Retrieval of Health Advice on the Web
AEHRC at ShARe/CLEF eHealth Evaluation Lab Task 3

G. Zuccon[1], B. Koopman[1,2], A. Nguyen[1]

[1] The Australian e-Health Research Centre (CSIRO), Brisbane, Australia
[2] Queensland University of Technology, Brisbane, Australia
{guido.zuccon, bevan.koopman, anthony.nguyen}@csiro.au

Abstract. This paper details the participation of the Australian e-Health Research Centre (AEHRC) in the ShARe/CLEF 2013 eHealth Evaluation Lab – Task 3. This task aims to evaluate the use of information retrieval (IR) systems to aid consumers (e.g. patients and their relatives) in seeking health advice on the Web.
Our submissions to the ShARe/CLEF challenge are based on language models generated from the web corpus provided by the organisers. Our baseline system is a standard Dirichlet smoothed language model. We enhance the baseline by identifying and correcting spelling mistakes in queries, as well as expanding acronyms using AEHRC's Medtex medical text analysis platform. We then consider the readability and the authoritativeness of web pages to further enhance the quality of the document ranking. Measures of readability are integrated in the language models used for retrieval via prior probabilities. Prior probabilities are also used to encode authoritativeness information derived from a list of top-100 consumer health websites.
Empirical results show that correcting spelling mistakes and expanding acronyms found in queries significantly improves the effectiveness of the language model baseline. Readability priors seem to increase retrieval effectiveness for graded relevance at early ranks (nDCG@5, but not precision), but no improvements are found at later ranks and when considering binary relevance. The authoritativeness prior does not appear to provide retrieval gains over the baseline: this is likely to be because of the small overlap between websites in the corpus and those in the top-100 consumer-health websites we acquired.

1 Introduction

Patients usually have limited medical knowledge and thus patient education can improve a patients understanding of their health condition, as well as adherence and compliance to a treatment. The use of web search engines to retrieve medical advice online is increasingly popular [1–3]. The ShARe/CLEF 2013 eHealth Evaluation Lab Task 3 [4] aims to evaluate search engines in the task of retrieving health advice on the web, and to uncover issues that the research community can address to improve the effectiveness of search engines technologies for this task.

The Australian e-Health Research Centre (AEHRC) contributed 4 runs to this year's challenge. Our methods are based on the language modelling framework for information retrieval [5]. Our baseline submission (teamAEHRC.1.3) implements a language model with Dirichlet smoothing [6]. The remaining submissions build upon this baseline approach. Specifically, we consider the contribution to retrieval effectiveness of query error correction and acronym expansion (teamAEHRC.5.3), readability measures (teamAEHRC.6.3), authoritativeness as derived from a list of top-100 consumer health websites (teamAEHRC.7.3). Our approaches are detailed in Section 2.

Empirical results obtained over the web corpus compiled by the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3 organisers highlight the importance of correcting typographic errors in health consumer queries, as well as normalising acronyms to their expanded form to increase the quality of the query representation. Considering the readability of web pages when providing information to health consumers provides improvements in retrieval effectiveness when considering the graded relevance results at early ranks. Web page authority as assessed from a list of top-100 consumer-health websites does not seem to improve retrieval quality: this may be due to the limited overlap between the website list and the ShARe/CLEF 2013 document corpus. Details of the results achieved by our submissions are given in Section 3.

2 Methods

The next sections describes the methods we used to address the problem of retrieving web pages for health consumers seeking for medical advice [3]. The document rankings generated by our methods were submitted to the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3. All methods are implemented using the Lemur/Indri information retrieval toolkit[1].

2.1 A baseline Language Model (teamAEHRC.1.3)

We used a language modelling approach with Dirichlet smoothing as the baseline retrieval method. Following this approach, documents are ranked according to the probability of a document d given the submitted query Q, i.e. $P(d|Q)$, computed as:

$$P(d|Q) \approx P(Q|d)P(d) \approx \prod_{q_i \in Q} \frac{P(q_i|d) + \mu P(q_i|C)}{|d| + \mu} \tag{1}$$

where the prior probability $P(d)$ is considered uniform over the document collection and can thus be ignored for ranking equivalence reasons, $|d|$ is the length in tokens of document d, $P(q_i|C)$ is the maximum likelihood estimate of q_i in the collection, and μ is the Dirichlet smoothing parameter. This parameter was set to 2,500 in all our submission; this is a common value for the smoothing

[1] http://lemurproject.org/

parameter. In our implementation (as it is common in IR), only documents that contain at least one of the query terms are considered for retrieval for each given query. Of these, only the top 1,000 documents, according to Equation 1, are used to form the submission.

2.2 Correcting spelling mistakes and expanding acronyms (teamAEHRC.5.3)

The analysis of the training set provided by the task organisers revealed that queries may contain (i) spelling mistakes or alternative spellings, e.g., groupo B for group B; (ii) acronyms and abbreviations, e.g., Cdiff for Clostridium difficile. Similar cases can in fact also be found in the test set, e.g. Hypothyreoidism for Hypothyroidism, ASA for acetylsalicylic acid (aspirin)[2]. The presence of spelling mistakes and the use of acronyms in queries may adversely affect retrieval of a standard keyword-based system like ours based on language modelling.

To overcome this issue, we use the AEHRC's Medtex medical text analysis platform [7] to individuate misspelled terms (and uncommon variants of medical terms), as well as acronyms. Medtex is part of the medical natural language processing technology that the AEHRC uses to deliver automated solutions to improve health service delivery, like cancer incidence and mortality reporting [8, 9] and radiology reconciliation [10].

To correct candidate misspellings and uncommon variants individuated by Medtex, we implemented a call to the Google web search engine[3] and extracted the query correction suggestion (i.e. "Showing results for") provided by the search engine.

To expand candidate acronyms individuated by Medtex, we parsed the list of common abbreviations used in medical prescriptions provided by Wikipedia[4]. This list contains triples <*abbreviation, latin, meaning*>, where *abbreviation* is the target acronym or abbreviation expression, *latin* is the Latin term that represents the abbreviation (if any) and *meaning* is the English expansion. If an English expansion was available for an abbreviation, then we ignored the latin term, otherwise we used the latin term as a translation of the abbreviation.

When spelling corrections and acronym expansions are produced for candidate terms of a query, we create a new query formulation that appends to the original query terms those from the spelling correction and acronym expansion. The new query formulation is then used to score documents against using the model of Equation 1. This forms the submission named teamAEHRC.5.3.

---

[2] Note, however, that the test set was not consulted when developing the approach described here.

[3] `http://www.google.com`

[4] `http://en.wikipedia.org/wiki/List_of_abbreviations_used_in_medical_prescriptions`

2.3 Taking readability into account: the readability prior (teamAEHRC.6.3)

Health consumers seeking medical advice on the web do not usually have expertise in the medical domain and are thus not familiar with the medical language. A page providing health information for health practitioners (e.g., doctors and nurses) is likely to be difficult to read for a health consumer, such as a patient. We follow this intuition and argue that web pages retrieved for providing advice to health consumer need to be easily understood by a non-expert reader. We thus enhance the approach used for building the submission teamAEHRC.5.3 (previous section) by considering document readability. We use a common measure of text readability to estimate how likely the content of a web page is understandable by health consumers. The selected readability measure is the Flesch Kincaid Reading Ease (FRES) formula [11]. This measure provides a score between 0 and 100. A high score indicates that the text is easy to read, while low scores suggest the text is complicated to understand. The Flesch Kincaid Reading Ease measure has been used in previous work on readability of health content, for example, to assess whether informed consent forms for participation in oncology research are readable by patients and their families [12]. The Flesch Kincaid Reading Ease is calculated according to the following formula:

$$206.835 - 1.015 \cdot \frac{\#(words)}{\#(sentences)} - 84.6 \cdot \frac{\#(syllable)}{\#(words)} \tag{2}$$

where the function $\#(x)$ provides the total count of item x in the document, e.g. #(syllable) is the total number of syllable in the document.

To consider the readability measure during the retrieval process, we compute a prior probability distribution over all documents in the collection, where the value of the prior probability assigned to a document is proportional to its Flesch Kincaid Reading Ease score. Thus, documents that are more readable according to the Flesch Kincaid Reading Ease measure would be more likely relevant according to our prior. The prior is integrated in the retrieval formula by modifying Equation 1 so that the readability prior is substituted to the uniform prior, $P(d)$, used for the previous runs. This method forms the submission named teamAEHRC.6.3.

2.4 Considering authoritativeness: a prior for the top-100 consumer health websites (teamAEHRC.7.3)

Health information presented to consumers should not only be easy to understand, but also reliable. To take reliability of the content into account during the retrieval process, we obtained a list of recommended health-consumer web sites. The list has been compiled by CAPHIS[5] and can be retrieved at `http://caphis.mlanet.org/consumer/`. This list contains 100 sites that have

[5] The Consumer and Patient Health Information Section (CAPHIS) is part of the Medical Library Association, an association of health information professionals.

been selected according to criteria that include currency, credibility, content and audience. We see these criteria as an overall measure of how authoritative the websites are.

To integrate authoritativeness information in the retrieval process, we took an approach similar to that used for the document prior in run teamAEHRC.6.3. An uniform prior was computed for all documents in the collection. For documents whose base URL is in the CAPHIS list, we boosted the corresponding prior by 10 times. The resulting score distribution was then normalised to resemble a probability distribution, this formed our authoritativeness prior. The prior was then applied together with the readability prior, transforming the retrieval formula to the following:

$$P(d|Q) \approx P(Q|d)P_r(d)P_a(d) \approx \prod_{q_i \in Q} P_r(d)P_a(d)\frac{P(q_i|d) + \mu P(q_i|C)}{|d| + \mu} \quad (3)$$

where $P_r(d)$ is the readability prior for document d and $P_a(d)$ is the authoritativeness prior. This formed the method used for generating teamAEHRC.7.3 submission.

3 Evaluation on the ShARe/CLEF 2013 Challenge

3.1 Evaluation Settings

Details of the collection used are given by Goeuriot et al. [3]. We indexed the document collection using Lemur; the INQUIRY stop list and the Porter stemmer as implemented in Lemur were used when indexing documents.

Table 1. Percentage indicating the coverage of the relevance assessments with respect to the top 10 results retrieved by each of our submissions for each query in the ShARe/CLEF 2013 retrieval challenge. A low percentage for a submission suggests that metric such as P@10 and nDCG@10 may be underestimated for that submission.

Run id	% coverage
teamAEHRC.1.3	100.00%
teamAEHRC.5.3	100.00%
teamAEHRC.6.3	71.40%
teamAEHRC.7.3	43.60%

Runs are evaluated according to the guidelines provided by the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3 organisers; Precision@10 and nDCG@10 are used as main evaluation measures. Organisers formed the pools used for relevance assessments by considering, for each query, the top 10 documents retrieved by only selected participants submissions. The selected submission include our teamAEHRC.1.3 and teamAEHRC.5.3 runs; thus the set of top 10

documents retrieved for each query by teamAEHRC.6.3 and teamAEHRC.7.3 may contain unjudged documents. Unjudged documents are considered irrelevant (according to standard IR practice); this may result in an underestimation of the Precision@10 and nDCG@10 for teamAEHRC.6.3 and teamAEHRC.7.3. Table 1 reports the percentage of the top 10 rank positions of each query for our submissions that are covered by the ShARe/CLEF 2013 relevance assessments. The table highlights that the top 10 rankings of each queries for submissions teamAEHRC.6.3 and teamAEHRC.7.3 have only partially been assessed: less than half of the documents retrieved by teamAEHRC.7.3 in the top 10 ranks have been judged. Thus, metrics such as precision and nDCG calculated at rank 10 for these submissions may be underestimating the quality of the submission itself. The extent to which the effectiveness of these runs are underestimated depends upon the breadth of the pools assessed in the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3.

Table 2. Official retrieval effectiveness of our approaches obtained on the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3.

(a) Retrieval effectiveness measured by Precision at rank 5 and 10 (P@5, P@10).

Run id	P@5	P@10
teamAEHRC.1.3	0.4440	0.4540
teamAEHRC.5.3	0.4560	0.4840
teamAEHRC.6.3	0.4440	0.4220
teamAEHRC.7.3	0.2080	0.2200

(b) Retrieval effectiveness measured by normalised discounted cumulative gain at rank 5 and 10 (nDCG@5, nDCG@10).

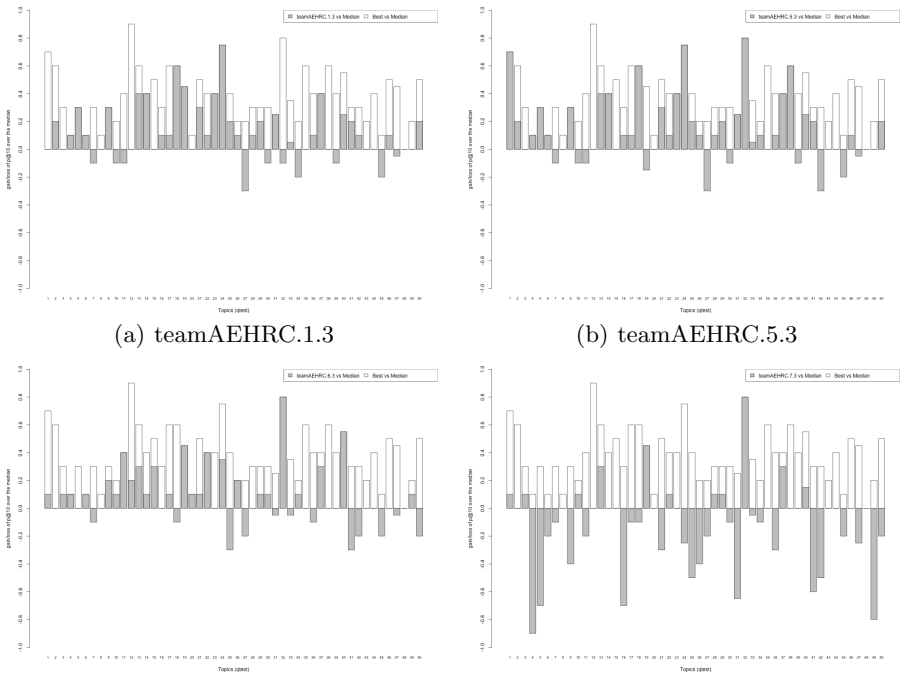
Run id	nDCG@5	nDCG@10
teamAEHRC.1.3	0.3825	0.3991
teamAEHRC.5.3	0.3946	0.4223
teamAEHRC.6.3	0.4128	0.3992
teamAEHRC.7.3	0.1937	0.1992

3.2 Results

Results of our submissions are summarised in Tables 2(a) and 2(b) for Precision@10 and nDCG@10 respectively.

The results suggest that expanding queries by including spelling corrections, common alternative spelling of medical terms, acronym and abbreviation expansions (submission teamAEHRC.5.3), does sensibly improve retrieval effectiveness across the whole set of evaluation metrics. The inclusion of readability measures as a prior of the language model used for retrieval does not provide the hoped improvement when considering binary relevance and precision. When graded relevance is considered, accounting for the readability of the content does provide better effectiveness at early ranks, as measured by the increase of nDCG@5 obtained by teamAEHRC.6.3 over the baseline teamAEHRC.1.3 (+8%) and teamAEHRC.5.3 (+5%). These gains are, however, not maintained at lower ranks. The inclusion of the authoritativeness prior actually degrades the retrieval effectiveness; this is likely to be a result of the limited overlap between the URLs in the CAPHIS list and those in the ShARe/CLEF 2013 corpus. An additional factor influencing the low performance of this run may be the limited overlap between retrieved and assessed documents. As discussed in Section 3.1, this factor

may cause an underestimation of the retrieval effectiveness of this submission, the extent of which can only be judged based on the completeness and breadth of the pool of documents and systems that have been assessed.



(a) teamAEHRC.1.3 (b) teamAEHRC.5.3

(c) teamAEHRC.6.3 (d) teamAEHRC.7.3

Fig. 1. Gains and losses in precision at 10 of our submissions with respect to the median system at ShARe/CLEF 2013. White bars identify the gains of the best system for each query.

Figure 1 reports the gains and the losses in precision at 10 of our submissions when compared to the median system at ShARe/CLEF 2013. These are compared to the highest gains measured on a query-basis in the ShARe/CLEF 2013 challenge.

Submission teamAEHRC.5.3 provides large gains over teamAEHRC.1.3 (and the median of ShARe/CLEF 2013 systems) in queries where error corrections and acronym expansion are fundamental. For example, the system used for teamAEHRC.1.3 retrieved only two documents for query qtest1, *'hypothyreoidism'*; this is because only two documents contain this uncommon spelling of this condition. Our query terms correction method added to the query the term *'hypothyroidism'*, which allowed for the retrieval of a large quantity of relevant documents. The submission formed by the system implementing our query terms correction (and acronym expansion) technique, i.e. teamAEHRC.5.3, performed

best for queries like query qtest1. Overall, the gains in retrieval effectiveness over the baseline provided by the technique used for teamAEHRC.5.3 are partially lost when introducing the readability prior in the retrieval method. However, the query-by-query analysis of Figure 1 highlights that losses do not affect all queries: the introduction of the readability prior delivers effectiveness gains for some of the queries in the ShARe/CLEF 2013 test set. This is the case for example for queries qtest10, *'dysplasia and multiple sclerosis'*, and qtest11, *'chest pain and liver transplantation'*. In the latter case, the system that uses the readability prior for retrieval results is, in fact, the best across all ShARe/CLEF 2013 systems.

Finally, we have observed that our submissions teamAEHRC.6.3 have no results for qtest50, and teamAEHRC.7.3 has no results for qtest49 and qtest50. An analysis of our code revealed an error in the generation of the final document rankings to submit to the challenge; this produced truncated results for submissions teamAEHRC.6.3 and teamAEHRC.7.3.

4 Conclusions

This paper outlined our submissions to the ShARe/CLEF 2013 eHealth Evaluation Lab Task 3. Our approaches are based on a Dirichlet smoothing language modelling framework, and investigate the effect of misspelling corrections and acronyms expansions in queries, and the inclusion of readability and authoritativeness information in the scoring function. A preliminary analysis of our results have highlighted the gains that can be obtained by correcting misspellings in queries and expanding acronyms and abbreviations. The inclusion of readability information have shown promise for retrieving information for health consumers seeking medical advice on the web. Further analysis is required to better gauge the impact of our approaches. Future work will also investigate alternative approaches for (i) computing readability of health content, (ii) encoding readability in the retrieval function.

References

1. Fox, S.: Health topics: 80% of internet users look for health information online. Technical report, Pew Research Center (February 2011)
2. White, R., Horvitz, E.: Cyberchondria: Studies of the escalation of medical concerns in web search. Technical report, Microsoft Research (2008)
3. Goeuriot, L., Kelly, L., Jones, G., Zuccon, G., Suominen, H., Hanbury, A., Muller, H., Leveling, J.: Creation of a new medical information retrieval evaluation benchmark targeting patient information needs. In: 5th International Workshop on Evaluating Information Access (EVIA). (2013)
4. Suominen, H., Salanterä, S., Velupillai, S., Chapman, W.W., Savova, G., Elhadad, N., Mowery, D., Leveling, J., Goeuriot, L., Kelly, L., Martinez, D., Zuccon, G.: Overview of the ShARe/CLEF eHealth Evaluation Lab 2013. In: CLEF 2013. Lecture Notes in Computer Science (LNCS), Springer (2013)

5. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM (1998) 275–281
6. Zhai, C., Lafferty, J.: A study of smoothing methods for language models applied to information retrieval. ACM Transactions on Information Systems (TOIS) **22**(2) (2004) 179–214
7. Nguyen, A., Moore, J., Lawley, M., Hansen, D., Colquist, S.: Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. In: Health Informatics Conference. (2011) 117–124
8. Zuccon, G., Nguyen, A., Bergheim, A., Wickman, S., Grayson, N.: The impact of ocr accuracy on automated cancer classification of pathology reports. In: Health Informatics: Building a Healthcare Future Through Trusted Information-Selected Papers from the 20th Australian National Health Informatics Conference (Hic 2012). Volume 178., Ios PressInc (2012) 250
9. Butt, L., Zuccon, G., Nguyen, A., Bergheim, A., Grayson, N.: Classification of cancer-related death certificates using machine learning. Australasian Medical Journal **6**(5) (2013) 292–299
10. Zuccon, G., Wagholikar, A., Nguyen, A., Chu, K., Martin, S., Lai, K., Greenslade, J.: Automatic classification of free-text radiology reports to identify limb fractures using machine learning and the snomed ct ontology. In: AMIA Clinical Research Informatics. (2013)
11. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, DTIC Document (1975)
12. Grossman, S.A., Piantadosi, S., Covahey, C.: Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? Journal of Clinical Oncology **12**(10) (1994) 2211–2215