A New Readability Measure for Web Documents and its Evaluation on an Effective Web Search Engine

Yume Sasaki
Department of Computer Science
Tokyo Institute of Technology
sasaki@lsc.cs.titech.ac.jp

Takuya Komatsuda
Hitachi, Ltd.
takuya.komatsuda.ur@hitachi.com

Atsushi Keyaki
Department of Computer Science
Tokyo Institute of Technology
keyaki@lsc.cs.titech.ac.jp

Jun Miyazaki
Department of Computer Science
Tokyo Institute of Technology
miyazaki@cs.titech.ac.jp

ABSTRACT

In this study, we propose a readability measure for Web documents and an information retrieval system that considers readability. Previous information retrieval systems aim to identify documents that are relevant to a given query; however, as information requirements of search system users becomes increasingly diverse and complicated, systems that take such new criteria into account are constantly being introduced. In particular, the focus of our present paper is on readability. Given that the population of non-native English speakers exceeds that of native English speakers, incorporating readability into an information retrieval system is crucial. Therefore, we propose (1) a readability measure that considers document simplicity and document structure as new features for readability and (2) a score fusion method that combines relevance and readability scores. In our experimental results, we found that our proposed readability measure outperformed an existing readability measure. Moreover, we found score fusion methods using a statistical framework called a copula improved overall accuracy as compared to such existing methods as linear combination.

CCS Concepts

•Information systems → Content analysis and feature selection; Combination, fusion and federated search; •Computing methodologies → Ranking;

Keywords

information retrieval; readability; text simplification; score fusion; copula

iiWAS '16, November 28-30, 2016, Singapore, Singapore
© 2016 ACM. ISBN 978-1-4503-4807-2/16/11...$15.00
DOI: http://dx.doi.org/10.1145/3011141.3011172

1. INTRODUCTION

As information requirements of users become increasingly diverse and complicated, information retrieval systems that consider new criteria, including complexity [7], freshness [11], and readability [21, 16], are being introduced. In this study, we focus on developing an information retrieval system that takes readability into account. From a user's perspective, readability is crucial as well as relevance to queries, especially given that the population of non-native English speakers exceeds that of native speakers[1]. Moreover, reading a document that contains difficult words and complex grammar is laborious even for native English speakers, thus considering readability helps both native and non-native English speakers. As such, it is important to return relevant and readable documents to users.

Text readability can be formally defined as the sum of all elements in textual material that affect a reader's understanding, reading speed, and level of interest in the given material [21]. Readability can be determined via many features, including semantic and syntactic difficulty levels. In general, various readability measures have been proposed [6]. Although a Web-specific readability measure has also been proposed in [20], collecting such a corpus to be used in such a study is costly and impractical.

To address the above problem, we propose a new readability measure for Web documents that considers, in addition to existing features, document simplicity and complexity of document structure. More specifically, unigram and bigram part-of-speech (POS) tags and the depth of heading tags are used as specific features, each of these features approximating the simplicity of a document and the complexity of document structure.

There are some possible an information retrieval system that considers readability via several approaches. As examples, we might (1) display retrieval results with readability scores to assist in user browsing efforts, (2) filter out documents with readability score below a certain threshold, or (3) integrate readability scores into the overall scoring method. Given (1), the burden on the user may actually increase; and given (2), defining an acceptable threshold to classify each document as either relevant or irrelevant may prove to be

[1] `https://hbr.org/2012/05/global-business-speaks-english`

difficult. Therefore, in this study, we adopt (3). More specifically, our proposed system combines a readability score and a relevant score via a statistical framework called a copula. Although linear combination is a well-known score fusion method, it cannot capture complex dependencies that are not linear. In contrast, copulas can capture such dependencies by considering dependences of scores. In fact, some studies have reported that score fusion methods using copulas yielded higher levels of accuracy than methods using linear combination when several scores were combined [14, 22].

In our own experiments, we evaluated our proposed readability measure and corresponding information retrieval system. Our proposed readability measure outperformed an existing readability measure in terms of precision, recall, and F-measure. Further, our proposed information retrieval system outperformed a system using linear combination, thereby showing the effectiveness of our fusion method using a copula.

In addition to this introductory section, the remainder of this paper is structured as follows. In Section 2, we describe related work involving readability measures, scoring functions, and copulas. Next, in Section 3, we define our proposed readability measure and information retrieval system. In Section 4, we discuss our evaluation experiments and results. Finally, in Section 5, we conclude our paper and describe avenues of future work.

2. RELATED WORK

2.1 Readability Measures

The readability of documents has actually been studied since the 1920s [16]. It has since been well-established that the key features affecting readability are legibility, vocabulary, semantics, syntax, discourse, the use of idioms, pragmatic semantics, user interest, and background [6]. These features can be classified into three groups or levels, i.e., the vocabulary level, the sentence level, and the document level.

In [23], Lorge proposed a readability measure that only uses three variables, i.e., average sentence length x_1, the number of prepositional phrases per 100 words x_2, and the number of words not included in the '*769 easy words list*' by Dale[2] x_3, as encompassed by Equation (1) below.

$$\text{Readability}_{\text{Lorge}} = 0.06x_1 + 0.10x_2 + 0.10x_3 + 1.99 \quad (1)$$

In [17], Flesch proposed a readability measure that uses average word length per sentence x_1 and average number of syllables per word x_2, as described by Equation (2). Unlike Lorge's readability measure, this readability measure does not introduce any external knowledge, such as an easy word list.

$$\text{Readability}_{\text{flesch}} = -1.015x_1 - 84.6x_2 + 206.835 \quad (2)$$

Defined in [5], the New Dale-Chall (NDC) approach uses average sentence length and a percentage of difficulty words. To distinguish between easy and difficult words, an easy word list[3] containing 3,000 common words is used. To validate the NDC approach, the correlation between the number of words not included in the word list and the average school grades of students who correctly answered a reading comprehension test more than 50% of the time (i.e., C_{50}) was calculated; this correlation was larger than the correlation between the *affixed morphemes* feature proposed by Flesch and C_{50}, thus showing that a word list is a good readability variable. NDC is defined as

$$\text{NDC} = \begin{cases} 0.1579 \times \text{PDW} + 0.0496 \times \text{ASL} + 3.6365 \\ \hfill (\text{if PDW} > 5\%) \\ 0.1579 \times \text{PDW} + 0.0496 \times \text{ASL} \\ \hfill (\text{otherwise}) \end{cases}, \quad (3)$$

which includes the percentage of difficult words (PDW) and the average sentence length (ASL).

Readability measures that use linear sums of shallow features of the given language are easy to calculate, thus these measures have largely been used in the educational domain; however, Bormuth has reported that the relationship between a feature and readability is not necessarily linear [1]. More specifically, Bormuth proposed a non-linear readability measure in which features are the average of characters per word, the frequency of words included in the word list using NDC, the average word length, and the frequency of modal verbs. The correlation between this readability measure and a reading comprehension test showed to be over 0.9. Note that the features introduced into the readability measure are essentially the length of a sentence and the number of easy words, which matches the features of NDC. As such, Bormuth concluded that shallow language features and the familiarity of vocabulary are the only factors for determining readability.

Recent studies using natural language processing (NLP) techniques, such as syntactic analysis and statistical language modeling, show these techniques are able to capture more complex language features. In [8], Collins-Thompson et al. constructed statistical language models for each grade based on each grade's document collection. Unlike the aforementioned studies, this model accurately measured readability for Web documents; however, this approach has an inherent shortcoming in that Web documents classified into grades are required, which is expensive. In their experiment, models were evaluated via a correlation between readability measure scores and document difficulty levels, which were defined in advance. A model using the word list of NDC showed a stronger correlation than their statistical language model, though this depended on the document set.

Finally, we refer to text simplification hereinafter, which has strong relevance to document readability. Text simplification is a task that converts given text into readable text to enhance document readability. In [25], Napoles et al. verified that Simple Wikipedia can be used as a corpus of text simplification. Their experimental evaluations showed that their classifier, which used a bag-of-words approach and ratios of part-of-speech (POS) tags, accurately classified documents in terms of whether a document was from the ordinary English Wikipedia or from Simple Wikipedia. Note that the POS tags are classified into six types, i.e., *noun*, *verb*, *adjective*, *adverb*, *determiner*, and *relative*. Next, they measured unigram (i.e., same as original frequency) and bigram frequencies of the POS tags. As a concrete example, suppose we are given the sentence, "**Readability**/*noun* **measure**/*noun* **is**/*verb* **important**/*adjective*." Then, bigram tags are *noun:noun* (POS tags of **"Readability"** and **"measure"**), *noun:verb* (POS tags of **"measure"** and **"is"**), and *verb:adjective* (POS tags of **"is"** and **"important"**).

[2] `http://www.betterendings.org/homeschool/words/769words1.htm`

[3] `http://www.rtp-templates.com/Research-Articles/Dale-Chall-3000-Simple-Word-List`

2.2 Scoring Functions

Many information retrieval systems use scoring functions that calculate relevance scores for a given document. Described in [35], BM25 is one of the classical probability scoring functions and has demonstrated a high level of effectiveness. In [28], Ponte and Croft proposed a probabilistic language model, developed in a mathematical framework. Further, vector space models [30] and classical probability models have also been proposed as heuristic approaches.

Although many scoring functions do exist, it is difficult to identify an appropriate scoring function that always performs the best, because information requirements of each user are diverse and complex. To address this challenge, various studies have focused on the fusion of multiple relevance scores calculated via several scoring functions. As an example, some meta search engines have attempted to improve accuracy by combining results from multiple engines. These studies are known to calculate score fusions of relevance scores.

In general, score fusion is often achieved by obtaining the sums or products of results from individual systems (scoring functions) [18]; probabilistic approaches also exist [10].

In [35], Vogt et al. incorporated linear combination into information retrieval, and the suitability of this approach has been demonstrated. In [19], Gerani et al. applied a nonlinear transformation to relevance scores before applying the linear combination approach, showing that their method outperformed the standard linear combination approach. Their results demonstrate the requirement for a model that can capture complex and nonlinear dependencies.

In [4], Burges et al. present a ranking model based on gradient descent that uses machine learning to enable easy unification to obtain one composite score from a large number of document features. This approach extracts features of relevant documents from a set of documents labeled as either relevant or irrelevant. The disadvantage of this approach is the difficulty in understanding the acquired model.

In general, copulas are widely used in quantitative finance and portfolio management [2, 15]; further, recent studies have applied copulas to other research fields [31, 29, 27]. As an example, in [36], Vrac et al. applied a mixture copula to a global climate dataset, showing that a mixture copula can group locations of the world based on climate.

As another example, in [14], Eickhoff et al. applied a unimodal copula to score fusion, and, in some cases, their proposed method outperformed the baselines as a result of combining two relevant features. They verified the effectiveness of their approach by increasing the number of relevant features increased from two to 136, thereby showing that their proposed method is more effective than linear combination as the number of relevant criteria increases [13].

Finally, in [22], Komatsuda et al. proposed a score fusion method using a mixture copula that can model multiple copulas when relevant documents are derived from multiple distributions. They combined multiple query relevance measures, and their proposed method outperformed existing methods that use linear combination and a unimodal copula [14].

2.3 Copulas

In this subsection, we provide an overview of copulas. For score fusion with copulas, the following conditions must be defined: (1) the marginal distribution functions of each dimension; (2) the type of copula; (3) parameters of the copula if the type of copula indeed has parameters; and (4) the number of clusters and weights if a mixture copula is used as a scoring function.

2.3.1 *Definitions and Properties*

Copulas are models that describe the relationship between a multivariate distribution and marginal distributions. Let X be a k-dimensional random vector, i.e., $X = (x_1, x_2, ..., x_k)$. Further, let function $F_k(x)$ be a marginal cumulative distribution function for element x_k of random vector X, where $F_k(x) = P[X_k \geq x]$. Then, X can be mapped to k-dimensional unit cube $[0, 1]^k$ as $U = (u_1, u_2, ..., u_k) = (F_1(x_1), F_2(x_2), ..., F_k(x_k))$. Then, k-dimensional copula C is described as a joint cumulative distribution function of normalized random vector U. Most importantly, in [26], Nelsen proved that there exists a copula C that satisfies $F(x_1, x_2, ..., x_k) = C(F_1(x_1), F_2(x_2), ..., F_k(x_k))$ in any k-dimensional joint cumulative distribution function $F(x_1, x_2, ..., x_k)$, thus showing the high applicability of copulas. In addition, copulas facilitate our analysis of the structure of joint distribution, because we separately estimate each marginal distribution $F_k(.)$ and the dependency structure between the marginal distributions.

2.3.2 *Typical Families of Copulas*

There are various types of copulas, including elliptical copulas, Archimedean copulas, and empirical copulas, each of which is described below.

- **Elliptical Copulas:** An elliptical copula is a copula derived from a standard distribution, such as a Gaussian distribution or a t distribution. Equation (4) below shows the formula for a Gaussian copula.

$$C_{Gaussian}(U) = \Phi_\Sigma(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_k)) \quad (4)$$

  Φ_Σ denotes a cumulative distribution function of a standard normal distribution and Φ^{-1} denotes its inverse function. A Gaussian copula requires parameter $\Sigma \in R^{k \times k}$, which shows the observed covariance matrix.

- **Archimedean Copulas:** Let ϕ be a continuous, strictly decreasing function from $\mathbf{I}$ to $[0, \infty]$ such that $\phi(1) = 0$. Then, we have

$$C_\phi(U) = \phi^{-1}(\phi(u_1) + \phi(u_2) + ... + \phi(u_k)), U \in (0, 1]^k,$$

  which represents a k-dimensional Archimedean copula, where ϕ is a generator of C_ϕ. For $\phi(t) = \frac{t^{-\theta}-1}{\theta}, (-\log t)^\theta, -\log\frac{e^{\theta t}-1}{e^{\theta}-1}$, the copulas are called Clayton copulas, Gumbel copulas, and Frank copulas, respectively. Further, these copulas are defined below by Equations (5), (6), and (7), respectively.

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k} \frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \quad (5)$$

$$C_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k}(-\log(u_i))^{\theta})^{\frac{1}{\theta}}) \quad (6)$$

$$C_{Frank}(U) = \frac{1}{\theta}\log(1 + \frac{\prod_{i=1}^{k}(\exp(-\theta\ u_i) - 1)}{\exp((-\theta) - 1)^{k-1}}) \quad (7)$$

  Different copulas have different features. For example, we assume that for a Clayton copula, the dependency of the lower region is strong, whereas the dependency of the upper region is independent. The use of a Clayton copula is effective if the dependency of the relevance scores is strong in cases where the relevance scores are low.

- **Empirical Copulas:** An empirical copula refers to a copula that is derived from an empirical joint distribution whose marginal distributions are estimated by empirical distribution, thus an empirical copula is a nonparametric joint distribution based on observations without assuming any specific distribution. Therefore, k-dimensional empirical copula $\hat{C}(U)$ is described as

$$\hat{C}(U) = \frac{1}{N} \sum_{n=1}^{N} \prod_{i=1}^{k} \mathbf{1}\{t_i^n \leq u_i\}, \qquad (8)$$

where N denotes the number of observations required to estimate the empirical copula and t_i^n represents a score on the i-axis of the n^{th} observation. The probability of a k-dimensional joint cumulative distribution derived from an empirical copula is calculated by dividing the number of training data such that $(t_1^n \leq u_1, t_2^n \leq u_2, ..., t_k^n \leq u_k)$ by the total number of training data N.

Appropriate criteria are required to be selected from the various types of copulas. A model of copulas can be selected based on certain criteria, such as tail dependence coefficient and rank correlation coefficients [15]. The tail dependence coefficient is an indicator of the dependence structure at the endpoints of the probability, i.e., for probabilities close to zero and one. If the tail dependence coefficient is used for the selection of a model, it implies that we are focusing on the dependency on either high relevance or low relevance part of the distributions. The rank correlation coefficient is an indicator of the dependence structure in the entire distribution. If a model based on the rank correlation coefficient is selected, it implies that we are focusing on the average dependency in the overall distribution.

2.3.3 Mixture Copula

A mixture copula is a copula composed of several copulas [22]. By using a mixture copula, a multimodal joint distribution can be built that enables us to capture a complex dependency.

More specifically, a mixture copula is described as the weighted sum of k copulas, i.e,

$$C_{mix}(U) = \sum_{i=1}^{k} p_i C_i(U). \qquad (9)$$

To construct a mixture copula, each copula C_i and its weight p_i should be estimated. These parameters can be estimated by using approaches based on clustering [32, 12], thus a mixture copula can be constructed by first splitting the training data used to estimate the mixture copula into k clusters, then fitting the data in each cluster to a unimodal copula. Note that number of clusters k is determined in advance. One of the methods to decide the value of k involves the use of an information criterion, such as Akaike's Information Criterion (AIC) [3]. The product of a mixture copula and its likelihood are sometimes used as a scoring function, i.e.,

$$C_{mixprod}(U) = C_{mix}(U) \prod_{i=1}^{k} u_i. \qquad (10)$$

3. PROPOSED METHOD

In this section, we describe our two proposed methods. The first proposal is a readability measure that captures the complexity of vocabulary and the structure of sentences and documents. Document simplicity is also examined. In Section 3.1 below, we describe features used in our readability measure and our calculation method. Our second proposal is an information retrieval system that considers readability. Our proposed system combines both the readability and relevance of a document, then calculates its score. The methods for constructing and using this system are detailed in Section 3.2.

3.1 Readability Measure

3.1.1 Readability Features

Our proposed method examines the complexity of a document via features that affect readability, i.e., features at the vocabulary level, sentence level, and document level. Each of these is further described below.

Vocabulary level: For vocabulary features, we used the average number of syllables per sentence, the percentage of difficult words, and ratios of the POS tags. Note that these features has been used in previous studies [5, 25].

We calculated the percentage of difficult words was based on the word list from [5] that contains 3,000 easy words. A key characteristics of Web documents is that they often contain many proper nouns. Because proper nouns are not in the easy word list, all proper nouns are regarded as difficult words; however, as a result of our preliminary survey, we found that proper nouns generally do not decrease readability. Therefore, all proper nouns were regarded as easy words in our study. Note that the Stanford Log-linear Part-Of-Speech Tagger[4] [34] was used for POS tagging.

Finally, we counted the frequencies of unigram and bigram POS tags. These POS tags were classified into six types, i.e., *nouns*, *verbs*, *adjectives*, *adverbs*, *determiners*, and *relatives*, in the same manner as that of [25]. Note that the number of different types of bigram POS tags is 36 (= 6 × 6).

Sentence level: We used the average sentence length as a sentence level feature. Documents must be divided into sentences to obtain average sentence length. To achieve this, the simplest approach is to just use terminators, such as periods and question marks. Sequential words between terminators can then be regarded as a sentence; however, it is not always true that Web sentences end with a terminator. As an example, each item in a list tag may often end without a terminator even though it is a sentence, thus we introduced terminator tags. A terminator tag is defined as the tag for which a closing tag is regarded as the end of sentence. As examples, `H3` and `LI` tags serve as terminator tags. The HTML document in Figure 1 can be divided into five sentences with terminator tags.

In general, we used the following tags as terminator tags: `TD`, `TH`, `TR`, `UL`, `OL`, `LI`, `DT`, `DL`, `FORM`, `OPTION`, `SELECT`, `FIELDSET`, `TITLE`, `P`, `BR`, `H1`, `H2`, `H3`, `H4`, and `H5`.

Further, sentences with less than three words were discarded, because such Web sentences are often simply noise and not well-formed. Consequently, four sentences can be extracted from the HTML document in Figure 1, because

[4] http://nlp.stanford.edu/software/tagger.shtml

Table 1: Features used in our proposed measure and their classification

Classification	Features	Dimensions	NDC [5]
Vocabulary	Average number of syllables per sentence	1	-
	Percentage of difficult words	1	○
	Rates of unigram POS tags	6	-
	Rates of bigram POS tags	36	-
Sentence	Average sentence length	1	○
Document	Depth of heading tags	1	-
	Document length	1	-

```
<H3>Considering a Mac</H3>
<UL>
  <LI>Why you'll love a Mac</LI>
  <LI>Which Mac are you?</LI>
  <LI>FAQs</LI>
  <LI>Watch the ads</LI>
</UL>
```

Figure 1: An example HTML document in which closing tags serve as terminator tags

the fourth sentence *"FAQs"* is eliminated.

Document level: For document level features, we used the depth of heading tags and document length. In [24], Manabe and Tajima reported that logical structures of documents roughly match document structures expressed by heading tags. In our preliminary survey, we noted a tendency that well-structured documents had high levels of readability, thus we used the depth of heading tags. As examples, the depth of heading tags is two when the document includes only `H2` and `H3` tags, whereas the depth is six when heading tags appear in order from `H1` to `H6`.

Table 1 shows features used in our proposed method, as well as classification of features. In the table, features used from the NDC approach are marked accordingly.

3.1.2 Readability Estimation

The probability of readability of a document is calculated via logistic regression [9], which is a method that estimates an objective variable probability from explanatory variables. More specifically, logistic regression can evaluate contributions of each explanatory variable from an obtained model. Let p, $x = \{x_1, x_2, \ldots, x_n\}$, and $\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$ be an objective variable, explanatory variables, and regression coefficients respectively; then, logistic regression is defined as

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n. \qquad (11)$$

Explanatory variables used in our proposed measure are automatically selected from the best model by changing the explanatory variables. Specifically, AIC [3] is used to identify the best model.

More concretely, among the features listed in Table 1, we reveal a combination of the features that can most accurately classify a document between readable and unreadable. If each feature's coefficient of the obtained model is positive, a document is readable when the corresponding feature's value is high. Similarly, if a coefficient is negative, a document is readable when the corresponding feature's value is low.

3.2 Retrieval System

Our proposed retrieval system combines readability and relevance scores via a copula. A complex dependency between the scores are captured using a copula as a score fusion function.

We construct the system using the following four steps: (1) cluster the training data; (2) estimate the marginal distribution functions for each cluster; (3) estimate copulas for each cluster from the training data and estimated marginal distribution functions; and (4) combined the estimated copulas by assigning weights to clusters. Each of these steps is further detailed below.

1) Clustering the training data: We split the training data into clusters to calculate copulas at each cluster. The number of clusters can be identified via information criteria such as AIC [3].

2) Estimating the marginal distribution functions: The system uses dimensions of readability and relevance, thus we estimate the two marginal distribution functions at each cluster from the training data. In this study, we assume normal and empirical distributions; the parameters are then estimated for each distribution.

3) Estimating the copula: The copula is estimated from the training data and estimated marginal distribution functions for each cluster. In this study, we use a Gaussian copula, a Clayton copula, a Gumbel copula, a Frank copula, and an Empirical copula. As described in Section 2.3.2, these copulas are the copula families expressed by a parameter, thus the parameter is estimated from the training data and estimated marginal distribution functions.

4) Combining copulas: Estimated copulas are combined while considering the weight of each cluster. In our present study, the weight of a cluster is the ratio of the number of data included in the cluster to the number of training data.

After constructing a mixture copula via the above, score fusion is performed by assigning a document to the copula. There are two scoring functions, i.e., one uses the constructed copula as is (i.e., Equation (9)), while the other is the product of the copula and its likelihood (i.e., Equation (10)) [22].

4. EVALUATION

In this section, we describe our two experiments. First, we evaluate the performance of our proposed readability measure, then we evaluate the accuracy of our score fusion methods to combine relevance and readability scores.

4.1 Overall Settings

In this subsection, we describe the common settings of the two experiments. We used TREC category B of Clueweb09[5] as the dataset. This dataset contains approximately 50 million English Web documents. Next, five queries were chosen at random from 50 queries of the Web Track in TREC2011. Note that relevant judgment is provided by TREC Web track organizers.

To identify ground truth labels in terms of readability, documents were examined by three annotators, i.e., two graduate students and an office worker. All of them were non-native English speakers who have been learning English for over 10 years. Documents were labeled as readable or unreadable, where a readable label is assigned to a document judged to be readable based on such features as the difficulty of vocabulary and complexity of document structure, whereas an unreadable label is assigned to a document that is not judged as such. After all documents are individually judged by each annotator, all three annotators discussed the documents, in particular conflicts in assigning ground truth labels.

4.2 Evaluation of Readability Measure

4.2.1 Settings

For our evaluation, we prepared a dataset that contained 137 relevant and readable labeled documents for the five given queries. We measured precision, recall, and the F-measure on the classification task using fivefold-cross-validation. In our cross-validation, five sub-datasets were virtually created from the original dataset. Four sub-datasets were used as training data, while the other was used as test data. We generated a model with the training data, then the model was evaluated with the test data. The accuracy of the model was calculated by averaging the results of the five tests. Our proposed measure produces as an output a probability as to how readable the document is. The threshold of the classification between readable and unreadable was set to intermediate, i.e., 0.5. Note that models were reconstructed at each cross-validation.

From [5], we used NDC as a baseline, because NDC has shown comparable accuracy to the state-of-the-art [33]. In our experiment, the NDC's threshold of document classification was set to 10, which is the college graduate level. In other words, documents with NDC scores less than 10 were regarded as readable.

4.2.2 Results

Table 2 shows our accuracy results. Our proposed measure outperformed the baseline in terms of precision, recall, and the F-measure.

Table 2: Readability results

	Proposed	NDC (baseline)
Precision	**0.8889**	0.8312
Recall	**0.8865**	0.8668
F-measure	**0.8839**	0.8438

We created five models because a model is created for each cross-validation. Useful features in the accurate classification cases are listed in Table 3, i.e., we list the features that were selected from three or more models of the five models with signs (i.e., positive or negative) of the coefficients being consistent. Readability decreased when the percentage of difficult words increased. Documents with many *adjective:adverb* bigram tags were less readable. Similarly, documents with many *noun:relative* bigram tags were also less readable. These bigram tags may indicate complex sentence structure. Conversely, documents with many *noun:noun* bigram tags were more readable, suggesting that using bigrams of POS tags impacts readability more than using unigrams. Meanwhile, it seems that the depth of heading tags cannot properly represent the complexity of document structure, because the feature was not chosen.

Table 3: Variables used in our models

Variables	Effects on readability
Percentage of difficult words	Decrease
Rate of *adjective:adverb* bigram tag	Decrease
Rate of *noun:relative* bigram tag	Decrease
Rate of *noun:noun* bigram tag	Increase

4.3 Score Fusion Combining Relevance and Readability

4.3.1 Settings

In this experiment, we measured the accuracy of our proposed information retrieval system that incorporates readability. Correct search results must satisfy both conditions, i.e., relevance to a query and readable. As noted above the size of the dataset was 137. We constructed this experiment using fivefold-cross-validation. Copulas were computed on the training data, then a new instance from the test data was evaluated via the copulas. Further, we used BM25 to calculate relevant scores for documents.

Our evaluation measures were a normalized Discounted Cumulative Gain(nDCG@k) in top-k documents and interpolated Precision(iP@i), where i is the recall level. In this study, nDCG is calculated at $k = 5, 10, 15, 20$, while iP is calculated at $i = 0.0, 0.1, \ldots, 0.5$.

4.3.2 Score fusion methods

We used the following six score fusion methods: $C_{mix}(U)$; $C_{mixprod}(U)$, which is the product of likelihood and the mixture copula; $LIN(X)$, which is a linear combination of vector X; $SUM(X)$, which is the sum of vector X; $PROD(X)$, which is the product of vector X; and $HM(X)$, which is the harmonic mean of vector X

The x_i component of vector X denotes a normalized score, and component u_i of vector U denotes a score to which a cumulative distribution function $F_i(.)$ maps the x_i. The following equations are formulas of score fusion methods. Further, we optimized parameter λ_i of linear combination approach for training data in advance.

$$LIN(X) = \sum_{i=1}^{k} \lambda_i x_i \quad (\sum_{i=1}^{k} \lambda_i = 1, \lambda_i \neq 0) \tag{12}$$

$$SUM(X) = \sum_{i=1}^{k} x_i \tag{13}$$

$$PROD(X) = \prod_{i=i}^{k} x_i \tag{14}$$

$$HM(X) = \frac{k \cdot \prod_{j=1}^{k} x_j}{\sum_{i=1}^{k} \frac{\prod_{j=1}^{k} x_j}{x_i}} \tag{15}$$

Score fusion methods using a copula must be given a type of copula as a parameter. Moreover, marginal distributions

[5] http://www.lemurproject.org/clueweb09.php

can be estimated by both Gaussian and empirical distributions. The types of copulas used in our experiments here were the Gaussian copula, the Clayton copula, the Gumbel copula, the Frank copula, and the empirical copula. Since the distribution of relevant documents was approximately divided into two clusters, we let the number of clusters used in copula methods C_{mix} and $C_{mixprod}$ to be two. Note that the number of clusters can be determined automatically as described in Section 2.3.3.

Finally, we also evaluated BM25 and the readability score to confirm the improvement of score fusion.

4.3.3 Results

Table 4 and Figures 2a and 2b show the results of our experiment. Score fusion methods using copula were optimized and set to the best combination of a marginal distribution and a copula type.

Initially, fusion methods and single scores (BM25 and the readability score) were compared. One of the single scores, $READABILITY$, showed significantly low accuracy in all measures as compared with other methods; however this result is natural because $READABILITY$ does not consider the relevance to queries. Regarding $BM25$, in terms of IP@0.0, C_{mix}, $C_{mixprod}$, SUM, $PROD$, and HM outperformed $BM25$. Meanwhile, in terms of IP@0.1, C_{mix} and $C_{mixprod}$ outperformed $BM25$. Conversely, in terms of IP@$k(\leq 0.2)$, $BM25$ showed the best performance. It was found that fusion methods are better than $BM25$ in interpolated precision, because users are usually interested in only top results.

Finally, in terms of nDCG@$k(= 5, 10, 15)$, C_{mix} and $C_{mixprod}$ outperformed $BM25$ while in terms of nDCG@20, C_{mix} outperformed $BM25$. Consequently, fusion methods of copulas are better in terms of nDCG.

Further, C_{mix} and $C_{mixprod}$, fusion methods using copula, outperformed $BM25$ in terms of interpolated precision and nDCG. Thus, it can be concluded that the proposed fusion methods using copula are effective models.

5. CONCLUSION

In this work, we proposed a readability measure for Web documents and information retrieval methods that take into account readability. In the results of our experiments, our proposed readability measure improved the accuracy in comparison with existing measures. In the experiment of score fusion methods, combining readability and relevance using mixture copula improved the accuracy in comparison with single scores and a linear combination.

In future work, we plan to use a larger dataset to obtain more reliable results, since only a small dataset was used in this study.

Acknowledgment

This work was partly supported by JSPS KAKENHI Grant Numbers 15H02701, 15K20990, 16H02908, and 26280115.

6. REFERENCES

[1] John Bormuth. Readability: A new approach. In *Readabing Researh Quarterly*, volume 1, pages 79–132, 1966.

[2] Jean-Philippe Bouchaud and Marc Potters. Theory of Financial Risk and Derivative Pricing. *Theory of Financial Risk and Derivative Pricing From Statistical Physics to Risk Management*, page 379, 2003.

[3] W. Breymann, A. Dias, and Paul Embrechts. Dependence structures for multivariate high-frequency data in finance. *Quantitative Finance*, 3(1):1–14, 2003.

[4] Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning ICML 05*, pages 89–96, 2005.

[5] Jeanne S. Chall. *Readability Revisited: The New Dale–Chall Readability Formula.* Brookline Books, 1995.

[6] Kevyn Collins-Thompson. Computational Assessment of Text Readability: A Survey of Current and Future Research. *Recent Advances in Automatic Readability Assessment and Text Simplification*, 165(2):97–135, 2014.

[7] Kevyn Collins-Thompson, Paul N. Bennett, Ryen W. White, Sebastian de la Chica, and David Sontag. Personalizing web search results by reading level. In *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11*, page 403, 2011.

[8] Kevyn Collins-Thompson and Jamie Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the HLT/NAACL*, 2004.

[9] D. R. Cox. The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, 20(2):215–242, 1958.

[10] Ronan Cummins. Measuring the ability of score distributions to model relevance. *Information Retrieval Technology*, pages 25–36, 2011.

[11] Na Dai, Milad Shokouhi, and Brian D. Davison. Learning to rank for freshness and relevance. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information - SIGIR '11*, page 95, 2011.

[12] Edwin Diday, A Schroeder, and Y Ok. The Dynamic Clusters Method in Pattern Recognition. In *IFIP Congress*, pages 691–697, 1974.

[13] Carsten Eickhoff and Arjen P de Vries. Modelling Complex Relevance Spaces with Copulas. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management - CIKM '14*, pages 1831–1834, 2014.

[14] Carsten Eickhoff, Arjen P. de Vries, and Kevyn Collins-Thompson. Copulas for information retrieval. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13*, page 663, 2013.

[15] Paul Embrechts, Filip Lindskog, and Alexander Mcneil. Ch.8 Modelling dependence with copulas and applications to risk management. *Handbook of Heavy Tailed Distributions in Finance*, pages 329–384, 2003.

[16] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. Cognitively motivated features for readability assessment. In *EACL 2009 - 12th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings*, pages 229–237, 2009.

[17] Rudolf Flesch. A new readability yardstick. In *Journal of Applied Psychology*, volume 32, pages 221–233, 1948.

[18] Edward A Fox and Joseph A Shaw. Combination of Multiple Searches. In *The 2nd Text Retrieval Conference TREC2 NIST SP 500215*, volume 500-215, pages 243–252, 1994.

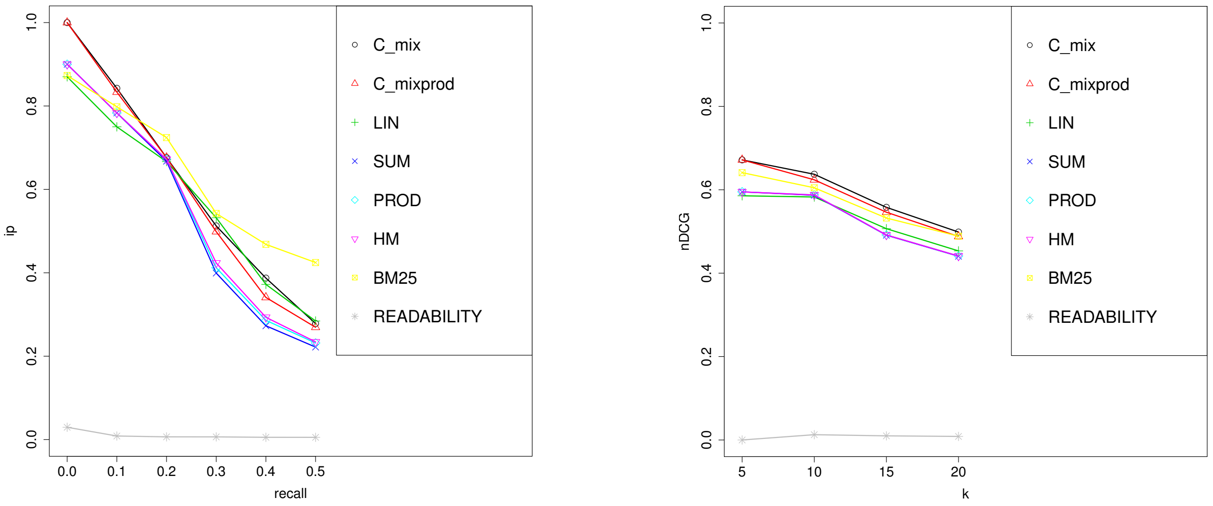
[19] Shima Gerani, Chengxiang Zhai, and Fabio Crestani. Score transformation in linear combination for multi-criteria relevance ranking. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7224 LNCS, pages 256–267, 2012.

[20] Kari Gyllstrom and Marie-Francine Moens. Wisdom of the Ages: Toward Delivering the Children's Web with the Link-based Agerank Algorithm. In *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, pages 159–168, 2010.

[21] George R. Klare. *The Measurement of Readability.* Iowa

Table 4: Results of score fusion methods

Method	C_{mix}	$C_{mixprod}$	LIN	SUM	$PROD$	HM	$BM25$	$READABILITY$
IP@0.0	**1.0000**	**1.0000**	0.8700	0.9000	0.9000	0.9000	0.8731	0.0296
IP@0.1	**0.8424**	0.8333	0.7500	0.7833	0.7833	0.7833	0.7981	0.0086
IP@0.2	0.6762	0.6762	0.6667	0.6676	0.6723	0.6723	**0.7243**	0.0066
IP@0.3	0.5124	0.4987	0.5320	0.3993	0.4121	0.4236	**0.5423**	0.0065
IP@0.4	0.3873	0.3409	0.3722	0.2729	0.2860	0.2933	**0.4682**	0.0055
IP@0.5	0.2783	0.2693	0.2844	0.2218	0.2316	0.2343	**0.4246**	0.0054
nDCG@5	**0.6717**	**0.6717**	0.5856	0.5950	0.5950	0.5950	0.6411	0.0000
nDCG@10	**0.6372**	0.6239	0.5827	0.5868	0.5876	0.5876	0.6048	0.0127
nDCG@15	**0.5579**	0.5462	0.5067	0.4911	0.4917	0.4917	0.5323	0.0099
nDCG@20	**0.4986**	0.4880	0.4534	0.4406	0.4412	0.4412	0.4887	0.0085



(a) IP@recall

(b) nDCG

Figure 2: Results of score fusion methods

State University Press, 1963.

[22] Takuya Komatsuda, Atsushi Keyaki, and Jun Miyazaki. A Score Fusion Method Using a Mixture Copula. In *Proceedings of the 27th International Conference on Database and Expert Systems Applications - DEXA*, 2016, to appear.

[23] Irving Lorge. Predicting Readability. In *Teachers College Record*, volume 45, pages 404–419, 1944.

[24] Tomohiro Manabe and Keishi Tajima. Extracting logical hierarchical structure of HTML documents based on headings. In *Proceedings of the 41st International Conference on Very Large Data Bases*, pages 1606–1617, 2015.

[25] Courtney Napoles and Mark Dredze. Learning Simple Wikipedia: A Cogitation in Ascertaining Abecedarian Language. In *Proceedings of the HLT/NAACL*, pages 42–50, 2010.

[26] Roger B Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013.

[27] Arno Onken, Steffen Grünewälder, Matthias H J Munk, and Klaus Obermayer. Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation. *PLoS Computational Biology*, 5(11), 2009.

[28] Jay M Ponte and W Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, volume 98, pages 275–281, 1998.

[29] B. Renard and M. Lang. Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology. *Advances in Water Resources*, 30(4):897–912, 2007.

[30] G. Salton. Automatic text processing: the transformation. *Analysis and Retrieval of Information by Computer*, 14:15, 1989.

[31] Christian Schoelzel, Petra Friederichs, et al. Multivariate non-normally distributed random variables in climate research–introduction to the copula approach. *Nonlin. Processes Geophys.*, 15(5):761–772, 2008.

[32] A J Scott and M J Symons. Clustering Methods Based on Likelihood Ratio Criteria. *Biometrics*, 27(2):387–397, 1971.

[33] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, and Katsumi Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the sixth ACM international conference on Web search and data mining - WSDM '13*, pages 475–484, 2013.

[34] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North America Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, pages 173–180, 2003.

[35] Christopher C Vogt and Garrison W Cottrell. Fusion Via a Linear Combination of Scores. *Information Retrieval*, 1(3):151–173, 1999.

[36] Mathieu Vrac, Lynne Billard, Edwin Diday, and Alain Chedin. Copula analysis of mixture models. *Computational Statistics*, pages 427–457, 2011.