Learning to Combine Representations for Medical Records Search

Nut Limsopatham[1], Craig Macdonald[2], Iadh Ounis[2]
nutli@dcs.gla.ac.uk[1], firstname.lastname@glasgow.ac.uk[2]

School of Computing Science
University of Glasgow, Glasgow, UK

ABSTRACT

The complexity of medical terminology raises challenges when searching medical records. For example, 'cancer', 'tumour', and 'neoplasms', which are synonyms, may prevent a traditional search system from retrieving relevant records that contain only synonyms of the query terms. Prior works use *bag-of-concepts* approaches, to deal with this by representing medical terms sharing the same meanings using concepts from medical resources (e.g. MeSH). The relevance scores are then combined with a traditional bag-of-words representation, when inferring the relevance of medical records. Even though the existing approaches are effective, the predicted retrieval effectiveness of either the bag-of-words or bag-of-concepts representation, which may be used to effectively model the score combination and hence improve retrieval performance, is not taken into account. In this paper, we propose a novel learning framework that models the importance of the bag-of-words and the bag-of-concepts representations, combining their scores on a per-query basis. Our proposed framework leverages retrieval performance predictors, such as the clarity score and AvIDF, calculated on both representations as learning features. We evaluate our proposed framework using the TREC Medical Records track's test collections. As our proposed framework can significantly outperform an existing approach that linearly merges the relevance scores, we conclude that retrieval performance predictors can be effectively leveraged when combining the relevance scores.

Categories and Subject Descriptors: H.3.3 [Information Search & Retrieval]: Search process

General Terms: Experimentation, Performance

Keywords: Medical Records Search; Regression; Controlled Vocabulary; Retrieval Performance Predictors

1. INTRODUCTION

Medical terminology, which can be complex, inconsistent, and ambiguous, poses an important challenge when searching in the medical domain [9, 10, 12, 15, 16]. For example, 'heart disease' can be referred to as 'coronary artery disease', 'coronary heart disease', or 'CHD'. This means that traditional search systems may not be able to retrieve medical documents relevant to a query, if those documents contain only synonyms of the query terms. To tackle this, prior works (e.g. [2, 8]) proposed *bag-of-concepts* (BoC) approaches to represent medical documents and queries using concepts from medical resources, such as MeSH[1] and UMLS Metathesaurus[2]. Under these approaches, 'heart disease', 'coronary artery disease', 'coronary heart disease', and 'CHD', which share the similar meaning, are represented with the same concept. For instance, Aronson [2] deployed MetaMap [3] to identify medical concepts in medical records and queries and represented them in the form of the UMLS Concept Unique Identifiers (CUIs). Intuitively, such approaches should alleviate the terminology mismatch problem. However, empirical studies [15, 16] have shown that the BoC performance can be inconsistent, sometimes underperforming the traditional bag-of-words representation (BoW), since not all documents and queries could be effectively represented using medical concepts. For example, medical concepts may not be found in some queries. To cope with such a challenge, other works (e.g. [9, 15, 16]) combined the relevance scores of both BoW and BoC when inferring the relevance of a document. In particular, Srinivasan [15] proposed the so-called *score combination* approach that linearly combines the relevance scores from both BoW and BoC, when inferring the relevance of a document d towards a query Q, as follows [15]:

$$score(d, Q) = \delta \cdot score_{BoW}(d, Q) + score_{BoC}(d, Q) \tag{1}$$

where δ is a parameter to emphasise the relevance score computed using BoW, which is set to 2.00 for all queries, as suggested in [9, 15].

In the context of medical records search, Limsopatham et al. [9] improved retrieval performance markedly by using the aforementioned score combination to merge the relevance scores from the BoW and their proposed task-specific representation (i.e. a BoC). They showed that combining the relevance scores from BoW and BoC is effective for searching in the medical domain. Importantly, these score combination approaches merge the relevance scores computed from both BoW and BoC representations by fixing a particular weight irrespective of the query.

We hypothesise that by learning a weight for BoW and BoC on a per-query basis, we can rank medical records more effectively. In this paper, we propose a novel learning

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

[1] http://www.ncbi.nlm.nih.gov/mesh

[2] http://www.nlm.nih.gov/research/umls/

framework to model the importance of BoW and BoC, when inferring the relevance of a medical record. Our proposed regression-based learning framework leverages retrieval performance predictors, such as the clarity score [5] and query scope [7], computed on both BoW and BoC as features, to learn an effective combination model on a per-query basis.

We evaluate our proposed framework in the context of the TREC 2011 [19] and 2012 [18] Medical Records track. Our results show that our learning framework is effective. Indeed, it significantly outperforms an existing strong score combination baseline.

The main contributions of this paper are threefold:

1. We show that some particular queries benefit more from a bag-of-words (BoW) representation, while the others profit from a bag-of-concepts (BoC) representation.
2. We propose a novel regression-based learning framework to model the importance of BoW and BoC using retrieval performance predictors, when inferring the relevance of medical records.
3. We thoroughly evaluate our proposed framework using standard collections provided by the TREC 2011 and 2012 Medical Records track.

The remainder of this paper is structured as follows. Section 2 introduces our novel regression framework that leverages retrieval performance predictors to learn an effective score combination model. Our experimental setup and results are presented in Sections 3 and 4. Finally, we provide concluding remarks in Section 5.

2. OUR PROPOSED FRAMEWORK

In this section, we describe our novel learning framework that models the combination of the relevance scores from the bag-of-words (BoW) and the bag-of-concepts (BoC) representations, for medical records search. The central idea is that queries may benefit differently from BoW and BoC; hence, we propose to learn a weight for BoW and BoC on a per-query basis. To do so, we use retrieval performance predictors as learning features to estimate the predicted retrieval effectiveness of each representation, when estimating the relevance scores of a medical record. In particular, we deploy a regression technique to learn the importance of the two representations when combining their relevance scores.

Our framework consists of four components:

1. A score combination model.
2. A procedure to estimate the model parameter for a query.
3. A set of learning features to learn the model.
4. A regression procedure to infer the model using the learning features.

In the remainder of this section, we describe each of these four components.

2.1 A Score Combination Model

To take advantage of both BoW and BoC, we follow [15] and combine the relevance scores of a medical record d towards a query Q as follows:

$$\begin{aligned} score(d,Q) = &\lambda_Q \cdot score_{BoW}(d,Q) \\ &+ (1-\lambda_Q) \cdot score_{BoC}(d,Q) \end{aligned} \tag{2}$$

where λ_Q ($0 \leq \lambda_Q \leq 1$) is a per-query parameter to estimate the importance of the relevance scores computed using the bag-of-words (BoW) and bag-of-concepts (BoC) representations. The higher the λ_Q, the more the relevance score depends on BoW. Indeed, to generalise the model, we introduce a modification to Equation (1) of [15] with respect to the weighting between the relevance scores of BoW and BoC, so that our combination model can take into account the situation where only BoW ($\lambda_Q = 1$) or BoC ($\lambda_Q = 0$) is individually effective. In addition, when $\lambda_Q = 0.667$, our model could produce the same list of medical records as Equation (1) with the recommended setting (i.e. $\delta = 2.00$), since the proportion of relevance scores from BoW and BoC computed by Equations (1) and (2) are equal.

Table 1: List of learning features used to predict the importance of the relevance scores from the bag-of-words (BoW) and bag-of-concepts (BoC) representations.

ID	Feature – Ratio (BoW/BoC)
1	Clarity Score [5]
2	SCQ [21]
3	MAXCQ [21]
4	NSCQ [21]
5	AvICTF [4]
6	AvIDF [4]
7	EnIDF [4]
8	Query Scope (ω) [7]
9	AvPMI [4]
10	γ_1 [7]
11	γ_2 [7]
12	Query length [7]

2.2 Estimating the Combination Model

Next, in order to estimate an effective λ_Q of the combination model, described in Section 2.1 (Equation (2)), on the training set, we identify the best λ_Q that achieves the optimal retrieval effectiveness in terms of a particular retrieval measure (e.g. infNDCG) for each training query. Indeed, for each query, we sweep the λ_Q parameter between 0 and 1 to find the best combination model in terms of the retrieval performance for that query. The identified effective λ_Q parameter is used as the weight for the learning component of our framework to learn an effective combination model from the retrieval performance prediction features.

2.3 Learning Features

We next identify the features that we will use to choose the weight for an unseen query. These features should generalise across queries and correlate well with the λ_Q that could result in the optimal retrieval performance. Table 1 lists our features. In particular, as previously discussed in Section 1, we propose to use existing retrieval performance predictors to estimate the retrieval performance of BoW and BoC. Hence, we use the ratio between the retrieval performance predictors computed on BoW and BoC, as the learning features. Specifically, the first set of features (Features 1-4), including the clarity score [5], SCQ [21], MaxSCQ [21] and NSCQ [21], consider the ambiguity of a query by measuring the coherence of the language used in each medical record. The more similar the query model is to the collection model, the better the retrieval performance would be expected. The next set of features (Features 5-8) measure the specificity of each query within a representation approach. Indeed, queries with explicit intents could result in a better performance than queries with general terms. The features include Average Inverse Collection Term Frequency (AvICTF) [4], Average Inverse Document Frequency (AvIDF) [4], EnIDF [4], and the query scope (ω) [7]. Next, Feature 9, the Average of the Pointwise Mutual Information over all query term pairs (AvPMI) [4], focuses on the relationship between query terms. The more co-occurrences among query terms, the better the chance that the relevant documents are being retrieved. Features 10-11 measure the distribution of informativeness among the query terms (i.e. γ_1 and γ_2 [7]), as a query with informative terms could attain an effective retrieval performance. Finally, Feature 12 is the number of non-stopword query terms, which could impact the normalisation methods of the probabilistic retrieval models, and hence affect retrieval performance [7].

2.4 Inferring the Combination Model using Regression Trees

We view the task of estimating the importance of different representation approaches as a supervised regression problem, where the objective is to predict a proper weight (λ_Q) for each query, based on effective weights for similar training queries. By doing so, we would benefit from the fact that several retrieval performance predictors of the representation approaches can be used as learning features, when combining the relevance scores.

While any regression learners could be used here, we deploy the Gradient Boosted Regression Trees (GBRT) [17] (as implemented in the jforests package [6][3]) to learn the combination model discussed in Section 2.1, as it has been shown to be effective in several search and regression tasks (e.g. [17, 20]). We use the root-mean-square error (RMSE) as the loss function when learning a combination model. Our proposed framework leverages retrieval performance predictors, introduced in Section 2.3, as learning features for the GBRT learner.

3. EXPERIMENTAL SETUP

In this section, we discuss our experimental setup when evaluating our proposed framework. In particular, Section 3.1 describes the used test collections and Section 3.2 discusses our ranking strategies.

3.1 Test Collection

We evaluate our framework using the 34 and 47 queries from the TREC 2011 and 2012 Medical Records track [18, 19], respectively. The task is to retrieve patient visits relevant to a given query. Indeed, a patient visit is identified by the medical records associated with a particular visit to a hospital by a patient. The collection contains about 102k medical records, which are associated with 17,265 patient visits [18, 19].

TREC deployed various measures to cope with the possible incompleteness of the gold-standard relevance judgements. In particular, bpref is used as the official measure for TREC 2011 [19], while infNDCG and infAP are used for TREC 2012 [18].

3.2 Ranking Approaches

We index the medical records using Terrier [14]. For the bag-of-words (BoW) representation, we apply Porter's English stemmer and remove stopwords. For the bag-of-concepts (BoC) representation, we follow [9] and apply the so-called *task-specific representation* to represent medical records and queries using only medical concepts related to the medical decision criteria (namely, symptom, diagnostic test, diagnosis, and treatment), as it has been shown to be effective for medical records search. In all experiments, the effective parameter-free DPH term weighting model [1] is used to rank medical records. To learn the combination model, when ranking medical records, we use the default setting of GBRT from the jforests package. We use a 5-fold cross validation across the 34 topics of TREC 2011 and 47 topics of TREC 2012, where each fold has separate training and test query sets. When training the combination model, we target the bpref and infNDCG retrieval measures for TREC 2011 and 2012 topics sets, respectively. Finally, to rank patient visits based on the relevance scores of their associated medical records, we use the expCombSUM voting technique [13], which gives more importance to the highly relevant medical records. Following [11], the number of medical records voting for the relevance of patient visits is limited to 5,000.

4. EXPERIMENTAL RESULTS

We evaluate the retrieval effectiveness of our proposed framework to learn an effective combination model of the bag-of-words (BoW) and the bag-of-concepts (BoC) representations using the retrieval performance predictors discussed in Section 2. Table 2 compares the retrieval performance of our framework on the TREC 2011 and 2012 Medical Records track test collection with three baselines, including a traditional bag-of-words representation (BoW), a task-specific representation [9] (BoC), and an existing score combination approach [15] (i.e. Equation (1)) with the suggested setting from [9, 15]. In addition, to evaluate the optimal potential effectiveness, the best retrieval performances that our proposed framework and the existing score combination could achieve are also reported (denoted oracle).

From Table 2, we observe the following. First, we see that for both TREC 2011 and TREC 2012 topics sets, both our proposed framework and the existing score combination approach markedly outperform the baselines where either of the representations are taken into account. This shows that combining the relevance scores from BoW and BoC is effective for medical records search. Next, for the TREC 2012 topics set, the retrieval performances of our framework (5-fold) markedly outperform those of the score combination baseline ($\delta = 2$). In particular, in terms of the infNDCG retrieval performance, our framework (infNDCG 0.4723) significantly outperforms (paired t-test, $p < 0.05$) the existing score combination baseline (infNDCG 0.4557). For the infAP measure, our proposed framework performs markedly better than the score combination baseline (+6.5% improvement, from 0.1975 to 0.2133). In addition, our proposed framework (5-fold) also results in a markedly better retrieval effectiveness than the best possible setting of the score combination baseline (oracle). Indeed, in terms of infNDCG, our proposed framework significantly outperforms the score combination with the best setting for upto 3.21% ($p < 0.05$). For the infAP retrieval measure, our regression-based framework performs +4.50% better than the best setting of the score combination. However, for the TREC 2011 topics set, our framework (5-fold) could not outperform the score combination ($\delta = 2$) baseline (bpref 0.5078 vs. 0.5118). This is partially due to the fact that the TREC 2011 topics set contains only 34 queries; hence, with a small number of queries, when we conduct a 5-fold cross validation, the training and test sets could not generalise.

Finally, we discuss the optimal retrieval performance that our proposed framework could achieve to evaluate the po-

[3] http://code.google.com/p/jforests/

Table 2: The retrieval performances of different representation approaches on TREC 2011 and 2012 Medical Records track test collections. Statistical significance (paired t-test) at $p < 0.05$, at $p < 0.01$, and at $p < 0.001$ over a baseline are denoted ^{a}, ^{aa} and ^{aaa}, respectively. ^{a} is ^{1}, ^{2}, ^{3}, ^{4} or ^{5} to represent the bag-of-words representation (BoW), the task-specific representation (BoC), the score combination ($\delta = 2$), our learning framework (5-fold), or the score combination (oracle) baselines, respectively.

Approaches	2011	2012	
	bpref	infNDCG	infAP
Bag-of-words representation (BoW)	0.4871	0.4167	0.1703
Task-specific representation (BoC)	0.4929	0.4218	0.1920
Score Combination [15] ($\delta = 2$)	**0.5118**	0.4557^{11}	0.1975^{1}
Our learning framework (5-fold)	0.5078	**0.4723**^{11,22,3}	**0.2133**^{1,2}
Score Combination [15] (oracle)	0.5121	0.4604^{1,2,4}	0.2048^{1}
Our learning framework (oracle)	**0.5796**^{111,222,333,444,555}	**0.5130**^{111,222,333,444,555}	**0.2381**^{111,222,33,444,555}

tential effectiveness of our framework, if more training data were available. As expected, we observe that, with the best setting, our framework (oracle) significantly ($p < 0.01$) outperforms all of the approaches discussed in this paper. This supports our hypothesis that some particular queries differently benefit from BoW and BoC. In particular, the retrieval performance of our framework with the best setting is upto +17.06% better than the 5-fold cross validation. Importantly, we find that the mean of the effective weights (λ_Q with the best possible setting) across the two collections is 0.48459 ($0 \leq \lambda_Q \leq 1$), while the standard deviation is 0.38085, which suggests that the effective weight should indeed vary across topics. For example, to attain an effective retrieval performance when a query contains multiple complex concepts (e.g. topic#106: patients who had positron emission tomography (PET), magnetic resonance imaging (MRI), or computed tomography (CT) for staging or monitoring of cancer), λ in the combination model (Equation (2)) should be low, if all the concepts in the query can be effectively identified. From this, we conclude that there is no one combination of BoW and BoC that is effective for all queries. Hence, per-query prediction approaches, like the ones deployed here, have great potential to improve medical records search. However, there is still an open research area to explore effective features and learners to close the performance gap between the cross-validation and oracle regimes, even though by deploying the existing learner and features, our framework could in general markedly and significantly outperform the existing score combination approach [15].

5. CONCLUSIONS

We have tackled the challenge of dealing with the complex and ambiguous terminology in medical records search by modelling the combination of the relevance scores from both bag-of-words (BoW) and bag-of-concepts (BoC) representations. We have proposed a regression-trees-based learning framework that can effectively handle this combination using the Gradient Boosted Regression Trees to learn an effective combination model via retrieval performance predictors, such as the clarity score [5] and the query scope [7]. We have shown that our proposed framework is effective for the medical records search, as it could markedly and significantly outperform an effective score combination approach [15].

6. REFERENCES

[1] G. Amati, E. Ambrosi, M. Bianchi, C. Gaibisso, and G. Gambosi. FUB, IASI-CNR and University of Tor Vergata at TREC 2007 Blog Track, In *TREC* 2007.

[2] A. R. Aronson. Exploiting a Large Thesaurus for Information Retrieval. In *RIAO* 1994.

[3] A. R. Aronson and F. Lang. An Overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3), 2010.

[4] D. Carmel and E. Yom-Tov. Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1), 2010.

[5] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting Query Performance. In *SIGIR* 2002.

[6] Y. Ganjisaffar, R. Caruana, and C. V. Lopes. Bagging Gradient-Boosted Trees for High Precision, Low Variance Ranking Models. In *SIGIR* 2011.

[7] B. He and I. Ounis. Query Performance Prediction. *Inf. Syst.*, 31(7), 2006.

[8] W. Hersh, D. Hickam, R. Haynes, and K. McKibbon. A Performance and Failure Analysis of SAPHIRE with a MEDLINE Test Collection. *J. Am. Med. Inform. Assoc.*, 1(1), 1994.

[9] N. Limsopatham, C. Macdonald, and I. Ounis. A Task-Specific Query and Document Representation for Medical Records Search. In *ECIR* 2013.

[10] N. Limsopatham, C. Macdonald, and I. Ounis. Inferring Conceptual Relationships to Improve Medical Records Search. In *OAIR* 2013.

[11] N. Limsopatham, C. Macdonald, I. Ounis, G. McDonald, M. Bouamrane. University of Glasgow at Medical Records track 2011: Experiments with Terrier. In *TREC* 2011.

[12] N. Limsopatham, R. L. T. Santos, C. Macdonald, and I. Ounis. Disambiguating Biomedical Acronyms using EMIM. In *SIGIR* 2011.

[13] C. Macdonald and I. Ounis. Voting for Candidates: adapting data fusion techniques for an expert search task. In *CIKM* 2006.

[14] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *OSIR at SIGIR* 2006.

[15] P. Srinivasan. Optimal Document-Indexing Vocabulary for MEDLINE. *Inf. Process. Manage.*, 32(5), 1996.

[16] D. Trieschnigg, D. Hiemstra, F. de Jong, and W. Kraaij. A Cross-Lingual Framework for Monolingual Biomedical Information Retrieval. In *CIKM* 2010.

[17] S. Tyree, K. Q. Weinberger, K. Agrawal, and J. Paykin. Parallel Boosted Regression Trees for Web Search Ranking. In *WWW* 2011.

[18] E. Voorhees and W. Hersh. Overview of the TREC 2012 Medical Records Track. In *TREC* 2012.

[19] E. Voorhees and R. Tong. Overview of the TREC 2011 Medical Records Track. In *TREC* 2011.

[20] Y. Wang, B. Wei, J. Yan, Y. Hu, Z. H. Deng, and Z. Chen. A Novel Local Patch Framework for Fixing Supervised Learning Models. In *CIKM* 2012.

[21] Y. Zhao, F. Scholer, and Y. Tsegay. Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In *ECIR* 2008.