**A Study of Web Page Understandability for Consumer Health Search**

Joao Palotti1,3; Guido Zuccon2; Allan Hanbury3

1Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar

2Queensland University of Technology, Brisbane, Australia

3Vienna University of Technology, Vienna, Austria

# *Abstract*

**Background:** Understandability plays a key role in ensuring that people accessing health information are capable of gaining insights that can assist them with their health concerns and choices. The access to unclear or misleading information has been shown to negatively impact on the health decisions of the general public.

**Objective:** We investigated methods to estimate the understandability of health Web pages and used these to improve the retrieval of information for people seeking health advice on the Web.

**Methods:** Our investigation considered methods to automatically estimate the understandability of health information in Web pages, and it provided a thorough evaluation of these methods using human assessments as well as an analysis of pre-processing factors affecting understandability estimations, and associated pitfalls. Furthermore, lessons learnt for estimating Web page understandability were applied to the construction of retrieval methods with specific attention to retrieving information understandable by the general public.

**Results:** We found that machine learning techniques were more suitable to estimate health Web page understandability than traditional readability formulas, which are often used as guidelines and benchmarking by health information providers on the Web (larger difference found for Pearson correlation of .602 using Gradient Boosting regressor compared to .438 using SMOG

Index with CLEF 2015 collection). Learning to rank effectively exploited these estimates to provide the general public with more understandable search results (*HR*∗ *BP* reached 29.20, 22% higher than a BM25 baseline and 13% higher than the best system at CLEF 2016, both *P ≤ .*001).

**Conclusions:** The findings reported in this article are important for specialised search services tailored to support the general public in seeking health advice on the Web, as they document and empirically validate state-of-the-art techniques and settings for this domain application.

**KEYWORDS:** Consumer Health Search; Document Understandability; Document Readability; Learning to Rank

# *Introduction*

in which dimensions other than topicality have an important

role in the information seeking and decision-making process.

The seeking of health information and advice on the Web by

Search engines are concerned with retrieving relevant informa- tion to support a user’s information seeking task. Commonly, signals about the topicality or aboutness of a piece of infor- mation with respect to a query are used to estimate relevance, with other relevance dimensions like understandability, trust- worthiness, etc. [1] being relegated to a secondary position, or completely neglected. While this may be a minor problem for many information seeking tasks, there are some specific tasks

the general public is one such task.

A key problem when searching the Web for health informa- tion is that this can be too technical, unreliable, generally mis- leading, and can lead to unfounded escalations and poor deci- sions [2]. Where correct information exists, it can be hard to find and digest amongst the noise, spam, technicalities, and ir- relevant information. In *high-stakes search tasks* such as this, access to poor information can lead to poor decisions which ul-

timately can have a significant impact on our health and well- being [2, 3]. In this work we are specifically interested in the understandability of health information retrieved by search en- gines, and in improving search results to favour information understandable by the general public. We leave addressing reliability and trustworthiness of the retrieved information to future work; however, this can be achieved by extending the framework we investigate here.

The use of general purpose Web search engines like Google, Bing and Baidu for seeking health advice has been largely anal- ysed, questioned and criticised [4–9] , despite the commend- able efforts these services have put into providing increasingly better health information, e.g., the Google Health Cards [10] .

Ad-hoc solutions to support the general public in searching and accessing health information on the Web have been imple- mented, typically supported by government initiatives or medi- cal practitioner associations, e.g., *HealthOnNet.org* (HON) and *HealthDirect.gov.au*, among others. These solutions aim to provide *better* health information to the general public. For example, HON’s mission statement is "to guide Internet users to reliable, understandable, accessible and trustworthy sources of medical and health information". But, do the solutions these services currently employ actually provide this type of infor- mation to the health-seeking general public?

As an illustrative example, we analysed the top 10 search results retrieved by HON on 01/10/2017 in answer to 300 health search queries generated by regular health consumers in health forums. These queries are part of the CLEF 2016 eHealth collection, which shall be extensively used in this article. The understandability score of the retrieved pages was estimated with the most effective readability formula and preprocessing settings analysed in this article (low scores cor- respond to easy to understand Web pages). Figure 1 reports the cumulative distribution of understandability scores for these search results (note, we did not assess their topical relevance here). We report also the scores for the "optimal" search results (Oracle), as found in CLEF 2016 (relevant results that have the highest understandability scores), along with the scores for the baseline method (BM25) and our best retrieval method (XGB). The results clearly indicate that, despite solutions like HON being explicitly aimed at supporting access to understandable health information, they often fail to do so.

In this article, we aim to establish methods and best practice for developing search engines that retrieve *relevant and under- standable* health advice from the Web. The overall contribu- tions of this article can be summarised as:

1. We propose and investigate methods for the estimation of the understandability of health information in Web pages:

a large number of medically-focused features are grouped in meaningful categories and their contribution to the un- derstandability estimation task is carefully measured;

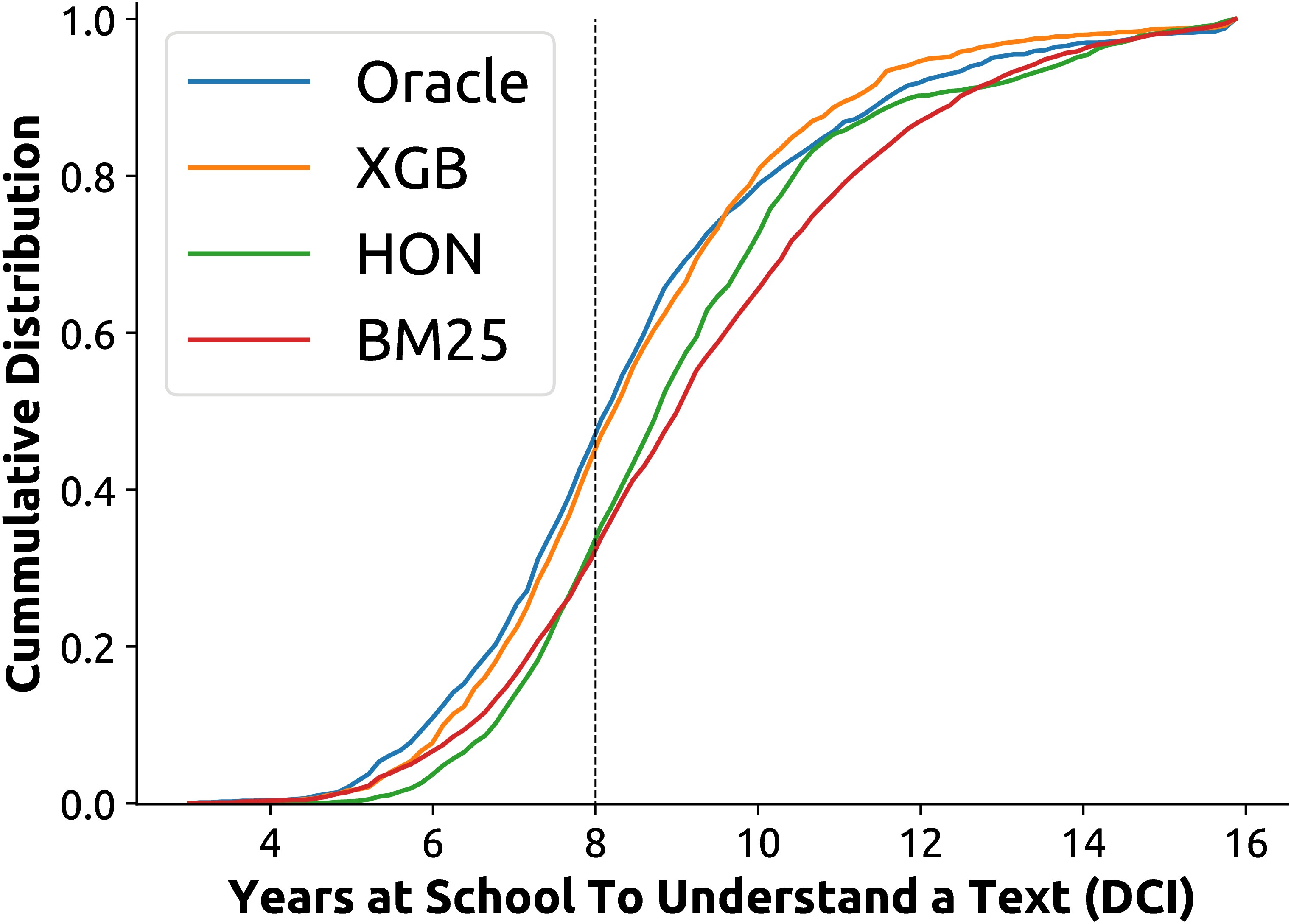
1. We further study the influence of HTML processing meth- ods on these estimations and their pitfalls, extending our previous work that has shown how this often ignored as- pect greatly impacts effectiveness [16] ;
2. We further investigate how understandability estimations can be integrated into retrieval methods to enhance the quality of the retrieved health information with particu- lar attention to its understandability by the general public. New models are explored in this article, also extending our previous work [17];

This paper makes concrete contributions to practice, as it in- forms health search engines specifically tailored to the general public (for example the HON or HealthDirect services referred to above) about the best methods they should adopt, but they currently don’t. These are novel and significant contributions, as no previous work has systematically analysed the influence of the components at play in this study and we show that these greatly influence retrieval effectiveness and thus delivery of rel- evant and understandable health advice.

## Related Work

Understandability refers to the ease of comprehension of the information presented to a user. Put in other words, health in- formation is understandable "when consumers of diverse back- grounds and varying levels of health literacy can process and explain key messages" [18]. Often the terms understandability and readability are used interchangeably: we use readability to refer to formulas that estimate how easy is to understand a text, usually based on its words and sentences. We use understand- ability to refer to the broader concept of ease of understand- ing: this is affected by text readability (as increasing readabil- ity tends to improve understanding), but may also be influenced by how legible a text is and its layout, including e.g., the use of images to explain difficult concepts.

There is a large body of literature that has examined the un- derstandability of Web health content when the information seeker is a member of the general public. For example, Becker reported that the majority of health Web sites are not well de- signed for the elderly [19], while Stossel et al. found that health education material on the Web is not written at an ad- equate reading level [14]. Zheng and Yu have reported on the readability of electronic health records compared to Wikipedia pages related to diabetes and found that readability measures



**Figure 1.** Distribution of Dale-Chall Index (DCI) of search results. DCI measures the years of schooling required to understand a document. The average US resident reads at or below an 8th grade level (dashed line) [11–14] , which is the level suggested by NIH for health information on the Web [15] . The distribution for HON is similar to that of the baseline used in this article (BM25). Our best method (XGB) re-ranks documents to provide more understandable results; its distribution is similar to that of an "Oracle" system.

often do not align with user ratings of readability [20]. A com- mon finding of these studies is that, in general, health content available on Web pages is often hard to understand by the gen- eral public; this includes content that is retrieved in top-ranked positions by current commercial search engines [4–8].

Previous Linguistics and Information Retrieval research has attempted to devise computational methods for the automatic estimation of text readability and understandability, and for the inclusion of these within search methods or their evaluation. Computational approaches to understandability estimations in- clude (1) *readability formulas*, which generally exploit word surface characteristics of the text, (2) *machine learning* ap- proaches, (3) matching with specialised *dictionaries or termi- nologies*, often compiled with information about understand- ability difficulty.

Measures such as Coleman-Liau Index (CLI) [21], Dale- Chall Index (DCI) [22] and Flesch Reading Easy (FRE) [23]

belong to the first category. These measures generally rely on surface-level characteristics of text, such as characters, sylla- bles and word counts [24] . While these measures have been widely used in studies investigating the understandability of health content retrieved by search engines (e.g. [4, 5, 7, 8, 14, 19]), our preliminary work found that these measures are heavily affected by the methods used to extract text from the HTML source [16]. We were able to identify specific settings of an HTML preprocessing pipeline that provided consistent estimates, but due to the lack of human assessments, we were not able to investigate how well each HTML preprocessing pipeline correlated with human assessments. In this article, we revisited and extended this work in more details, as we further investigated this problem by comparing the effect of HTML preprocessing on text understandability estimations in light of explicit human assessments.

The use of machine learning to estimate understandabil-

ity forms an alternative approach. Earlier research explored

# *Methods*

the use of statistical natural language processing and language

modelling [25–27] as well as linguistic factors, such as syn- tactic features or lexical cohesion [28] . While we replicated here many of the features devised in these works, they focus on estimating readability of general English documents rather than medical ones. In the medical domain, Zeng et al. explored features such as word frequency in different medical corpora to estimate concept familiarity, which prompted the construction of the Consumer Health Vocabulary (CHV) [29–31].

The actual use of CHV or other terminologies such as the Medical Subject Headings (MeSH) belongs to the third cat- egory of approaches. The CHV is a prominent medical vo- cabulary dedicated to mapping layperson vocabulary to tech- nical terms [31]. It attributes a score for each of its concepts with respect to their difficulty, with lower/higher scores for harder/easier concepts. Researchers have evaluated CHV in tasks such as document analysis [32] and medical expertise prediction [33]. The hierarchy of MeSH was previously used in the literature to identify hard concepts, assuming that a con- cept deep in the hierarchy is harder than a shallow one [34]. Other approaches combined vocabularies with word surface characteristics and syntactic features, like part of speech, into a unique readability measure [35].

In this work, we investigated approaches to estimate un- derstandability from each of these categories, including mea- sure the influence of HTML preprocessing on automatic un- derstandability methods and establish best practices.

Some prior work has attempted to use understandability estimations for improving search results in consumer health search; as well as methods to evaluate retrieval systems that do account for understandability along with topical relevance. Palotti et al. have used learning to rank with standard retrieval features along with features based on readability formulas and medical lexical aspects to determine understandability [17]. Van Doorn et al. have shown that learning a set of rankers that provide trade-offs across a number of relevance criteria, including readability/understandability, increases overall sys- tem effectiveness [36] . Zuccon and Koopman [37], and later Zuccon [38], have proposed and investigated a family of mea- sures based on the gain-discount framework, where the gain of a document is influenced by both its topical relevance and its understandability. They showed that, although generally corre- lated, topical-relevance evaluation alone provides differing sys- tem rankings compared to understandability-biased evaluation measures. In this work, we further explored the development of retrieval methods that combine signals about topical relevance and understandability.

## Data Collection

In this article, we investigated methods to estimate Web page understandability, including the effect HTML preprocessing pipelines and heuristics have, and their search effectiveness when employed within retrieval methods. To obtain both top- ical relevance and understandability assessments, we used the data from the CLEF 2015 and 2016 eHealth collections. (We refer to topical relevance simply as relevance in the reminder of the paper, when this does not cause confusion.)

The CLEF 2015 collection contains 50 queries and 1,437 documents that have been assessed relevant by clinical ex- perts and have an assessment for understandability [39]. Doc- uments in this collection are a selected crawl of health Web sites, of which the majority are certified HON Web sites. The CLEF 2016 collection contains 300 queries and 3,298 rele- vant documents that also have been assessed with respect to understandability [40]. Documents in this collection belong to the ClueWeb12 B13 corpus, and thus are general English Web pages, not necessarily targeted to health topics, nor of a controlled quality (as are instead HON certified pages). Un- derstandability assessments were provided on a 5-point Likert scale for CLEF 2015, and on a [0,100] range for CLEF 2016 (0 indicates the highest understandability).

To support the investigation of methods to automatically es- timate the understandability of Web pages, we further con- sidered correlations between multiple human assessors (inter- assessor agreement). For CLEF 2015, we used the publicly available additional assessments made by unpaid medical stu- dents and health consumers collected by Palotti et al. in a study of how medical expertise affects assessments [41]. For CLEF 2016 we collected understandability assessments for 100 doc- uments. Three members of our research team, which did not author this article and were not medical experts, were recruited to provide the assessments (the correlation of these additional assessments and CLEF’s ground-truth is examined further in this article). The Relevation tool [42] was used to assist with the assessments, mimicking the settings used in CLEF.

## Evaluation Measures

In the experiments, we used Pearson, Kendall and Spearman correlations to compare the understandability assessments of human assessors with estimations obtained by the considered approaches, under all combinations of pipelines and heuristics. Pearson correlation is used to calculate the strength of the linear relationship between two variables, while Kendall and Spear- man measure the rank correlations between the variables. We opted to report all three correlation coefficients to allow for a

thorough comparison to other work, as they are equally used in the literature.

For the retrieval experiments, we used evaluation mea- sures that rely on both (topical) relevance and understand- ability. The uRBP measure [38] extends rank biased pre- cision (RBP) to situations where multiple relevance dimen-

sions are used. The measure is formulated as *uRBP* (*ρ*) = (1 *− ρ*) *K ρk*−1*r*(*d*@*k*)*u*(*d*@*k*), where *r*(*d*@*k*) is the gain for retrieving a relevant document at rank *k* and *u*(*d*@*k*) is the

*k*=1

Σ

gain for retrieving a document of a certain understandability at rank *k*; *ρ* is the RBP persistence parameter. This measure was an official evaluation measure used in CLEF (we also set *ρ* = 0*.*8).

A drawback of uRBP is that relevance and understandabil-

ity are combined into a unique evaluation score, thus making it difficult to interpret whether improvements are due to more understandable or more topical documents being retrieved. To overcome this, we first separately calculated an RBP value for relevance and another for understandability, and then combined them into a unique effectiveness measure:

* *RBPr*@*n*(*ρ*): uses the relevance assessments for the top *n* search results (i.e. this is the common RBP). We re- garded a document as topically relevant if assessed as somewhat relevant or highly relevant..
* *RBPu*@*n*(*ρ*): uses the understandability assessments for the top *n* search results. We regarded a document as un- derstandable (1) for CLEF 2015 if assessed easy or some- what easy to understand; (2) for CLEF 2016 if its assessed understandability score was smaller than a threshold *U* . We used*U* = 40, based on the distribution of understand- ability assessments. This distribution can be found in the appendix.}
* *HRBP* @*n*(*ρ*) = 2 *× RBPr* @*n*×*RBPu*@*n* : combines the previous two RBP values into a unique measurement us- ing the harmonic mean (in the same fashion that the *F*1 measure combines recall and precision).

*RBPr* @*n*+*RBPu*@*n*

For all measures, we set *n* = 10 because shallow pools were used in CLEF along with measures that focused on the top 10 search results (including *RBPr*@10).

Along with these measures of search effectiveness, we also recorded the number of unassessed documents, the RBP residu- als, *RBPr*∗@10, *RBPu*∗@10 and *HR*∗ *BP* , i.e. the corresponding

measures calculated by ignoring unassessed documents. These

latter measures implement the condensed measures approach proposed by Sakai as a way to deal with unassessed docu- ments [43]. We did this to minimise pool bias since the pools built in CLEF were of limited size, and the investigated meth- ods retrieved a substantial number of unassessed documents.

## Understandability Estimators

Several methods have been used to estimate the understand- ability of health Web pages, with the most popular methods (at least in the biomedical literature) being readability formu- las based on surface level characteristics of the text. Next, we outline the categories of methods to estimate understandability used in this work; an overview is shown in Table 1 . Some of these methods further expand measures used in the literature.

**Table 1.** Methods used to estimate understandability. ?: raw values were used. ◆: values normalised by number of words in a

document were used. †: values normalised by number of sentences in a document were used. **TODO: Need to replace ? by another symbol**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Cat. | Method | Cat. | Method | Cat. | Method |
| **RF** | Automated Readability Index (ARI) [44]  Coleman-Liau Index (CLI) [21] | **WFF** | 25th percentil English Wikipedia  50th percentil English Wikipedia | **HF** | # of Ab  # of A t |

Dale-Chall Index (DCI) [22] 75th percentil English Wikipedia # of Blo

Flesch-Kincaid Grade Level (FKGL) [23] Mean Rank English Wikipedia # of Bo Flesch Reading Ease (FRE) [23] Mean Rank English Wikip. - Includes OV # of Cit

Gunning Fog Index (GFI) [45] 25th percentil Medical Reddit # of Div

Lasbarhetsindex (LIX) [46] 50th percentil Medical Reddit # of For

Simple Measure of Gobbledygook (SMOG) [47] 75th percentil Medical Reddit # of H1

**CRF** # of Characters ?◆† Mean Rank Medical Reddit # of H2

# of Words ?† Mean Rank Medical Reddit - Includes OV # of H3

# of Sentences ?◆ 25th percentil Pubmed # of H4

# of Difficult Words (Dale-Chall list [22] ) ?◆† 50th percentil Pubmed # of H5

# of Words Longer than 4 chars ?◆† 75th percentil Pubmed # of H6

# of Words Longer than 6 chars ?◆† Mean Rank Pubmed # of Hs

# of Words Longer than 10 chars ?◆† Mean Rank Pubmed - Includes OV # of Im

# of Words Longer than 13 chars ?◆† 25th p. Wikipedia+Reddit+Pubmed # of Inp

# of Number of Syllables ?◆† 50th p. Wikipedia+Reddit+Pubmed # of Lin

# of Polysyllable Words (>3 Syllables) ?◆† 75th p. Wikipedia+Reddit+Pubmed # of DL

**NL** # of Entities ?◆† Mean R. Wiki.+Reddit+Pubmed # of UL

# of verbs ?◆† Mean R. Wiki.+Reddit+Pubmed - w. OV # of OL

# of nouns ?◆† **GMV** # of Words with Medical Prefix ?◆† # of Lis

# of pronouns ?◆† # of Words with Medical Suffix ?◆† # of Q t

# of adjectives ?◆† # of Acronyms ?◆† # of Scr

# of adverbs ?◆† # of ICD Concepts ?◆† # of Spa

# of adpositions ?◆† # of Drugbank ?◆† # of Tab

# of conjunctions ?◆† # of Words in medical dict. (OpenMedSpel) ?◆† # of P t

# of determiners ?◆† **CMV** CHV Mean Score for all Concepts ?◆† **MLR** Linear

# of cardinal numbers ?◆† # of CHV Concepts ?◆† Multi-la

# of particles or other function words ?◆† CHV Mean Score for Symptom Concepts ?◆† Random

# of other POS (foreign words, typos) ?◆† # of CHV Symptom Concepts ?◆† Support

# of punctuation ?◆† CHV Mean Score for Disease Concepts ?◆† Gradien Height of part-of-speech parser tree ?◆† # of CHV Disease Concepts ?◆† Logistic

# of Stopwords ?◆† **EMV** # of MeSH Concepts ?◆† **MLC** Multi-la

# of words not found in Aspell Eng. dict. ?◆† Average Tree of MeSH Concepts ?◆† Random Positive Words ?◆† # of MeSH Symptom Concepts ?◆† Support

Negative Words ?◆† Average Tree of MeSH Symptom Concepts ?◆† Multino

Neutral Words ?◆† # of MeSH Disease Concepts ?◆† Gradien Average Tree of MeSH Disease Concepts ?◆†

*Traditional Readability Formulas (RF):* These include the most popular readability formulas [21–23], as well as other, less popular ones [44–46]. A full list is provided in surveys by Collins-Thompson [48] and Dubay [24] .

*Raw Components of Readability Formulas (CRF):* These are formed by the "building blocks" used in the traditional read- ability formulas. Examples include the average number of characters per word and the average number of syllables in a sentence. Words are divided into syllables using the Python package Pyphen [49].

*General Medical Vocabularies (GMV):* These include meth- ods that count the number of words with a medical prefix or suffix, i.e. beginning or ending with Latin or Greek particles (e.g., amni-, angi-, algia-, arteri-), and text strings included in lists of acronyms or in medical vocabularies such as the In- ternational Statistical Classification of Diseases and Related Health Problems (ICD), Drugbank and the OpenMedSpel dic- tionary [50]. An acronym list from the ADAM database [51] was used. Methods in this category were matched with docu- ments using simple keywords matching.

*Consumer Medical Vocabulary (CMV):* The popular MetaMap [52] tool was used to map the text content of Web pages to entries in CHV [31]. We used the MetaMap semantic types to retain only concepts identified as symptoms or dis- eases. Similar approaches have been commonly used in the literature (e.g., [53–56]).

*Expert Medical Vocabulary (EMV):* Similarly to the CHV features, we used MetaMap to convert the content of Web pages into MeSH entities, studying symptom and disease con- cepts separately.

*Natural Language Features (NLF):* These included com- monly used natural language heuristics such as the ratio of part- of-speech (POS) classes, the height of the POS parser tree, the number of entities in the text, the sentiment polarity and the ra- tio of words found in English vocabularies. The Python pack- age NLTK [57] was employed for sentiment analysis, POS tag- ging and entity recognition. The GNU Aspell [58] dictionary was used as a standard English vocabulary and a stop word list was built by merging those of Indri [59] and Terrier [60]. Discourse features, such as the distribution of POS classes and density of entity in a text, were previously studied in the task of understandability prediction [61] and found superior to com- plex features such as entity co-reference and entity grid [62]. To the best of our knowledge, sentiment polarity was never in- vestigated in this task. Our intuition is that the content pro- duced by laypeople in patient forums or blogs (easy-to-read) is potentially more emotional than scientific publications (hard- to-read).

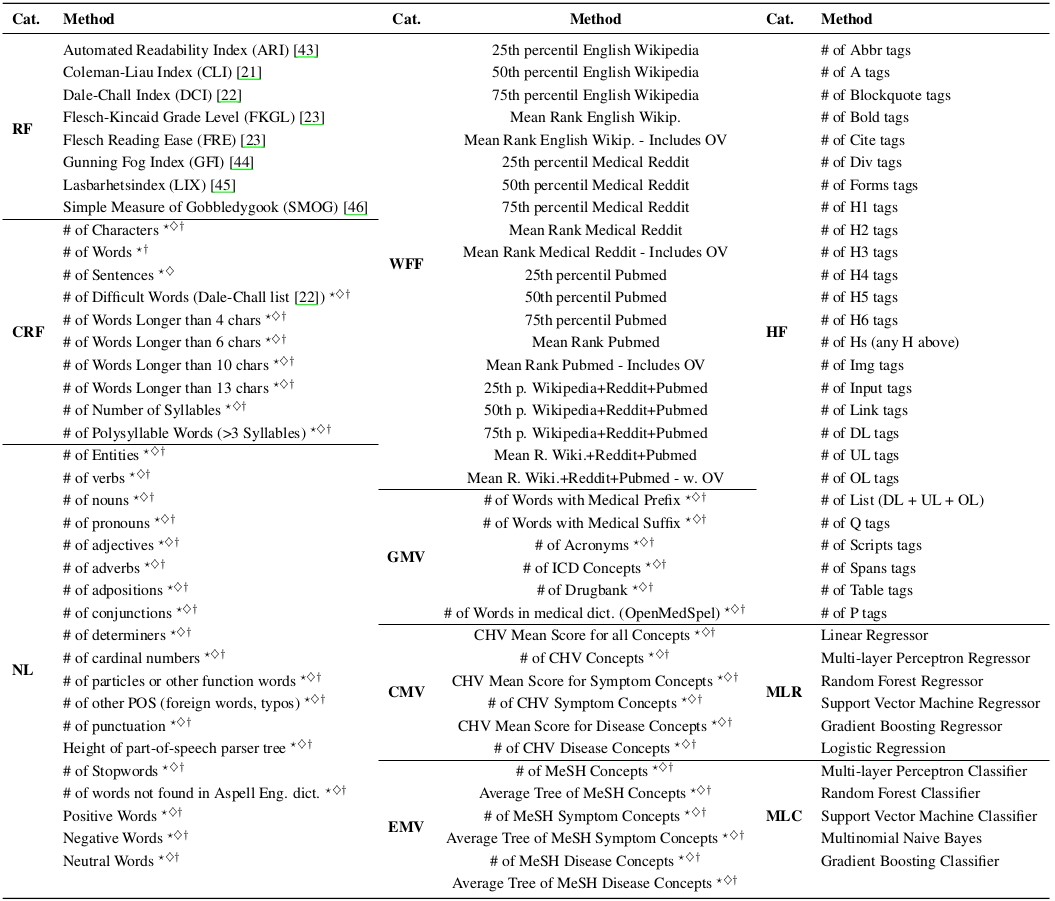
*HTML Features (HF):* These included the identification of a large number of HTML tags, which were extracted with the

Python library BeautifulSoap [63]. The intuition for these fea- tures is that Web pages with many images and tables may ex- plain and summarise health content better, thus providing more understandable content to the general public.

*Word Frequency Features (WFF):* Generally speaking, com- mon and known words are usually frequent words, while un- known and obscure words are generally rare. This idea is im- plemented in readability formulas such as the DCI, which uses a list of common words and counts the number of words that fall outside this list (complex words) [22] and has shown suc- cess in other recent approaches [64, 65]. We extended these observations by studying corpus-wide word frequencies. We modelled word frequencies in a corpus in a straightforward manner: we sorted the word frequencies and normalised word rankings such that values close to 100 are attributed to common words and values close to 0 to rare words. Three corpora were analysed to extract word frequencies:

* Medical Reddit: Reddit [66] is a Web forum with a size- able user community which is responsible for generat- ing and moderating its content. This forum is inten- sively used for health purposes: for example in the Red- dit community AskDocs [67], licensed nurses and doc- tors (subject to user identity verification) advise help seek- ers free of charge. We selected six of such communities (medical, AskDocs, AskDoctorSmeeee, Health, Women- sHealth, Mens\_Health) and downloaded all user interac- tions available until September 1st 2017 using the Python library PRAW [68]. In total 43,019 discussions were col- lected.
* Medical English Wikipedia: after obtaining a recent Wikipedia dump [69] (May 1st 2017), we filtered ar- ticles to only those containing an Infobox in which at least one of the following words appeared as a prop- erty: ICD10, ICD9, DiseasesDB, MeSH, MeSHID, Mesh- Name, MeshNumber, GeneReviewsName, Orphanet, eMedicine, MedlinePlus, drug\_name, Drugs.com, Daily- MedID, LOINC. A Wikipedia infobox is a structured tem- plate that appears on the right of Wikipedia pages sum- marising key aspects of articles. This process followed the method by Soldaini et al. [70], which favours preci- sion over recall when identifying a health-related article. This resulted in a collection of 11,868 articles.
* PubMed Central: PubMed Central [71] is an online database of biomedical literature. We used the collection distributed for the TREC 2014 and 2015 Clinical Decision Support Track [72, 73], consisting of 733,191 articles.

A summary of the statistics of the corpora is reported in Ta- ble 2 . Unless explicitly stated otherwise, we ignored out of



**Figure 2.** Methods used to estimate understandability. ?: raw values were used. ◆: values normalised by number of words in adocument were used. †: values normalised by number of sentences in a document were used. **TODO: Need to replace ? by another symbol. In case we go for the figure, we need to replace the ? and update reference numbers.**

vocabulary words in the feature calculations.

*Machine Learning on Text - Regressors (MLR) and Classi- fiers (MLC):* These include machine learning methods for esti- mating Web page understandability. While Collins-Thompson highlighted the promise of estimating understandability using machine learning methods, a challenge is identifying the back- ground corpus to be used for training [48]. To this aim, we used the three corpora detailed above, and assumed understandabil- ity labels according to the expected difficulty of documents in these collections:

* Medical Reddit (label 1): Documents in this corpus are expected to be written in a colloquial style, and thus the easiest to understand. All the conversations are in fact explicitly directed to assist inexpert health consumers;
* Medical English Wikipedia (label 2): Documents in this corpus are expected to be less formal than scientific arti- cles, but more formal than a Web forum like Reddit, thus somewhat more difficult to understand;
* PubMed Central (label 3): Documents in this corpus are expected to be written in a highly formal style, as the target audience are physicians and biomedical researchers.

Models were learnt using all documents from these corpora after features were extracted using Latent Semantic Analysis (LSA) with 10 dimensions (this number of dimensions was chosen based on preliminary experiments with the Random

**Table 2.** Statistics for the corpora used as background models for understandability estimations.Statistic

|  |  |  |  |
| --- | --- | --- | --- |
| Statistic | Medical Wikipedia | Medical Reddit | PubMed Central |
| Number of Documents | 11*,* 868 | 43*,* 019 | 733*,* 191 |
| Number of Words | 10*,* 655*,* 572 | 11*,* 978*,* 447 | 144*,* 024*,* 976 |
| Number of Unique Words | 467*,* 650 | 317*,* 106 | 2*,* 933*,* 167 |
| Avg. Words per Document | 898*.*90 *±* 1351*.*76 | 278*.*45 *±* 359*.*70 | 227*.*22 *±* 270*.*44 |

Avg. Char per Document 5*,* 107*.*81 *±* 7*,* 618*.*57 1*,* 258*.*44 *±* 1*,* 659*.*96 1*,* 309*.*11 *±* 1*,* 447*.*31

Avg. Char per Word 5*.*68 *±* 3*.*75 4*.*52 *±* 3*.*52 5*.*76 *±* 3*.*51

Forest algorithm; we leave as future work a detailed study on the impact of different number of dimensions on other machine learning algorithms). We modelled a classification task as well as a regression task using these corpora. Thus, after applying the same LSA transformation to test documents from CLEF, a continuous score was assigned to each document by a regres- sor, while each classifier assigned the documents to one of the three classes.

## Preprocessing Pipelines and Heuristics

As part of our study, we investigated the influence the prepro- cessing of Web pages has on the estimation of understandabil- ity computed using the methods described above. We did so by comparing the combination of a number of preprocessing pipelines, heuristics, and understandability estimation methods with human assessments of Web page understandability. Our experiments extended those by Palotti et al. [16] and provided a much thorough analysis, as they only evaluated surface level readability formulas and did not compare their results against human assessments.

To extract the content of a Web page from the HTML source we tested: BeautifulSoup [63] (*Naive*), which just naively re- moves HTML tags, Boilerpipe [74] (*Boi*) and Justext [75] (*Jst)*, which eliminates boilerplate text together with HTML tags. Palotti et al.’s data analysis highlighted that the text in HTML fields like titles, menus, tables and lists often missed a correct punctuation mark and thus the text extracted from them could be interpreted as many short sentences or few very long sen- tences, depending on whether a period was forced at the end of fields/sentences. We thus implemented the same two heuristics devised by Palotti et al. to deal with this: *ForcePeriod (FP)* and *DoNotForcePeriod (DNFP)*. The FP heuristic forces a pe- riod at the end of each extracted HTML field, while the DNFP does not.

## Integrating Understandability into Retrieval

We then investigated how understandability estimations can be integrated into retrieval methods to increase the quality of search results. Specifically, we considered three retrieval methods of differing quality for the initial retrieval. These in- cluded the best two runs submitted to each CLEF task, and a plain BM25 baseline (default Terrier parameters: *b* = 0*.*75 and *k*1 = 1*.*2). As understandability estimators we used the eXtreme Gradient Boosting (XGB) regressor [76], as well as SMOG for CLEF 2015 and DCI for CLEF 2016. These were selected as they were the best performing readability formu- las and machine learning method for each collection (details in the evaluation of understandability estimators in the Results section). Note that for XGB, for assessed documents we used 10-fold cross validation, training XGB on 90% of the data, and used its predictions for the remaining 10%. For unassessed documents, we trained XGB on all assessed data and applied this model to generate predictions. Different machine learn- ing methods and feature selection schemes were experimented with; results are available in the appendix. XGB was selected because its results were the best, although other methods fol- lowed similar trends.

To integrate understandability estimators into the retrieval process, we first investigated *re-ranking* search results retrieved by the initial runs purely based on the understandability esti- mations. If all the search results from a run were to be consid- ered, then such a re-ranking method may place at early ranks Web pages highly likely to be understandable, but possibly less likely to be topically relevant. To balance relevance and un- derstandability, we only re-ranked the first *k* documents. We explored rank cut-offs *k* = 15*,* 20*,* 50. Because evaluation was performed with respect to the first *n* = 10 rank positions, the setting *k* = 15 provided a conservative re-ranking of search results, while, *k* = 50 provided a less conservative re-ranking approach.

As an alternative to the previous two-step ranking strategy for combining topical relevance and understandability, we ex- plored the *fusion* of two search results lists separately obtained

for relevance and understandability. For this, we used the Re- ciprocal Rank Fusion (RRF) method [77], which was shown ef- fective for combining two lists of search results based on their documents *ranks*, rather than scores. This approach was se- lected above score-based fusion methods because of the differ- ent scoring strategies and distributions employed when scoring for relevance compared to for understandability. For relevance, we used, separately, the three methods used for re-ranking (ECNU [78] and KISTI [79] for CLEF2015, GUIR [80] and

ECNU [81] for CLEF 2016, and BM25 for both collections). For understandability, we used, separately, the estimations from SMOG/DCI and XGB. Also for this approach, we studied limiting the ranking of results to be considered by the methods

across the cut-offs *k* = 15*,* 20*,* 50.

Finally, we considered a third alternative to combine rele- vance and understandability: using *learning to rank* with fea- tures derived from retrieval methods (IR features) and under- standability estimators. With the CLEF 2016 collection, we explored five combinations of label attribution and feature sets, maintaining the same pairwise learning to rank algorithm based on tree boosting (XGB). These combinations are listed in Ta- ble 3, with *R* being the relevance of documents and *U* their un- derstandability estimation. While the definitions of LTR 1 and 2 are straightforward, the other methods deserve some further explanation. In LTR 3, a penalty was proportionally assigned to documents according to how far their understandability score was from a target score *U* (we used *U* = 40). For example, a document with understandability 100 received no penalty, as 100 was the easiest level of understanding, while another with understandability 50 received a 50% penalty, meaning that its relevance score was halved. LTR 4 and 5 were based on a fixed threshold applied to the understandability score: if the score was higher than the threshold *U* = 40, then the original rel- evance score (for LTR 4) or a boosted value (for LTR 5) was assigned to the corresponding document.

# *Results*

In order to keep this article succinct, in the following we only report a subset of the results. The remaining results (which show similar trends to those reported here) are made available in the appendix material for completeness. *All data and code will be shared on GitHub upon acceptance.*

## Evaluation of understandability estimators

Using the CLEF eHealth 2015 and 2016 collections, we studied the correlations of methods to estimate Web page understand- ability (Table 1 ), compared with human assessments. For each

category of understandability estimation, Table 4 reports the methods with highest Pearson, Spearman or Kendall correla- tions. For each method, we used the best preprocessing set- tings; a study of the impact of preprocessing is reported in the next subsection.

Overall, Spearman and Kendall correlations obtained similar results (in terms of which methods exhibited the highest corre- lations): this was expected as, unlike Pearson, they are both rank-based correlations.

For traditional readability measures, SMOG had the highest correlations for CLEF 2015 and DCI for CLEF 2016, regard- less of correlation measure. These results resonate with those obtained for the category of raw components of readability for- mulas. In fact, the polysyllable words measure, which is the main feature used in SMOG, had the highest correlation for CLEF 2015 among methods in this category. Similarly, the number of difficult words, which is the main feature used in DCI, had the highest correlation for CLEF 2016 among meth- ods in this category.

When examining the expert vocabulary category, we found that the number of MeSH concepts obtained the highest corre- lations with human assessments; however, its correlations were substantially lower than those achieved by the best method from the consumer medical vocabularies category, i.e. the scores of CHV concepts. For the natural language category, we found that the number of pronouns, the number of stop words and the number of out of vocabulary words had the highest correlations – and these were even higher than those obtained with MeSH and CHV based methods. In turn, the methods that obtained the highest correlations among the HTML cate- gory (counts of P tags and list tags) exhibited overall the low- est correlations compared to methods in the other categories. P tags are used to create paragraphs in a Web page, being thus a rough proxy for text length. Among methods in the word frequency category, the use of Medical Reddit (but also of PubMed) showed the highest correlations, and these were comparable to those obtained by the readability formulas.

Finally, regressors and classifiers exhibited the highest cor- relations amongst all methods: in this category, the eXtreme Gradient Boosting (XGB) regressor and the multinomial Naive Bayes best correlated with human assessments.

## Evaluation of Preprocessing Pipelines and Heuristics

Results from experiments with different preprocessing pipelines and heuristics are shown in Figure 4 (top: CLEF 2015; bottom: CLEF 2016). For each category of methods and combination of preprocessing and heuristics, we report their variability in terms of Spearman rank correlation with human assessments. Results for Pearson and Kendall correlations are

**Table 3.** Learning to rank (LTR) settings.

|  |  |  |
| --- | --- | --- |
| Name | Feature Set | Labeling Function |
| LTR1 | IR features | *F* (*R, U* ) = *R* |
| LTR2 | IR + Unders. features | *F* (*R, U* ) = *R* |
| LTR3 | IR + Unders. features |  |

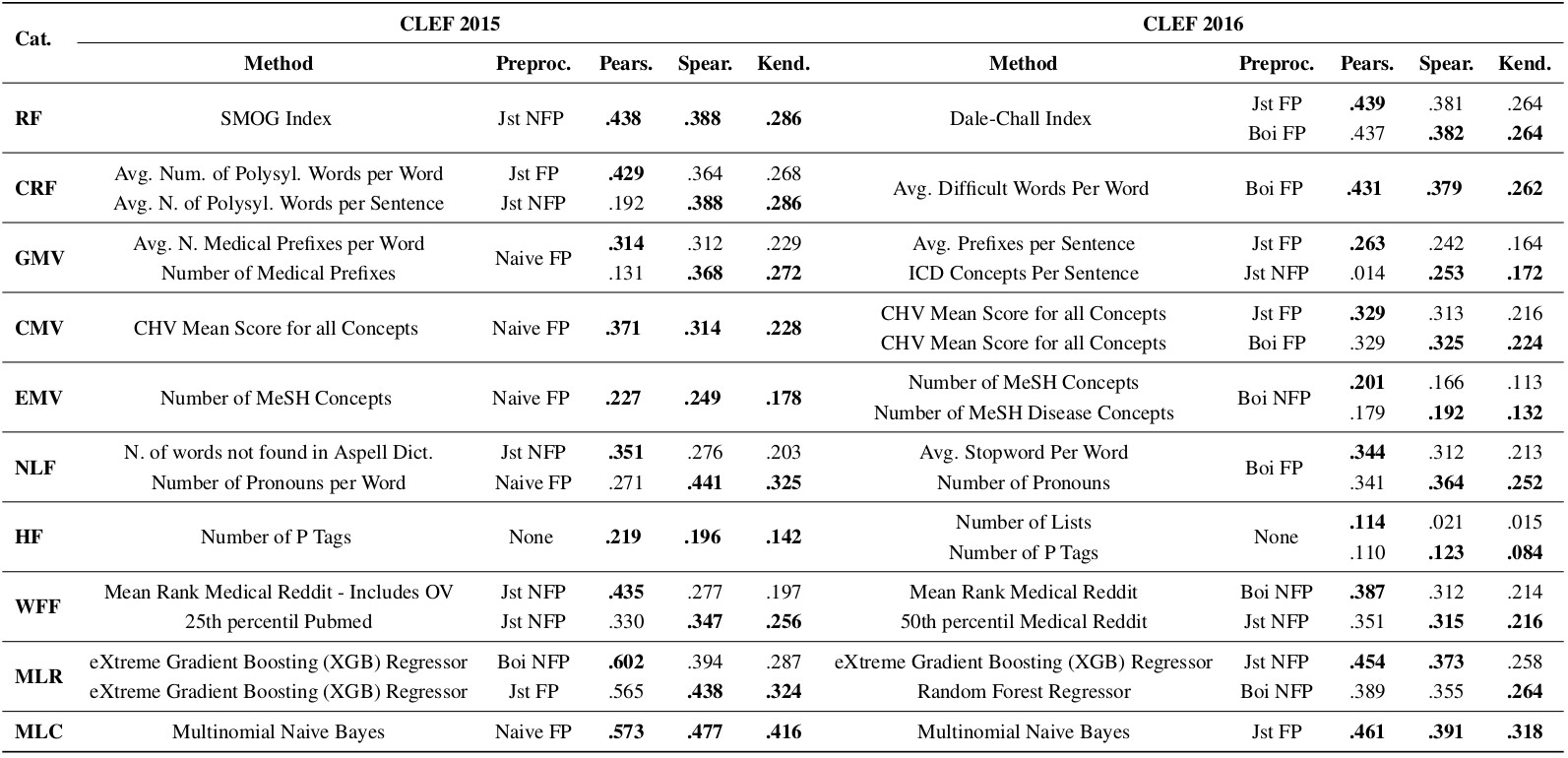
*F* (*R, U* ) = *R ×* (100 *− U* )*/*100



LTR4 IR + Unders. features *F* (*R, U* ) = *R* if *U ≤* 40

0 otherwise

LTR5 IR + Unders. features *F* (*R, U* ) = 2 *× R* if *U ≤* 40



*R*

otherwise

**Figure 3.** Methods with the highest correlation per category. Bold is used to highlight the best result of each group.

reported in the appendix, but showed similar trends. We fur- ther report the summary results across all understandability as- sessment methods and sentence-ending heuristics for each of the preprocessing pipelines. Finally, we also report the inter- assessor correlation (last box) when multiple assessors pro- vided judgements about the understandability of Web pages. This provides an indication of the range of variability and sub- jectiveness when assessing understandability, along with the highest correlation we measured between human assessors.

We first examined the correlations between human assess- ments and readability formulas. We found that the *Naive* pre- processing resulted in the lowest correlations, regardless of readability formula and heuristic (although *DoNotForcePeriod* performed better than *ForcePeriod*). Using Justext or Boiler- plate resulted in higher correlations with human understand- ability assessments, and the *ForcePeriod* heuristic was shown to be better than *DoNotForcePeriod*. These results confirm the speculations of Palotti et al. [16] : they found these settings to

produce lower variances in understandability estimations and thus hypothesised that they were better suited to the task.

Overall, among readability formulas, the best results (high- est correlations) were obtained by SMOG and DCI (see also Table 4). Although no single setting outperformed the others in both collections, we found that the use of CLI and FRE with *Justext* provided the most stable results across the collections, with correlations as high as the best ones in both collections. These results confirmed the advice put forward by Palotti et al. [16], i.e. in general, if using readability measures, then CLI is to be preferred, along with an appropriate HTML extraction pipeline, regardless of the heuristic for sentence ending. We provide detailed plots to compare our results with Palotti’s in the appendix.

When considering methods beyond those based on read- ability formulas, we found that the highest correlations were achieved by the regressors (MLR) and classifiers (MLC), in- dependently of the preprocessing method used. There is little

**Table 4.** Methods with the highest correlation per category. Bold is used to highlight the best result of each group.

Cat. CLEF 2015 CLEF 2016

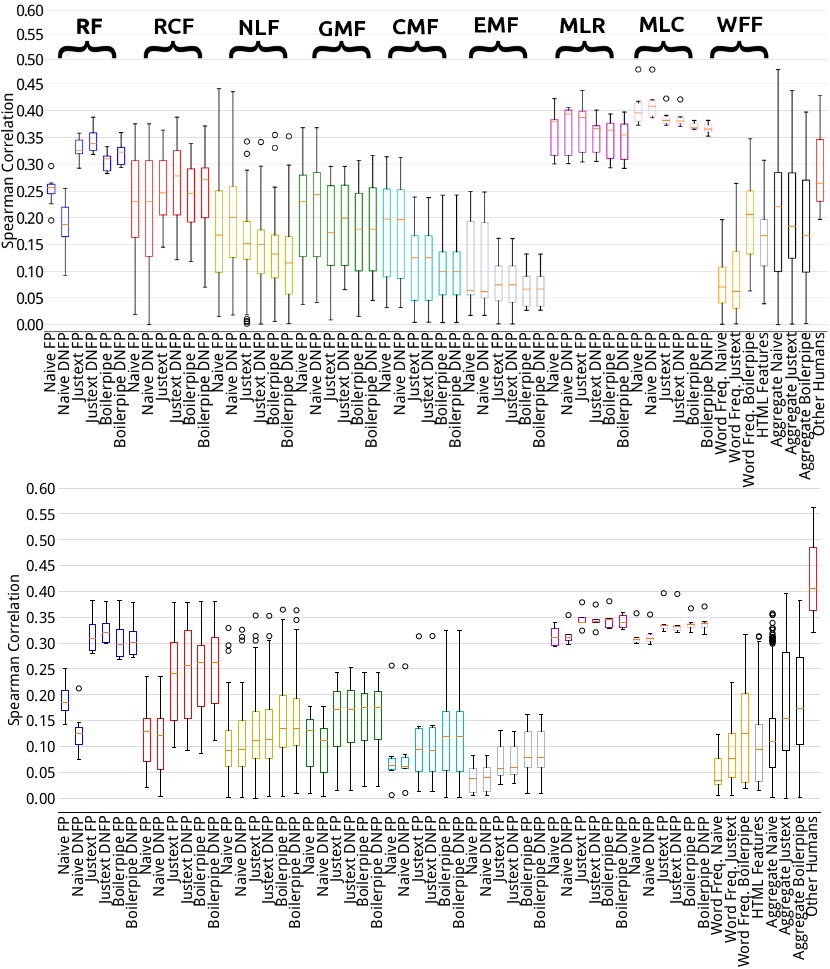
Method Pre-

proc.

Pears. Spear. Kend. Method Pre- proc.

Pears. Spear. Kend.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **RF** | SMOG Index | Jst  NFP | **.438** | **.388** | **.286** | Dale-Chall Index | Jst  FP | **.439** | .381 | **.264** |
|  |  |  |  |  |  |  | Boi  FP | .437 | **.382** | **.264** |
| **CRF** | Avg. Num. of Polysyl.  Words per Word | Jst  FP | .**429** | .364 | **.286** | Avg. Difficult Words Per  Word | Boi  FP | **.431** | **.379** | .**262** |
|  | Avg. N. of Polysyl. Words  per Sentence | Jst  NFP | .192 | **.388** | **.286** |  |  |  |  |  |
| **GMV** | Avg. N. Medical Prefixes  per Word | Naive  FP | **.314** | .312 | .229 | Avg. Prefixes per Sentence | Jst  FP | .**263** | .242 | .164 |
|  | Number of Medical Prefixes |  | .131 | **.368** | **.272** | ICD Concepts Per Sentence | Jst  NFP | .014 | **.253** | **.172** |
| **CMV** | CHV Mean Score for all  Concepts | Naive  FP | **.371** | .**314** | **.228** | CHV Mean Score for all  Concepts | Jst  FP | .**329** | .313 | .216 |
|  |  |  |  |  |  | CHV Mean Score for all  Concepts | Boi  FP | **.329** | **.325** | **.224** |
| **EMV** | Number of MeSH Concepts | Naive  FP | .227 | .249 | .178 | Number of MeSH Concepts | Boi  NFP | .**201** | .166 | .113 |
|  |  |  |  |  |  | Number of MeSH Disease  Concepts | Boi  NFP | .179 | .**192** | **.132** |
| **NLF** | N. of words not found in  Aspell Dict. | Jst  NFP | **.351** | .276 | .203 | Avg. Stopword Per Word | Boi  FP | **.344** | .312 | .213 |
|  | Number of Pronouns per  Word | Naive  FP | .271 | **.441** | **.325** | Number of Pronouns | Boi  FP | .341 | **.364** | **.252** |
| **HF** | Number of P Tags | None | **.219** | **.196** | **.142** | Number of Lists | None | **.114** | .021 | .015 |
|  |  |  |  |  |  | Number of P Tags | None | .110 | **.123** | **.084** |
| **WFF** | Mean Rank Medical Reddit  - Includes OV | Jst  NFP | **.435** | .277 | .197 | Mean Rank Medical Reddit | Boi  NFP | **.387** | .312 | .214 |
|  | 25th percentil Pubmed | Jst  NFP | .330 | **.347** | **.256** | 50th percentil Medical  Reddit | Jst  NFP | .351 | **.315** | .**216** |
| **MLR** | eXtreme Gradient Boosting  (XGB) Regressor | Boi  NFP | **.602** | .394 | .287 | eXtreme Gradient Boosting  (XGB) Regressor | Jst  NFP | **.454** | **.373** | .258 |
|  | eXtreme Gradient Boosting  (XGB) Regressor | Jst  FP | .565 | **.438** | **.324** | Random Forest Regressor | Boi  NFP | .389 | .355 | **.264** |
| **MLC** | Multinomial Naive Bayes | Naive  FP | **.573** | **.477** | .**416** | Multinomial Naive Bayes | Jst  FP | **.461** | **.391** | **.318** |



**Figure 4.** Correlations between understandability estimators and human assessments for CLEF 2015 (top) and 2016 (bottom). For example, the first boxplot on the top represents the distribution of Spearman correlations with human assessments across all features in the category Readability Features (Table 1), obtained with the *Naive ForcePeriod* preprocessing, for CLEF 2015. Each box extends from the lower to the upper quartile values, with the red marker representing the median value for that category. Whiskers show the range of the data in each category and circles represent values considered outliers for the category (e.g., Spearman correlation for SMOG index was 0.296 and for ARI was 0.194: these were outliers for that category.)

difference in terms of effectiveness of methods in these cate- gories, with the exception of regressors on CLEF 2015 that ex- hibited not negligible variances: while for the Neural Network Regressor the Pearson correlation was 0.44, for the Support Vector Regressor it was only 0.30.

A common trend when comparing preprocessing pipelines is that the Naive pipeline provided the weakest correlations with human assessments for CLEF 2016, regardless of estimation methods and heuristics. This result, however, was not con- firmed for CLEF 2015, where the Naive preprocessing nega- tively influenced correlations for the readability formula cate- gory (RF), but not for other categories, although it was gener- ally associated with larger variances for the correlation coeffi- cients.

## Evaluation of Understandability Retrieval

Results for the considered retrieval methods are reported in Ta- ble 5. We report only the results for CLEF 2016 for brevity; those for CLEF 2015 exhibited similar trends and are included in the appendix. As both the RBP residuals and the column *Unj* quantify the effect unassessed documents have on evaluation, we opt to show the RBP residuals only in the supplementary appendix and show *Unj* here, as its interpretation is straightfor- ward: *Unj* is the average number of documents without assess- ment in the top 10 results. Larger values of *Unj* entail larger room for improvements. Here we also show the values for the condensed measures.

In Table 5 , statistical significance was measured with re- spect to the best CLEF 2016 run, GUIR (p-values indicated with *Pg* ), and the BM25 baseline (p-values indicated with *Pbl*). A two-tailed paired t-test was used to compute statistical sig- nificance; note that the baseline BM25 is significantly worse than GUIR across all measures.

The effectiveness of the top two submissions to CLEF 2016 and the BM25 baseline are reported at indices 1-3 of Table 5. In turn, we report the results of each sub-experiment: *Simple re-ranking* (indices 4-21), *Fusion Experiments* (indices 22-30), *Learning to rank* (indices 31-35).

***Simple Re-ranking:*** Indices 4-12 of Table 5 report the re- sults of re-ranking methods applied to the runs listed at indices 1-3. Re-ranking was applied based on the DCI score of each document calculated using the preprocessing combination of Boilerpipe and ForcePeriod (best according to Pearson correlation, from Table 4). We found that the relevance

of the re-ranked runs (as measured by *RBPr* and *RBPr*∗)

significantly decreased, compared to the original runs: e.g., re-ranking the top 15 search results using DCI made *RBPr* decreasing from 25.28 to 21.58. However, as expected, these re-ranked results were significantly more understandable: for

the previous example, *RBPu* passed from 42.08 to 47.09.

In the experiments, we also studied the influence of the num- ber of documents considered for re-ranking (cut-off). Indices 4-6 refer to re-ranking only the top *k* = 15 documents from the original runs; 7-9 refer to the first *k* = 20; and 10-12 to the first *k* = 50. The results show that the more documents are considered for re-ranking, the more degradation in *RBPr* effectiveness. Considering understandability-only in the eval- uation shows mixed results. Similar trends were observed for evaluation measures that consider understandability (*RBP* and *RBPu*), however with some exceptions. For example, an in- crease in *uRBP* was observed when re-ranking ECNU using the top 50 results.

Note that with the increase of the number of documents considered for re-ranking, there is an increase in the number of unassessed documents being considered by the evaluation measures. Nevertheless, we note that if unassessed documents are excluded from the evaluation, similar trends are observed, e.g., compare findings with those for the condensed measures *uRBP* ∗, *RBPr*∗, *RBPu*∗ and *HR*∗ *BP* .

Indices 13-21 refer to using the XGB regressor trained with all features listed in Table 1 to estimate understandability. Sim- ilarly to when using DCI, as the cut-off increased, e.g., from *k* = 15to *k* = 50, the documents returned were more under- standable but less relevant. For the same cut-off value, e.g., *k* = 15, the machine learning method used for estimating understandability consistently yielded more understandable re- sults than DCI (higher *RBPu* and *RBPu*∗).

Overall, statistically significant improvements over the base- lines were observed for most configurations and measures.

***Rank Fusion:*** Next, we report the results of automatically combining topical relevance and understandability through rank fusion (indices 22 to 30). We used the XGB method for estimating understandability, as it was the one yielding highest effectiveness for the re-ranking method. Runs were thus pro- duced by fusing the re-ranking with XGB and the original run. (Results for DCI are reported in the appendix and confirm the superiority of XGB.)

Like as for re-ranking, also for the rank fusion approaches we found that, in general, higher cut-offs were associated to higher effectiveness in terms of understandability measures on one hand, but higher losses in terms of relevance-oriented mea- sures on the other. Overall, results obtained with rank fusion were superior to those obtained with re-ranking only, though most differences were not statistically significant. Statistically significant improvements over the baselines were instead ob- served for most configurations and measures.

***Learning to Rank:*** Last, we analyse the results obtained by the learning to rank methods (indices 31-35). Unlike with the

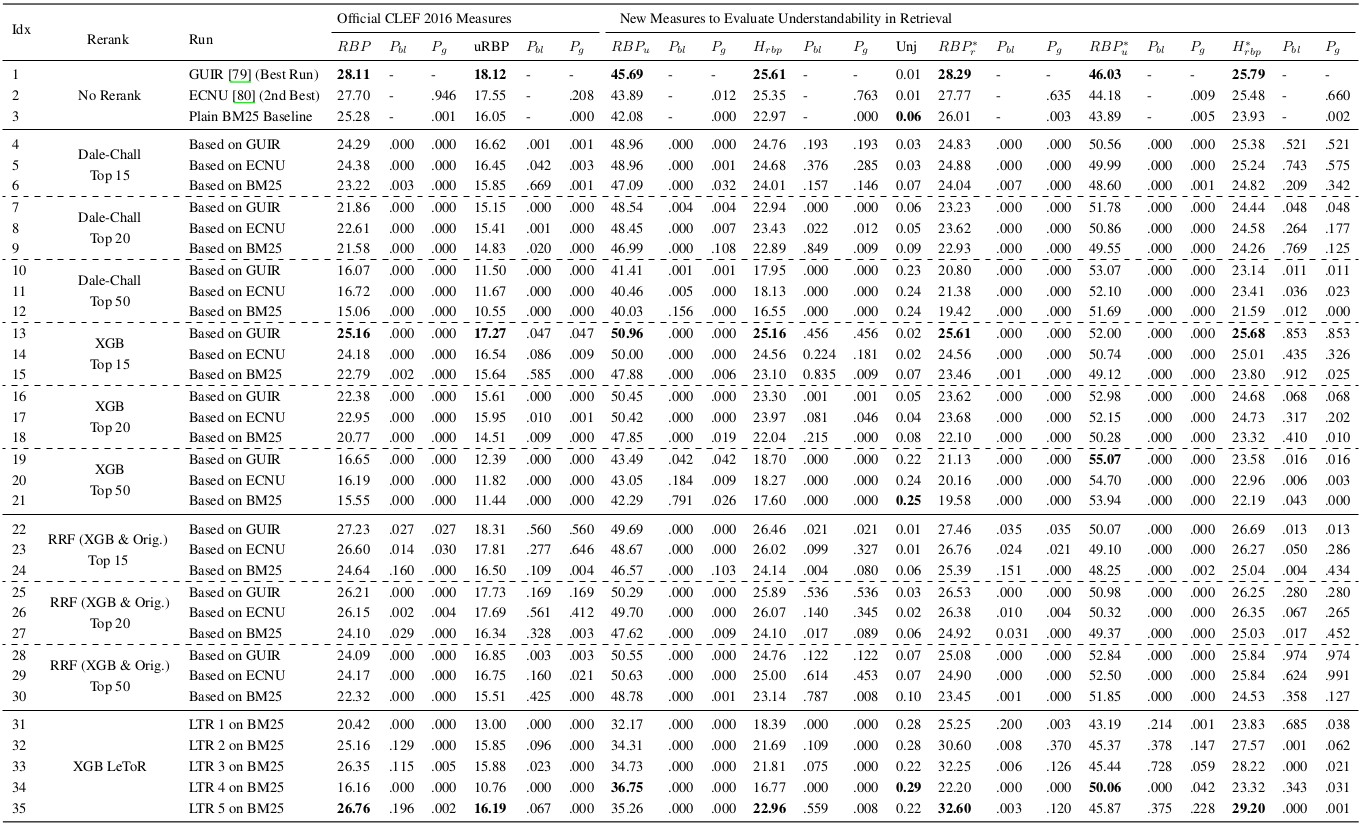
**Table 5.** Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1-3 (the index column is labelled Idx). Re-ranking experiments are reported at indices 4-21. Fusion experiments are reported at indices 22-30. Learning to rank experiments are reported at indices 31-35. All measures were calculated up to rank n = 10. P-values are rounded to the third decimal digit and reported w.r.t GUIR run (*Pg* ) and the BM25 baseline (*Pbl*). The highest result of each set of experiments is reported in bold face.

previous methods, we did not impose a rank cut-off on learn- ing to rank. Learning to rank was only applied to the BM25 baseline, as we had no access to the IR features for the runs submitted at CLEF (i.e. GUIR and ECNU for CLEF 2016).

When considering *RBPr* and *uRBP* , learning to rank ex- hibited effectiveness that was significantly inferior to that of the GUIR and ECNU baseline runs, though higher than those for the BM25 baseline (for some configurations). The examination of the number of unassessed documents (and the RBP residu- als, see appendix) revealed that this may have been because measures were affected by the large number of unassessed doc- uments retrieved in the top 10 ranks. For example, the *RBPr*

residual for learning to rank methods was about double than that of the baselines or other approaches (see appendix). In fact, among the documents retrieved in the top 10 results by learning to rank, there were 20% that were unassessed, com- pared to an average of 3% for the other methods. (Excluding XGB with cut-off 50, which also exhibited high residuals).

We thus should carefully account for unassessed documents through considering the residuals of RBP measures, as well as the condensed measures. When this was done, we observed that learning to rank methods overall provided substantial gains over the original runs and other methods (when considering *RBPr*∗, *RBPu*∗ and *HR*∗ *BP* ), or large potential gains over these



**Figure 5.** Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1-3 (the index column is labelled Idx). Re-ranking experiments are reported at indices 4-21. Fusion experiments are reported at indices 22-30. Learning to rank experiments are reported at indices 31-35. All measures were calculated up to rank n = 10. P-values are rounded to the third decimal digit and reported w.r.t GUIR run (*Pg* ) and the BM25 baseline (*Pbl*). The highest result of each set of experiments is reported in bold face. **TODO: update references**

methods (when considering the residuals). Next, we analyse these results in more detail.

No improvements over the baselines were found for LTR

only documents that were easy-to-read (understandability la- bel *≤ U* ), LTR 5 considered all documents, but boosted the relevance score. LTR 4 reached the highest understandability

1 (index 31), and the high residuals for *RBP* \_*r* were not

score for the learning-to-rank approaches (*RBPu*∗

= 50*.*06),

matched by other residuals or by considering only assessed documents (see appendix). LTR 1 was a simple method that used only IR features and was trained only on topical rele- vance. Specifically, we devised 24 IR features using the Ter- rier framework. The score of various retrieval models was ex- tracted from a multi-field index composed of title, body and whole document. Although simple, this is a typical learning to rank setting.

Compared to LTR 1, LTR 2 (index 32) included the un- derstandability features listed in Table 1. This inclusion was as beneficial to the understandability measures as to the rele- vance measures, with *RBPr*∗, *RBPu*∗ and *HR*∗ *BP* all showing

but it failed to retrieve a substantial number of relevant docu- ments (*RBPr*∗ = 22*.*20). In turn, LTR 5 reached the highest understandability-relevance trade-off (*HR*∗ *BP* = 29*.*20). Com- pared to the BM25 baseline (on which it was based), LTR 5

largely increased both relevance (*RBPr*∗ from 26.01 to 32.60

- a 25% increase, *Pbl* = 0*.*003) and understandability (*RBPu*∗ from 43.89 to 45.87 - a 4% increase, *Pbl <* 0*.*000). Note that LTR 5 was also significantly better than the best run submitted to CLEF 2016 for both *RBPr*∗ (15% increase, *Pg* = 0*.*120) and *HR*∗ *BP* (13% increase, *Pg* = 0*.*001).

# *Discussion*

gains over the baselines. LTR 3 obtained similar *HR*∗ *BP* val-

ues, though with higher effectiveness for relevance measures (*RBPr*∗) than for understandability (*RBPu*∗).

LTRs 4 and 5 were devised based on a set understandabil- ity threshold *U* = 40. While LTR 4 took into consideration

## Principal Findings

The empirical experiments suggested that:

1. Machine learning methods based on regression are best

suited to estimate the understandability of health Web pages;

1. Preprocessing does affect effectiveness (both for under- standability prediction and document retrieval), although, compared to other methods, ML-based methods for un- derstandability estimation are less subject to variability caused by poor preprocessing;
2. Learning to rank methods can be specifically trained to promote more understandable search results, while still providing an effective trade-off with topical relevance.

## Limitations

In this work, we relied on data collected through the CLEF 2015 and CLEF 2016 evaluation efforts to evaluate the effec- tiveness of methods that estimate the understandability of the Web pages. These assessments were obtained by asking rat-

ple seeking health information. We found that machine learn- ing methods are better suited than traditionally employed read- ability measures for assessing the understandability of health related web pages and that learning to rank is the most effec- tive strategy to integrate this into retrieval. We also found that HTML and text pre-processing do affect the effectiveness of both understandability estimations and of the retrieval process, although machine learning methods are less sensitive to this issue.

This article contributes to improving search engines tailored to consumer health search because it thoroughly investigates promises and pitfalls of understandability estimations and their integration into retrieval methods. The article further high- lights which methods and settings should be used to provide better search results to health information seekers. As shown in Figure 1, these methods would clearly improve current health- focused search engines.

# *References*

ings to medical experts and practitioners; although they were

asked to estimate the understandability of the content as if they were the patients they treat, there may have been noise and imprecisions in the collection mechanism due to the subjectiv- ity of the task. Figure 4 highlights this by showing that the agreement between assessors is relatively low. A better setting may have been to directly recruit health consumers: the task would still have been subjective, but would have captured real ratings, rather than inferred or perceived ratings. Despite this, our previous work has shown that no substantial differences were found in the downstream evaluation of retrieval systems, when we acquired understandability assessments from health consumers for a subset of the CLEF 2015 collection [41].

Relevance assessments on the CLEF 2015 and 2016 col- lections are incomplete [39, 40], i.e. not all top ranked web pages retrieved by the investigated methods have an explicit relevance assessment. This is often the case in information re- trieval, where the validity of experiments based on incomplete assessments has been thoroughly investigated [82]. Nonethe- less, we carefully controlled for the impact unassessed docu- ments had in our experiments by measuring their number and using measures like RBP that account for residuals and con- densed variants. The residuals analysis has been reported in the appendix.

## Conclusions

We have examined approaches to estimate the understandabil- ity of health Web pages, including the impact of HTML prepro- cessing techniques, and how to integrate these within retrieval methods to provide more understandable search results for peo-

1. Zhang Y, Zhang J, Lease M, Gwizdka J. Multidimen- sional relevance modeling via psychometrics and crowd- sourcing. In: Proceedings of the 37th international ACM SIGIR conference on Research & development in informa- tion retrieval; 2014. p. 435–444. Available from: 10.1145/ 2600428.2609577.
2. White RW, Horvitz E. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. ACM Transactions on Information Systems. 2009 November;27(4):23:1–23:37. Available from: 10.1145/1629096.1629101.
3. White R. Beliefs and biases in web search. In: Proceed- ings of the 36th international ACM SIGIR conference on Research and development in information retrieval. New York, NY, USA: ACM; 2013. p. 3–12. Available from: 10.1145/2484028.2484053.
4. Graber MA, Roller CM, Kaeble B. Readability levels of patient education material on the World Wide Web. Jour- nal of Family Practice. 1999;48(1):58–59. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/9934385.](http://www.ncbi.nlm.nih.gov/pubmed/9934385)
5. Fitzsimmons PR, Michael BD, Hulley JL, Scott GO. A readability assessment of online Parkinson’s disease in- formation. The journal of the Royal College of Physi- cians of Edinburgh. 2010;40(4):292–296. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/21132132.](http://www.ncbi.nlm.nih.gov/pubmed/21132132)
6. Wiener RC, Wiener-Pla R. Literacy, pregnancy and poten- tial oral health changes: The internet and readability levels.

Maternal and child health journal. 2014;18(3):657–662. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/](http://www.ncbi.nlm.nih.gov/pubmed/) 23784613.

1. Patel CR, Cherla DV, Sanghvi S, Baredes S, Eloy JA. Readability assessment of online thyroid surgery patient education materials. Head & neck. 2013;35(10):1421– 1425. Available from: [https://www.ncbi.nlm.nih.go](http://www.ncbi.nlm.nih.gov/)v/ pubmed/22972634.
2. Meillier A, Patel S. Readability of Healthcare Litera- ture for Gastroparesis and Evaluation of Medical Termi- nology in Reading Difficulty. Gastroenterology Research. 2017;10(1):1–5. Available from: [https://www.ncbi.nlm.](http://www.ncbi.nlm/) nih.gov/pubmed/28270870.
3. Ellimoottil C, Polcari A, Kadlec A, Gupta G. Read- ability of websites containing information about prostate cancer treatment options. The Journal of urology. 2012;188(6):2171–2176. Available from: [https://www](http://www/). ncbi.nlm.nih.gov/pubmed/23083852.
4. Evgeniy G. Cura Te Ipsum: answering symptom queries with question intent. In: Second WebQA workshop, SI- GIR 2016 (invited talk); 2016. Available from: [http://plg2.](http://plg2/) cs.uwaterloo.ca/avtyurin/WebQA2016/.
5. Cowan CF. Teaching patients with low literacy skills. Jones & Bartlett Learning; 2004.
6. Wallace LS, Lennon ES. American Academy of Family Physicians patient education materials: can patients read them? Family medicine. 2004;36(8):571–574. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/15343418.](http://www.ncbi.nlm.nih.gov/pubmed/15343418)
7. Davis TC, Wolf MS. Health literacy: implications for family medicine. Family Medicine. 2004;36(8):595–598. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/](http://www.ncbi.nlm.nih.gov/pubmed/) 15343422.
8. Stossel LM, Segar N, Gliatto P, Fallar R, Karani R. Readability of patient education materials available at the point of care. Journal of general internal medicine. 2012;27(9):1165–1170. Available from: [https://www](http://www/). ncbi.nlm.nih.gov/pubmed/22528620.
9. National CI. Clear & Simple: Developing Effective Print Materials for Low-literate Readers. National Institutes of Health; Accessed: 2017-09. Available from: [https://www](http://www.nih.gov/institutes-nih/nih-office-).nih.go[v/institutes-nih/nih-office-](http://www.nih.gov/institutes-nih/nih-office-) director/office-communications-public-liaison/clear- communication/clear-simple.
10. Palotti J, Zuccon G, Hanbury A. The Influence of Pre- processing on the Estimation of Readability of Web Docu- ments. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management. New York, NY, USA: ACM; 2015. p. 1763–1766. Avail- able from: 10.1145/2806416.2806613.
11. Palotti J, Goeuriot L, Zuccon G, Hanbury A. Ranking health web pages with relevance and understandability. In: Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval; 2016. p. 965–968. Available from: 10.1145/ 2911451.2914741.
12. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient educa- tion and counseling. 2014;96(3):395–403. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/24973195.](http://www.ncbi.nlm.nih.gov/pubmed/24973195)
13. Becker SA. A study of web usability for older adults seeking online health resources. ACM Transactions on Computer-Human Interaction (TOCHI). 2004;11(4):387– 406. Available from: 10.1145/1035575.1035578.
14. Zheng J, Yu H. Readability formulas and user perceptions of electronic health records difficulty: a corpus study. Jour- nal of medical Internet research. 2017;19(3). Available from: 10.2196/jmir.6962.
15. Coleman M, Liau TL. A Computer Readability Formula Designed for Machine Scoring. Journal of Applied Psy- chology. 1975;Available from: 10.1037/h0076540.
16. Edgar D, Jeanne SC. A Formula for Predicting Read- ability: Instructions. Educational Research Bulletin. 1948;27(2):37–54. Available from: <http://www.jstor.org/> stable/1473669.
17. Kincaid J, Fishburne R, Rogers R, Chissom B. Derivation of New Readability Formulas for Navy Enlisted Personnel. National Technical Information Service; 1975. Available from: [http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf.](http://www.dtic.mil/dtic/tr/fulltext/u2/a006655.pdf)
18. William HD. The Principles of Readability. Costa Mesa, CA: Impact Information. 2004;Available from: 10.1.1.91. 4042.
19. Liu X, Croft WB, Oh P, Hart D. Automatic Recogni- tion of Reading Levels from User Queries. In: Pro- ceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Informa- tion Retrieval. ACM; 2004. p. 548–549. Available from: 10.1145/1008992.1009114.
20. Collins-Thompson K, Callan J. Predicting reading dif- ficulty with statistical language models. Journal of the Association for Information Science and Technology. 2005;56(13):1448–1462. Available from: 10.1002/asi. 20243.
21. Heilman M, Collins-Thompson K, Callan J, Eskenazi M. Combining lexical and grammatical features to improve readability measures for first and second language texts. In: Human Language Technologies 2007: The Confer- ence of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference; 2007. p. 460–467. Available from: 10.1.1.70. 1391.
22. Pitler E, Nenkova A. Revisiting readability: A unified framework for predicting text quality. In: Proceedings of the conference on empirical methods in natural lan- guage processing; 2008. p. 186–195. Available from: [http://dl.acm.org/citation.cfm?id=1613715.1613742.](http://dl.acm.org/citation.cfm?id=1613715.1613742)
23. Zeng Q, Kim E, Crowell J, Tse T. A text corpora-based es- timation of the familiarity of health terminology. Biolog- ical and Medical Data Analysis. 2005;p. 184–192. Avail- able from: 10.1007/11573067\\_19.
24. Zeng-Treitler Q, Goryachev S, Tse T, Keselman A, Boxwala A. Estimating consumer familiarity with health terminology: a context-based approach. Jour- nal of the American Medical Informatics Association. 2008;15(3):349–356. Available from: https://www.ncbi. nlm.nih.gov/pmc/articles/PMC2409994/.
25. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. Journal of the American Medical In- formatics Association. 2006;13(1):24–29. Available from: [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380193.](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1380193)
26. Leroy G, Helmreich S, Cowie JR, Miller T, Zheng W. Evaluating online health information: Beyond readabil- ity formulas. In: AMIA Annual Symposium Proceed- ings. vol. 2008; 2008. p. 394. Available from: https:

[//www.ncbi.nlm.nih.gov/pmc/articles/PMC2656067.](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2656067)

1. Joao P, Allan H, Henning M. Exploiting Health Related Features to Infer User Expertise in the Medical Domain. In: Proceedings of WSCD Workshop on Web Search and Data Mining. John Wiley & Sons, Inc.; 2014. Available from: <http://publications.hevs.ch/index.php/publications/> show/1632.
2. Yan X, Lau RYK, Song D, Li X, Ma J. Toward a seman- tic granularity model for domain-specific information re- trieval. ACM Transactions on Information Systems. 2011

July;29(3):15:1–15:46. Available from: 10.1145/1993036. 1993039.

1. Kim H, Goryachev S, Rosemblat G, Browne A, Kesel- man A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. In: AMIA Annual Symposium Proceedings. vol. 2007; 2007.

p. 418. Available from: [https://www.ncbi.nlm.nih.go](http://www.ncbi.nlm.nih.gov/)v/ pmc/articles/PMC2655856.

1. van Doorn J, Odijk D, Roijers DM, de Rijke M. Bal- ancing relevance criteria through multi-objective optimiza- tion. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in In- formation Retrieval; 2016. p. 769–772. Available from: 10.1145/2911451.2914708.
2. Guido Z, Bevan K. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In: MedIR; 2014. Available from: https://eprints.qut.edu.au/ 72854/.
3. Zuccon G. Understandability biased evaluation for in- formation retrieval. In: European Conference on Infor- mation Retrieval; 2016. p. 280–292. Available from: 10.1007/978-3-319-30671-1\\_21.
4. João P, Guido Z, Lorraine G, Liadh K, Allan H, Gareth JFJ, et al. ShARe/CLEF eHealth Evaluation Lab 2015, Task 2: User-centred Health Information Retrieval. In: Working Notes for CLEF 2015 Conference, Toulouse, France, September 8-11, 2015.; 2015. Available from: ceur-ws.org/Vol-1391/inv-pap9-CR.pdf.
5. Zuccon G, Palotti J, Goeuriot L, Kelly L, Lupu M, Pecina P, et al. The IR Task at the CLEF eHealth evaluation lab 2016: user-centred health information retrieval. In: CLEF 2016-Conference and Labs of the Evaluation Forum. vol. 1609; 2016. p. 15–27. Available from: <http://ceur-ws.org/> Vol-1609/16090015.pdf.
6. Joao P, Guido Z, Johannes B, Allan H, Lorraine G. As- sessors Agreement: A Case Study across Assessor Type, Payment Levels, Query Variations and Relevance Dimen- sions. In: Experimental IR Meets Multilinguality, Mul- timodality, and Interaction: 7th International Conference of the CLEF Association, CLEF’16 Proceedings. Springer International Publishing; 2016. Available from: 10.1007/ 978-3-319-44564-9\\_4.
7. Koopman B, Zuccon G. Relevation!: An open source sys- tem for information retrieval relevance assessment. In: Proceedings of the 37th international ACM SIGIR confer- ence on Research & development in information retrieval;

2014. p. 1243–1244. Available from: 10.1145/2600428.

2611175.

1. Sakai T. Alternatives to Bpref. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, NY, USA: ACM; 2007. p. 71–78. Avail- able from: [10.1145/1277741.1277756;http://doi.acm.or](http://doi.acm.org/)g/ 10.1145/1277741.1277756.
2. E AS, R JS. Automated Readability Index. Aerospace Medical Research Laboratories; 1967. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/5302480.](http://www.ncbi.nlm.nih.gov/pubmed/5302480)
3. Robert G. The Technique of Clear Writing. McGraw-Hill; 1952.
4. C HB. Readability of Newspapers in 11 Languages. Read- ing Research Quarterly. 1983;18(4):480–497. Available from: [http://www.jstor.org/stable/747382.](http://www.jstor.org/stable/747382)
5. G HM. SMOG Grading - a New Readability Formula. Journal of Reading. 1969;Available from: [https://www](http://www/). jstor.org/stable/40011226.
6. Collins-Thompson K. Computational assessment of text readability: A survey of current and future re- search. ITL-International Journal of Applied Linguistics. 2014;165(2):97–135. Available from: 10.1075/itl.165.2. 01col.
7. PyPhen. Python module to hyphenate text; 2017.
8. OpenMedSpel. OpenOffice Medical Dictionary Extension; 2017.
9. Zhou W, Torvik V, Smalheiser N. ADAM: Another Database of Abbreviations in MEDLINE. Bioinformatics. 2006;22(22):2813–2818. Available from: [https://www](http://www/). ncbi.nlm.nih.gov/pubmed/16982707.
10. Alan RA, François-Michel L. An overview of MetaMap: historical perspective and recent advances. JAMIA. 2010;17(3):229–236. Available from: 10.1136/jamia. 2009.002733.
11. Pang CI. Understanding Exploratory Search in Seeking Health Information; 2016. Available from: [http://hdl.](http://hdl/) handle.net/11343/115239.
12. Christopher A, Avi A. Augmenting Medical Queries with UMLS Concepts via MetaMap. In: Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016; 2016. Available from: https://trec.nist.gov/pubs/ trec25/papers/DUTH-CL.pdf.
13. Palotti J, Hanbury A, Müller H, Kahn CE. How users search and what they search for in the medical domain. Information Retrieval Journal. 2016 Apr;19(1):189–224. Available from: 10.1007/s10791-015-9269-8.
14. Yates A, Goharian N. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: European Conference on In- formation Retrieval; 2013. p. 816–819. Available from: 10.1007/978-3-642-36973-5\\_92.
15. NLTK V. Python Natural Language Toolkit Library; 2017.
16. GNU A. GNU English Dictionary Aspell; 2017.
17. Strohman T, Metzler D, Turtle H, Croft WB. Indri: A language model-based search engine for complex queries. In: Proceedings of the International Conference on In- telligent Analysis. vol. 2; 2005. p. 2–6. Available from: [http://ciir.cs.umass.edu/pubfiles/ir-407.pdf.](http://ciir.cs.umass.edu/pubfiles/ir-407.pdf)
18. Ounis I, Amati G, Plachouras V, He B, Macdonald C, Johnson. Terrier Information Retrieval Platform. In: Proceedings of the 27th European Conference on IR Re- search (ECIR 2005). vol. 3408. Springer; 2005. p. 517– 519. Available from: 10.1007/978-3-540-31865-1\\_37.
19. Feng L, Jansche M, Huenerfauth M, Elhadad N. A com- parison of features for automatic readability assessment. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters; 2010. p. 276–284.
20. Barzilay R, Lapata M. Modeling Local Coherence: An Entity-based Approach. Comput Linguist. 2008 March;34(1):1–34. Available from: 10.1162/coli.2008.34. 1.1.
21. BeautifulSoap V. BeautifulSoap; 2017.
22. Elhadad N. Comprehending technical texts: Predicting and defining unfamiliar terms. In: AMIA annual sympo- sium proceedings. vol. 2006; 2006. p. 239. Available from: [https://www.ncbi.nlm.nih.gov/pubmed/17238339.](http://www.ncbi.nlm.nih.gov/pubmed/17238339)
23. Wu DT, Hanauer DA, Mei Q, Clark PM, An LC, Proulx J, et al. Assessing the readability of ClinicalTrials.gov. Journal of the American Medical Informatics Association. 2016;23(2):269–275.
24. Reddit. Reddit Webstie; 2017.
25. Reddit. Reddit Ask A Doctor Community; 2017.
26. Python RAV. PRAW: The Python Reddit API Wrapper; 2017.
27. Wikimedia D. English Wikipedia Dumps; 2017.
28. Soldaini L, Cohan A, Yates A, Goharian N, Frieder O. Retrieving Medical Literature for Clinical Decision Sup- port. In: European Conference on Information Re- trieval. Springer International Publishing; 2015. p. 538– 549. Available from: 10.1007/978-3-319-16354-3\\_59.
29. PubMed C. National Center for Biotechnology Informa- tion PubMed Central; 2017.
30. Roberts K, Simpson M, Demner-Fushman D, Voorhees E, Hersh W. State-of-the-art in biomedical literature re- trieval for clinical cases: a survey of the TREC 2014 CDS track. Information Retrieval Journal. 2016;19(1):113–148. Available from: 10.1007/s10791-015-9259-x.
31. Kirk R, Matthew SS, Ellen MV, William RH. Overview of the TREC 2015 Clinical Decision Support Track. In: Proceedings of The Twenty-Fourth Text REtrieval Confer- ence, TREC 2015, Gaithersburg, Maryland, USA, Novem- ber 17-20, 2015; 2015. Available from: https://trec.nist. gov/pubs/trec24/papers/Overview-CL.pdf.
32. Kohlschütter C, Fankhauser P, Nejdl W. Boilerplate de- tection using shallow text features. In: Proceedings of the third ACM international conference on Web search and data mining; 2010. p. 441–450. Available from: 10.1145/1718487.1718542.
33. Jan P. Removing Boilerplate and Duplicate Content from Web Corpora; 2011. Available from: https://theses.cz/id/ nqo9nn/.
34. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22Nd ACM SIGKDD In- ternational Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2016. p. 785–794. Available from: 10.1145/2939672.2939785.
35. Cormack GV, Clarke CLA, Buettcher S. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In: Proceedings of the 32Nd Interna- tional ACM SIGIR Conference on Research and Devel- opment in Information Retrieval. New York, NY, USA: ACM; 2009. p. 758–759. Available from: 10.1145/ 1571941.1572114.
36. Yang S, Yun H, Qinmin H, Liang H, E MH. ECNU at 2015 eHealth Task 2: User-centred Health Information Re- trieval. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, Septem- ber 8-11, 2015.; 2015. Available from: <http://ceur-ws.org/> Vol-1391/80-CR.pdf.
37. Heung-Seon O, Yuchul J, Kwang-Young K. KISTI at CLEF eHealth 2015 Task 2. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.; 2015. Available from: ceur-ws.org/Vol-1391/17-CR.pdf.
38. Luca S, Will E, Nazli G. Team GU-IRLAB at CLEF eHealth 2016: Task 3. In: Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Por- tugal, 5-8 September, 2016.; 2016. p. 143–146. Available from: ceur-ws.org/Vol-1609/16090143.pdf.
39. Yang S, Yun H, Hongyu L, Yueyao W, Qinmin H, Liang

H. ECNU at 2016 eHealth Task 3: Patient-centred Infor- mation Retrieval. In: Working Notes of CLEF 2016 - Con- ference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016.; 2016. p. 157–161. Available from: [http://ceur-ws.org/Vol-1609/16090157.pdf.](http://ceur-ws.org/Vol-1609/16090157.pdf)

1. Sanderson M, et al. Test collection based evaluation of in- formation retrieval systems. Foundations and Trends®in Information Retrieval. 2010;4(4):247–375. Available from: 10.1561/1500000009.