Ginsburg to Trump: Measuring Semantic Complexity

Pamela Toman
353 Serra Mall
Stanford, CA 94305
ptoman@stanford.edu

Pujun Bhatnagar
353 Serra Mall
Stanford, CA 94305
pujun@cs.stanford.edu

Zhiyang He
353 Serra Mall
Stanford, CA 94305
hzyjerry@stanford.edu

Abstract

Integrative complexity is a construct from political psychology that measures semantic complexity in discourse. Although this metric has been shown useful in predicting violence and understanding elections, it is very time-consuming for analysts to assess. We describe a theory-driven automated system that improves the state-of-the-art for this task from Pearson's $r = 0.57$ to $r = 0.73$ through framing the task as ordinal regression, leveraging dense vector representations of words, and developing syntactic and semantic features that go beyond lexical phrase matching. Our approach is less labor-intensive and more transferable than the previous state-of-the-art for this task. The success of this system demonstrates the usefulness of word vectors in transferring context into new problems with limited available data.

1 Introduction

Although Supreme Court justices and political candidates are known to use language differently from each other (e.g., [4, 7]), it is not immediately obvious how to measure their discourse complexity on a semantic rather than syntactic level. Political psychology, however, has a well-vetted and operationalized construct for this purpose. The "integrative complexity" metric has been studied for 40 years and has been shown useful in realms like predicting military attacks [23], violence [31], and winners of elections [4].

Integrative complexity was introduced by Tetlock and Suedfeld (see [22, 24]) to measure the extent to which an author entertains multiple perspectives and integrates them into coherence. Discourse with low integrative complexity admits the existence and legitimacy of only a single perspective, whereas discourse with high integrative complexity allows the simultaneous correctness of multiple perspectives. However, manual scoring of integrative complexity requires substantial training and time.

Previous attempts to automate the measuring of integrative complexity use surface lexical features combined heuristically [5] or classification techniques from machine learning [9]. We improve on the state-of-the-art with three main contributions. First, we formulate the task as an ordinal regression optimization problem. Second, we leverage dense vector representations of words. Third, we develop additional syntactic and semantic features. Through imposing this additional structure, we raise state-of-the-art performance on the 30-question test that assesses expert coders from Pearson's $r = 0.57$ to $r = 0.73$, and we show gains on four of five larger datasets. In this paper we discuss our approach and experiments on semantic features, providing insight into the structure of human complex thought.

2 Related work

We are aware of two published approaches to automatically score integrative complexity. Conway et al. [5] describes a hierarchical system that measures the presence of multiple ideas and then the integration of those ideas using a set of human-developed weights. Kannan-Ambili [9] explores machine learning methods and proposes a semantic coherence feature based in WordNet.

Conway et al. report the best performance on this problem to date. However, the model used is quite simple. Its use of surface features and hand-estimated weights for keywords require a substantial level of human involvement that means the approach cannot easily be transferred to new languages or genres. In fact, the authors note that language change over time poses a challenge to the system.

Kannan-Ambili's thesis explores a variety of

machine learning algorithms for classification, including logistic regression, support vector machines, and multi-class classifiers. The work proposes a "semantic coherence" measurement that averages a function of the path length and path depth in WordNet between words in the first sentence and the words in each succeeding sentence.

With this paper, we bring a machine learning and linguistics perspective to this problem that primarily engages psychologists and political scientists.

3 Data

Integrative complexity is measured on a 7-point scale. Scores of 1 reflect single-idea thinking, and scores of 7 are assigned to paragraphs with complicated integration of multiple ideas. Benchmark descriptions fall in between these poles. The metric is well described in Baker-Brown et al. [1]. In an appendix, we include example data that motivates the idea that semantic complexity differs from syntactic complexity.

Thanks to Lucien Conway, Associate Professor at the University of Montana, we obtained a dataset of 1108 paragraphs attached with the rounded average of multiple expert human coder scores. He and his lab hand-coded these paragraphs for previous projects. The dataset includes the training and testing materials used in a well-known seminar for training human annotators [6], as well as political debates, self-reflections, and writings on Christianity.

Much of the data has had capitalization and punctuation stripped, and in many datasets, the named entities are replaced with values like "[agent]". Additionally, the data lacks examples at high levels of complexity (see Figure 2) due to the infrequency with which people convey complex ideas.

4 Approach

We formulate an ordinal regression problem and evaluate lexical, syntactic, and semantic features, focusing specifically on alternative formulations of semantic complexity. Our implementation is available on GitHub.

4.1 Approach to modeling

We use all-threshold ordinal regression as generalized from logistic regression by Rennie and Srebro in 2005 [20] and implemented in `mord` [18]. To perform k-way classification, Rennie and Srebro's method transforms the input data to lie on a unidimensional line and learns $k-1$ distinguishing thresholds on that line. The loss function bounds the mean absolute error through penalizing each threshold that is crossed upon a mistaken classification.

Letting $s(l;y) = \begin{cases} -1, & \text{if } l < y \\ +1, & \text{if } l \geq y \end{cases}$, the all-threshold ordinal regression loss function $\mathcal{L}$ is:

$$\mathcal{L}(z;y) = \sum_{l=1}^{k-1} f\left(s(l;y)(\theta_l - z)\right)$$

In this equation, z is the number-line value assigned to an example, y is the target k-point value, θ_l is the threshold associated with class l, and $f(\cdot)$ is the logistic loss margin penalty function.

Since the data are ordinal, the use of ordinal regression is theoretically appropriate. We observe empirical support for this approach: this method produces 300% improvement over vanilla maximum entropy classification (from $r = 0.11 \pm 0.03$ to $r = 0.30 \pm 0.02$ in an early experiment) and 25% improvement over class-weighted maximum entropy classification ($r = 0.24 \pm 0.02$). We expect class-weighted ordinal regression would produce further gains.

4.2 Approach to features

We develop length, lexical, syntatic, and semantic features.

4.2.1 Length features

We extract the following length-related features: word counts, number of characters, mean and median word length, and number of words with length greater than 6.

4.2.2 Lexical features

In keeping with the political psychology literature, we pre-define lexical patterns whose presence indicates low and high levels of complexity. In particular, we extract the counts of types and tokens from string matching on word lists in the following categories: transitional phrases on fourteen themes from a GRE study guide [30] (e.g., *as a result*, *on the other hand*), comparatives (e.g., *-est*, *-ly*, *less*), modals (e.g., *will*, *could*), hedges (e.g., *practically*, *would argue*), conjunctives (e.g., *accordingly*, *because*), and punctuation (e.g., commas, quotation marks).

4.2.3 Syntactic features

We include basic syntatic features: the count of S and SBAR units heading phrases, and the mean / median / max / min / spread of syntactic tree heights in paragraph.

We also include the count of instances in which a definite determiner appears in the predicate. This feature draws on the insight from discourse analysis that people move from "old information" to "new information": the subject of each sentence is expected by the recipient, but the predicate may be new information. We pair this insight with the insight that authors use definite determiners only when they expect the reader is familiar with the content. As a result, we expect high numbers of definite determiners in the predicate when the author perceives only a single valid viewpoint. As motivating examples, we find this pattern in low-complexity texts in the training data, such as *We observe the depravity of our age*, and *Abortion threatens the moral and Christian character of this nation*.

4.2.4 Semantic complexity: Sentiment feature

Following suggestions in the literature (e.g., [8, 25, 32]), we hypothesize that complex writing is more likely to express both positive and negative aspects of a topic. To capture this intuition, we measure the min-max spread of sentiment in a paragraph using Turney and Littman's semantic orientation method [28], which calculates a sentiment score for each word through creating a scale from two opposing seed-sets. We experimented also with using the variance, but the min-max spread performed better. Due to the lack of punctuation in our dataset, we calculate sentence level spread by arbitrarily dividing documents into units of empirical sentence length 15.

4.2.5 Semantic complexity: Kannan-Ambili's feature

Kannan-Ambili [9] hypothesizes that simple arguments use more synonyms and closely related words than complex arguments.

We reimplement Kannan-Ambili's semantic coherence measure, which treats complexity as taxonomic distance between word pairs. She identifies whether variants on the same theme recur through averaging a function of the path length and path depth in WordNet [15] between words in the first sentence and the words in each succeeding sentence.

Specifically, Kannan-Ambili defines the semantic similarity between words w_i and w_j as the value $s_{ij} = f_1(\ell_{ij}) * f_2(h_{ij})$. In this equation, $f_1(\ell_{ij}) = 1$ if w_i and w_j are part of the same concept in WordNet or $f_1(\ell_{ij}) = e^{-\alpha \ell_{ij}}$ if they are part of different concepts in WordNet, where $\alpha = 0.2$ is a hyperparameter and ℓ_{ij} is the length of the path between the words. Additionally, $f_2(h_{ij}) = \frac{e^{\beta h_{ij}} - e^{-\beta h_{ij}}}{e^{\beta h_{ij}} + e^{-\beta h_{ij}}}$, where $\beta = 0.6$ is a hyperparameter and h_{ij} is the height of the lowest common subsumer of the word pairs.

Then for each word w_j in the first sentence, Kannan-Ambili estimates its semantic similarity $g(w_j)$ to the rest of the paragraph as the average of its pairwise similarities with each of the m other words w_i:

$$g(w_j) = \frac{1}{m} \sum_{i}^{m} s_{ij}$$

The total paragraph coherence, scaled to fall between 0 and 1, is the sum of the n word-level scores across each word in the first sentence:

$$P = \exp(-\sum_{j}^{n} g(w_j))$$

Our implementation uses the primary synset for each word, and returns 0 similarity when a word does not appear in WordNet or when the pair has no common subsumers.

4.2.6 Semantic complexity: PCA on GloVe features

We hypothesize that simple arguments lie in lower-dimensional semantic spaces than complex arguments. Given this intuition, we perform Principal Components Analysis on a matrix composed of the GloVe-50 [19] vectors for each unique word in a paragraph. Using the cumulative variance explained by each of the leading ten singular values, we predict each paragraph's integrative complexity score. We chose 10 singular values empirically: additional singular values do not improve performance, and they increase the proportion of paragraph matrices whose dimensionality is driven by paragraph length rather than by its underlying semantic dimensionality.

To calculate this measure for each paragraph, we tokenize the paragraph and form an m-by-n matrix M that vertically stacks the 50-dimensional

GloVe vectors for each unique token. We zero-center the columns of M to produce M', and find the matrices U, Σ, and V such that $M' = U\Sigma V$. We then calculate the amount of variance $v(s_i)$ explained by the leading 10 singular values s_i in Σ:

$$v(s_i) = \frac{s_i^2}{m^2 \cdot n \cdot \sum_{j=1}^{n} s_j^2}$$

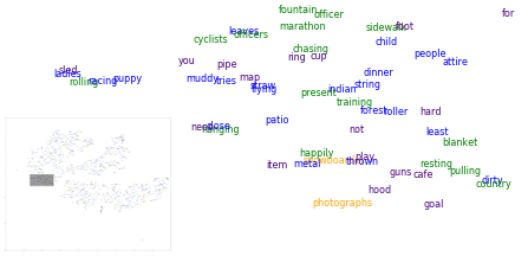
Our features are the cumulative amount of variance explained by each of the largest 10 singular values in Σ: $f(s_i) = \sum_{j=1}^{i} v(s_j)$.

4.2.7 Semantic complexity: LSTM entailment features

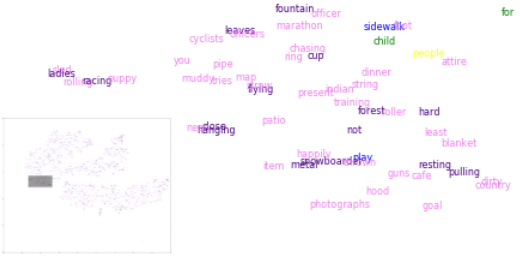
As our final semantic feature, we hypothesize that a paragraph's internal entailment indicates lack of diversity in viewpoints. To engage this intuition, we design a transfer learning task: we train a Long-Short Term Memory network to predict whether a sentence from an entailment dataset is the "more" or "less" complex member,[1] and then we feed integrative complexity data forward through the trained model and use the model's final state to predict integrative complexity scores. Through leveraging the 383,252 entailment examples in the SNLI [2] and SICK [13] corpora, we are able to surmount the limited amount of integrative complexity data available.

Our initial experiments initialized the LSTM with GloVe [19] word embeddings. Held-out accuracies under this approach, however, are no better than chance (50%). Initializing with the GloVe vectors and updated the embeddings using backpropogation produces only 53% accuracy even running the model for large number of epochs. After trying to make sense of what led to such low accuracy and looking into similar experiments that were run by others [10], we conclude that the GloVe vector space is not suitable for discerning the complexity of a sentence measured as entailment.

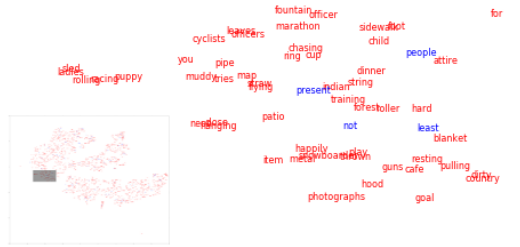
We instead allow the model to learn custom word representations during backpropagation. With custom embeddings, the entailment LSTM achieves 88% accuracy in distinguishing entailer from entailee – a substantial improvement. However, with custom embeddings, any words in the integrative complexity data that do not appear in the entailment training data back off to random vectors.



(a) Word lengths (warm: longer, cool: shorter).



(b) Word frequency on log scale (warm: more common, cool: less common).



(c) Entailment (red: more likely to be the entailer, blue: more likely to be the entailee.)

Figure 1: t-SNE plot of the embeddings of the most common 1000 words, colored by possible explanatory variables. No clear patterns.

To gain insight into what the custom embeddings are capturing that GloVe vectors do not capture, we visualize the learned embeddings using t-SNE [29]. We hypothesize that the word embeddings might be identifying "genericness" in a way that GloVe vectors do not capture, such that "child" may be far from "boy" but close to "kid" and "furniture". Within the visualization, however, the only pattern we see is that the "entailer" relations contain a wider variety of top words than the "entailee" relations.

We consider two approaches to using the model's final state to tackle predicting integra-

[1] For example, SNLI provides the following pair: *(A soccer game with multiple males playing.), (Some men are playing a sport)*. In this pair, the more detailed and complex left-hand side sentence entails the right-hand side sentence.

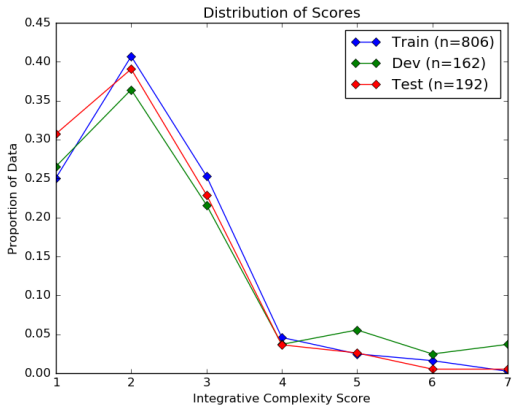


Figure 2: Distribution of human scores in data. Most paragraphs have low complexity.

tive complexity. In the first approach, we use the LSTM's softmax assessment that the input is the entailer vs. the entailee to estimate the likelihood that the input is complex. In the second approach, we introduce a new fully-connected layer of 128 nodes prior to the softmax, and we use the activations in this layer as the feature vector for predictions.

We test both the approaches and find that the softmax approach yields similar and, in some cases, better results when used as a feature. In most cases, the fully connected layer is functionally equivalent to the softmax value, but leads to poorer results when paragraph contains words of radically different length. We suspect that this happens because of noise introduced by 128 different signals in the fully-connected layer.

4.3 Approach to evaluation

We split the data such that approximately 70% of the data is in train (806 units), 15% is in dev (162 units), and 15% is in test (192 units) (see Figure 2). We control the splits in two ways: all units from a single research study are in the same set to facilitate out-of-sample extrapolation, and all sets contain a similar proportion of genres, split as general purpose, political debate, and human training data material.

Given the limited size of the dataset, we regularize all models and we retrain on the combined train+dev set after tuning the hyperparameters on the training set.

We follow Conway et al. and the qualitative political psychology literature on integrative complexity in reporting Pearson's correlation coefficient r and Cronbach's internal consistency measure α, both of which are frequently used for ordinal data.

The Pearson correlation r between two zero-centered vectors v_1 and v_2 is calculated as:

$$r = \frac{v_1^T v_2}{\sqrt{(v_1^T v_1)(v_2^T v_2)}}$$

Cronbach's α is a commonly used measure of psychometric scale reliability that is calculated as:

$$\alpha = \frac{\bar{c} \cdot n}{\bar{\sigma} + \bar{c}(n-1)}$$

where $\bar{c}$ is the average inter-item covariance, $\bar{\sigma}$ is the average variance, and n is the number of items.

Both metrics rely on relative relationships rather than absolute correspondence, which means they may over-represent model performance. For instance, although a system that predicts (1, 1, 2) when truth is (6, 6, 7) has perfect correlation, its performance is clearly imperfect. For this reason, we also explored average F_1 score, a classification metric that performs intuitively even in cases of unbalanced classes. Unfortunately, our limited amount of data means there are often classes with zero predicted examples, which renders average F_1 meaningless. Although we opt for comparison with the existing literature through r and α, future work may wish to use a statistic of ordinal association like Kendall's τ.

5 Results

We achieve state-of-the-art performance in automatically measuring integrative complexity. On the coding test that assesses human performance, we achieve $r = 0.73$ and $\alpha = 0.72$, which exceeds and ties respectively the best reported performance on this task in the literature. However, we do not achieve parity with expert human coders, which requires scores of $r \geq 0.85$ [1]. Table 1 provides the comparative performance of our system and the system described in Conway et al. [5].

Our system, like Conway et al.'s, performs best on the training materials provided by Baker-Brown et al. [1]. This suggests that those materials may be more straightforward to score than paragraphs in the wild.

We confirm Conway et al.'s finding that genre matters: our performance is worse on the corpus

Dataset (Genre)	Units	r Us	r C	α Us	α C
Train	*806*	*0.72*	*0.56*	*0.82*	*0.72*
Practice Sets (T)	155	**0.66**	0.61	**0.72**	0.76
Christian (C)	173	0.47	**0.59**	0.57	**0.74**
Heritability (H)	309	**0.60**	0.49	**0.75**	0.65
Nixon/Kennedy (P)	95	0.38	0.18	0.47	0.30
2004 Primaries (P)	74	0.56	0.42	0.70	0.55
Dev	*162*	*0.80*	*NR*	*0.86*	*NR*
Coding Manual (T)	68	0.73	NR	0.79	NR
Bush/Kerry (P)	94	0.40	0.34	0.51	0.54
Test	*192*	*0.61*	*0.41*	*0.63*	*0.58*
Coding Test (T)	30	**0.73**	0.57	**0.72**	**0.72**
Obama/McCain (P)	162	**0.47**	0.46	0.53	**0.63**

Table 1: Comparison of our "all features" model to that of Conway et al. [5]. "NR" indicates unreported datasets. Grey cells warn of unfair comparisons between models; these are datasets that Conway et al. used in test. "Genre" reflects the training materials (T), political debates (P), early Christian writings (C), and responses to hot-button issue prompts (H).

of early Christian writings and 1960s-era Nixon-Kennedy debates, likely because we are unfamiliar with the contexts and wrote no features specifically for them.

Given the variance in performance that we and Conway et al. observe, we suspect that genre has a substantial effect on the computational task as currently formulated. Additional data would assist in fully understanding and characterizing the ways in which semantic complexity manifests.

6 Analysis and discussion

Ablation tests (in Table 2), although limited by the inherent independence assumption, indicate that lexical features are a prime source of good performance. Syntactic features and length-related features are also useful, though less useful than lexical phrase matching.

The semantic approaches we test alternatively provide small improvements (PCA) or no improvements (Kannan-Ambili); some are actively harmful (sentiment features and LSTM transfer features). We suspect the harmful features are the result of the curse of dimensionality. By including the marginally useful features, we substantially increase the search space and thus decrease the chances of finding good parameterizations.

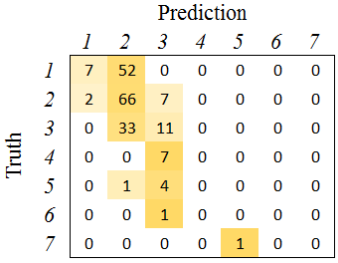


Figure 3: Confusion matrix on the test set using "all features" model.

6.1 Success analysis

Because lexical features are a prime source of good performance, we are unsurprised that Conway et al.'s rule-based scoring approach using only lexical features performed well. We expect combining our ordinal regression approach and syntactic features with their more extensive lexical features would produce further improvements in the realm of automating integrative complexity.

Our system succeeds primarily in cases where there are one or two ideas written in simple prose and on cases that display lexical features indicative of single mindedness like "ever" and "will", such as *i will eliminate capital gains taxes for the small businesses and the start ups that will create the high wage high tech jobs of tomorrow* . (Prediction: 2, Truth: 2).

6.2 Error analysis

In general, our system can distinguish "low" (scores of 1-3, no integration) from "high" (scores of 4+, some integration) complexity. A confusion matrix appears in Figure 3. Because the low/high structure of the predictions is prominent, it may be helpful for future modeling efforts to

	Feature Ablation	r	α
	All Features Model	0.612	0.633
⇓	Lexical Features	0.440	0.514
	Syntactic Features	0.596	0.614
	PCA Features	0.610	0.609
⇑	Kannan-Ambili	0.612	0.634
	Length Features	0.625	0.639
	Sentiment Features	0.634	0.634
	LSTM Transfer Features	0.638	0.652
	Removing Harmful Features	0.658	0.658

Table 2: Performance on test set given ablation of individual feature types.

follow the two-stage approach taken in Conway et al., in which the presence of multiple ideas is identified independently and prior to identifying whether those ideas are integrated, with a distinct model designed for each task.

The system rarely predicts high scores. We expect that the pull toward scores of 2, the most common class, is the result of limited data. Without sufficient data to learn a robust set of weights, the model is limited in what it can confidently learn.

Recurring errors fall into four categories: short but complex, political but simple, theological but simple, and missing real-world insight. We estimate preponderances by examining a set of 40 errors made on the test set, and 40 examples chosen at random from the cross-validated train set.

Short but Complex (~5% train, ~0% test) We fail to predict high complexity on some short syntactically simple texts. Fixing this type of error may require explicit reasoning about relations between ideas. It does not suffice to include features for argument structure as drawn from the literature (e.g., [12, 16, 21, 26]).

I like to seek the help of the people around me. Sometimes I gain a lot of valuable information this way and sometimes it is more confusing. Even if I do become a little more confused at first, it is worth seeking advice. Information, like doubt, holds possibilities. (Prediction: 3, Truth: 7)

Political but Simple (~8% train, ~8% test) We fail to predict simplicity when politicans use big words with complex syntax to say little. Progress on this type of error might occur with correlates of off-the-cuff political speech (e.g., "uh", "my opponent") that signal a need for a lower score.

now as far as president [name] is concerned i have often heard him discuss this question as i uh related a moment ago the president has always indicated that we must not make the mistake in dealing with the dictator of indicating that we are going to make a concession at the point of a gun whenever you do that inevitably the dictator is encouraged to try it again so first it will be [region a] and [region b] next it may be [region c] . (Prediction: 5, Truth: 1)

Theological but Simple (~15% train, N/A test) We fail to correctly predict theological statements as simple. We expect this is because believers see complex and contradiction-rife systems of faith as coherent entities. Fixing this type of error requires context knowledge and computational theory of mind.

stop your ears therefore when any one speaks to you at variance with jesus christ the son of god who was descended from david and was also of mary who was truly begotten of god and of the virgin but not after the same manner for indeed god and man are not the same he truly assumed a body ... (Prediction: 3, Truth: 1)

Real-World Insight (~100% train, ~100% test) On all examples, we find that the inability to reason about real-world relationships causes performance decreases. Addressing such errors requires background knowledge, possibly obtainable through dataset mining and additional modeling.

I like sweets but I don't really eat them that much because they tend to make me fat. (Prediction: 2, Truth: 5)

These results suggest that in addition to more data, this task requires the development of methods that can understand and reason about the structures of texts.

6.3 Sensitivity to preprocessing

We explore sensitivity to the data constraint of no punctuation by comparing (1) training our model with all punctuation removed, and (2) training our model with all punctuation available. We use three datasets from Suedfeld's training workshop [6] for which both punctuated and unpunctuated data are available: Practice Sets, Manual, Coding Test Sets.

With an independent test of our model using 10-fold cross-validation on these data, we find correlation and Cronbach's α are statistically unaffected by the presence of punctuation.

	Punctuated dataset	Unpunctuated dataset
Pearson's r	0.54 ± 0.07	0.55 ± 0.07
Cronbach's α	0.68 ± 0.07	0.70 ± 0.06

Table 3: Performance is statistically unaffected by the presence of punctuation.

Intuition / Hypothesis	r / α	Upheld?
Simple arguments lie in lower-dimensional semantic spaces than complex arguments.	0.38 / 0.46	✓
Complex writing is more likely to express both positive and negative aspects of an issue.	0.14 / 0.16	?
Simple arguments use more synonyms and closely related words than complex arguments.	0.00 / 0.00	–
Internal entailment indicates lack of diversity in viewpoints.	0.00 / 0.00	–

Table 4: Our results suggests only that semantically complex paragraphs lie in higher dimensional spaces than simple paragraphs. (Pearson's r and Cronbach's α provide predictive performance of each type of feature alone; 0.00s indicate across-the-board predictions of the most common class.)

6.4 Discussion of semantic featurs

To gain insight into the structure of the integrative complexity task, we tested four intuitions about semantic complexity in human prose. On these hypotheses, we are surprised to see that only the hypothesis that simple texts lie in low-dimensional semantic spaces had predictive power (see Table 4) – and that this approach works even though the GloVe dense word representations were not designed for this task.

Possible explanations for the lack of support for the other three hypotheses include:

- The sentiment measurements may be less informative than hoped because the dispassionate genres of the test data offered little opportunity for emotional awareness to shine, and/or because the training data included hot-button issues and the training did not transfer.

- The Kannan-Ambili metric may be unsuccessful at measuring integrative complexity because much of our data lacks the sentence boundaries on which the metric is based, and/or because synonymy is not a good predictor of complexity.

- The LSTM transfer learning may be unsuccessful at measuring integrative complexity because 43.71% of the vocabulary in the train+test data were out of sample for the LSTM and backed off to random embeddings, and/or because the entailment task as formulated does not support the integrative complexity task.

On the basis of these results, we suggest that future work on features continue to engage only with sentiment and the amount of variance explained in low-dimensional spaces, as well as length, lexical and syntactic features.

7 Conclusion

Our system improves on the state-of-the-art for measuring integrative complexity, raising correlations from 0.57 to 0.73 on the official 30-question test and improving scores on four of five larger datasets. The approach is less labor-intensive and more transferable to new languages and genres than the previous state-of-the-art.

Our improvements come from matching the limited data with stronger theoretical assumptions like ordinal regression, word vector semantic spaces, and linguistically driven syntactic features. We suspect there is room for additional gains in other areas driven by theory, including explicit feature selection, ordinal regression with class-weighting, two-stage modeling in which weights are learned to optimize scoring success, and features based on theory and systems for detecting discourse structures within text.

Although it is challenging to develop useful semantic features for this task, the only semantic hypotheses that had any predictive power used meaningful word vector representations. The integrative complexity task, then, is another unenvisioned area in which dense word vector representations work well.

Finally, the lack of substantial and diverse data is a major impediment to building successful systems for measuring integrative complexity. For automation of integrative complexity to succeed, additional data needs to be collected and made publicly available.

References

[1] G. Baker-Brown et al. "Coding Manual for Conceptual/Integrative Complexity". In: *Cambridge University Press* (1992).

[2] Samuel R. Bowman et al. "A large annotated corpus for learning natural language inference". In: *Proceedings of the*

2015 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2015.

[3] Lars Buitinck et al. "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 2013, pp. 108–122.

[4] Lucian Gideon Conway III et al. "Does complex or simple rhetoric win elections? An integrative complexity analysis of US presidential campaigns". In: *Political Psychology* 33.5 (2012), pp. 599–618.

[5] L. Conway et al. "Automated Integrative Complexity". In: *Political Psychology* (2014), pp. 603–624.

[6] *Electronic Complexity Workshop*. Personal Website of P. Suedfeld. Mar. 2005. URL: http://www2.psych.ubc.ca/~psuedfeld/index2.html.

[7] Deborah H Gruenfeld. "Status, ideology, and integrative complexity on the US Supreme Court: Rethinking the politics of political decision making." In: *Journal of Personality and Social Psychology* 68.1 (1995), p. 5.

[8] S. C. Houck, L. C. Conway, and L. J. Gornick. "Automated Integrative Complexity: Current Challenges and Future Directions". In: *Political Psychology* 35 (2014), pp. 647–659.

[9] A. Kannan-Ambili. "Automated Scoring of Integrative Complexity Using Machine Learning and Natural Language Processing". MA thesis. University of Georgia, Dec. 2014.

[10] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: *CoRR* abs/1405.4053 (2014). URL: http://arxiv.org/abs/1405.4053.

[11] Edward Loper and Steven Bird. "NLTK: The Natural Language Toolkit". In: *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*. ETMTNLP '02. Philadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 63–70. DOI: 10.3115/1118108.1118117. URL: http://dx.doi.org/10.3115/1118108.1118117.

[12] Nitin Madnani et al. "Identifying high-level organizational elements in argumentative discourse". In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics. 2012, pp. 20–28.

[13] Marco Marelli et al. "Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment". In: *SemEval-2014* (2014).

[14] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: http://tensorflow.org/.

[15] George A. Miller. "WordNet: A Lexical Database for English". In: *Commun. ACM* 38.11 (Nov. 1995), pp. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748. URL: http://doi.acm.org/10.1145/219717.219748.

[16] R. M. Palau and M. Moens. "Argumentation mining: the detection, classification and structure of arguments in text". In: *Proceedings of the 12th international conference on artificial intelligence and law*. ACM. 2009, pp. 98–107.

[17] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[18] F. Pedregosa-Izquierdo. "Feature extraction and supervised learning on fMRI: from practice to theory". Theses. Université Pierre et Marie Curie - Paris VI, 2015.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[20] Jason DM Rennie and Nathan Srebro. "Loss functions for preference levels: Regression

with discrete ordered labels". In: *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*. Kluwer Norwell, MA. 2005, pp. 180–186.

[21] C. Stab and I. Gurevych. "Identifying Argumentative Discourse Structures in Persuasive Essays." In: *EMNLP*. 2014, pp. 46–56.

[22] S. Streufert, P. Suedfeld, and M. J. Driver. "Conceptual structure, information search, and information utilization". In: *Journal of Personality and Social Psychology* (1965), pp. 736–740.

[23] P. Suedfeld and P. Tetlock. "Integrative complexity of communications in international crises". In: *Journal of Conflict Resolution* (1977), pp. 169–184.

[24] P. Suedfeld, P. Tetlock, and S. Streufert. "Conceptual/integrative complexity. In C. Smith (Ed.), Handbook of thematic content analysis". In: *Cambridge University Press* (1992), pp. 393–401.

[25] Philip E Tetlock et al. "Integrative complexity coding raises integratively complex issues". In: *Political Psychology* 35.5 (2014), pp. 625–634.

[26] S. Teufel. "Argumentative zoning: Information extraction from scientific text". PhD thesis. University of Edinburgh, 1999.

[27] *TFlearn: TFLearn: Deep learning library featuring a higher-level API for TensorFlow*. Software available from tflearn.org. 2016. URL: http://tflearn.org/.

[28] Peter D. Turney and Michael L. Littman. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association". In: *ACM Trans. Inf. Syst.* 21.4 (Oct. 2003), pp. 315–346. ISSN: 1046-8188. DOI: 10.1145/944012.944013. URL: http://doi.acm.org/10.1145/944012.944013.

[29] Laurens Van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.2579-2605 (2008), p. 85.

[30] "Vocabulary and spelling series - Transitional Words and Phrases". In: *GRE Study Guides And Strategies* (1996). URL: http://www.studygs.net/wrtstr6.html.

[31] D. A. Winter. "Slot rattling from law enforcement to law-breaking: A personal construct theory exploration of police stress". In: *International Journal of Personal Construct Psychology* 6 (1993), pp. 253–267.

[32] M. D. Young and M. G. Hermann. "Increased Complexity Has Its Benefits". In: *Political Psychology* 35 (2014), pp. 30–43.

Syntactic - Semantic Complexity	Example
Low - Low	Soviet agriculture is a disaster and for an obvious reason. Fifty years ago they collectivized all their farms and made farmers work not for themselves but for the government. Individual incentives were lost. Farmers had to work for the glory of the state. And ever since, the Soviets have not been able to produce enough food to feed their people. This dismal performance will continue as long as the leaders in the Kremlin remain committed to the silly notion that people will work as hard for others as for themselves.
Low - High	Some view abortion as a civil liberties issue; others see abortion as murder. How you view abortion depends on a complicated mixture of legal, moral, philosophical and perhaps scientific judgments. For example, is there a constitutional right to abortion? If there is, what criteria should be used to determine when human life begins? And, a question that must be answered before any of the others can be, who possesses the authority to resolve these issues?
High - Low	Renunciation of thinking is a declaration of spiritual bankruptcy. Where there is no longer a conviction that men can get to know the truth by their own thinking, skepticism begins. Those who work to make our age skeptical in this way, do so in the expectation that, as a result of denouncing all hope of self-discovered truth, men will end by accepting as truth what is forced upon them by authority and by propaganda.
High - High	Their experiences with war and depression during the thirties created in many members of our parents' generation a drive to create some form of security for the future that was not available for them to enjoy in earlier years. By continuously building upon their gradually increasing assets while still maintaining the conservative lifestyles they had been pressed to follow during hard times, they created economic stability for themselves. This economic stability, enjoyed by many approaching old age, lends greater power to seniors' increasingly vocal demands for an improved quality of life for the elderly. Their offspring, not having faced the same hardships as their parents, have had opportunity and cause to be somewhat reflective about issues pertaining to the quality of life in general, including the plight of the elderly.

Table 5: Examples of varying levels of syntactic and semantic complexity, taken from Baker-Brown et al. [1]. Assessments of complexity are based on gold standard scores and intend to illustrate the space of integrative complexity.