A Study of Web Page Understandability for Consumer Health Search

Anonymous Author(s)

ABSTRACT

In this paper we investigate methods to estimate the understandability of health Web pages, and use these to improve the retrieval of information for people seeking health advice on the Web. 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.

Our investigation considers methods to automatically estimate the understandability of health information in Web pages, and it provides 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 are applied to the construction of retrieval methods with specific attention to retrieving information understandable by the general public.

We found that machine learning techniques are 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. Learning to rank effectively exploits these estimates to provide the general public with more understandable search results. These results are important for specialised search services tailored to support the general public in seeking health advice on the Web.

ACM Reference Format:
Anonymous Author(s). 2017. A Study of Web Page Understandability for Consumer Health Search. In *Proceedings of The Web Conference (WWW'18).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Search engines are concerned with retrieving relevant information to support a user's information seeking task. Commonly, signals about the topicality or aboutness of a piece of information with respect to a query are used to estimate relevance, with other relevance dimensions like understandability, trustworthiness, etc. [73] 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 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 the general public is one such task.

A key problem when searching the Web for health information is that this can be too technical, unreliable, generally misleading, and can lead to unfounded escalations and poor decisions [66]. Where correct information exists, it can be hard to find and digest amongst the noise, spam, technicalities, and irrelevant information. In *high-stakes search tasks* such as this, access to poor information can lead to poor decisions which ultimately can have a significant impact on our health and well-being [65, 66]. In this work we are specifically interested in the understandability of health information retrieved by search engines, and in improving search results to favour information understandable by the general public.

The use of general purpose Web search engines like Google, Bing and Baidu for seeking health advice has been largely analysed, questioned and criticised [4, 18, 19, 21, 32, 43, 67], despite the commendable efforts these services have put into providing increasingly better health information, e.g., the Google Health Cards [20].

Ad-hoc solutions to support the general public in searching and accessing health information on the Web have been implemented, typically supported by government initiatives or medical 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 information to the health-seeking general public? As an illustrative example, we analysed the top 10 search results retrieved by HON[1] in answer to 300 search queries from CLEF 2016 eHealth (see Section 3). Figure 1 reports the cumulative distribution of understandability scores for these search results (note, we did not assess their topical relevance). Understandability scores were computed with the most effective readability formula and settings from Section 6 (Dale-Chall Index). We report also the scores for the "optimal" search results (Oracle), as found from CLEF 2016 (relevant results that have the highest understandability scores), along with the scores for the best retrieval method from Section 8. 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 paper we propose and investigate methods for the estimation of the understandability of health information in Web pages. In doing so, we also study the influence of HTML processing methods on these estimations, and their pitfalls. Then, we investigate how understandability estimations can be integrated into retrieval methods to enhance the quality of the retrieved health information, with particular attention to its understandability by the general public. This paper makes a concrete contribution to practice, as it informs health search engines specifically tailored to the general public about the best methods they should adopt.

2 RELATED WORK

Understandability refers to the ease of comprehension of the information presented to a user. Put in other words, health information is understandable "when consumers of diverse backgrounds and

WWW'18, April 2018, Lyon, France
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] Results retrieved on 01/10/2017.

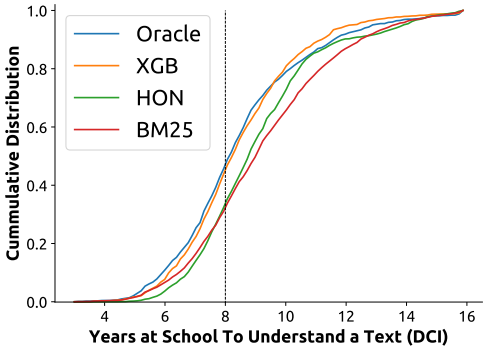

Figure 1: Distribution of Dale-Chall Index (DCI) of search results. DCI measures the years of schooling required to understand a document. The average US resident reads at or below an 8th grade level (dashed line)[13, 15, 57, 64], which is the level suggested by NIH for health information on the Web [59]. The distribution for HON is similar to that of the baseline used in this paper (BM25). Our best method (XGB) re-ranks documents to provide more understandable results; its distribution is similar to that of an "Oracle" system.

varying levels of health literacy can process and explain key messages" [51]. 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 understandability to refer to the broader concept of ease of understanding: this is affected by text readability[2], 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 understandability 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 designed for the elderlies [5], while Stossel et al. found that health education material on the Web is not written at an adequate reading level [57]. A common finding of these studies is that, in general, health content available on Web pages is often hard to understand by the general public; this includes content that is retrieved in top ranked positions by current commercial search engines [4, 19, 21, 32, 43, 67].

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 include (1) readability formulas, which generally exploit word surface characteristics of the text, (2) machine learning approaches, (3) matching with specialised dictionaries or terminologies, often compiled with information about understandability difficulty.

Measures such as Coleman-Liau Index [9], Dale-Chall Index [14] and Flesch Reading Easy [25] belong to the first category. These measures generally rely on surface-level characteristics of text, such as characters, syllables and word counts [16]. While these measures have been widely used in studies investigating the understandability of health content retrieved by search engines (e.g., [4, 5, 19, 21, 32, 43, 57, 67]), Palotti et al. found that these measures are heavily affected by the methods used to extract text from the HTML source [41]. They were able to identify specific settings of an HTML preprocessing pipeline that provided consistent estimates. We shall revisit this work in more details in Section 6, as we further investigate 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 understandability forms an alternative approach. Earlier research explored the use of statistical natural language and language modeling [11, 23, 30] as well as linguistic factors, such as syntactic features or lexical cohesion [44]. While we replicate 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) [70–72].

The actual use of CHV or other terminologies such as the Medical Subject Headings (MeSH) belongs to the third category of approaches. The CHV is a prominent medical vocabulary dedicated to mapping layperson vocabulary to technical terms [71]. 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 [28] and medical expertise prediction [37]. The hierarchy of MeSH was previously used in the literature to identify hard concepts, assuming that a concept deep in the hierarchy is harder than a shallow one [68]. Other approaches combined vocabularies with word surface characteristics and syntactic features, like part of speech, into a unique readability measure [24].

In this work, we investigate approaches to estimate understandability from each of these categories. We further extend Palotti et al.'s work to understand the influence of HTML preprocessing on automatic understandability 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. [36] have used learning to rank with standard retrieval features along with features based on readability measures and medical lexical aspects to determine understandability. Van Doorn et al. [63] have shown that learning a set of rankers that provide trade-offs across a number of relevance criteria, including readability/understandability, increases overall system effectiveness. Zuccon and Koopman [76], and later Zuccon [75], have proposed and investigated a family of measures 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 correlated, topical-relevance evaluation alone provides differing system rankings compared to understandability-biased evaluation measures. In this work we further explore the development of retrieval methods that combine signals about topical relevance and understandability.

[2]Increasing readability tends to improve understanding [29].

3 DATA AND RESOURCES

3.1 Data Collections

In this paper we investigate 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 topical relevance[3] and understandability assessments, we used the data from the CLEF 2015 and 2016 eHealth collections.

The CLEF 2015 collection contains 50 queries and 1,437 documents that have been assessed relevant by clinical experts and have an assessment for understandability [40]. Documents in this collection are a selected crawl of health websites, of which the majority are certified HON websites. The CLEF 2016 collection contains 300 queries and 3,298 relevant documents that also have been assessed with respect to understandability [77]. 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). Understandability assessments were provided on a 5-point Likert scale for CLEF 2015, and on a $[0, 100]$ range for CLEF 2016 (0 indicates highest understandability).

To support the investigation in Section 6 (evaluation of preprocessing pipelines and heuristics), we further considered correlations between multiple human assessors (inter-assessor agreement). For CLEF 2015, we used the publicly available additional assessments made by unpaid medical students and health consumers collected by Palotti et al. [39] in a study of how medical expertise affects assessments. For CLEF 2016 we collected understandability assessments for 100 documents. Three members of our research team, which did not author this paper, were recruited to provide the assessments. The Relevation tool [27] was used to assist with the assessments, mimicking the settings used in CLEF.

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 relation between two variables, while Kendall and Spearman 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.

3.2 Evaluation Measures

For the retrieval experiments in Section 8, we use evaluation measures that use both relevance and understandability. The uRBP measure [75] extends rank biased precision (RBP) to scenarios where multiple relevance dimensions are used. Formally, the measure is formulated as $uRBP(\rho) = (1-\rho)\sum_{k=1}^{K}\rho^{k-1}r(d@k)u(d@k)$, where $r(d@k)$ is the gain for retrieving a relevant document at rank k and $u(d@k)$ is the gain for retrieving a document of a certain understandability at rank k; ρ is the RBP persistence parameter. This measure was an official metric used in CLEF (we also set $\rho = 0.8$).

[3]We refer to this simply as relevance in the reminder of the paper, when this does not cause confusion.

A drawback of uRBP is that relevance and understandability 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 calculate an RBP value for relevance and another for understandability, and then combine them into a unique effectiveness measure:

- $RBP_r@n(\rho)$: uses the relevance assessments for the top n search results (i.e. this is the common RBP). In this paper, we regarded a document as topically relevant if assessed as somewhat relevant or highly relevant.
- $RBP_u@n(\rho)$: uses the understandability assessments for the top n search results. In this paper, we regarded a document as understandable (1) for CLEF 2015 if assessed easy or somewhat easy to understand; (2) for CLEF 2016 if its assessed understandability score was smaller than a threshold U (we used $U = 40$ [4]).
- $HRBP@n(\rho) = 2 \times \frac{RBP_r@n \times RBP_u@n}{RBP_r@n + RBP_u@n}$: combines the previous two RBP values into a unique measurement using the harmonic mean (in the same fashion that the F_1 measure combines recall and precision).

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 $RBP_r@10$).

Along with these measures of search effectiveness, we also reported the number of unassessed documents, the RBP residual, and $RBP_r@10^*$, $RBP_u@10^*$ and $HRBP^*$, i.e. the corresponding measures calculated by ignoring unassessed documents. We did this to minimise pool bias since the pools built in CLEF were of limited size, and the investigated methods retrieved a substantial number of unassessed documents.

3.3 Preprocessing Pipelines and Heuristics

We study the influence the preprocessing of Web pages have on the estimation of understandability when using the evaluated methods present in Section 4. We do 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 extend those by Palotti et al. [41], who 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 test: BeautifulSoup [61] (*Naive*), which just naively removes HTML tags, Boilerpipe [26] (*Boi*) and Justext [45] (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 sentences, depending on whether a period was forced at the end of fields/sentences. We thus implement the same two heuristics devised by Palotti et al. to deal with this: *ForcePeriod (FP)* and *DoNotForcePeriod (DNFP)*. The FP heuristic forces a period at the end of each extracted HTML field, while the DNFP does not.

[4]This choice for U is based on the distribution of understandability assessments. This distribution can be found in the online appendix.

3.4 Additional Resources

Because of space limitations, in this paper we only report a subset of the results; the remaining results (which show similar trends to those reported here) are made available in an online appendix for completeness: https://sites.google.com/view/www2018-sub341. All data and code will be shared on GitHub upon acceptance.

4 UNDERSTANDABILITY ESTIMATORS

As reviewed in Section 2, several methods have been used to estimate the understandability of health Web pages, with the most popular methods (at least in the biomedical literature) being readability formulas based on surface level characteristics of text. Next, we outline the categories of methods to estimate understandability used in this work; an overview is shown in Table 2. Some of these methods further expand measures used in the literature.

Traditional Readability Formulas (RF): These include the readability formulas mentioned in Section 2, as well as other, less popular ones. A full list is provided in surveys by Collins-Thompson [10] and Dubay [16].

Raw Components of Readability Formulas (CRF): These are formed by the "building blocks" used in the traditional readability formulas; examples of such building blocks include the average number of characters per word and the average number of syllables in a sentence.

General Medical Vocabularies (GMV): These include methods 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-), text strings included in lists of acronyms or in medical vocabularies such as the International Statistical Classification of Diseases and Related Health Problems (ICD), Drugbank and the OpenMedSpel dictionary [34]. An acronym list from the ADAM database [74] is used and methods in this group are matched with documents using simple keywords matching.

Consumer Medical Vocabulary (CMV): The popular MetaMap [2] tool is used to map the text content of Web pages to entries in CHV [71]. We use the MetaMap semantic types to retain only concepts identified as symptoms or diseases. Similar approaches have been commonly used in the literature (e.g., [1, 38, 42, 69]).

Expert Medical Vocabulary (EMV): Similarly to the CHV features, we use MetaMap to convert the content of Web pages into MeSH entities, studying symptoms and disease concepts separately.

Natural Language Features (NLF): These include commonly used natural language heuristics such as the ratio of part-of-speech (POS) classes, the sentiment polarity and the ratio of words found in English vocabularies. The Python package NLTK [60] is employed for sentiment analysis and POS tagging. The GNU Aspell [3] dictionary is used as a standard English vocabulary and a stop word list is built by merging those of Indri [58] and Terrier [35].

HTML Features (HF): These include the identification of a large number of HTML tags, which are extracted with the Python library BeautifulSoap [61]. The intuition for these features is that Web pages with many images and tables may explain and summarise health content better, thus providing more understandable content to the general public.

Word Frequency Features (WFF): Generally speaking, common and known words are usually frequent words, while unknown and obscure words are generally rare. This idea is implemented 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) [14]. We extend these observations by studying corpus-wide word frequencies. We model word frequencies in a corpus in a straightforward manner: we sort the word frequencies and normalize word rankings such that values close to 100 are attributed to common words and values close to 0 to rare words. Three corpora are analysed to extract word frequencies:

Table 1: Statistics for the collections used as background models for understandability estimations.

Statistic	Medical Wikipedia	Medical Reddit	PubMed Central
Number of Docs.	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 Doc.	898.90 ± 1351.76	278.45 ± 359.70	227.22 ± 270.44
Avg. Char per Doc.	5107.81 ± 7618.57	1258.44 ± 1659.96	1309.11 ± 1447.31
Avg. Char per Word	5.68 ± 3.75	4.52 ± 3.52	5.76 ± 3.51

- Medical Reddit: Reddit [48] is a Web forum with a sizeable user community which is responsible for generating and moderating its content. Any user can start a discussion or reply to a discussion. This forum is intensively used for health purposes: for example in the Reddit community AskDocs [47], licensed nurses and doctors (subject to user identity verification) advise help seekers free of charge. We selected six of such communities (medical, AskDocs, AskDoctorSmeeee, Health, WomensHealth, Mens_Health) and downloaded all user interactions available until September 1st 2017 using the Python library PRAW [62]. In total 43,019 discussions were collected.
- Medical English Wikipedia: after obtaining a recent Wikipedia dump [17] (May 1st 2017), we filtered articles to only those containing an Infobox[5] in which at least one of the following words appeared as a property: ICD10, ICD9, DiseasesDB, MeSH, MeSHID, MeshName, MeshNumber, GeneReviewsName, Orphanet, eMedicine, MedlinePlus, drug_name, Drugs.com, DailyMedID, LOINC. In doing so, we followed the method by Soldaini et al. [53], which favours precision over recall when identifying a health-related article. This resulted in a collection of 11,868 articles.
- PubMed Central: PubMed Central [7] is an online database of full-text biomedical literature. We used the collection distributed for the TREC 2014 and 2015 Clinical Decision Support Track [49, 50], consisting of 733,191 articles.

A summary of the statistics of these three collections is reported in Table 1. Unless explicitly stated otherwise, we ignored out of vocabulary words in our feature calculations.

Machine Learning on Text - Regressors (MLR) and Classifiers (MLC): These include machine learning methods for estimating Web page understandability. While Collins-Thompson highlighted the promise of estimating understandability using machine learning methods, a challenge is identifying the background corpus to be used for training [10]. To this aim, we use the three corpora detailed above, and assume understandability labels according to the expected difficulty of documents in these collections:

- Medical Reddit (label 1): Documents in this collection are expected to be written in a colloquial style, and thus the easiest to

[5] A Wikipedia infobox is a structured template that appears on the right of Wikipedia pages summarizing key aspects of articles.

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

Models are learned using all documents from these collections after features are extracted using Latent Semantic Analysis (LSA) with 10 dimensions (empirically set based on document word counts in the three collections). We model a classification task as well as a regression task using these collections. Thus, after applying the same LSA transformation to test documents from CLEF, a continuous score is assigned to each document by a regressor[6], while each classifier assigns the documents to one of the three classes.

5 EVALUATION OF UNDERSTANDABILITY ESTIMATORS

Using the CLEF eHealth 2015 and 2016 collections, we studied the correlations of methods to estimate Web page understandability (Table 2) and human assessments. For each category of understandability estimation method, Table 3 reports the methods with highest Pearson, Spearman or Kendall correlations.

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

For surface level readability measures, SMOG had the highest correlations for CLEF 2015 and DCI for CLEF 2016, regardless of correlation measure. These results resonated with those obtained for the category of raw components of readability formulas. In fact, the polysyllable words measure and the number of difficult words are, respectively, parts of the SMOG and DCI formulas, which had the highest correlation for CLEF 2015 and 2016 among these methods.

When examining the expert vocabulary category, we found that the number of MeSH concepts obtained the highest correlations with human assessments; however its correlations were significantly 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 category (counts of P tags and list tags) exhibited overall the lowest 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.

[6]In principle, regressors should output a continuous value between 1 and 3, but no restrictions are set and potentially any value can be assigned to a document.

Table 2: 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.

Cat.	Method
RF	Automated Readability Index (ARI) [52]
	Coleman-Liau Index (CLI) [9]
	Dale Chall Index (DCI) [14]
	Flesch-Kincaid Grade Level (FKGL) [25]
	Flesch Reading Ease (FRE) [25]
	Gunning Fog Index (GFI) [22]
	Lasbarhetsindex (LIX) [6]
	Simple Measure of Gobbledygook (SMOG) [31]
CRF	# of Characters ⋆⋄†
	# of Words ⋆†
	# of Sentences ⋆⋄
	# of Difficult Words (Dale Chall list [14]) ⋆⋄†
	# of Words Longer than 4 chars ⋆⋄†
	# of Words Longer than 6 chars ⋆⋄†
	# of Words Longer than 10 chars ⋆⋄†
	# of Words Longer than 13 chars ⋆⋄†
	# of Number of Syllables ⋆⋄†
	# of Polysyllable Words (>3 Syllables) ⋆⋄†
GMV	# of Words with Medical Prefix ⋆⋄†
	# of Words with Medical Suffix ⋆⋄†
	# of Acronyms ⋆⋄†
	# of ICD Concepts ⋆⋄†
	# of Drugbank ⋆⋄†
	# of Words in medical dict. (OpenMedSpel) ⋆⋄†
CMV	CHV Mean Score for all Concepts ⋆⋄†
	# of CHV Concepts ⋆⋄†
	CHV Mean Score for Symptom Concepts ⋆⋄†
	# of CHV Symptom Concepts ⋆⋄†
	CHV Mean Score for Disease Concepts ⋆⋄†
	# of CHV Disease Concepts ⋆⋄†
EMV	# of MeSH Concepts ⋆⋄†
	Average Tree of MeSH Concepts ⋆⋄†
	# of MeSH Symptom Concepts ⋆⋄†
	Average Tree of MeSH Symptom Concepts ⋆⋄†
	# of MeSH Disease Concepts ⋆⋄†
	Average Tree of MeSH Disease Concepts ⋆⋄†
NL	Positive Words ⋆⋄†
	Negative Words ⋆⋄†
	Neutral Words ⋆⋄†
	# of verbs ⋆⋄†
	# of nouns ⋆⋄†
	# of pronouns ⋆⋄†
	# of adjectives ⋆⋄†
	# of adverbs ⋆⋄†
	# of adpositions ⋆⋄†
	# of conjunctions ⋆⋄†
	# of determiners ⋆⋄†
	# of cardinal numbers ⋆⋄†
	# of particles or other function words ⋆⋄†
	# of other POS (foreign words, typos) ⋆⋄†
	# of punctuation ⋆⋄†
	Height of part-of-speech parser tree ⋆⋄†
	# of Entities ⋆⋄†
	# of Stopwords ⋆⋄†
	# of words not found in Aspell Eng. dict. ⋆⋄†

Cat.	Method
HF	# of Abbr tags
	# of A tags
	# of Blockquote tags
	# of Bold tags
	# of Cite tags
	# of Div tags
	# of Forms tags
	# of H1 tags
	# of H2 tags
	# of H3 tags
	# of H4 tags
	# of H5 tags
	# of H6 tags
	# of Hs (any H above)
	# of Img tags
	# of Input tags
	# of Link tags
	# of DL tags
	# of UL tags
	# of OL tags
	# of List (DL + UL + OL)
	# of Q tags
	# of Scripts tags
	# of Spans tags
	# of Table tags
	# of P tags
WFF	25th percentil English Wikipedia
	50th percentil English Wikipedia
	75th percentil English Wikipedia
	Mean Rank English Wikip.
	Mean Rank English Wikip. - Includes OV
	25th percentil Medical Reddit
	50th percentil Medical Reddit
	75th percentil Medical Reddit
	Mean Rank Medical Reddit
	Mean Rank Medical Reddit - Includes OV
	25th percentil Pubmed
	50th percentil Pubmed
	75th percentil Pubmed
	Mean Rank Pubmed
	Mean Rank Pubmed - Includes OV
	25th p. Wikipedia+Reddit+Pubmed
	50th p. Wikipedia+Reddit+Pubmed
	75th p. Wikipedia+Reddit+Pubmed
	Mean R. Wiki.+Reddit+Pubmed
	Mean R. Wiki.+Reddit+Pubmed - w. OV
MLR	Linear Regressor
	Gradient Boosting Regressor
	Multi-layer Perceptron Regressor
	Random Forest Regressor
	Support Vector Machine Regressor
MLC	Logistic Regression
	Gradient Boosting Classifier
	Multi-layer Perceptron Classifier
	Random Forest Classifier
	Support Vector Machine Classifier
	Multinomial Naive Bayes

Finally, regressors and classifiers exhibited the highest correlations across all categories: in this category, the Neural Network regressor and the multinomial Naive Bayes best correlated with human assessments.

6 EVALUATION OF PREPROCESSING PIPELINES AND HEURISTICS

Results from experiments with different preprocessing pipelines and heuristics are shown in Figure 2 (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 the human assessments[7]. We further report summary results across all understandability assessment methods and sentence ending heuristics for each of the preprocessing

[7]Results for Pearson and Kendall correlations are reported in the online appendix, but showed similar trends.

Table 3: Methods with the highest correlation per category.

Cat.	CLEF 2015					CLEF 2016				
	Method	Preproc.	Pears.	Spear.	Kend.	Method	Preproc.	Pears.	Spear.	Kend.
RF	SMOG Index	Jst NFP	**0.438**	**0.388**	**0.286**	Dale Chall Index	Jst FP	**0.439**	0.381	0.264
							Boi FP	0.437	**0.382**	**0.264**
CRF	Avg. Num. of Polysyl. Words per Word	Jst FP	**0.429**	0.364	0.268	Avg. Difficult Words Per Word	Boi FP	**0.431**	**0.379**	**0.262**
	Avg. N. of Polysyl. Words per Sentence	Jst NFP	0.192	**0.388**	**0.286**					
GMV	Avg. N. Medical Prefixes per Word	Naive FP	**0.314**	0.312	0.229	Avg. Prefixes per Sentence	Jst FP	**0.263**	0.242	0.164
	Number of Medical Prefixes		0.131	**0.368**	**0.272**	ICD Concepts Per Sentence	Jst NFP	0.014	**0.253**	**0.172**
CMV	CHV Mean Score for all Concepts	Naive FP	**0.371**	**0.314**	**0.228**	CHV Mean Score for all Concepts	Jst FP	**0.329**	0.313	0.216
						CHV Mean Score for all Concepts	Boi FP	0.329	**0.325**	**0.224**
EMV	Number of MeSH Concepts	Naive FP	**0.227**	**0.249**	**0.178**	Number of MeSH Concepts	Boi NFP	**0.201**	0.166	0.113
						Number of MeSH Disease Concepts		0.179	**0.192**	**0.132**
NLF	N. of words not found in Aspell Dict.	Jst NFP	**0.351**	0.276	0.203	Avg. Stopword Per Word	Boi FP	**0.344**	0.312	0.213
	Number of Pronouns per Word	Naive FP	0.271	**0.441**	**0.325**	Number of Pronouns		0.341	**0.364**	**0.252**
HF	Number of P Tags	None	**0.219**	**0.196**	**0.142**	Number of Lists	None	**0.114**	0.021	0.015
						Number of P Tags		0.110	**0.123**	**0.084**
WFF	Mean Rank Medical Reddit - Includes OV	Jst NFP	**0.435**	0.277	0.197	Mean Rank Medical Reddit	Boi NFP	**0.387**	0.312	0.214
	25th percentil Pubmed	Jst NFP	0.330	**0.347**	**0.256**	50th percentil Medical Reddit	Jst NFP	0.351	**0.315**	**0.216**
MLR	Neural Network Regressor	Boi NFP	**0.602**	0.394	0.287	Neural Network Regressor	Jst NFP	**0.454**	**0.373**	0.258
	Neural Network Regressor	Jst FP	0.565	**0.438**	**0.324**	Random Forest Regressor	Boi NFP	0.389	0.355	**0.264**
MLC	Multinomial Naive Bayes	Naive FP	**0.573**	**0.477**	**0.416**	Multinomial Naive Bayes	Jst FP	**0.461**	**0.391**	**0.318**

pipelines. Finally, we also report the inter-assessor correlation (last box) when multiple assessors provided judgements about the understandability of Web pages (details about this data in Section 3). This provides an indication of the range of variability and subjectiveness when assessing understandability, along with the highest correlation we measured between human assessors.

We first examined the correlations between human assessments and readability formulas. We found that the *Naive* preprocessing resulted in the lowest correlations, regardless of readability formula and heuristics (although *DoNotForcePeriod* performed better than *ForcePeriod*). Using Justext or Boilerplate resulted in higher correlations with human understandability assessments, and the *ForcePeriod* heuristic was shown to be better than *DoNotForcePeriod*. These results confirm the speculations of Palotti et al. [41]: 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 (highest correlations) were obtained by SMOG and DCI (see Table 3). 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. [41], i.e. if using readability measures, then CLI should be preferred, along with an appropriate HTML extraction pipeline, regardless of the heuristic for sentence ending[8].

When considering methods beyond those based on readability formulas, we found that the highest correlations were archived by the regressors (MLR) and classifiers (MLC), independent of the preprocessing method used. There is little difference in terms of effectiveness of methods in these categories, with the exception of regressors on CLEF 2015 that exhibited not negligible variances: while the Neural Network Regressor Pearson correlation was 0.44, the Support Vector Regressor 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 confirmed for CLEF 2015, where the Naive preprocessing negatively influenced correlations for the readability formula category (RF), but not for other categories, although it was generally associated with larger variances on the correlation coefficients.

[8]We provide detailed plots to compare our results with Palotti's in the online appendix.

7 INTEGRATING UNDERSTANDABILITY INTO RETRIEVAL

We 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 included the best two runs submitted to each CLEF task (2015 and 2016), and a plain BM25 baseline (default Terrier parameters: $b = 0.75$ and $k1 = 1.2$). As understandability estimators we used the eXtreme Gradient Boosting (XGB) regressor[9][8], as well as SMOG for CLEF 2015 and DCI for CLEF 2016. These were the best performing approaches from Section 5.

To integrate understandability estimators into the retrieval process, we first investigated *re-ranking* search results from the initial runs purely based on the understandability estimations. If all the search results from a run were considered, 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 understandability, 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. Results are presented in Section 8.1.

[9]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 learning methods and feature selection schemes were experimented with; results are available in the online appendix. XGB was selected because its results were the best, although other methods followed similar trends.

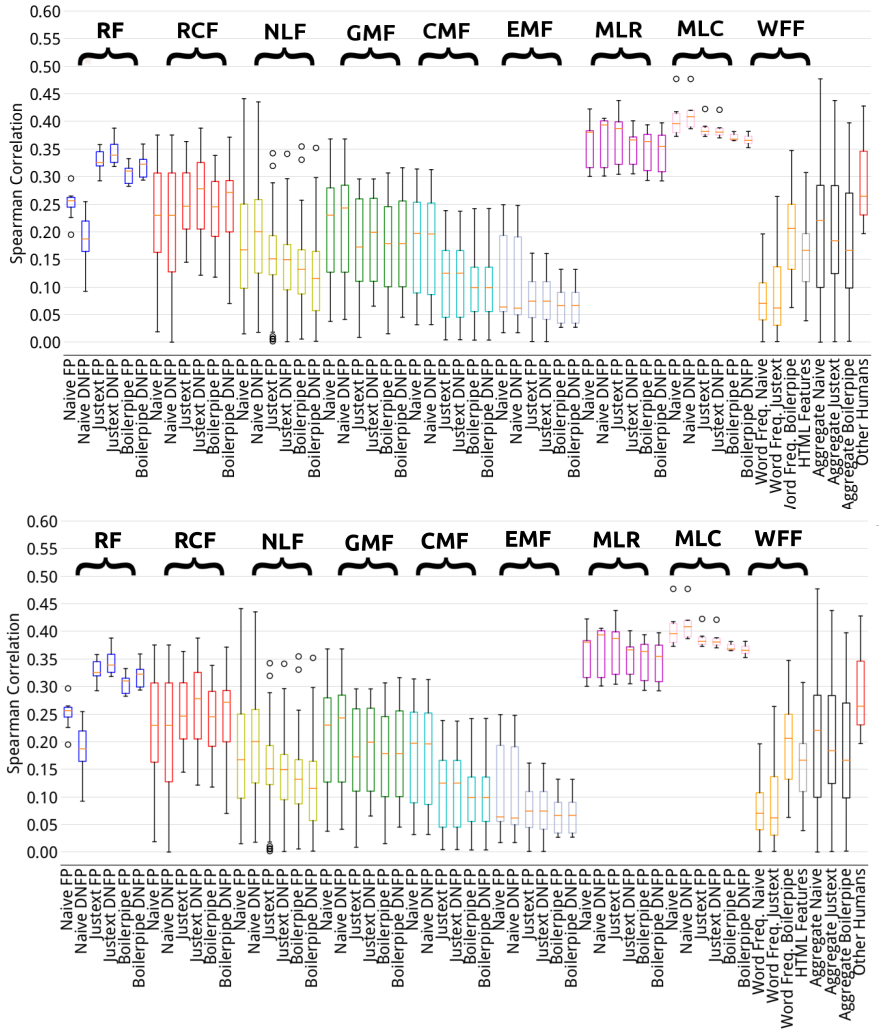

Figure 2: Correlations between understandability estimators and human assessments for CLEF 2015 (top) and 2016 (bottom).

As an alternative to the previous two-step ranking strategy for combining topical relevance and understandability, we explore the fusion of two search results lists separately obtained for relevance and understandability. For this, we use the Reciprocal Rank Fusion (RRF) method [12], which has proven effective for combining two lists of search results based on their document *ranks*, rather than scores. This approach was selected above score-based fusion methods because of the different scoring strategies and distributions employed when scoring for relevance compared to for understandability. For topical relevance, we use, separately, the three methods used for re-ranking (ECNU [55] and KISTI [33] for CLEF2015, GUIR [54] and ECNU [56] for CLEF 2016, and BM25 for both collections). For understandability, we use, separately, the estimations from SMOG/Dale Chall and XGB. Also for this approach we study limiting the ranking of results to be considered by the methods across the cut-offs $k = [15, 20, 50]$. Results are presented in Section 8.2.

Finally, we consider a third alternative to combine topical relevance and understandability: using learning to rank with features derived from retrieval methods and the understandability estimators. With the CLEF 2016 collection, we explore five combinations of label attribution and feature sets, keeping the same pairwise learning to rank algorithm based on tree boosting (XGB). These combinations are listed in Table 4, with R being the topical relevance of documents and U their understandability assessments.

Table 4: Learning to rank experimental settings.

Name	Feature Set	Labeling Function
Combo 1	IR features	$F(R,U) = R$
Combo 2	IR + Unders. features	$F(R,U) = R$
Combo 3	IR + Unders. features	$F(R,U) = R \times (100 - U)/100$
Combo 4	IR + Unders. features	$F(R,U) = \begin{cases} R & \text{if } U \leq 40 \\ 0 & \text{otherwise} \end{cases}$
Combo 5	IR + Unders. features	$F(R,U) = \begin{cases} 2 \times R & \text{if } U \leq 40 \\ R & \text{otherwise} \end{cases}$

Table 5: Results obtained by integrating understandability estimations within retrieval methods on CLEF 2016. Baseline runs are reported at table indices 1-3. 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 $n = 10$.

Index	Rerank	Run	Official CLEF 2016 Metrics					New Metrics to Evaluate Underst. in Retrieval - Sec. 3					
			RBP	RBP Res.	uRBP	uRBP Res.	*uRBP**	RBP_u	*HRBP*	Unj	RBP^*_r	RBP^*_u	*HRBP**
1		GUIR [54] (Best Run)	28.11	7.65	18.12	7.19	18.22	45.69	25.61	0.01	28.29	46.03	25.79
2	No Rerank	ECNU [56] (Runner Up)	27.70	7.37	17.55	7.34	17.61	43.89$^\diamond$	25.35	0.01	27.77	44.18$^\diamond$	25.48
3		Plain BM25 Baseline	25.28$^\diamond$	8.24	16.05$^\diamond$	6.94	16.46$^\diamond$	**42.08**$^\diamond$	22.97$^\diamond$	0.06	**26.01**$^\diamond$	**43.89**$^\diamond$	**23.93**$^\diamond$
4	Dale-Chall Top 15	Based on GUIR	24.70$^{\dagger\diamond}$	8.70	16.83$^{\dagger\diamond}$	7.27	17.18$^{\dagger\diamond}$	49.10$^{\dagger\diamond}$	24.94	0.03	25.24$^{\dagger\diamond}$	50.33$^{\dagger\diamond}$	25.54
5		Based on ECNU	24.78$^{\dagger\diamond}$	7.83	16.64$^\diamond$	7.16	16.87	48.88$^{\dagger\diamond}$	24.80	0.02	25.12$^{\dagger\diamond}$	49.64$^{\dagger\diamond}$	25.21
6		Based on BM25	23.22$^{\dagger\diamond}$	8.78	15.85 $^\diamond$	6.94	16.34	**47.09**$^{\dagger\diamond}$	24.01	0.07	24.04$^{\dagger\diamond}$	48.60$^{\dagger\diamond}$	24.82
7	Dale-Chall Top 20	Based on GUIR	22.19$^{\dagger\diamond}$	9.37	15.36$^{\dagger\diamond}$	6.98	16.10$^{\dagger\diamond}$	48.71$^{\dagger\diamond}$	23.21$^{\dagger\diamond}$	0.06	23.26$^{\dagger\diamond}$	51.39$^{\dagger\diamond}$	24.45$^{\dagger\diamond}$
8		Based on ECNU	23.01$^{\dagger\diamond}$	8.93	15.70$^{\dagger\diamond}$	6.91	16.24$^\diamond$	48.99$^{\dagger\diamond}$	23.73$^{\dagger\diamond}$	0.05	23.84$^{\dagger\diamond}$	51.00$^{\dagger\diamond}$	24.66
9		Based on BM25	21.58$^{\dagger\diamond}$	9.51	14.83$^{\dagger\diamond}$	7.02	15.65$^\diamond$	46.99†	22.89$^\diamond$	0.09	22.93$^{\dagger\diamond}$	49.55$^{\dagger\diamond}$	24.26
10	Dale-Chall Top 50	Based on GUIR	16.18$^{\dagger\diamond}$	15.24	11.56$^{\dagger\diamond}$	6.80	14.78$^{\dagger\diamond}$	41.79$^{\dagger\diamond}$	18.10$^{\dagger\diamond}$	0.22	20.90$^{\dagger\diamond}$	53.28$^{\dagger\diamond}$	23.27$^{\dagger\diamond}$
11		Based on ECNU	16.88$^{\dagger\diamond}$	17.37	11.78$^{\dagger\diamond}$	7.30	14.84$^{\dagger\diamond}$	40.76$^{\dagger\diamond}$	18.30$^{\dagger\diamond}$	0.24	21.34$^{\dagger\diamond}$	52.07$^{\dagger\diamond}$	23.33$^{\dagger\diamond}$
12		Based on BM25	15.06$^{\dagger\diamond}$	15.35$^{\dagger\diamond}$	10.55	6.62	13.52$^{\dagger\diamond}$	40.03 $^\diamond$	16.55$^{\dagger\diamond}$	0.24	19.42$^{\dagger\diamond}$	51.69$^{\dagger\diamond}$	21.59$^{\dagger\diamond}$
13	XGB Top 15	Based on GUIR	25.16$^{\dagger\diamond}$	8.09	17.27$^{\dagger\diamond}$	7.12	17.58	50.96$^{\dagger\diamond}$	25.16	0.02	25.61$^{\dagger\diamond}$	52.00$^{\dagger\diamond}$	25.68
14		Based on ECNU	24.18$^{\dagger\diamond}$	7.69	16.54 $^\diamond$	7.09	16.78$^\diamond$	50.00$^{\dagger\diamond}$	24.56	0.02	24.56$^{\dagger\diamond}$	50.74$^{\dagger\diamond}$	25.01
15		Based on BM25	22.33$^{\dagger\diamond}$	8.14	15.46	6.76	15.89$^\diamond$	47.90$^{\dagger\diamond}$	22.89$^\diamond$	0.07	23.11$^{\dagger\diamond}$	49.43$^{\dagger\diamond}$	23.69$^\diamond$
16	XGB Top 20	Based on GUIR	22.38$^{\dagger\diamond}$	9.49	15.61$^{\dagger\diamond}$	7.05	16.48$^{\dagger\diamond}$	50.45$^{\dagger\diamond}$	23.30$^{\dagger\diamond}$	0.05	23.62$^{\dagger\diamond}$	52.98$^{\dagger\diamond}$	24.68
17		Based on ECNU	22.95$^{\dagger\diamond}$	8.82	15.95$^{\dagger\diamond}$	7.02	16.41$^\diamond$	50.42$^{\dagger\diamond}$	23.97$^\diamond$	0.04	23.68$^{\dagger\diamond}$	52.15$^{\dagger\diamond}$	24.73
18		Based on BM25	20.65$^{\dagger\diamond}$	9.42	14.46$^{\dagger\diamond}$	6.84	15.26$^{\dagger\diamond}$	47.74$^{\dagger\diamond}$	21.93$^\diamond$	0.09	21.98$^{\dagger\diamond}$	50.28$^{\dagger\diamond}$	23.27$^\diamond$
19	XGB Top 50	Based on GUIR	16.65$^{\dagger\diamond}$	15.73	12.39$^{\dagger\diamond}$	6.84	15.45$^{\dagger\diamond}$	43.49$^{\dagger\diamond}$	18.70$^{\dagger\diamond}$	0.22	21.13$^{\dagger\diamond}$	55.07$^{\dagger\diamond}$	23.58$^{\dagger\diamond}$
20		Based on ECNU	16.19$^{\dagger\diamond}$	17.01	11.82$^{\dagger\diamond}$	7.27	14.52$^{\dagger\diamond}$	43.05$^\diamond$	18.27$^{\dagger\diamond}$	0.24	20.16$^{\dagger\diamond}$	54.70$^{\dagger\diamond}$	22.96$^{\dagger\diamond}$
21		Based on BM25	15.43$^{\dagger\diamond}$	15.37	11.33$^{\dagger\diamond}$	6.48	14.16$^{\dagger\diamond}$	41.93$^\diamond$	17.43$^{\dagger\diamond}$	0.26	19.58$^{\dagger\diamond}$	54.04$^{\dagger\diamond}$	**22.17**$^{\dagger\diamond}$
22	RRF (XGB & Orig.) Top 15	Based on GUIR	27.23$^{\dagger\diamond}$	7.76	18.31	7.23	18.44	49.69$^{\dagger\diamond}$	26.49$^{\dagger\diamond}$	0.01	27.46$^{\dagger\diamond}$	50.07$^{\dagger\diamond}$	26.69$^{\dagger\diamond}$
23		Based on ECNU	26.60$^{\dagger\diamond}$	7.41	17.81	7.19	17.91	48.67$^{\dagger\diamond}$	26.02	0.01	26.76$^{\dagger\diamond}$	49.10$^{\dagger\diamond}$	26.27†
24		Based on BM25	24.57$^\diamond$	8.15	16.51$^\diamond$	6.91	16.91$^\diamond$	46.76†	24.16†	0.06	25.32$^\diamond$	48.52$^{\dagger\diamond}$	25.08†
25	RRF (XGB & Orig.) Top 20	Based on GUIR	26.21$^{\dagger\diamond}$	7.96	17.73	7.19	17.94	50.29$^{\dagger\diamond}$	25.89	0.03	26.53$^{\dagger\diamond}$	50.98$^{\dagger\diamond}$	26.25
26		Based on ECNU	26.15$^{\dagger\diamond}$	7.64	17.69	7.09	17.85	49.70$^{\dagger\diamond}$	26.07	0.02	26.38$^{\dagger\diamond}$	50.32$^{\dagger\diamond}$	26.35
27		Based on BM25	24.04$^{\dagger\diamond}$	8.24	16.32$^\diamond$	6.87	16.77$^\diamond$	47.69$^{\dagger\diamond}$	24.08$^{\dagger\diamond}$	0.06	24.82$^{\dagger\diamond}$	49.52$^{\dagger\diamond}$	25.01†
28	RRF (XGB & Orig.) Top 50	Based on GUIR	24.09$^{\dagger\diamond}$	9.44	16.85$^{\dagger\diamond}$	7.02	17.49	50.55$^{\dagger\diamond}$	24.76	0.07	25.08$^{\dagger\diamond}$	52.84$^{\dagger\diamond}$	25.84
29		Based on ECNU	24.17$^{\dagger\diamond}$	8.67	16.75$^\diamond$	7.12	17.22$^\diamond$	50.63$^{\dagger\diamond}$	25.00	0.07	24.90$^{\dagger\diamond}$	52.50$^{\dagger\diamond}$	25.84
30		Based on BM25	22.28$^{\dagger\diamond}$	8.87	15.50	6.76	16.25$^\diamond$	48.79$^{\dagger\diamond}$	23.13$^{\dagger\diamond}$	0.10	23.46$^{\dagger\diamond}$	51.89$^{\dagger\diamond}$	**24.57**
31	XGB LeToR	Combo 1 on BM25	20.42$^{\dagger\diamond}$	17.61	13.00$^{\dagger\diamond}$	7.41	15.94$^\diamond$	32.17$^{\dagger\diamond}$	18.39$^{\dagger\diamond}$	0.28	25.25$^\diamond$	43.19$^\diamond$	23.83$^\diamond$
32		Combo 2 on BM25	24.98$^{\dagger\diamond}$	19.83	15.30$^{\dagger\diamond}$	8.09	18.71	35.09$^{\dagger\diamond}$	22.26$^\diamond$	0.24	30.41	46.09	28.28$^{\dagger\diamond}$
33		Combo 3 on BM25	26.35†	20.48	15.88$^{\dagger\diamond}$	8.16	19.57	34.73$^{\dagger\diamond}$	21.81†	0.22	32.25$^\diamond$	45.44	28.22$^{\dagger\diamond}$
34		Combo 4 on BM25	16.16$^{\dagger\diamond}$	19.48	10.76$^{\dagger\diamond}$	7.27	14.59$^{\dagger\diamond}$	36.75$^{\dagger\diamond}$	16.77$^{\dagger\diamond}$	0.29	**22.20**$^{\dagger\diamond}$	**50.06**$^{\dagger\diamond}$	23.32$^\diamond$
35		Combo 5 on BM25	26.76$^\diamond$	20.48	16.19$^\diamond$	8.34	19.83†	35.26$^{\dagger\diamond}$	22.96	0.22	**32.60**†	**45.87**	**29.20**$^{\dagger\diamond}$

8 EVALUATION OF UNDERSTANDABILITY RETRIEVAL

Results for the considered retrieval methods are reported in Table 5. We reported only the results for CLEF 2016 for brevity; those for CLEF 2015 exhibited similar trends and are included in the online appendix. The effectiveness of the top two submissions to CLEF 2016 and the BM25 baseline are reported at indices 1-3 of Table 5. Statistically significant differences compared to the best CLEF 2016 run, GUIR, are reported with $\diamond$; differences between an original run (indices 1-3) and its modifications are reported with $\dagger$ (paired, two-tail t-test, $p < 0.05$). Note that the baseline BM25 is significantly worse than GUIR across all measures.

8.1 Re-ranking

Indices 4-12 of Table 5 report the results 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 3). We found that the relevance of the re-ranked runs (as measured by RBP_r and RBP^*_r) significantly decreased, compared to the original runs: e.g., re-ranking the top 15 search results using DCI made RBP_r decreasing from 25.28 to 21.58. However, these re-ranked results were significantly more understandable: for the previous example, RBP_u passed from 42.08 to 47.09.

In the experiments, we also studied the influence of the numbers of documents considered for re-ranking (cut-off). Indices 4-6 refer to re-ranking only the top $n = 15$ documents from the original runs; 7-9 refer to the first $n = 20$; and 10-12 to the first $n = 50$. The results show that the more documents are considered for re-ranking, the more degradation in RBP_r effectiveness. Considering understandability-only in the evaluation, shows mixed results. Similar trends are observed for evaluation measures that consider understandability (*RBP* and RBP_u), however with some exceptions. For example, an increase in *uRBP* is 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 number of unassessed documents being considered by the evaluation measures. Both the RBP residuals and the column *Unj* quantify the effect unassessed documents have on evaluation. Nevertheless, we note that if unassessed

documents are excluded from the evaluation, similar trends are observed, e.g. compare findings with those for $uRBP^*$, RBP^*_r, RBP^*_u and $HRBP^*$.

Indices 13-21 refer to using the XGB regressor trained using all features listed in Table 2 to estimate understandability. Similarly to when using DCI, as the cut-off increases, e.g., from $n = 15$ to $n = 50$, documents returned are more understandable but less relevant. For the same cut-off value, e.g., $n = 15$, the machine learning method used for estimating understandability consistently yielded more understandable results than DCI (higher RBP_u and RBP^*_u).

Overall, statistical significant improvements over the baselines are observed for most configurations and measures.

8.2 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 producing by fusing the re-ranking with XGB and the original run. (Results for DCI are in the appendix and confirm the superiority of XGB.)

Like as for re-ranking, also for 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 measures on the other. Overall, results obtained with rank fusion are superior to those obtained with re-ranking only, though most differences are not statistically significant. Statistical significant improvements over the baselines are instead observed for most configurations and measures.

8.3 Learning to Rank

Last, we analyse the results obtained by the learning to rank methods (indices 31-35). Unlike with the previous methods, we did not impose a rank cut-off on learning to rank. Also, recall that learning to rank was only applied to the BM25 baseline.

When considering RBP_r and $uRBP$, the results for learning to rank exhibit effectiveness that is 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 RBP residuals (and the number of unassessed documents) reveals that this may be because measures may be affected by the large number of unassessed documents retrieved in the top 10 ranks. For example, the RBP_r residual for learning to rank methods is about double than that of the baselines or other approaches. In fact, among the documents retrieved in the top 10 results by learning to rank, there are 20% that are unassessed, compared 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 measures that ignore unassessed documents. When this is done, we observe that learning to rank methods overall provide substantial gains over the original runs and other methods (when considering RBP^*_r, RBP^*_u and $HRBP^*$), or large potential gains over these methods (when considering the residuals). Next, we analyse these results in more detail.

No improvements over the baselines were found for Combo 1 (index 31), and the high residuals for RBP_r are not matched by other residuals or by considering only assessed documents. Combo 1 was a simple method that used only IR features[10] and was trained only on topical relevance. Although simple, this is a typical learning to rank setting.

Compared to Combo 1, Combo 2 (index 32) included the understandability features listed in Table 2. The inclusion of understandability features was as beneficial to the understandability measures as for the relevance measures, with RBP^*_r, RBP^*_u and $HRBP^*$ all showing gains over the baselines. Combo 3 showed similar $HRBP^*$ values, though with higher effectiveness with respect to relevance (RBP^*_r) than to understandability (RBP^*_u)

Combos 4 and 5 were devised based on a pre-defined understandability threshold $U = 40$. While Combo 4 took into consideration only documents that are easy-to-read (understandability label $\leq U$), Combo 5 considered all documents, but boosted the relevance score of easy-to-read documents. While Combo 4 reached the highest understandability score for the learning-to-rank approaches ($RBP^*_u = 50.06$), it failed to retrieve relevant documents ($RBP^*_r = 22.20$). In turn, Combo 5 reached the highest understandability-relevance trade-off ($HRB^* = 29.20$). Compared to the BM25 baseline (on which it is based), Combo 5 largely increased both relevance (RBP^*_r from 26.01 to 32.60 – a 25% increase) and understandability (RBP^*_u from 43.89 to 45.87 – a 4% increase). Note that Combo 5 is also significantly better than the best run submitted to CLEF 2016 for both RBP^*_r (15% increase) and $HRBP^*$ (13% increase).

9 CONCLUSION

In this paper we have examined approaches to estimate the understandability of health Web pages, including the impact of HTML preprocessing techniques, and how to integrate these within retrieval methods to provide more understandable search results for people seeking health information.

The empirical experiments suggested that: (1) machine learning methods based on regression are best suited to estimate the understandability of health Web pages; (2) preprocessing does affect effectiveness (both for understandability prediction and document retrieval), although, compared to other methods, ML-based methods for understandability estimation are less subject to variability due to poor preprocessing; (3) learning to rank methods can be specifically trained to promote more understandable search results.

This paper makes a clear contribution 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 paper further highlights which methods and settings do provide better search results to health information seekers. As shown in Figure 1, these methods would clearly improve current health-focused search engines.

[10] We devised 24 IR features using the Terrier framework. The score of various retrieval models were extracted from a multi-field index composed of title, body and whole document.

REFERENCES

[1] Christopher Agrafiotes and Avi Arampatzis. 2016. 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.*

[2] Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *JAMIA* 17, 3 (2010), 229–236. https://doi.org/10.1136/jamia.2009.002733

[3] GNU Aspell. 2017. GNU English Dictionary Aspell. http://www.aspell.net/. (2017). [Online: accessed 21-October-2017].

[4] Samuel R Atcherson, Ashley E DeLaune, Kristie Hadden, Richard I Zraick, Rebecca J Kelly-Campbell, and Carlos P Minaya. 2014. A Computer-Based Readability Analysis of Consumer Materials on the American Speech-Language-Hearing Association Website. *Contemporary Issues in Communication Science & Disorders* 41 (2014).

[5] Shirley Ann Becker. 2004. A study of web usability for older adults seeking online health resources. *ACM Transactions on Computer-Human Interaction (TOCHI)* 11, 4 (2004), 387–406.

[6] C. H. Björnsson. 1983. Readability of Newspapers in 11 Languages. *Reading Research Quarterly* 18, 4 (1983), 480–497. http://www.jstor.org/stable/747382

[7] PubMed Central. 2017. National Center for Biotechnology Information PubMed Central. https://www.ncbi.nlm.nih.gov/pmc/. (2017). [Online: accessed 21-October-2017].

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[9] Meri Coleman and T. L. Liau. 1975. A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology* (1975).

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

[11] Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the Association for Information Science and Technology* 56, 13 (2005), 1448–1462.

[12] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '09)*. ACM, New York, NY, USA, 758–759. http://doi.acm.org/10.1145/1571941.1572114

[13] Connie F Cowan. 2004. Teaching patients with low literacy skills. *Fuszard's Innovative Teaching Strategies in Nursing* (2004), 278.

[14] Edgar Dale and Jeanne S. Chall. 1948. A Formula for Predicting Readability: Instructions. *Educational Research Bulletin* 27, 2 (1948), 37–54. http://www.jstor.org/stable/1473669

[15] Terry C Davis and Michael S Wolf. 2004. Health literacy: implications for family medicine. *Family Medicine* 36, 8 (2004), 595–598.

[16] William H. Dubay. 2004. The Principles of Readability. *Costa Mesa, CA: Impact Information* (2004).

[17] Wikimedia Dump. 2017. English Wikipedia Dumps. https://dumps.wikimedia.org/enwiki/. (2017). [Online: accessed 21-October-2017].

[18] Chandy Ellimoottil, Anthony Polcari, Adam Kadlec, and Gopal Gupta. 2012. Readability of websites containing information about prostate cancer treatment options. *The Journal of urology* 188, 6 (2012), 2171–2176.

[19] PR Fitzsimmons, BD Michael, JL Hulley, and GO Scott. 2010. A readability assessment of online Parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh* 40, 4 (2010), 292–296.

[20] Evgeniy Gabrilovich. 2016. Cura Te Ipsum: answering symptom queries with question intent. In *Second WebQA workshop, SIGIR 2016 (invited talk)*.

[21] Mark A Graber, Cathy M Roller, and Betsy Kaeble. 1999. Readability levels of patient education material on the World Wide Web. *Journal of Family Practice* 48, 1 (1999), 58–59.

[22] Robert Gunning. 1952. *The Technique of Clear Writing*. McGraw-Hill.

[23] Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*. 460–467.

[24] Hyeoneui Kim, Sergey Goryachev, Graciela Rosemblat, Allen Browne, Alla Keselman, and Qing Zeng-Treitler. 2007. Beyond surface characteristics: a new health text-specific readability measurement. In *AMIA Annual Symposium Proceedings*, Vol. 2007. American Medical Informatics Association, 418.

[25] J. Kincaid, Robert Fishburne, Richard Rogers, and Brad Chissom. 1975. *Derivation of New Readability Formulas for Navy Enlisted Personnel*. National Technical Information Service.

[26] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. 2010. Boilerplate detection using shallow text features. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 441–450.

[27] Bevan Koopman and Guido Zuccon. 2014. Relevation!: An open source system for information retrieval relevance assessment. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 1243–1244.

[28] Gondy Leroy, Stephen Helmreich, James R Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *AMIA Annual Symposium Proceedings*, Vol. 2008. American Medical Informatics Association, 394.

[29] Philip Ley and Tony Florio. 1996. The use of readability formulas in health care. *Psychology, Health & Medicine* 1, 1 (1996), 7–28.

[30] Xiaoyong Liu, W. Bruce Croft, Paul Oh, and David Hart. 2004. Automatic Recognition of Reading Levels from User Queries. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04)*. ACM, 548–549.

[31] G. Harry McLaughlin. 1969. SMOG Grading - a New Readability Formula. *Journal of Reading* (1969).

[32] Andrew Meillier and Shyam Patel. 2017. Readability of Healthcare Literature for Gastroparesis and Evaluation of Medical Terminology in Reading Difficulty. *Gastroenterology Research* 10, 1 (2017), 1–5.

[33] Heung-Seon Oh, Yuchul Jung, and Kwang-Young Kim. 2015. 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.*

[34] OpenMedSpel. 2017. OpenOffice Medical Dictionary Extension. http://extensions.openoffice.org/en/project/openmedspel-en-us. (2017). [Online: accessed 21-October-2017].

[35] I. Ounis, G. Amati, Plachouras V., B. He, C. Macdonald, and Johnson. 2005. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005) (Lecture Notes in Computer Science)*, Vol. 3408. Springer, 517–519.

[36] Joao Palotti, Lorraine Goeuriot, Guido Zuccon, and Allan Hanbury. 2016. Ranking health web pages with relevance and understandability. In *Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 965–968.

[37] Joao Palotti, Allan Hanbury, and Henning Muller. 2014. 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.

[38] João Palotti, Allan Hanbury, Henning Müller, and Charles E. Kahn. 2016. How users search and what they search for in the medical domain. *Information Retrieval Journal* 19, 1 (Apr 2016), 189–224. https://doi.org/10.1007/s10791-015-9269-8

[39] Joao Palotti, Guido Zuccon, Johannes Bernhardt, Allan Hanbury, and Lorraine Goeuriot. 2016. Assessors Agreement: A Case Study across Assessor Type, Payment Levels, Query Variations and Relevance Dimensions. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF'16 Proceedings*. Springer International Publishing.

[40] João Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanbury, Gareth J. F. Jones, Mihai Lupu, and Pavel Pecina. 2015. 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.*

[41] João Palotti, Guido Zuccon, and Allan Hanbury. 2015. The Influence of Preprocessing on the Estimation of Readability of Web Documents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)*. ACM, New York, NY, USA, 1763–1766.

[42] Cheong Iao Pang. 2016. *Understanding Exploratory Search in Seeking Health Information*. Ph.D. Dissertation. The University of Melbourne.

[43] Chirag R Patel, Deepa V Cherla, Saurin Sanghvi, Soly Baredes, and Jean Anderson Eloy. 2013. Readability assessment of online thyroid surgery patient education materials. *Head & neck* 35, 10 (2013), 1421–1425.

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

[45] Jan Pomikálek. 2011. *Removing Boilerplate and Duplicate Content from Web Corpora*. Ph.D. Dissertation. Masaryk University, Czech Republic.

[46] PyPhen. 2017. Python module to hyphenate text. http://www.pyphen.org/. (2017). [Online: accessed 21-October-2017].

[47] Reddit. 2017. Reddit Ask A Doctor Community. https://www.reddit.com/r/AskDocs/. (2017). [Online: accessed 21-October-2017].

[48] Reddit. 2017. Reddit Webstie. https://www.reddit.com. (2017). [Online: accessed 21-October-2017].

[49] Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2016. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal* 19, 1 (2016), 113–148.

[50] Kirk Roberts, Matthew S. Simpson, Ellen M. Voorhees, and William R. Hersh. 2015. Overview of the TREC 2015 Clinical Decision Support Track. In *Proceedings of The Twenty-Fourth Text REtrieval Conference, TREC 2015, Gaithersburg, Maryland, USA, November 17-20, 2015.*

[51] Sarah J Shoemaker, Michael S Wolf, and Cindy Brach. 2014. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. *Patient education and counseling* 96, 3 (2014), 395–403.
[52] E. A. Smith and R. J. Senter. 1967. *Automated Readability Index*. Aerospace Medical Research Laboratories.
[53] Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. 2015. *Retrieving Medical Literature for Clinical Decision Support*. Springer International Publishing, 538–549.
[54] Luca Soldaini, Will Edman, and Nazli Goharian. 2016. Team GU-IRLAB at CLEF eHealth 2016: Task 3. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*. 143–146.
[55] Yang Song, Yun He, Qinmin Hu, Liang He, and E. Mark Haacke. 2015. ECNU at 2015 eHealth Task 2: User-centred Health Information Retrieval. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015*.
[56] Yang Song, Yun He, Hongyu Liu, Yueyao Wang, Qinmin Hu, and Liang He. 2016. ECNU at 2016 eHealth Task 3: Patient-centred Information Retrieval. In *Working Notes of CLEF 2016 - Conference and Labs of the Evaluation forum, Évora, Portugal, 5-8 September, 2016*. 157–161.
[57] Lauren M Stossel, Nora Segar, Peter Gliatto, Robert Fallar, and Reena Karani. 2012. Readability of patient education materials available at the point of care. *Journal of general internal medicine* 27, 9 (2012), 1165–1170.
[58] Trevor Strohman, Donald Metzler, Howard Turtle, and W Bruce Croft. 2005. Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, Vol. 2. Amherst, MA, USA, 2–6.
[59] National Cancer Institute (U.S.). Accessed: 2017-09. Clear & Simple: Developing Effective Print Materials for Low-literate Readers. https://www.nih.gov/institutes-nih/nih-office-director/office-communications-public-liaison/clear-communication/clear-simple. (Accessed: 2017-09).
[60] NLTK V3.2. 2017. Python Natural Language Toolkit Library. http://www.nltk.org/. (2017). [Online: accessed 21-October-2017].
[61] BeautifulSoap V4.4. 2017. BeautifulSoap. https://www.crummy.com/software/BeautifulSoup/. (2017). [Online: accessed 21-October-2017].
[62] Python Reddit API V5.1. 2017. PRAW: The Python Reddit API Wrapper. https://praw.readthedocs.io/. (2017). [Online: accessed 21-October-2017].
[63] Joost van Doorn, Daan Odijk, Diederik M Roijers, and Maarten de Rijke. 2016. Balancing relevance criteria through multi-objective optimization. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 769–772.
[64] Lorraine Silver Wallace and Elizabeth S Lennon. 2004. American Academy of Family Physicians patient education materials: can patients read them? *Family medicine* 36, 8 (2004), 571–574.
[65] Ryen White. 2013. Beliefs and biases in web search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13)*. ACM, New York, NY, USA, 3–12.
[66] Ryen W. White and Eric Horvitz. 2009. Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search. *ACM Transactions on Information Systems* 27, 4, Article 23 (Nov. 2009), 37 pages.
[67] R Constance Wiener and Regina Wiener-Pla. 2014. Literacy, pregnancy and potential oral health changes: The internet and readability levels. *Maternal and child health journal* 18, 3 (2014), 657–662.
[68] Xin Yan, Raymond Y.K. Lau, Dawei Song, Xue Li, and Jian Ma. 2011. Toward a semantic granularity model for domain-specific information retrieval. *ACM Transactions on Information Systems* 29, 3, Article 15 (July 2011), 46 pages.
[69] Andrew Yates and Nazli Goharian. 2013. ADRTrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In *European Conference on Information Retrieval*. Springer, 816–819.
[70] Qing Zeng, Eunjung Kim, Jon Crowell, and Tony Tse. 2005. A text corpora-based estimation of the familiarity of health terminology. *Biological and Medical Data Analysis* (2005), 184–192.
[71] Qing T Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association* 13, 1 (2006), 24–29.
[72] Qing Zeng-Treitler, Sergey Goryachev, Tony Tse, Alla Keselman, and Aziz Boxwala. 2008. Estimating consumer familiarity with health terminology: a context-based approach. *Journal of the American Medical Informatics Association* 15, 3 (2008), 349–356.
[73] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 435–444.
[74] W. Zhou, V. Torvik, and N. Smalheiser. 2006. ADAM: Another Database of Abbreviations in MEDLINE. *Bioinformatics* 22, 22 (2006), 2813–2818.
[75] Guido Zuccon. 2016. Understandability biased evaluation for information retrieval. In *European Conference on Information Retrieval*. Springer, 280–292.
[76] Guido Zuccon and Bevan Koopman. 2014. Integrating Understandability in the Evaluation of Consumer Health Search Engines. In *MedIR*.
[77] Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. 2016. 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. 15–27.