Exploring Understandability Features to Personalize Consumer Health Search
TUW at CLEF 2017 eHealth

Joao Palotti and Navid Rekabsaz

Vienna University of Technology (TUW)
Favoritenstrasse 9-11/188 1040
Vienna, Austria
[palotti,rekabsaz]@ifs.tuwien.ac.at

1 Introduction

This paper describes the participation of Technical University of Vienna (TUW) at CLEF eHealth 2017 Task 3 [5,9]. This track runs annually since 2013 (see [3,4,7,12]) and this year's challenge is a continuation of 2016's one. The Information Retrieval task of CLEF eHealth Lab aims to foster research on search for health consumers, emphasizing crucial aspects of this domain such as document understandability and trustworthiness.

In 2016, fifty topics were extracted from real user posts/interactions in the AskDocs section of Reddit[1]. Each topic was presented to six query creators with different medical expertise. Their job was to read a post (usually with a medical question) and formulate a query using their medical background knowledge, if any. In total 300 queries were created.

This year, this track has 4 subtasks (named IRTasks, see [9] for a full description of each task) and TUW submitted runs for two of them, IRTask 1 and 2. IRTask 1 is the Ad-Hoc task with the same topics of last year, aiming to increase the number of assessed documents for the collection. IRTask 2 is a new task, and the goal is to personalize the results for each query creator according to his/her medical expertise.

The experiments conducted by TUW aim to investigate two research directions:

1. *IRTask 1*: Can understandability metrics be used to improve retrieval?
2. *IRTask 2*: How to personalize retrieval in a learning to rank setting, according to different reading profiles and user expertise?

For IRTask 1, a previous study conducted in the context of CLEF eHealth 2014 and 2015 ([6]) showed promising improvements when using a small set of understandability estimators in a learning to rank context. Here we expand the set of understandability features used as well as non-understandability features (see Section 2.2). Our aim is to investigate if the improvements first seen in [6]

[1] https://www.reddit.com/r/AskDocs/

would also occur in this dataset. For IRTask 2, we propose to explicitly define learning to rank features based on different user profiles. We study the effect of the suggested features in the system effectiveness.

2 Methodology

In this section we describe our learning to rank approach, the feature set devised, and our functions to map topical and understandability assessments into a single relevance label.

2.1 Learning to Rank

Our learning to rank approach is based on 3 items: (1) a set of features, (2) a set of <document, label> pairs, and (3) a learning to rank algorithm. The set of features is described in Section 2.2. We consider in this work three different functions to label documents: for Subtask 1, we only use the pure topical relevance as judged in 2016; for Subtask 2, we define two understandability-biased function (named *boost* and *float*). Given a document with topical relevance T and understandability score U, and an user with a reading goal G, we define *boost* and *float* as:

$$boost(T, U) = \begin{cases} 2 * T & \text{if } |G - U| \leq 0.2 \\ T & \text{if } |G - U| > 0.2 \end{cases} \tag{1}$$

$$float(T, U) = T * (1.0 - |G - U|) \tag{2}$$

As topical relevance scores are either 0, 1 or 2, and the understandability scores are float numbers from 0.0 to 1.0, the possible values for function *boost* are the integers 4, 2, 1 and 0, while the possible values for function *float* are any float precision number between 0.0 and 2.0. All experiments used the pairwise learning to rank algorithm based on gradient boosting implemented in XGboost[2] with NDCG@20 as goal to be optimized. Differently from past work [10,6], we do consider up to 1000 documents when re-ranking documents.

2.2 Features

We devised 91 features from 3 distinct groups: information retrieval traditional features, understandability related features, and the modified output of regression algorithms made to estimate the understandability of a document. Elaborated features based on recent advance on semantic similarity, as made in [10], are left as future work. A comprehensive list of all features used in this work is shown in Table 1.

[2] `https://github.com/dmlc/xgboost/tree/master/demo/rank`

Feature Type	Feature Category	Feature Name
IR Features (12)	Common IR Models (7)	BM25 PL2 DirichletLM LemurTF_IDF TF_IDF DFRee Hiemstra_LM
	Query Independ. (3)	Document Length Document Spam Scores Document Page Rank
	Doc. Score Modifier (2)	Divergence from Randomness Markov Random Field
Understandability Features (72)	Traditional Formulas (8)	ARI Index Coleman Liau Index Dale-Chall Score Flesch Kincaid Grade Flesch Reading Ease Gunning Fog Index LIX Index SMOG Index
	Surface Measures (25)	# Characters $^{\diamond\dagger}$ # Sentences $^{\diamond}$ # Syllables $^{\diamond\dagger}$ # Words † # (| Syllables(Word) | > 3) $^{\diamond\dagger}$ # (| Word | > 4) $^{\diamond\dagger}$ # (| Word | > 6) $^{\diamond\dagger}$ # (| Word | > 10) $^{\diamond\dagger}$ # (| Word | > 13) $^{\diamond\dagger}$
	General Vocabulary Related Features (12)	Numbers $^{\diamond\dagger}$ English Dictionary $^{\diamond\dagger}$ Dale-Chall List $^{\diamond\dagger}$ stopwords $^{\diamond\dagger}$
	Medical Vocabulary Related Features (27)	Acronyms $^{\diamond\dagger}$ Mesh $^{\diamond\dagger}$ DrugBank $^{\diamond\dagger}$ ICD10 (International classification of Diseases) $^{\diamond\dagger}$ Medical Prefixes $^{\diamond\dagger}$ Medical Suffixes $^{\diamond\dagger}$ Consumer Health Vocabulary $^{\diamond\dagger}$ Sum(chv Score) $^{\diamond\dagger}$ Mean(chv Score) $^{\diamond\dagger}$
Regression Features (7)	Modified Regression Scores (7)	Ada Boosting Regressor Extra Tree Regressor Gradient Boosting Regressor K-Nearest Neighbor Regressor Linear Regression Support Vector Machine Regressor Random Forest Regressor

Table 1: Features used in the learning to rank process; the number of features for each group is reported in parenthesis. $\diamondsuit$: raw feature values and values normalised by number of words in a documents are used. $\dagger$: raw feature values and values normalised by number of sentences in a document are used.

IR Features: Regularly used information retrieval features are considered in this work. This list includes many commonly used retrieval models and document specific values, such as Spam scores[1] and PageRank scores[3].

Understandability Features: All HTML pages were preprocessed with Boilerpipe[4] to remove the undesirable boilerplate content as suggested in [8]. Then, a series of traditional readability metrics was calculated [2], as well as a number of basic syntactic and lexical features that are important components of such readability metrics. Finally, we measure the occurrence of words in different vocabularies, both medical and non-medical ones.

Regression Features: We adapted the output of regression algorithms to create personalized features. The 2016's judgements were used as labels for a num-

[3] `http://www.lemurproject.org/clueweb12/PageRank.php`

[4] `https://pypi.python.org/pypi/boilerpipe`

ber of regression algorithms (the list of algorithms used is shown in Table 1). Models were trained on a Latent Semantic Analysis (LSA) applied on words from 3.549 documents marked as topical relevant in the QRels from 2016, which understandability label varied from 0 (easy to understand) to 100 (hard to understand). LSA dimensions vary from 40 to 240 according to the best result of a 10-fold cross validation experiment. In order to avoid interference from the training set in the learning to rank algorithm, scores for the documents in the training set were predicted also in a 10-fold cross validation fashion. The personalization step consisted in calculating the absolute difference between the estimated score and the goal score, which is defined by user. For example, if the score estimated by a regression algorithm for a document D was 0.45 and the reading goal of a user U was 0.80, we used as feature the value 0.35 (the absolute difference between 80 and 45). We want to evaluate if features like these ones can help the learning to rank model to adapt according to the reading skills of a user.

3 Experiments

3.1 Evaluation Metrics

We consider a large number of evaluation metrics in this work. As topical relevance centred evaluation metrics, we consider Precision at 10 (P@10) and Rank Biased Precision with μ parameter set to 0.8 (RBP(0.8)). Due to the fact that a learning to rank algorithm has the potential to bring many unjudged documents to the top of the ranking list, we consider also a modified version of P@10, Only Judged P@10, which will calculate P@10 considering only the first 10 judged documents of each topic.

As modified metrics that take into account understandability scores, we consider understandability-biased Rank Biased Precision, also with μ parameter set to 0.8 (uRBP(0.8)) as proposed by [11], and propose three new metrics for personalized search.

The first personalization-aware metric is a specialization of uRBP, auRBP, which takes advantage of an α parameter to model the kind of documents a user wants to read. We assume that α is a parameter that models understandability profiles of an entity. A low α is assigned to items/documents/users that are experts, while a high α are the opposite. We assume that a user with a low α wants to read specialized documents to the detriment of easy and introductory documents, while laypeople want the opposite. We model in auRBP a penalty for the case in which a low α document is presented to a user that wants high α documents and vice versa. While we are still investigating which one is the best function to model this penalty, we assume a normal penalty. Figure 1 shows an example in which a user is looking forward to reading documents with α=20 and other values for α would have a penalty associated to them according to this normal curve with mean 20 and standard deviation of 30. We use the standard deviation of 30 in all of our experiments, but it is left as future work ways to estimate a right value for it.

The second and third personalization-aware metrics are simple modifications of Precision at depth X. For the relevant documents found in the top X, we inspect how far is the understandability label of each document to the expected value required by a user. We could penalize the absolute difference linearly (LinUndP@X) or using the same Gaussian curve as in auRBP (GaussianUndP@10). Note that lower values are better for LinUndP@10, meaning that the distance from the required understandability value is small, and higher values are better for GaussianUndP@10, as a value of 100 is the best value one could reach.

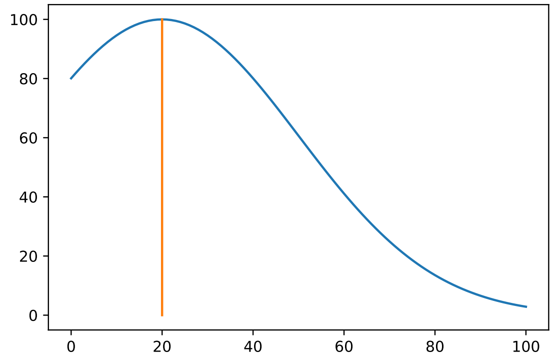

Fig. 1: A Gaussian model for penalty. This example of normal curve has its peak (mean) at 20 and standard deviation of 30. A document with α=60 would be worth only 41% of a document with the desired α=20.

3.2 Runs Description

Seven runs were submitted to IRTasks 1 and another seven were submitted to IRTask 2. Tables 2 and 3 present a summary of each approach, submissions for IRTask 1 and 2, respectively, and the results using 2016 Qrels.

4 Discussion and Conclusion

As shown in Table 2 and 3, we based our runs on the BM25 implementation from a Terrier 4.2 index of ClueWeb 12-B. The results of using relevance feedback are high because the judged as relevant documents appear at the top of the ranking list of each topic, but it does not necessarily means that these approach will be much better than a plain BM25 for 2017, as the already judged documents will be discarded by the organizers.

Run ID	Run description	Results on CLEF eHealth 2016 QRels						
		P@10	Only Jud. P@10	LinUnd P@10	GaussianUnd P@10	RBP(0.80)@10	uRBP(0.8)@10	auRBP(0.8)@10
TUW1	Baseline: Terrier 4.2 BM25	26.46	27.63	33.39	55.73	27.32	17.34	14.14
TUW2	Baseline2: Terrier 4.2 BM25 with Rel. Feedback	39.83	44.87	33.02	56.76	41.73	25.76	21.36
TUW3	LTR on top 1000 from TUW1 - All Features	25.46	29.37	33.19	56.07	27.23	16.98	13.91
TUW4	LTR on top 1000 from TUW2 - All Features	41.36	50.07	33.55	56.04	41.94	25.69	21.26
TUW5	LTR on top 1000 from TUW1 - IR only	25.26	27.57	33.80	55.04	26.44	16.59	13.68
TUW6	LTR on top 1000 from TUW2 - IR only	44.36	51.50	33.71	56.22	46.43	27.41	23.64
-	LTR on top 1000 from TUW1 - IR + Underst.	25.00	29.50	33.20	56.37	26.56	16.67	13.64
TUW7	LTR on top 1000 from TUW2 - IR + Underst.	41.97	49.57	33.58	56.32	42.90	25.96	21.68

Table 2: Results on CLEF eHealth 2016 QRels and runs submitted to CLEF eHealth 2017 IRTask 1.

Run ID	Run description	Results on CLEF eHealth 2016 QRels						
		P@10	Only Jud. P@10	LinUnd P@10	GaussianUnd P@10	RBP(0.80)@10	uRBP(0.8)@10	auRBP(0.8)@10
BL1	Baseline: Terrier 4.2 BM25	26.46	27.63	33.39	55.73	27.32	17.34	14.14
BL2	Baseline2: Terrier 4.2 BM25 with Rel. Feedback	39.83	44.87	33.02	56.76	41.73	25.76	21.36
TUW1	LTR on top 1000 from BL1 - All Features. Labels w. Boost	24.90	28.90	32.95	56.49	26.84	16.95	13.90
TUW2	LTR on top 1000 from BL2 - All Features. Labels w. Boost	42.60	49.73	34.24	55.44	42.88	26.34	21.70
TUW3	LTR on top 1000 from BL1 - All Features. Labels w. Float	25.43	29.30	32.88	56.96	27.16	17.11	13.87
TUW4	LTR on top 1000 from BL2 - All Features. Labels w. Float	42.23	50.20	33.92	55.91	43.59	26.44	22.17
-	LTR on top 1000 from BL1 - IR only w. Boost	25.23	27.53	33.57	55.35	26.39	16.56	13.52
TUW5	LTR on top 1000 from BL2 - IR only w. Boost	43.20	51.13	33.34	56.34	45.67	27.10	23.01
-	LTR on top 1000 from BL1 - IR only w. Float	25.57	27.60	33.45	55.47	26.31	16.53	13.54
TUW6	LTR on top 1000 from BL2 - IR only w. Float	43.53	51.33	33.50	56.51	45.42	26.55	22.95
-	LTR on top 1000 from BL1 - IR + Regres. Labels w. Boost	25.30	27.83	33.77	55.07	26.27	16.36	13.50
-	LTR on top 1000 from BL2 - IR + Regres. Labels w. Boost	42.56	50.70	33.49	56.14	45.22	26.20	22.85
-	LTR on top 1000 from BL1 - IR + Regres. Labels w. Float	25.10	27.90	33.79	55.10	26.06	16.41	13.33
TUW7	LTR on top 1000 from BL2 - IR + Regres. Labels w. Float	43.63	51.50	33.64	55.90	45.83	26.78	23.11
-	LTR on top 1000 from BL1 - IR + Unders. Labels w. Boost	24.70	29.40	33.30	56.25	25.39	17.47	13.55
-	LTR on top 1000 from BL1 - IR + Unders. Labels w. Boost	43.03	50.97	34.15	55.23	42.93	27.16	21.84
-	LTR on top 1000 from BL1 - IR + Underst. Labels w. Float	25.50	30.60	33.42	56.10	25.80	17.86	13.91
-	LTR on top 1000 from BL2 - IR + Underst. Labels w. Float	41.00	50.03	33.82	55.30	41.95	26.39	21.41

Table 3: Results on CLEF eHealth 2016 QRels and runs submitted to CLEF eHealth 2017 IRTask 2.

Considering our experiments with IRTask1, shown in Table 2, we noticed a degradation of P@10 for runs TUW3 and TUW5 if compared to the baseline TUW1. This, however, is not the case for the modified version of P@10 which considers only judged documents. This is the same of RBP(0.8). Our expectation is that TUW3 and TUW5 are going to be more effective than TUW1, as well as, TUW4, TUW6 and TUW7 are going to be more effective than TUW2. Note that higher RBP(0.8) are followed by higher uRBP(0.8) and auRBP(0.8), while higher P@10 are not followed by higher LinUnd.P@10 or GausianUnd.P@10. This means that our efforts to retrieve more topical relevant documents also increases uRBP and auRBP, but does not affect LinUnd. and GausianUnd. metrics.

Table 3 shows our experiments with different labelling function (*Boost* and *Float*). Again there is no approach that could beat the best P@10 value for BL1 (which is TUW1 in IRTask1), while several approaches beat P@10 of BL2. When looking at understandability biased metrics, in especial to LinUndP@10 and GaussianP@10, we can see much more variance than in Table 3. The best result found was TUW3 in Table 3 with 32.88 for LinUnd.P@10 (the smaller the better) and 56.96 for GausianUnd.P@10 (the higher the better).

We are looking forward to evaluating our results with 2017 QRels, but as the 2017's assessments are still being conducted, an analysis of the official results will be posted online at `https://github.com/joaopalotti/tuw_at_clef_ehealth_2017`.

References

1. Gordon V. Cormack, Mark D. Smucker, and Charles L. A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *CoRR*, abs/1004.5168, 2010.
2. William H. Dubay. The principles of readability. *Costa Mesa, CA: Impact Information*, 2004.
3. Lorraine Goeuriot, Gareth JF Jones, Liadh Kelly, Johannes Leveling, Allan Hanbury, Henning Müller, Sanna Salantera, Hanna Suominen, and Guido Zuccon. ShARe/CLEF eHealth Evaluation Lab 2013, Task 3: Information retrieval to address patients' questions when reading clinical reports. *CLEF 2013 Online Working Notes*, 8138, 2013.
4. Lorraine Goeuriot, Liadh Kelly, Wei Lee, Joao Palotti, Pavel Pecina, Guido Zuccon, Allan Hanbury, and Henning Mueller Gareth J.F. Jones. ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In *CLEF 2014 Evaluation Labs and Workshop: Online Working Notes*, Sheffield, UK, 2014.
5. Lorraine Goeuriot, Liadh Kelly, Hanna Suominen, Aurélie Névéol, Aude Robert, Evangelos Kanoulas, Rene Spijker, Joao Palotti, and Guido Zuccon. Clef 2017 ehealth evaluation lab overview. In *Proceedings of CLEF 2017 - 8th Conference and Labs of the Evaluation Forum.* Lecture Notes in Computer Science (LNCS), Springer, September 2017.
6. Joao Palotti, Lorraine Goeuriot, Guido Zuccon, and Allan Hanbury. Ranking health web pages with relevance and understandability. In *Proceedings of the 39th international ACM SIGIR conference on Research and development in information retrieval*, pages 965–968. ACM, 2016.
7. Joao Palotti, Guido Zuccon, Lorraine Goeuriot, Liadh Kelly, Allan Hanburyn, Gareth J.F. Jones, Mihai Lupu, and Pavel Pecina. CLEF eHealth Evaluation Lab 2015, Task 2: Retrieving Information about Medical Symptoms. In *CLEF 2015 Online Working Notes.* CEUR-WS, 2015.
8. Joao Palotti, Guido Zuccon, and Allan Hanbury. The influence of pre-processing on the estimation of readability of web documents. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1763–1766. ACM, 2015.
9. Joao Palotti, Guido Zuccon, Jimmy, Pavel Pecina, Mihai Lupu, Lorraine Goeuriot, Liadh Kelly, and Allan Hanburyn. Clef 2017 task overview: The ir task at the ehealth evaluation lab. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum.* CEUR Workshop Proceedings. Proceedings of CLEF 2017 - 8th Conference and Labs of the Evaluation Forum, 2017.
10. Luca Soldaini and Nazli Goharian. *Learning to Rank for Consumer Health Search: A Semantic Approach*, pages 640–646. Springer International Publishing, 2017.
11. Guido Zuccon. Understandability biased evaluation for information retrieval. In *Proc. of ECIR*, 2016.
12. 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. In *CLEF 2016 Evaluation Labs and Workshop: Online Working Notes, CEUR-WS*, September 2016.