Fine-Grained Readability Estimation using Language Modeling

Venkat Arun 130101085
v.arun@iitg.ernet.in

Desh Raj 130101018
r.desh@iitg.ernet.in

Mrinal Tak 130101049
m.tak@iitg.ernet.in

Sumeet Ranka
130101062
sumeet.ranka@iitg.ernet.in

ABSTRACT

We conjecture that predictability of a text is a viable metric of its readability. By using modern language models as predictors, we believe this metric may provide an automated, fine-grained measure of readability. It also provides a natural mechanism to combine scores from different language models, and hence the ability to generalize to a diverse set of texts. Individual language models encode the specific linguistic background that a reader may have, hence providing customized scores for each type of reader. Our work provides authors with a valuable tool to (1) assess the readability of their content for readers with different linguistic backgrounds and, (2) identify pain points at a word-level granularity in their text in order to improve it. Our evaluations support our conjecture and show that the resulting scores work across a wide range of scenarios.

1. INTRODUCTION

Readability is a measure of the ease with which a reader understands a piece of written text [?]. It is different from legibility in that it is not dependent upon clear and distinguishable characters, typography, presentation, and so on. The definition of readability suggests that it is characterized by the textual content as well as the reader, which is primarily the reason why it has been considered challenging to defined a unified metric to score the readability of a text. The challenge lies in attempting to quanitfy an inherently subjective and opinionated value.

Text readability as a quantifiable metric has prospects in predicting magazine and newspaper circulation, journal impact, and author popularity. It can help authors fine-tune their writing, search engines target their results to the user's background and organizations to ensure a minimum standard of writing in their documents. Linguists and historians have also been able to use readability of a text to determine its authorship and period. L.A.Sherman [?] established in the late 19th century that literature was a subject for statistical analysis, and over time, the readability of a written text increases if it begins to appropriate speech.

It is useful to be able to automatically assess the readability of a piece of text. It can help authors fine-tune their writing, search engines target their results to the user's background and organizations ensure quality of writing. The problem is well studied, dating back at-least to Flesh's paper [Fle48] in 1948. There are two primary classes of approaches to solve this problem. Early methods use simple formulae of metrics such as words per sentence and syllables per word. The advent of computers enabled more complicated features such as word familiarity (as determined from a training corpus), number of verb phrases and, depth of a parse tree corpus to be used. These features are then combined using machine learning methods trained on some ground truth.

The technique we propose is predicated on the hypothesis that, *text predictability is a direct measure of readability*. This metric is designed to measure how fluently a mature reader may read the text. Intuitively, it makes sense that the easier it is for a human to predict a text, the faster they can read it. This relationship has been examined by linguists, who have found a strong correlation between text predictability (as scored by human experts) and the speed of eye movement[KGRE04]. We posit that modern language models are mature enough to match humans at assesing predictability, and may serve as part of an automated mechanism to predict how fast a human would read the text.

Our approach is simple. We train a generative model to predict the next word or character based on previous text. This model serves as a predictor in a predictive compression algorithm. The average number of bits required to encode a sentence in the text is a measure of its predictability and functions as a measure of the text's readability. Typically we'd use a model that provides low perplexity as they can model many complex features of language. Today such models are based on deep neural networks. This may change as the field evolves. The key difference from prior methods is the complete abstinence from using text annotated with readability scores. Our approach offers several advantages:

- Readability is a function of the textual content as well as the reader's linguistic background. For instance, readability requirements for a medical text is different depending on whether it is targeted at an average reader or a medical professional. In our method, the corpus on which the model is trained serves as a proxy for the reader's linguistic background. Thus we can report different scores based on the reader's background.

- Data driven methods require text annotated with readability scores. Such data are expensive to create and often not freely available. Since our method does not require annotation, it can be used for multiple languages and for generating fine-grained scores targeted at readers with specific backgrounds. In each case, we only require a large un-annotated corpus of text which are easily available in the public domain.

- The number of bits required to encode each word can also be provided to the user, hence pointing out the pain-points in a text which the author can use as hints to improve the writing.

- The method provides a unified framework such that models which differ in their algorithms and/or corpora may score on a common scale.

- Languages are complex distributions and it may be difficult to account for all of their features. We tackle this challenge by "outsourcing" this task to extensive research that has already been done on language modeling. We are able to do this because we do not need annotation and can use standard large corpora for training.

2. RELATED WORK

People have worked on scientifically assessing the readability of text for about a century. A good review of both early and modern methods can be found in [CT14]. Early methods used simple metrics such as sentence length, number of syllables per word, number of connectives etc. [Fle48, ML69, Fry68, DC48] to estimate complexity of vocabulary and sentence structure. These metrics often correspond to popular guidelines for effective writing. The focus was on keeping the estimation simple enough to be quickly performed by hand. The advent of computers and large data-sets allowed more complete estimation. Word frequency indicates familiarity of vocabulary used and sentence complexity and cohesion can be measured more directly. Some metrics such as word frequency can also be adjusted to account for the reader's linguistic background.

Modern methods use machine learning methods to map a set of features to ground truth using training data. The ground truth may be in the form of grade-level assignments or pairs of texts where one has been determined to be more readable than the other. [Cal04] uses unigram models determined from texts written for various grade levels (1 through 12), to determine the most likely grade level of an unseen text. [PN08a] study a number of other features such as the number of word phrases per sentence and word overlap across sentences. Then they use standard linear regression and SVMs to predict readability. Others use more features such as syntactic structure and cross-sentence coherence [PN08b, KLP$^+$10] in addition to vocabulary. Aluisio et al. [ASGS10] target a slightly orthogonal problem of using readability metrics for simplifying complex text for semi-literate readers. Others use language models to obtain linguistic features that they use in their supervised learning methods [CTC04].

While the use of language models alleviates the issue of parsing large amounts of text to a certain extent, the approaches are still limited because of their supervised nature. These methods essentially reduce the task of quanitfying readability to a classficiation problem, wherein a model is trained on documents that have been annotated with grade labels, and prediction involves classifying unseen texts to the most appropriate grade. This method entails three disadvantages: (1) it requires large amounts of annotated text to train the language model, (2) the labels used for classifying the training data may not be appropriate for the test data, and (3) the method is not reader-specific since the annotations are provided by a human expert.

There are two primary challenges to measuring readability. The first challenge arises because languages are complex distributions and it may be difficult to account for all of their features. Initial approaches used simple features that were deemed relevant to readability. However, with increase in computing resources, the number and complexity of these features could be increased [KLP$^+$10]. Others use simple unigram-based language models [CTC04, PN08b] for this purpose. We tackle this challenge by "outsourcing" this task to extensive research that has already been done on language modeling.

The second challenge stems from having to precisely define what is meant by readability. Early methods use heuristics to come up with scores. Later, people began using 'supervised learning' to scale readability against labeled training data [CTC04]. Alternatively, [PN08b] defines readability as the product of likelihood predictions of the text determined by several different approaches. While this method does not require the use of training data to calibrate the scale, it is not clear why the likelihoods should be combined by taking a product. Our novelty lies in having several language models to compress the text. The best compression ratio may provide a canonical measure of its predictability, and hence its readability.

In this work, we propose a novel language model based metric that is both unsupervised and generic. Our technical objectives may be summarized thus.

1. To exploit the sophisticated language models available at present to provide a fine-grained assessment of readability.

2. To provide a unified framework such that models which differ in their algorithms and/or corpora may score on a common scale.

3. To effectively utilize large corpora for training by eliminating the need for annotated data.

3. PROPOSED METHOD

Our system consists of a "quorum" of models, each trained on a separate corpus representing the target reader's linguistic background. We use each model's predictions to "compress" the text as much as possible. The extent of compression serves as the readability score given by the model. In addition we also compute the minimum of all the scores (entropies). Given enough number of relevant models, it represents how readable the text will be to someone familiar with the area being discussed.

For the model representing a general reader who is fluent in English but does not have any domain specific expertise, we use Google's network [JVS$^+$16] trained on a news data-set. We also train specialized models trained on corpora having roughly 10,000 Wikipedia articles each on topics in Biology, CS, Phsycology and Geography. For this we used a simple Character-CNN model [Kar15] because it has fewer parameters and can train well on the relatively smaller corpus. In both cases, we use character-level models because, to score on the same scale, each model must be given the same task: compress the given text to the maximum extent possible.

The corpus on which a model is trained serves to represent the reader's linguistic background. This is important because readability of a text is a function of both the text *and* the reader. For instance, even a very lucidly written text for medical professionals may be incomprehensible to a computer scientist. To account for this, we output two scores as follows.

1. The entropy produced by the best performing model on that text. This represents readability when the text is being read by a specialist (assuming one of the corpora includes this speciality).

2. The entropy produced by a model trained on a generic corpus such as news articles or Wikipedia. This indicates how readable the text is for a human reader who is proficient in the language, but does not possess any domain-specific knowledge.

Mathematically, suppose we have a language model M which has been pretrained on a corpus C. For an unseen text $W =< w_1, \ldots, w_N >$ consisting of N tokens, suppose our language model assigns probabilities $p_1, \ldots, p_N$ to each token, then the likelihood of the sample under the given model is

$$L = \prod_{i=1}^{N} p_i. \tag{1}$$

The log-likelihood is then defined as

$$\log L = \sum_{i=1}^{N} \log p_i, \tag{2}$$

and the normalized log-likelihood of the sample is given as

$$R = -\frac{1}{N} \log L \tag{3}$$

We propose the quantity R as a measure of the perplexity per word of the sample, and conjecture that this quantity adequately denotes the readability of the text with respect to the language model.

We mentioned earlier that our model "outsources" the language modeling task to existing sophisticated methods that may be genre-specific. For this purpose, we train several models $L_1, \ldots, L_m$ on the same corpus C, and define our metric as the maximum of all the readability values obtained using each model, i.e.,

$$R = \max\{R_1, \ldots, R_m\}. \tag{4}$$

We argue that this "group" of models may be considered equivalent to a panel of human readers. Since the models are trained on the same corpus, we conjecture that the readers have similar qualifications. A limitation of a single model-based approach is that the readability value is overly dependent on the performance of the model. Such a group-based approach nullifies this dependence considerably and ensures that the "reader" aspect of readability is only represented in the corpus on which the models have been trained.

This method may further be extended to domain-specific readability, wherein a language model trained on a non-generic corpus such as biomedical journal articles is used to appropriate a subject expert. This approach provides a convenient way to define readability for subject readers, which may be important for niche publications.

4. EXPERIMENTS

4.1 Baselines

We compare our approach to six formula-based metrics and one ML based scheme. The variables used in the formulae below are as follows. L - number of letters, W - number of words, S - number of sentences, s - number of syllables, L_{100} - average number of letters per 100 words, S_{100} - average number of sentences per 100 words, W_c - number of 'complex words' defined as the words outside a fixed word-list and, P - number of polysyllables.

1. Automated readability index [SS67] (ARI) outputs a score in the range 0 to 14 using the formula

   $$\text{ARI} = 4.71(\frac{L}{W}) + 0.5(\frac{W}{S}) - 21.43$$

2. Coleman-Liau index (CLI) is calculated as

   $$\text{CLI} = 0.0588 L_{100} - 0.296 S_{100} - 15.8$$

3. Flesch-Kincaid grade level [KFJRC75] is calculated as

   $$\text{FK} = 0.39(\frac{W}{S}) + 11.8(\frac{s}{W}) - 15.59$$

   and computes a value between 0-14 which presents the score as a U.S. grade level, similar to the previous metrics.

4. In the Flesch reading-ease test [Fle48], higher scores indicate material that is easier to read; lower numbers mark passages that are more difficult to read. The formula for the Flesch reading-ease score (FRES) test is

   $$\text{FRES} = 206.835 - 1.015(\frac{W}{S}) - 84.6(\frac{s}{W})$$

5. The Gunning fog index [DuB15] estimates readability between 6 and 17 using the formula

   $$\text{GFI} = 0.4[(\frac{W}{S}) + 100(\frac{W_c}{W})]$$

6. SMOG (Simple Measure of Gobbledygook) [**?**] was developed as a more accurate and more easily calculated substitute for the GFI and is calculated as

   $$\text{SMOG} = 1.0430\sqrt{P \times \frac{30}{S}} + 3.1291$$

We use the data-driven model from [Cal04]. Here, 12 unigram models are derived from texts from each grade level. The perplexity of the given passage is calculated from this model and the grade level is predicted from whichever unigram model produces the least perplexity. We use unigram models provided by the author as well as those produced from our own dataset in our evaluations.

4.2 Grade Level Prediction

Grade levels are often readily available in the form of exceprts from text-books. They are also well defined. We evaluate on two data-sets. Hence the task of predicting the grade level of a given text has been well studied. They only provide resolution from grades 1-12 however. We use two grade-level datasets.

WSDM Dataset.

The first data-set is the one presented in [Cal04]. It consists of short excerpts (1-2 sentences) from school textbooks from grade levels 1-12 along with data from human annotators who compare readability of pairs of passages. They don't provide the books themselves due to copyright issues. Instead, the authors also provide unigram counts of words present in the book.

NCERT Dataset.

We create a new dataset from NCERT books. NCERT books are text-books published by the government of India for school childeren. Since these are intended to be in the public domain, copyright is not an issue. Since, they are textbooks, grade levels are well defined. We choose mathematics as our subject since the textbooks are available for grades 6-12. By parsing the .epub files from these books, we extract 50-100 1000 word passages from the books and also prepare unigram word counts from it. Hence we evaluate performance both on very short excerpts and on larger passages. The difference in performance between these two types of passages is striking and interesting.

4.3 Simple vs. Standard Wikipedia

We create a new dataset that consists of articles from simple Wikipedia and the corresponding article from the "normal" Wikipedia. Simple Wikipedia (http://simple.wikipedia.com) has articles which have been written specifically with the stated aim of keeping the language simple for "people with different needs, such as students, children, adults with learning difficulties, and people who are trying to learn English". They use 'basic' and 'special' english sentences with a special emphasis of keeping the sentences short. Since these provide pairs of texts, each talking about the same topic with a clear demarcation of which article is simpler to read, this offers us a good opportunity to rigorously evaluate the various readability assessment techniques.

Since simple Wikipedia articles are typically shorter and less informative we choose 95 articles which have been identified by the community as “very good” or “good”. One of the criteria followed by the community to bestow such a distinction to an article is that the article should contain most of the information present in the original article and should be simple to read. Indeed we find that these articles have a similar information content as their standard counterparts and are prime subjects for comparison.

4.4 Historic Texts

Readability is a function of both the content of the text and, the reader’s expected linguistic background. Hence one would expect that for a modern reader, the older a text, the lower its readability. This is because language evolves with time and both vocabulary and grammatical constructions change with time. A modern reader, used to modern vocabulary and grammar, would find it older text difficult to read, even if the meaning is discernible to them. Of course, given enough time, language itself changes and readers will not be able to comprehend the sentences.

Our approach should capture this dependence on the reader’s linguistic background in the form of the training corpus. It is expected to reflect this drop in readability, because it would be expecting modern grammatical constructions and vocabulary and instead would find unexpected text. To test this, we use a dataset consisting of plays dating back upto 200 years. We then plot the readability scores obtained by each method against the date at which each text was written. We would also expect other methods to be largely insensitive to the reader’s linguistic background and remain largely invariant to it. Our expectations are not met however. No readability metric shows any discernible trends.

4.5 Age and Sex of Blog Writers

We take a dataset consisting of blogs and plot the readability scores obtained by the various methods against the age and sex of the writers. We do this purely out of academic curiosity and are surprised to find some trends. We find that the maximum readability for authors is between the ages of 30-40 years. Further, across all readability metrics, female bloggers rate a higher readability scores.

5. RESULTS AND DISCUSSION

5.1 WSDM Dataset

This data-set consists of short passages (1-2 sentences) from textbooks of various grades from 1-12. Here both our approach and some of the formula based approaches perform well. The unigram-based model trained on similar texts also performed well, but when trained on our textbook data it did not perform well. The unigram model, therefore, is expectedly dependent on the training corpus, since it directly leverages the information from counts of various words. The formula-based approaches performed well in those pairs of passages where simple metrics such as sentence or word lengths were sufficiently able to discriminate readability. However, our method performed well on all such pairs as well as the more difficult pairs where such a distinction could not be made by simple metrics alone. Further, it was able to achieve this performance even when it had been trained on a dissimilar corpus, which argues for the strength of our approach.

5.2 NCERT Textbook Dataset

This dataset consists of several passages extracted from NCERT Mathematics textbooks from grade levels 6-12. Each passage has about 1000 words. Here none of the formula-based methods did any better than random as the readability scores were more a function of text content than simple linguistic features. Of the two unigram-models, the model extracted from the same set of textbooks performed very well. However models extracted from the textbooks presented in the original paper did not perform so well. This is because the unigram-model based approach is very sensitive to the topics present in the training corpus and is unable to generalize well to other settings. We evaluated two of our expert models trained on a Biology corpus and a Computer Science corpus, respectively, on this dataset, and found that the Computer Science expert model performed significantly better. This can be attributed to the fact that a number of contexts used in mathematics are also frequently encountered in Computer Science.

5.3 Simple vs. Normal Wikipedia

Simple Wikipedia focuses on having short sentences, which is a feature all methods other than the unigram-based method use. The formula based methods directly incorporate this feature, while our approach computes bits/sentence, a quantity which increases with increasing sentence length. The performance of various models directly corresponds to whether or not they use sentence length as a feature. This explains the high accuracy obtained using the formula-based approaches, since all of them use sentence length as a major metric. Our model achieves even better performance because it uses this value in conjugation with language modeling based word prediction scores.

5.4 Effect of author’s gender on blog readability

Fig. 2 shows the average readability scores for male and female authors in the Blog Authorship Corpus. While the scores calculated using each index is relatively close for each gender, there is a general trend of higher readability for female bloggers across all scoring schemes. We do not explore the reason behind this observation.

5.5 Short Term Language Change

Fig. 3 shows the average readability scores obtained using the statistical formula-based approaches for novels published every year between 1881 and 1922 in the Oxford “Corpus of English Novels.” There appears no common trend in the curve, and we may conclude that short-term language change has little or no effect on the overall readability of text. However, from the range of values that the scores lie in, it may be argued that the average English novel written during this period would be readable for a modern U.S. grade level 6-8. It still remains to be seen whether readability decreases for even older text.

5.6 Effect of author’s age on blog readability

When we plot the average readability of blogs against the author’s age in Fig. 4, we find an interesting trend. We find that the readability increases until the age of roughly 25, after which it stabilizes. After the age of 45, the readability declines. This goes to support a widely held belief that an author’s most productive years are between their 30s.

6. CONCLUSION AND FUTURE WORK

We have presented the hypothesis that text predictability, as assessed by modern language models, can function as a direct measure of readability. We have shown that, in addition to several qual-

	WSDM	Textbook	Wikipedia
ARI	79%	52%	97%
FRE	75%	49%	94%
FKGL	75%	50%	96%
GFI	71%	53%	97%
SI	63%	54%	98%
CLI	79%	54%	94%
UNI-WSDM	91%	47%	60%
UNI-Textbook	54%	95%	55%
This Paper	81%	76%	95%

Figure 1: Accuracy of prediction of pairwise readability for various schemes.

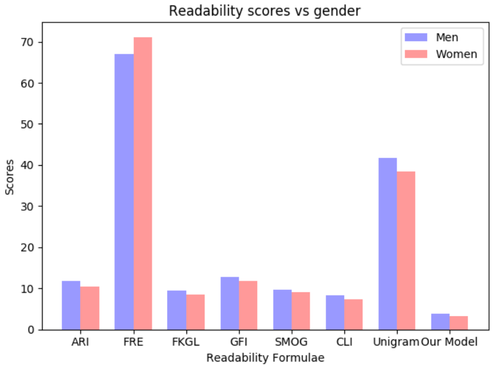


Figure 2: Effect of gender on blog readability. Comparison of average scores for blogs authored by male and female bloggers, obtained using formula-based approaches in the Blog Authorship Corpus. Except for Flesch reading score, in all metrics higher score implies lower readability.

itative advantages, it also offers better performance over a variety of scenarios. Further we have explored the relation of readability with age and sex of writers and shown interesting relationships. We have also shown that readability, as assessed by a modern reader, should decrease with increasing age of a text as expected.

For future work, we should evaluate against more ML based models. We could also include more expert language models to cater to a wider clientele. It would also be interesting to see if the readability of an article/paper/blog is correlated with its popularity.

A natural extension of this approach would be to provide authors with a tool that can point out the pain points in their text (ie. words with unusually high perplexity). It might then point out what word or sentence structure a reader might expect to follow.

In a more scientific vein, it would be interesting to replicate the results of [KGRE04] in studying the correlation between predictability and how fast humans read text (based on eye trackers). While they used human experts to compute predictability, we could try modern language models in a bid to evaluate how close they are to human level modeling of language.

7. CONTRIBUTIONS

The contributions of each of the four members are as follows:

Venkat Arun.
130101085

Conceptualized the problem and solution method. Created the NCERT and Wikipedia (simple vs. normal) datasets (these are new datasets). Preprocessed the WSDM dataset. Wrote the final report.

Desh Raj.
130101018

Trained and evaluated the CharCNN models on the various datasets. Prepared blog authorship dataset and CED, CEN and modern prose datasets. Wrote mid-term report.

Mrinal Tak.
130101049

Prepared the Wikipedia expert datasets (Computer Science, Biology, Phsycology, Geography). Implemented the formula based approaches and evaluated them on data.

Sumeet Ranka.
130101062

Prepared historic datasets, viz. CED, CEN and modern prose. Implemented the formula based approaches and evaluated them on data. Implemented the unigram model based approach.

8. REFERENCES

[ASGS10] Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9. Association for Computational Linguistics, 2010.

[Cal04] Kevyn Collins-Thompson Jamie Callan. A language modeling approach to predicting reading difficulty. 2004.

[CT14] Kevyn Collins-Thompson. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135, 2014.

[CTC04] Kevyn Collins-Thompson and James P Callan. A language modeling approach to predicting reading difficulty. In *HLT-NAACL*, pages 193–200, 2004.

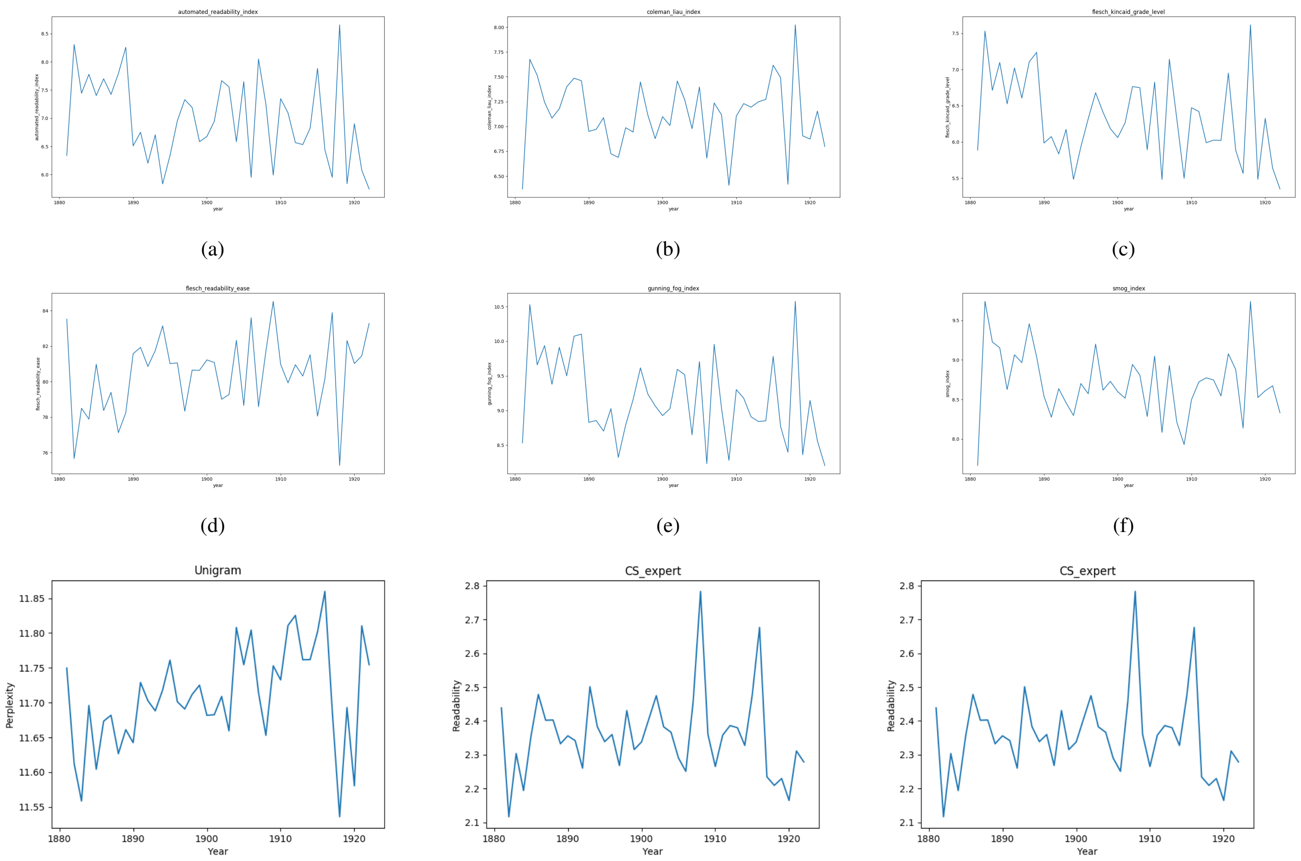
[DC48] Edgar Dale and Jeanne S Chall. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54, 1948.

[DuB15] William H DuBay. Judges scold lawyers for bad writing. *Plain Language At Work Newsletter (Impact Information 8). Accessed on March*, 13, 2015.

[Fle48] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[Fry68] Edward Fry. A readability formula that saves time. *Journal of reading*, 11(7):513–578, 1968.

[JVS$^+$16] Rafal Jozefowicz, Oriol Vinyals, Mike Schuster,



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 3: Short-term language change. Average readability scores of novels published every year between 1881 and 1922 calculated using: (a) automated readability index, (b) Coleman Liau index, (c) Flesch-Kincaid grade level, (d) Flesch reading ease, (e) Gunning fog index, (f) SMOG index, (g) unigram based model, (h) our score trained on Biology expert data and, (i) our score trained on CS expert data. Except for Flesch reading score, in all metrics higher score implies lower readability.

Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*, 2016.

[Kar15] Andrej Karpathy. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*, 2015.

[KFJRC75] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[KGRE04] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16(1-2):262–284, 2004.

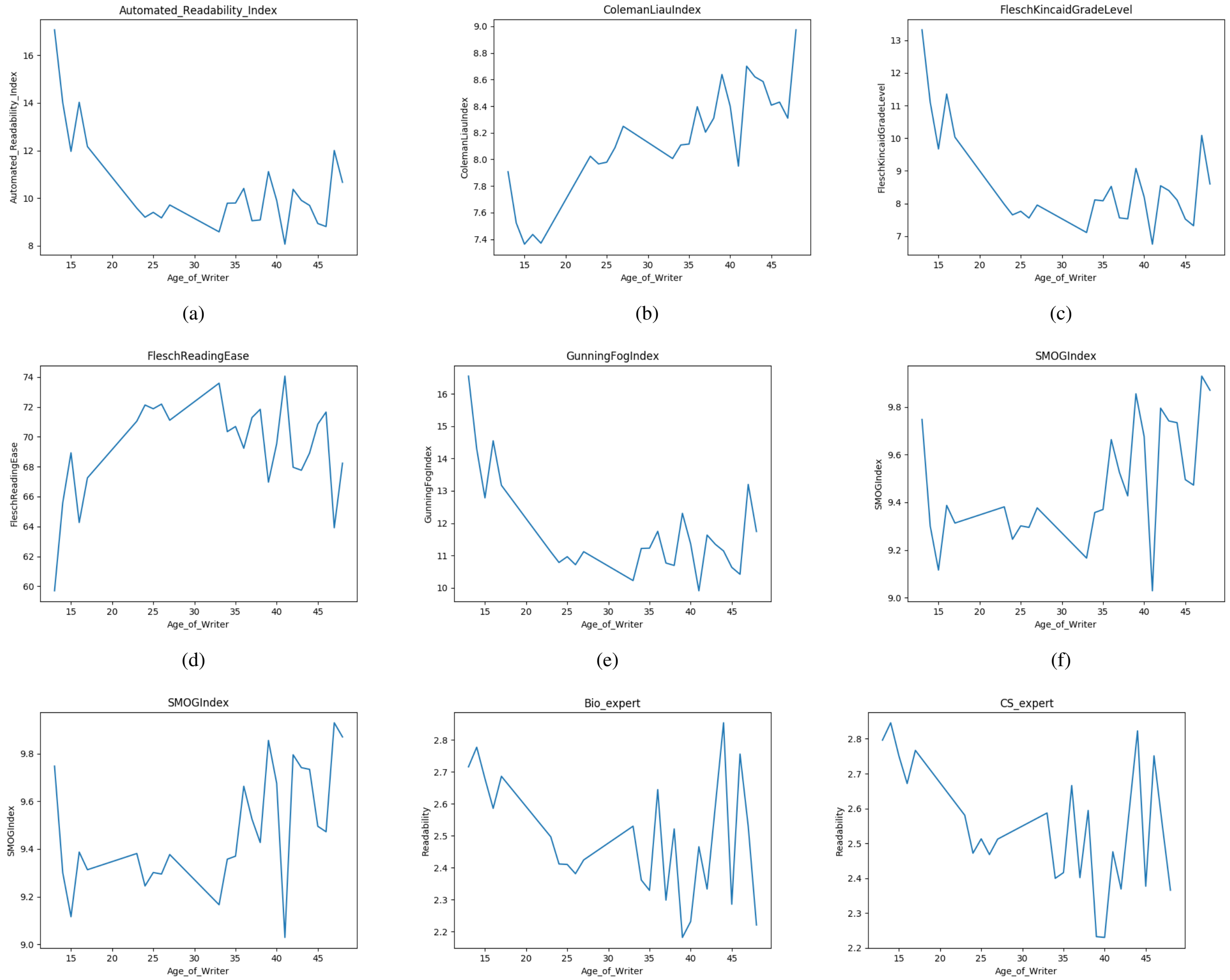
[KLP+10] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 546–554. Association for Computational Linguistics, 2010.

[ML69] G Harry Mc Laughlin. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[PN08a] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.

[PN08b] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*, pages 186–195. Association for Computational Linguistics, 2008.

[SS67] RJ Senter and Edgar A Smith. Automated readability index. Technical report, DTIC Document, 1967.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

Figure 4: Effect of age on blog readability. Average scores obtained using formula-based approaches plotted against age of author in the Blog Authorship Corpus: (a) automated readability index, (b) Coleman Liau index, (c) Flesch-Kincaid grade level, (d) Flesch reading ease, (e) Gunning fog index,(f) SMOG index, (g) our score when trained on Biology expert data and, (h) our score when trained on CS expert data. Except for Flesch reading score, in all metrics higher score implies lower readability.