latter as *sparse features*. We outline the features below:

3.3.1 Dense Features

- Character count (*char_count*)
- Word count (*word_count*)
- Number of exclamation marks
- Number of question marks
- Number of "difficult" words (*vocab*). We obtained a list of 5000 words that frequently appears on the SAT[2].
- Number of spelling mistakes (*spelling*). Spell checking was done using Enchant[3].
- Number of stopwords (*stopwords*). A list of 127-word stoplist was obtained from NLTK[4].

3.3.2 Sparse Features

- Part-of-speech n-grams. We considered unigrams, bigrams, and trigrams. Texts were tagged with the Stanford part-of-speech tagger (Toutanova et al, 2003). We only included feature instances that occurred in at least five different essays. Feature values were binarized: n-grams that were observed at least once in an essay have a corresponding feature value of 1, while unobserved n-grams have a value of 0.

4 Experiments

4.1 Evaluation Metrics

Results are evaluated upon Pearson's correlation, mean absolute error, and quadratic weighted kappa. The quadratic weighted kappa score is a measure of agreement of our scores and the human annotator's gold-standard. 0 represents only random agreement between the raters and 1 is full agreement. For N possible essay ratings, an $N \times N$ matrix O is constructed where $O_{i,j}$ represents the number of essays receiving grade i from the first grader and j from the second rater. Additionally, a matrix E is constructed the same way, but assuming there is no correlation. The matrices are normalized so that they have the same sum. An $N \times N$ matrix w is also calculated where:

$$w_{i,j} = \frac{(i-j)^2}{(N-1)^2}$$

The quadratic weighted kappa is calculated by:

$$\kappa = 1 - \frac{\sum_{i,j} w_{i,j} O_{i,j}}{\sum_{i,j} w_{i,j} E_{i,j}}$$

4.2 Results

Results are presented in Table 2. We present results for the baseline, dense and sparse feature settings independently, and the combination of the two. The baseline consists of character count and word

[2] http://freevocabulary.com/

[3] http://www.abisource.com/projects/enchant/

[4] http://www.nltk.org/

count, which were the two features used for the Kaggle competition's baseline. We combine dense and sparse features by taking a simple concatenation of the feature vectors. Dense, sparse, and combination feature settings were all significantly better than the baseline under all three metrics, evaluated using the Wilcoxon signed-rank test.

Table 2: Results on the test set, reported with Pearson's correlation (r), mean absolute error (MAE), and quadratic weighted kappa (kappa). Within a row, **boldface** shows the best result among the different feature settings.

		Feature Setting			
Essay Set	**Metric**	**Baseline**	**Dense**	**Sparse**	**Comb.**
1	r	0.791	0.804	0.780	**0.833**
	MAE	1.415	1.386	0.756	**0.685**
	kappa	0.658	0.672	0.713	**0.809**
2 (WA)	r	0.679	0.693	0.682	**0.716**
	MAE	0.800	0.769	0.476	**0.469**
	kappa	0.556	0.575	0.603	**0.658**
2 (LC)	r	0.504	0.512	**0.645**	0.582
	MAE	0.895	0.873	**0.495**	0.504
	kappa	0.404	0.407	0.466	**0.538**
3	r	0.732	0.732	0.624	**0.733**
	MAE	0.487	0.487	0.535	**0.457**
	kappa	0.675	0.675	0.393	**0.687**
4	r	0.757	**0.758**	0.701	**0.758**
	MAE	0.492	**0.491**	0.558	**0.491**
	kappa	**0.678**	**0.678**	0.570	**0.678**
5	r	**0.818**	0.816	0.758	**0.818**
	MAE	0.511	0.506	0.507	**0.466**
	kappa	0.770	0.771	0.672	**0.780**
6	r	0.703	0.710	**0.734**	0.717
	MAE	0.606	0.597	**0.541**	0.583
	kappa	0.660	0.668	0.593	**0.675**
7	r	0.657	0.666	**0.787**	**0.787**
	MAE	4.769	4.707	2.210	**2.190**
	kappa	0.554	0.563	0.761	**0.773**
8	r	0.539	0.598	0.680	**0.721**
	MAE	8.298	7.831	3.257	**3.103**
	kappa	0.407	0.450	0.648	**0.698**

4.3 Feature Analysis

In order to analyze which part of part-of-speech n-grams contributed to the predicted scores, we display the prominent features and their weights under the sparse feature setting in Table 3. The feature weights can be directly interpreted as score gain contributed to the predicted value $\hat{y}$ by the occurrence of the corresponding feature. We note that set 8 was the only set that had part-of-speech features with negative weights, which we show in Table 4.

These sparse features make intuitive sense of why they account for higher scores on essays. For instance, in essay set 5, the highest weighted feature is a bigram "VBZ VBG", which is a gerund or present participle followed by a 3rd person singular verb. This is a more complex grammatical structure than basic English and is likely representative of stronger writing ability. Also note the frequent presence of the POS tag "RB" which represents adverbs. Adverbs modify verbs, but also any other part of speech that is not a noun, including clauses, sentences, and other adverbs — again indicative of greater writing ability.

Feature weights can be misleading, as large weights do not necessarily correspond to large effects on the output. For example, our sparse features are binary, whereas our dense features can take integer values greater than 1[5]. Hence, it is unreasonable to directly compare weights of dense and sparse

[5]We note that regularization and scaling were tried on the dense features, but the raw integer values performed best when combined with sparse features.

Table 3: Highly weighted features from the sparse feature setting.

1		**2 (WA)**		**2 (LC)**		**3**		**4**	
VBP TO	0.764	, PRP$	0.612	RB VBG	0.425	VB RP	0.188	,	0.008
, PRP$	0.760	RB VBG	0.332	VBG TO	0.312	JJ	0.003	JJ	0.003
NN RB	0.430	VBP TO	0.165	, PRP$	0.268	-	-	-	-
VBZ .	0.379	IN WP	0.156	PRP ,	0.265	-	-	-	-
VB RP	0.341	VBG TO	0.101	VBG NN	0.131	-	-	-	-
IN WP	0.318	WP VBG	0.031	WRB VBD	0.109	-	-	-	-
WRB PRP$	0.304	,	0.022	VBP TO	0.095	-	-	-	-
NNS MD	0.234	PRP MD	0.013	IN WP	0.061	-	-	-	-
NN PRP	0.201	PRP	0.003	NN RB	0.049	-	-	-	-
PRP .	0.139	NN	0.001	VBP PRP IN	0.034	-	-	-	-

5		**6**		**7**		**8**	
VBZ VBG	0.134	NN NN IN	0.083	RP IN	1.732	, PRP$	9.503
VB RP	0.098	VB WRB	0.049	EX VBP	1.228	WRB PRP$	1.436
VBP	0.009	VBP PRP IN	0.012	IN WP	1.111	MD VB	1.424
JJ	0.002	,	0.006	MD VB	0.952	NNS IN NN	1.315
-	-	NN	0.004	VB RP	0.943	VBG TO	1.195
-	-	JJ	0.003	VBG NN	0.796	IN WRB	1.025
-	-	-	-	PRP .	0.656	MD PRP VB	0.993
-	-	-	-	NN VBG	0.621	VBG NN	0.991
-	-	-	-	NN WDT MD	0.566	NNP POS	0.989
-	-	-	-	JJ NN NN	0.564	PRP VBP NNS	0.924

Table 4: Negatively weighted features in Set 8 under the sparse feature setting.

8	
FW VB	-0.619
FW VBD	-0.542
MD VB PRP	-0.436
FW	-0.418
VBP PRP	-0.403
CD CC	-0.398
FW VBP	-0.340
NN PRP RB	-0.248
PRP PRP	-0.218
NN NNP VBD	-0.169

features. Instead, as presented in a recent work (Yano et al., 2012), we calculate the *impact* of each feature on the output $\hat{y}$ with respect to feature j as:

$$\frac{\beta_j}{n} \sum_{i=1}^{n} x_{ij}$$

where i indexes the test set (of which there are n examples), and x_{ij} denotes the jth feature value of the i the test example. Features with the highest impacts under the combined feature setting is shown in Table 5.

4.4 Examples

We present sample essays from the test set below. These examples are manually selected to illustrate the strengths and weaknesses of the model.

4.4.1 Essays with Low Scores

We first present examples essays that were manually annotated with relatively low scores. The following is an example in Set 8 which is manually annotated with a score of 21 (out of 60). Our system predicted a score of 22.60:

Table 5: High-impact features. *Italics* denote dense features.

1		**2 (WA)**		**2 (LC)**		**3**		**4**	
char_count	2.840	*char_count*	0.979	*char_count*	1.585	*char_count*	1.661	*char_count*	1.446
word_count	1.786	*word_count*	0.707	PRP$ VBG	0.441	RP NN	0.191	*vocab*	0.054
NNP ,	0.808	*vocab*	0.532	*vocab*	0.412	RB VBG	0.013	*word_count*	0.001
VBZ VBG	0.793	PRP$ VBG	0.342	VBG RP	0.325	*word_count*	0.003	*spelling*	-0.012
vocab	0.605	, NNP	0.305	, VBG	0.246	*spelling*	-0.008	*stopwords*	-0.062
CC JJ	0.460	NNP ,	0.294	VBZ VBG	0.236				
CC RB	0.367	VBZ VBG	0.222	NN RB	0.166				
RB VB	0.352	NN PRP	0.129	...	...				
NN PRP	0.284	VBG RP	0.122	*spelling*	-0.045				
...	...	...	...	*word_count*	-0.099				
stopwords	-0.689	*stopwords*	-0.313	*stopwords*	-0.347				

5		**6**		**7**		**8**	
char_count	1.814	*char_count*	2.720	*char_count*	7.710	*char_count*	12.389
stopwords	0.474	*word_count*	0.524	PRP .	1.585	NNP ,	9.391
, NNP	0.128	VBG JJ	0.063	NN PRP	1.250	RB VB	5.136
RB VBG	0.060	*vocab*	0.051	RP NN	1.046	*vocab*	4.542
RP NN	0.042	VBP PRP RB	0.049	RB PRP	0.869	VBG RP	2.396
vocab	-0.029	JJ ,	0.004	, MD	0.824	VBP CC	1.348
word_count	-0.209	*spelling*	-0.170	. PRP	0.652	...	...
		stopwords	-0.715	...	...	JJ RB	-0.299
				spelling	-0.620	*spelling*	-1.872
				word_count	-1.207	*word_count*	-2.442
				stopwords	-3.376	*stopwords*	-9.062

> I was selling some cookie dough for school and i waent to this olderly guys house. I went through the hole deal of raising money for my school and he said he was not interested. After that we got to talking i started making him laugh and eventaully he said "you make me laugh i like that i will buy some". I think that the laughter pklayed a key roll in him buying the cookie dough.

While the average length of essay in set 8 is 650 words, this essay consists of around only 80 words. As shown in Table 5, character count is the feature with the highest impact in essay set 8 and accordingly, this essay is given a low score.

We next give an example of an essay in Set 4 which is manually annotated with a score of 0 (out of 4), while our system predicted a score of 2.84:

> This story is all about overcoming hardships and disappointments as well as accepting and adapting to the things life throws at people, so it is appropriate that the story is ended with a goal and a determined attitude. Throughout the atire reading, Saeng sought comfort in the things most familiar to her, such as flower and the taste of bitter melon. She was hurting and disappointed, and those things were the only things that gave her peace. It was those things that save her the courage she needed to retake her drivers test in the spring, so it seems fitting that she end the story with the mention of the hibiscus. The end paragraph also shows that Saeng is adapting well to her new country. She says, "when they come back... in the spring, when the snow melts and the geese return..." At first the sounds of geese were alien to her, but now she has accepted their honking as a normal sound. This implies that she is learning to accept her new country, which is the perfect way to end the story.

Althought not evident from the text itself, this essay was off-topic from the given prompt. In terms of English grammar and mechanics, this writing shows no serious errors, and thus our system gives a non-zero, non-trivial score. Semantics is one area that must be taken into account in future[6].

[6] A naive approach would be to treat the prompt and essay as two word vectors, and calculate the similarity of the two. We can probably infer that near-zero similarity indicates that the essay is off-topic.

4.4.2 Essays with High Scores

We next present examples that were manually annotated with high scores.

The following is an example essay from Set 7 which is manually annotated with a score of 24 (out of 30). We predicted a score of 25.69:

> I walked into the big open room. The smell of crayon, animal crackers and dirt overcome the air and hit my nose like a big wave rolling in from the sea. I take my first nervous steps into the room... I never would have realized student teaching would have me doing so much teaching...Finally I get that light bulb going off in my head feeling I run to my purse grab my bag of candy and say, "when we finish learning the alphabet A-@CAPS4, anyone who can say ...

Looking at Table 5, we see that character count has the highest impact in this essay set, and though shortened in this quote this essay is very long, and accordingly our model gives a high score. Furthermore, we see that the part-of-speech tag "PRP", standing for Personal Pronoun, has a high impact on the test set; this essay has numerous sentences starting with the "PRP", "I". Looking beyond features with high impact to those with high weights, the POS Bigram "MD VB", or a modal verb such as "could" or "should", followed by a verb is one of the most highly weighted features in the sparse feature set. This essay contains many examples of this such as "would have" and "can say".

Contrasting the good performance in the previous essay, here is an example from Set 8 that is manually annotated with a perfect score of 60, while we predicted 43.01:

> Bell rings. Shuffle, shuffle. @CAPS1. Snap. EEEE. Crack. Slam. Click, stomp, @CAPS1. Tap tap tap. SLAM. Creak. Shoof, shoof. Sigh. Seventh class of the day. Here we go. "@CAPS2! Tu va ou pas? On a +tude cette class-l+. Tu peux aller au bibliotheque si tu veux...." @CAPS3 all blinked at me, @PERSON1, @NUM1le and @ORGANIZATION1, chocolate-haired and mocha skinned, impatiently awaiting my answer. The truth was, I knew @CAPS3 didn't really care if I came or not. It made no difference to them if I trailed a few feet behind like some pathetic puppy. I was silent but adorable, loved only because I was an @CAPS4. ...

It is evident that this essay contains more advanced grammatical constructs that most students did not include. It has irregular feature values, unlikely to have shown up in the training data, and thus received a low score. Furthermore, the vocabulary used contains quotes in French (likely to be classified as misspellings in our model's dense feature set). In addition, the POS tag, "FW", which represents a foreign word is in 4 of the negatively weighted features in essay set 8. As would be expected normally, using foreign language words in an essay written in English normally yields a low score. Here, it was used in a way that advanced the story and essay but our model did not capture this.

Through the previous discussion of essays of both high and low scores, we can see the robustness of this model. Lasso effectively reduces the number of significant features from a very high feature space. This is usually very useful in dealing with the large feature spaces encountered in natural language, but it can often fail at the extremes — either high or low. Length, one of the most salient features across all of the essay sets, is identified by l_1 regularization as being predictive but can cause artificially high scores on some essays. Additionally, often times in human language, the best works of written work are completely unlike the vast majority of other works in existence. These atypical data points are difficult for the vast majority of machine learning algorithms and make the study of natural language processing an interesting challenge.

4.5 Comparison with State-of-the-Art

Prior to the competition on Kaggle, the same dataset was distributed to several vendors and was experimented with existing commercial machine scoring systems (Shermis et al., 2012). In Figure 1, we show a comparison of our model's performance, labeled 10-701, against the commercial systems — including an open-source system developed here referred to as CMU. We see that our model

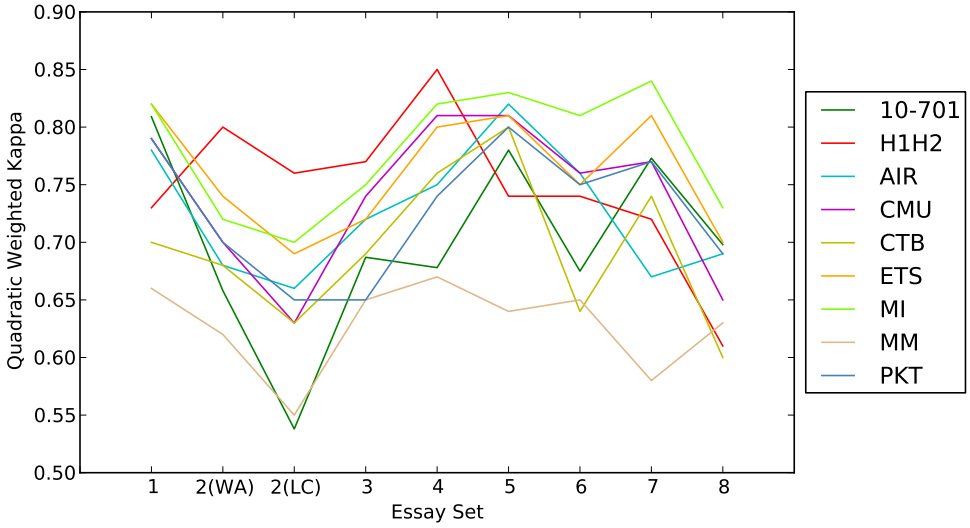

Figure 1: Comparison of our model against the State-of-the-Art

performs comparable to the state-of-the-art[7]. Additionally, H1H2 represents two human annotators' agreement with the same baseline human evaluation and performs at the same level as the ML methods. Essentially, the human κ score serves as a loose upper limit on performance, but is not a hard constraint as the quality of annotators may vary and could be higher if only the very best graders' scores were represented here.

5 Conclusion

We have greatly improved upon the baseline given for the competition and demonstrated the efficacy of using machine learning techniques to grade essays. We combined both sparse and dense feature sets using linear regression, and showed which features were effective in predicting the scores. We used the l_1 norm to make the weights sparse and easily interpretable, as the feature space for unigram, bigram, and trigram POS tags is very large. We obtained results that are comparable to the state-of-the-art and the inter-annotator agreement seen between humans. Noting that our method works well overall, we also demonstrated where our method breaks down when encountering outliers and why this is worthy of future exploration. Overall, machine learning methods can be used to evaluate and grade essays written in unstructured text, scaling and extending standardized tests beyond multiple choice.

References

[1] Finkel, J.R., Grenager T., & Manning, C. (2005) Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. of ACL*, pages 363-370.

[2] Shermis, M. D. & Hamner, B. (2012) Contrasting State-of-the-Art Automated Scoring of Essays: Analysis.

[3] Toutanova, K., Klein, D., Manning, C.G., & Singer, Y. (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proc. of NAACL-HLT*, pages 173-180.

[4] Yano, T., Smith, N., & Wilkerson, J.D. (2012) Textual Predictors of Bill Survival in Congressional Committees. In *Proc. of NAACL*.

[7] We note that our model's performance is an underestimate of the actual value. In the Kaggle competition, labels were only provided for the training data, and thus we split *their* training data into *our* training, dev, and test set. Accordingly, our model was trained on a much smaller dataset.