The effectiveness of query expansion when searching for health related content: InfoLab at CLEF eHealth 2016

Ricardo Silva[1] and Carla Teixeira Lopes[1,2]

[1] Department of Informatics Engineering, Faculty of Engineering, University of Porto, Portugal
[2] INESC TEC, Portugal
silva.ricardods@gmail.com,
ctl@fe.up.pt

Abstract. In this paper we describe the participation of InfoLab in the patient-centred information retrieval task of the CLEF eHealth 2016 lab. We analyse the performance of several query expansion strategies using different sources of terms and different methods to select the terms to be added to the original query. One of the strategies uses pseudo relevance feedback for term selection. The other strategies use external sources such as Wikipedia articles and definitions from the UMLS Metathesaurus for term selection. In the end, readability metrics such as SMOG, FOG and Flesch-Kincaid were used to re-rank the documents retrieved using the expanded queries. As the relevance and readability assessments are not available we can't make any conclusion regarding the results of our approaches.

Keywords: Health Information Retrieval, Query Expansion, Medical Text Indexer, Pseudo Relevance Feedback, Wikipedia, UMLS, Latent Dirichlet Allocation, Link Analysis, Readability

1 Introduction

A survey conducted by Susannah Fox [1] shows that 80% of internet users in the U.S. look online for health information, most frequently for information about a specific disease or medical problem. One of the most used techniques in this area of information retrieval is the query expansion. Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance [2]. This process might contribute to overcome one of the biggest difficulties for users in the search for health information: health consumers limited knowledge of medical terminology. This lack of knowledge influences the formulation of queries and the users expectations regarding the retrieved documents.

2 Methods

2.1 Task description

The 2016 CLEF eHealth Information Retrieval Task 3 Patient-Centred Information Retrieval[3] aims to evaluate systems that support people in understanding and searching for their health information [3]. This is possible by evaluating a ranked list of documents from the provided test collection as a response to patients queries. This is done through a TREC-style evaluation process, using the provided test collection. This task is split into three subTasks: ad-hoc search, query variation and multilingual search [4]. We participated in the ad-hoc search and query variation subTasks.

2.2 Document Collection

The collection for Task 3 is the ClueWeb12 B13 Dataset[4] [4]. This collection was generated by taking the 10 million ClueWeb09 URLs that had the highest PageRank scores, and then removing any page that was not in the top 90% of pages least likely to be spam, according to the Waterloo spam scores. These URLs were used as a starting point for a crawl which excluded any page that appeared in the blacklist provided by a URL blacklist service[5].

2.3 Queries

Queries from the Patient-Centred Information Retrieval task explore real health consumer posts from health web forums [4]. They were extracted from posts on the askDocs forum of Reddit, and presented to query generators. Query generators had to create queries based on what they read in the initial user post, making several variations for the same condition. The Linux program aspell was used to correct some misspellings on the English queries.

2.4 System description

The organization of the CLEF eHealth Lab used Terrier and Indri to index the provided collection [4]. In our work we chose to use Terrier. Terrier implements state-of-the-art indexing and retrieval functionalities, and provides an ideal platform for the rapid development and evaluation of large-scale retrieval applications. The Terrier index was provided to all the participants through a virtual machine on the Azure platform, doing so every team has access to same index allowing a easier comparison between different groups.

[3] `https://sites.google.com/site/clefehealth2016/task-3`
[4] `http://lemurproject.org/clueweb12/specs.php`
[5] `http://urlblacklist.com/`

3 Retrieval Approaches

Query expansion is the main approach used in this report. Query expansion (or term expansion) is the process of supplementing the original query with additional terms, and it can be considered as a method for improving retrieval performance [2]. Different sources and methodologies will be used to identify which terms will be added by the query expansion.

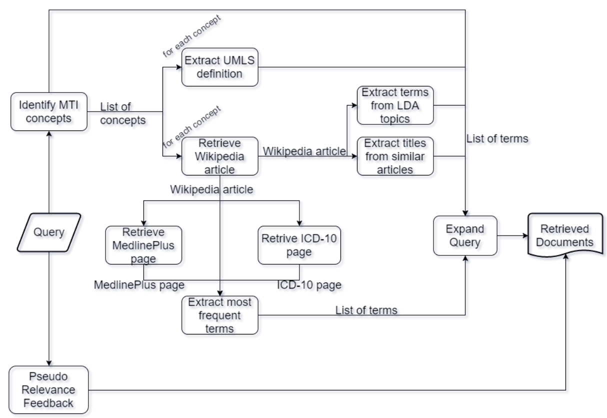

Fig. 1. Query expansion approaches.

In this section we present our approaches for Task 3: Patient-Centred Information Retrieval starting with the baseline, continuing with several query expansion methodologies and ending with the readability re-ranking. For all approaches we use the weighting model used in the baseline. In Figure 1 we present a summary of our query expansion approaches.

3.1 Baseline

We used BM25 term weighting model to score and rank medical documents. Okapi BM25 (BM stands for Best BM25 is a ranking function used by search engines to rank matching documents according to their relevance to a given search query. It is based on the probabilistic retrieval framework developed in the 1970s and 1980s by Stephen E. Robertson, Karen Sparck Jones, and others [5].

For a given query Q, the relevance score of a document D based on the BM25 term weighting model is expressed as:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{TF(q_i,D).(k1+1)}{TF(q_i,D)+k1.(1-b+b.\frac{|D|}{avgdl})} \tag{1}$$

where TF is the number of occurrences of a given term q_i in the document D. |D| the size of the document in words, *avgdl* the average size of a document. $k1$ anda b are free parameters, usually chosen, in absence of an advanced optimization, as [6]:

$$k1 \in [1.2; 2.0] \quad b = 0.75 \tag{2}$$

3.2 Pseudo Relevance Feedback

Pseudo Relevance Feedback is a method of query expansion that uses the document collection in which it runs as the source for its terms [2]. In this method the top documents returned by the baseline are used to modify the query by reweighting the existing query terms and by adding terms that appear useful and by deleting terms that do not. This process creates a new query which resembles the relevant documents more than the original query does [2].

Terrier provides two different models to apply the pseudo relevance feedback method: the Bose-Einstein and the Kullback-Leibler Divergence. The Bose-Einstein model calculates the weight of terms, as following [7]:

$$w(t) = tf_x.log_2(\frac{1 + P_n(t)}{P_n(t)}) + log_2(1 + P_n(t)) \tag{3}$$

$$P_n(t) = \frac{tf_c}{N} \tag{4}$$

where tf_x is the frequency of the query term t in the top-ranked documents, tf_c is the frequency of term t in the collection, and N is the number of documents in the collection [8]. The Kullback-Liebler Divergence computes the divergence between the probability distribution of terms in the whole collection and in the top ranked documents obtained using the original query [9]. The most likely terms to expand the query are those with a high probability in the top ranked set and low probability in the whole collection. For the term t this divergence is:

$$KLD(t) = [P_r(t) - P_c(t)].log \frac{\frac{f(t)}{NR}}{P_c(t)} \tag{5}$$

where $P_r(t)$ is the probability of t estimated from the top retrieved documents relative to a query (R). $P_c(t)$ is the probability of t estimated using the whole collection [8].

For this approach two runs were created to identify which one of this models provides better results.

3.3 Query expansion using the Medical Text Indexer

The NLM Medical Text Indexer (MTI) combines human NLM Index Section expertise and Natural Language Processing technology to curate the biomedical literature more efficiently and consistently. MTI is the main product of the

Indexing Initiative project and has been providing indexing recommendations based on the Medical Subject Headings (MeSH) vocabulary since 2002 [10].

Queries were processed by MTI which linked the text from the query to the MeSH vocabulary resulting in additional related concepts. The identified concepts are likely to be important for the retrieval process than other information in the query. However the MTI results are machine generated which, depending on the query, could result in irrelevant concepts.

In this approach, we appended all the concepts identified by the MTI to the original query.

3.4 Query expansion using Wikipedia

Wikipedia is a free encyclopedia, written collaboratively by the people who use it. Many people are constantly improving Wikipedia, making thousands of changes per hour[6]. This makes Wikipedia an enormous source of information likely to contain medical terms in lay language. As shown in the work of Laurent and Vickers [11], the English Wikipedia is a prominent source of online health information compared to other online health information providers like MedlinePlus. Using Wikipedia as a base, we defined two methods to get terms for the query expansion process. One of the methods extracts the most frequent terms from Wikipedia articles. The other uses Wikipedia as a directed graph to identify similar articles and then extracts terms from the titles of these articles.

Term Frequency The MediaWiki action API is a web service that provides a convenient access to wiki features, data, and meta-data over HTTP, via an URL[7]. This API was used to find the articles that best match the concepts obtained through the MTI.

An analysis of the obtained Wikipedia pages allowed us to identify some that were not health-related. To minimize this, we decided to exclude the pages not containing an infobox similar to the one presented in Figure 2[8] which contains information about the category of the page (e.g. anatomy, disease, drug).

This approach had several variants. We chose the 5, 10 and 15 most frequent terms of each article. In addition we considered (1) all articles found with the MTI concepts and (2) only the articles considered health-related using the strategy defined above.

Link Analysis As shown in the work of Almasari [12], Wikipedia is a hypertext network in which each article can refer to other Wikipedia article using hyperlinks. Considering only internal links, which are links that target other Wikipedia article it is possible to represent Wikipedia articles as a directed graph $G(A;L)$ of articles A connected by links L. L is the set of all the Incoming and Outgoing Links from the article A.

[6] `https://en.wikipedia.org/wiki/Wikipedia:Introduction`
[7] `https://www.mediawiki.org/wiki/MediaWiki`
[8] `https://en.wikipedia.org/wiki/Help:Infobox`

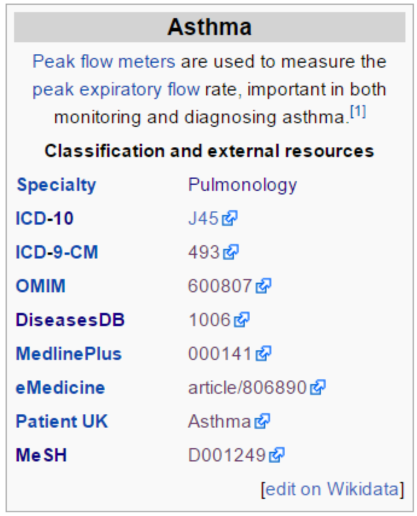

Fig. 2. Wikipedia Asthma Infobox.

Each concept from MTI was used to search for a Wikipedia article which served as a starting point. Using Wikipedia directed graph it is possible to retrieve the articles which referred and are referred by the first article. This method returns thousands of articles that aren't relevant to the expansion process because even if they're referred by the first article they might not be in the same category. To solve this issue we used the Jaccard similarity coefficient. This coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets [13]. The Incoming (I) and Outgoing (O) Links were used as the sets for the Jaccard coefficient.

$$J(I,O) = \frac{|I \bigcap O|}{|I \bigcup O|} \tag{6}$$

We used the Java Wikipedia Library (JWPL)[9] to compute this coefficient. The JWPL is a free, Java-based application programming interface that allows access to all the information in Wikipedia [14]. This library used a Wikipedia dump from March 2016.

For this approach we added to the original query all the titles of articles that had a Jaccard similarity coefficient greater than 0.25, 0.50 and 0.75. We considered one alternative using all articles found with the MTI concepts and another using only the articles considered health-related using the strategy defined above.

[9] `https://dkpro.github.io/dkpro-jwpl/`

3.5 Query expansion using MedlinePlus

MedlinePlus[10] is the National Institutes of Health Web site for patients, their families and friends. Produced by the National Library of Medicine, the worlds largest medical library, it brings information about diseases, conditions, and wellness issues in lay language.

Using the information on the infobox (Figure 2), obtained through the search of the MTI concepts on the Wikipedia, it is possible to access the corresponding MedlinePlus page. MedlinePlus pages are generally splitted in different sections with relevant information about the searched concept. The sections that were considered most relevant for the query expansion process were the Causes, Symptoms, Treatment, Possible Complications and Alternative Names sections.

We testes several variants of this method. We chose the top 5, 10 and 15 most frequent terms on each of the above mentioned sections. The structure and size of the Alternative Names section made us include all the terms included in this section. We also made a run with only the terms from the Alternative Names.

3.6 Query expansion using the ICD-10

ICD-10[11] is the 10th revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases. The code set allows more than 14,400 different codes and permits the tracking of many new diagnoses.

The approach using ICD-10 is similar to the one used in MedlinePlus taking advantage on the contents of the infobox from Wikipedia. The ICD-10 page contains not only information about the search concept but also about other diseases or symptoms related to the initial concept displayed in a hierarchy. This related information is used as a source of terms to use on the query expansion process.

We chose the top 5, 10 and 15 most frequent terms of an ICD-10 page to append to the original query.

3.7 Query expansion using Latent Dirichlet Allocation over Wikipedia

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words [15]. Figure 3 shows an example of LDA applied to a text, generating four topics for the text and a set of words which represent each topic. These distributions seem to capture some of the underlying topics in the corpus. Each word on the text

[10] `https://www.nlm.nih.gov/medlineplus/aboutmedlineplus.html`
[11] `http://www.who.int/classifications/icd/en/`

tends to peak towards one of the possible topic values, the words are color coded according to the topics they represent.

For this approach an implementation of LDA for Java (JGibbLDA) was used to generate topics from texts related to each query. These texts were the Wikipedia articles obtained through the MediaWiki API using the MTI concepts as search terms.

In this approach we tried different combinations of number of topics and number of words to to identify which one contributes the most. We chose a combination of 3 topics with 1, 5 and 10 words, and 1, 5 and 10 topics with 5 words.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Fig. 3. An example article from the TREC AP corpus. Each color codes a different factor from which the word is putatively generated [15].

3.8 Query expansion using the Unified Medical Language System

The Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine. The UMLS integrates over 2 million names for 900,000 concepts from more than 60 families of biomedical vocabularies, as well as 12 million relations among these concepts. Vocabularies integrated in the UMLS Metathesaurus include the NCBI taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), OMIM and the Digital Anatomist Symbolic Knowledge Base.

The UMLS Metathesaurus has a REST API that provides access to its information. Using this API it is possible to extract terms to be used in the expansion using the definitions in UMLS related to the MTI concepts.

We chose the top 5, 10 and 15 most frequent terms on the UMLS definitions to be added to the original query.

3.9 Readability

Readability metrics measure the difficulty of understanding a passage of text. Readability metrics are often based on features such as the average number of syllables per word, and words per sentence. These features ignore concept difficulty and are based on assumptions about writing style that may not hold in all environments.

There are many readability metrics. SMOG, FOG and Flesch-Kinciad are three of the most widely used readability metrics. They all estimate the educational grade level necessary to understand a document [16].

SMOG Readability Formula estimates the years of education a person needs to understand a piece of writing. McLaughlin created this formula as an improvement over other readability formulas [17] and defined it as:

$$SMOG = 3 + \sqrt{Number\ Of\ Polysyllable\ Words\ In\ 30\ Sentences} \quad (7)$$

If the document is longer than 30 sentences, the first 10 sentences, the middle 10 sentences, and the last 10 sentences are used. If the document has fewer than 30 sentences, some rules and a conversion table are used to calculate the grade level. The SMOG measure tends to give higher values than other readability metrics [17].

The Gunning Fog Index Readability Formula, or simply called FOG Index, is attributed to American textbook publisher, Robert Gunning. Gunning observed that most high school graduates were unable to read. Much of this reading problem was a writing problem. His opinion was that newspapers and business documents were full of fog and unnecessary complexity. In 1952, Gunning created an easy-to-use Fog Index defined as:

$$FOG = 0.4\ (ASL + PHW) \quad (8)$$

Where ASL is the Average Sentence Length (number of words divided by the number of sentences) and PHW is the Percentage of Hard Words (words of three or more syllables). The FOG metric is considered suitable for secondary and older primary age groups.

The Flesch-Kincaid readability metric was developed under contract to the U.S. Navy in 1975 by Rudolph Flesch and John Peter Kincaid. This Flesch-Kincaid formula was first used by the U.S. Army for assessing the difficulty of technical manuals in 1978 and soon after became the Department of Defense military standard. Pennsylvania was the first U.S. state to require that automobile insurance policies be written at no higher than a ninth-grade level of reading difficulty, as measured by the Flesch-Kincaid formula. This is now a common requirement in many other states and for other legal documents such as insurance policies[12]. The Flesch-Kincaid formula is defined as:

$$FK = (0.39 * ASL) + (11.8 * ASW) - 15.59 \tag{9}$$

where ASL is the Average Sentence Length (the number of words divided by the number of sentences) and the ASW is the Average number of Syllable per Word (the number of syllables divided by the number of words).

A re-rank method was developed to combine the readability metrics with the relevance scores from Terrier. Three different formulas were used to combine these values. The last formula was proposed by Zuccon and Koopman [18], where an user is characterized by a readability threshold th and every document that has a readability score below th is considered readable, while documents with readability above th are considered unreadable.

$$Score = Relevance/Readability \tag{10}$$

$$Score = Relevance * log(\frac{1}{Readability}) \tag{11}$$

$$Score = Relevance * (\frac{1}{2} - \frac{\arctan(Readability - th)}{\pi}) \tag{12}$$

In this approach we re-rank each run with one of the readability metrics and one of the above combination formulas (10, 11 and 12), which generates a total of 9 different variants for each of the previous runs.

4 Results

We participated in two subTasks: ad-hoc search and query variation. For the first subTask each query was treated individually ignoring all the query variations. For the second subTask each group of query variations was treated as one query. In this section we present the evaluation results of each approach compared with the baseline for the two subTasks. We also show the evaluation results for the submitted runs.

4.1 Submitted Runs

For Task 3: Patient-Centred Information Retrieval we could only submit up to 3 runs with the first one being the baseline run. For the other runs we chose to submit the Wikipedia Link Analysis run with a Jaccard similarity coefficient above 0.50 using health-related Wikipedia articles and the Latent Dirichlet Allocation run with 3 topics and 5 words. This runs were re-ranked using the SMOG readability metric and the formula 12.

[12] http://www.readabilityformulas.com/flesch-grade-level-readability-formula.php

Table 1. Evaluation of the submitted runs for subTask1.

Run	P@5	P@10	NDCG@5	NDCG@10	RBP	uRBP
(Run 1)Baseline	0.00	0.00	0.00	0.00	0.00	0.00
(Run 2)WikiLA Sim 0.50 Med	0.00	0.00	0.00	0.00	0.00	0.00
(Run 3)LDA T3 W5	0.00	0.00	0.00	0.00	0.00	0.00

Table 2. Evaluation of the submitted runs for subTask2.

Run	P@5	P@10	NDCG@5	NDCG@10	RBP	uRBP
(Run 1)Baseline	0.00	0.00	0.00	0.00	0.00	0.00
(Run 2)WikiLA Sim 0.50 Med	0.00	0.00	0.00	0.00	0.00	0.00
(Run 3)LDA T3 W5	0.00	0.00	0.00	0.00	0.00	0.00

5 Conclusion

At the time of writing this document the relevance and readability assessments for this test collection are not available. Because of this we can't make any conclusion regarding the results of our approaches.

6 Future Work

In future work, we will continue to explore new query expansion models to find an effective way of supporting patients to find useful medical information, and improve the current ones like the Latent Dirichlet Allocation that could be applied to other sources of texts such as the ones found through the Wikipedia Link Analysis or the MedlinePlus pages. In addition, we would like to incorporate readability metrics that weren't based on sentence lengths or polysyllabic words but were based in concepts.

References

1. Susannah Fox. Health topics: 80% of internet users look for health information online. *Pew Internet & American Life Project*, February 2011.
2. Efthimis N. Efthimiadis. Query expansion. *Annual Review of Information Systems and Technologys*, 31:121–187, 1996.
3. Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Aurlie Nvol, Joao Palotti, and Guido Zuccon. Overview of the clef ehealth evaluation lab 2016. *CLEF 2016 - 7th Conference and Labs of the Evaluation Forum, Lecture Notes in Computer Science (LNCS), Springer*, September 2016.
4. Guido Zuccon, Joao Palotti, Lorraine Goeuriot, Liadh Kelly, Mihai Lupu, Pavel Pecina, Henning Mueller, Julie Budaher, and Anthony Deacon. The ir task at the clef ehealth evaluation lab 2016: User-centred health information retrieval. *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, September 2016.
5. S. E. Robertson and K. Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.
6. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
7. Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002.
8. Fangmei Lu. Employing query expansion models to help patients diagnose themselves. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*, 2015.
9. Hazra Imran and Aditi Sharan. Selecting effective expansion terms for better information retrieval. *IJCSA*, 7(2):52–64, 2010.
10. Alan R. Aronson James G. Mork, Antonio J. Jimeno Yepes. The nlm medical text indexer system for indexing biomedical literature. 2013.
11. Michaël R. Laurent and Tim J. Vickers. Seeking health information online: Does wikipedia matter? *Journal of the American Medical Informatics Association*, 16(4):471–479, 2009.
12. Mohannad Almasri, Catherine Berrut, and Jean-Pierre Chevallet. Exploiting wikipedia structure for short query expansion in cultural heritage. In Marie-Francine Moens, Christian Viard-Gaudin, Hafa Zargayouna, and Oriol Ramos Terrades, editors, *CORIA-CIFED*, pages 287–302. ARIA-GRCE, 2014.
13. Loet Leydesdorff. On the normalization and visualization of author co-citation data: Salton's cosine versus the jaccard index. *Journal of the American Society for Information Science and Technology*, 59(1):77–85, 2008.
14. Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May 2008. electronic proceedings.
15. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.
16. Luo Si and Jamie Callan. A statistical model for scientific readability. In *Proceedings of the Tenth International Conference on Information and Knowledge Management*, CIKM '01, pages 574–576, New York, NY, USA, 2001. ACM.
17. Harry Mc Laughlin. SMOG Grading-a New Readability Formula. *Journal of Reading*, 12(8), 1969.

18. Guido Zuccon and Bevan Koopman. Integrating understandability in the evaluation of consumer health search engines. In *Medical Information Retrieval (MedIR) Workshop*, Gold Coast, Australia, July 2014.