LIMSI@CLEF eHealth 2018 Task 2: Technology Assisted Reviews by Stacking Active and Static Learning

Christopher Norman[1,2], Mariska Leeflang[2], and Aurélie Névéol[1]

[1] LIMSI, CNRS, Université Paris Saclay, F-91405 Orsay
firstname.lastname@limsi.fr
[2] Academic Medical Center, University of Amsterdam, Amsterdam, the Netherlands
m.m.leeflang@uva.nl

Abstract. This paper describes the participation of the LIMSI-MIROR team at CLEF eHealth 2018, task 2. The task addresses the automatic ranking of articles in order to assist with the screening process of Diagnostic Test Accuracy (DTA) Systematic Reviews. We ranked articles by stacking two models, one linear regressor trained on untargeted training data, and one model using active learning. The workload reduction to retrieve 95% of the relevant articles was estimated at 82.4%, and we observe a workload reduction less than 70% in only two topics. The results suggest that automatic assistance is promising for ranking the DTA literature.

Keywords: Evidence Based Medicine, Information Storage and Retrieval, Review Literature as Topic, Supervised Machine Learning

1 Introduction

Systematic reviews seek to gather all available published evidence for a given topic and provide an informed analysis of the results. This work constitutes some of the strongest forms of scientific evidence. Systematic reviews are an integral part of evidence based medicine in particular, and serve a key role in informing and guiding public and institutional decision-making. Systematic reviews for Diagnostic Test Accuracy (DTA) studies have been shown particularly challenging compared to other types of reviews because of the difficulty in defining search strategies offering acceptable recall [7]. For this reason, there is a need to investigate automation strategies to assist DTA systematic review writers, particularly in the time-consuming screening process.

Methods for automating the screening process in systematic reviews have been actively researched over the years [6], with promising results obtained using a range of machine learning methods. However, previous work has not addressed DTA studies.

This paper describes the work underlying our participation in the CLEF 2018 eHealth Task 2 [4, 8]. This work is part of an ongoing effort to provide automated

assistance in the screening process in systematic reviews addressing a variety of topics, including DTA studies.

The remainder of this paper is organized as follows; Section 2 presents the dataset used for system development. Section 3 provides an overview of our system and describes each component. Finally, section 4 reports our results and section 5 provides an analysis of our methods and participation in the task.

2 Material

In this work we have used the CLEF DATASET [3] as the gold standard for evaluation. The first iteration (2017) of the CLEF dataset [3] comprised 50 DTA systematic review topics (20 for training, 30 for testing) associated with the full list of articles retrieved by an expert query and assessed for inclusion based on title and abstract or full text. The second iteration (2018) uses the previous 50 topics for training, and supplies an additional 30 topics for testing.

For each of the datasets we know the inclusion decisions based on the abstracts, as well as the inclusion decisions based on the full text. We thus have two definitions of positive examples, depending on whether we use the abstract decisions or full text decisions as the gold standard.

We use a tripartite labeling to reflect this:

- **No (N)** is the set of articles that were excluded based on the abstract
- **Maybe (M)** is the set of articles that were preliminarily included based on the abstract, but later excluded based on the full text
- **Yes (Y)** is the set of articles that were included based on both the abstract and the full text, and later used in the meta-analysis

Table 1 shows a breakdown of the distribution of examples for each class in the CLEF dataset.

3 Methods

To rank candidate articles we construct three machine learning models:

3.1 Overview

cnrs static Our **static ranker** uses logistic regression trained on a large number (> 500,000) of features. This model is trained once on train split 1 (Table 1), and can then be used to rank candidate articles in any unseen DTA systematic review, without a provided search query or topic description. This model is intended to capture diagnostic test accuracy studies without considering whether the articles are topically relevant.

Split	Topic	Absolute number			Relative number		
		Y	M	N	Y	M	N
train split 1 (2017 train split)	CD008643	4	7	15065	0.0%	0.0%	99.9%
	CD009593	24	54	14844	0.2%	0.4%	99.5%
	CD011549	1	1	12699	0.0%	0.0%	100.0%
	CD010771	1	47	274	0.3%	14.6%	85.1%
	CD010438	3	36	3211	0.1%	1.1%	98.8%
	CD007427	17	106	1398	1.1%	7.0%	91.9%
	CD008686	5	2	3946	0.1%	0.1%	99.8%
	CD011548	5	108	12591	0.0%	0.9%	99.1%
	CD007394	47	48	2450	1.8%	1.9%	96.3%
	CD009323	9	113	3757	0.2%	2.9%	96.9%
	CD010632	14	18	1472	0.9%	1.2%	97.9%
	CD011975	60	559	7582	0.7%	6.8%	92.5%
	CD009944	64	53	1064	5.4%	4.5%	90.1%
	CD009591	41	103	7847	0.5%	1.3%	98.2%
	CD011134	49	166	1738	2.5%	8.5%	89.0%
	CD009020	12	150	1422	0.8%	9.5%	89.8%
	CD010409	41	35	43287	0.1%	0.1%	99.8%
	CD008691	20	53	1243	1.5%	4.0%	94.5%
	CD011984	28	426	7738	0.3%	5.2%	94.5%
	CD008054	41	233	2940	1.3%	7.2%	91.5%
train split 2 (2017 test split)	CD010783	11	19	10875	0.1%	0.2%	99.7%
	CD009135	19	58	714	2.4%	7.3%	90.3%
	CD009185	23	69	1523	1.4%	4.3%	94.3%
	CD010023	14	38	929	1.4%	3.9%	94.7%
	CD010653	0	45	7957	0.0%	0.6%	99.4%
	CD009647	17	39	2729	0.6%	1.4%	98.0%
	CD011145	48	154	10670	0.4%	1.4%	98.1%
	CD008760	9	3	52	14.1%	4.7%	81.2%
	CD010775	4	7	230	1.7%	2.9%	95.4%
	CD009925	55	405	6071	0.8%	6.2%	93.0%
	CD009372	10	15	2223	0.4%	0.7%	98.9%
	CD010896	3	3	163	1.8%	1.8%	96.4%
	CD010542	8	12	328	2.3%	3.4%	94.3%
	CD008803	99	0	5121	1.9%	0.0%	98.1%
	CD009519	46	58	5867	0.8%	1.0%	98.3%
	CD010386	1	1	623	0.2%	0.2%	99.7%
	CD008782	34	11	10462	0.3%	0.1%	99.6%
	CD009579	79	59	6317	1.2%	0.9%	97.9%
	CD010772	11	36	269	3.5%	11.4%	85.1%
	CD009551	16	30	1865	0.8%	1.6%	97.6%
	CD010173	10	13	5472	0.2%	0.2%	99.6%
	CD010339	9	105	12689	0.1%	0.8%	99.1%
	CD010633	3	1	1569	0.2%	0.1%	99.7%
	CD010705	18	5	91	15.8%	4.4%	79.8%
	CD012019	1	2	10314	0.0%	0.0%	100.0%
	CD007431	15	9	2050	0.7%	0.4%	98.8%
	CD010276	24	30	5441	0.4%	0.5%	99.0%
	CD009786	6	4	2055	0.3%	0.2%	99.5%
	CD008081	10	16	944	1.0%	1.6%	97.3%
	CD010860	4	3	87	4.3%	3.2%	92.6%
test split (2018 test split)	CD011602	1	7	6149	0.0%	0.1%	99.9%
	CD011515	1	126	7117	0.0%	1.7%	98.2%
	CD010864	3	41	2461	0.1%	1.6%	98.2%
	CD012083	5	6	311	1.6%	1.9%	96.6%
	CD010680	0	26	8379	0.0%	0.3%	99.7%
	CD011431	26	271	885	2.2%	22.9%	74.9%
	CD012216	1	10	206	0.5%	4.6%	94.9%
	CD012281	9	14	9853	0.1%	0.1%	99.8%
	CD011686	2	53	9388	0.0%	0.6%	99.4%
	CD009175	7	58	5579	0.1%	1.0%	98.8%
	CD010213	33	566	14599	0.2%	3.7%	96.1%
	CD010657	35	104	1720	1.9%	5.6%	92.5%
	CD012599	19	556	7473	0.2%	6.9%	92.9%
	CD011420	5	37	209	2.0%	14.7%	83.3%
	CD012009	4	33	499	0.7%	6.2%	93.1%
	CD009263	10	114	78679	0.0%	0.1%	99.8%
	CD011926	29	11	4010	0.7%	0.3%	99.0%
	CD008122	57	215	1639	3.0%	11.3%	85.8%
	CD008587	35	44	9073	0.4%	0.5%	99.1%
	CD011912	18	18	1370	1.3%	1.3%	97.4%
	CD009694	9	7	145	5.6%	4.3%	90.1%
	CD010296	38	15	4549	0.8%	0.3%	98.8%
	CD012165	47	261	9914	0.5%	2.6%	97.0%
	CD008759	42	18	872	4.5%	1.9%	93.6%
	CD012179	117	187	9528	1.2%	1.9%	96.9%
	CD010502	71	158	2756	2.4%	5.3%	92.3%
	CD008892	30	39	1430	2.0%	2.6%	95.4%
	CD012010	8	282	6540	0.1%	4.1%	95.8%
	CD011053	7	5	2223	0.3%	0.2%	99.5%
	CD011126	9	4	5987	0.1%	0.1%	99.8%

Table 1: The distribution of class labels in the dataset.

cnrs RF (uni-/bigram) We construct two **relevance feedback** (active learning) models uses logistic regression on a smaller number ($\approx$ 2,000) of features. These models are trained using relevance feedback on the target topic, starting with the topic description as an artificial seed document. The unigram model is a reimplementation of the CAL model by Cormack and Grossman [1, 2]. We also experiment on a model which uses bigrams in addition to unigrams. These models are intended to capture topicality, and to incrementally improve performance through the screening process.

cnrs combined Our **stacked metaclassifier** uses a three-layer feedforward dense neural network to estimate the optimal ranking based on the output of the **static** model and the **RF bigram** model.

We describe each system in detail in the remainder of this section.

3.2 Static Ranking Model

We here use a machine learning approach and train a classifier on the training split, largely identical to the implementation of our static model submitted in 2017 [5]. The decision function of the classifier can then be used to calculate probability scores for unseen candidate articles. This is a static model, intended to capture diagnostic test accuracy studies without considering whether the articles are topically relevant.

We use logistic regression trained using stochastic gradient descent (sklearn) on a sparse feature matrix consisting of a large number ($>$ 500,000) of features. We have tried using other classifiers, including SVMs, random forests, feed-forward neural networks, convolution networks and LSTMs, but logistic regression yields consistently better performance in our experiments with a fraction of the training time.

We handle class imbalance by class reweighting. We have implemented undersampling mechanisms, but these tend to decrease performance. We set the weight for the positive class to 80 for the initial intertopic classifier. We have determined this to be a reasonable weight experimentally in previous experiments on another dataset [5].

This model was trained on the 2017 training split.

3.3 Active Learning

We here use an active learning approach, where we at each timestep train a classifier (ranker) on the relevant articles screened so far. We start the process using the topic description as an artificial seed document. The model is intended to capture topical relevance, and to use the data collected through the screening process, which is generally more targeted than the data we have available in the training split.

The model largely follows the continuous active learning approach of Cormack and Grossman [1, 2], except for using bigrams in addition to unigrams. We repeat the procedure for clarity.

At each timestep we rank the candidate articles and show the top B articles to the oracle, and the oracle labels these as Y, M, or N. The number of articles B is initially set to 1 and is incremented by $\lfloor B \rfloor$ at each timestep.

We use the following process to construct positive training data:

- **if Y have been encountered:**
  Then we use all encountered Y as positive training data. The synthetic seed document and any encountered M are discarded.
- **else if M have been encountered, but no Y:**
  Then we use all encountered M as positive training data. The synthetic seed document is discarded.
- **else (no Y or M have been encountered):**
  We use the synthetic seed document as positive training data.

To construct negative training data we sample 100 articles (or as many as remains) from the unseen candidates and temporarily label these N, irrespective of their true labels. Any articles already shown to the oracle are not considered for use as negative data.

We train our model on using the above positive and negative data to re-rank the candidate articles and repeat the process until all articles have been shown to the oracle.

This model only uses the candidate articles and the topic description as training data, and thus do not depend on other training data, such as the topics in the training split.

3.4 Stacked Model

We use a three-layer dense neural network as a function approximator to estimate the joint score for a candidate document given the scores from our static and active models. We use 16 nodes in each layer, apply 30% dropout after each layer and use softmax activation on the final layer to simulate two-class logistic regression.

The model is trained by sampling training data uniformly from recorded active learning output. We have tried using uncertainty sampling, but this has yielded inferior results.

As input to the model we use the score values we get from the static and active learning models, along with meta-level features. The full set of features is as follows:

1. Static model document score (**static**)
2. Active model document score (**RF bigram**)
3. Number of Y found
4. Amount of relevance feedback (absolute number)
5. Amount of relevance feedback (percentage)
6. Relevance feedback stage (whether using seed, M or Y as positive training data)

Features 3 and 4 are normalized using the following log transform

$$\text{sgn}(x) \times \frac{\log_2(1 + |x|)}{8}$$

to keep numbers in mainly in the range $[0, 1]$. We do not truncate large numbers. Feature 6 take discrete values in $\{-1, 0, 1\}$

However, we observe that features 5 and 6 decrease model performance and we therefore excluded these in the model used in our officially submitted runs.

This model is trained on data generated from training split 2 (Table 1) to avoid overfitting. We generate the training data for the stacked model by letting the active model run on the training data, and at each step in the process we record the score generated by the active learning model, as well as the above features. We do this 100 times for each topic. One data point thus consists of the score from the static model (feature 1), and features 2–6 from this pre-generated data.

We train the stacked model on data sampled randomly from this pool of data points, by sampling 50 runs in each iteration, and sampling an equal number of positive and negative training examples from each run (with a minimum of 20 total). The model is trained on a batch of size 32. The training data is resampled every training iteration.

4 Results

We present our results for average precision in table 2, WSS@95 in table 3, WSS@100 in table 4, Last Rel in table 5, as well as the aggregate scores in table 6. For comparison, we also calculate a baseline by evaluating each metric on the data ordered randomly. The baseline values are calculated using the average and the standard deviation of 1000 repetitions.

The **RF unigram**, and **RF bigram**, and the **combined** model were submitted as our official runs.

The results omit one topic with no Y (CD010680).

5 Discussion

5.1 Datasets

One of the topics in the CLEF dataset, CD010653, has no Y. While we can still calculate performance scores relative to M, this topic might arguably have been omitted from the test data. One of the topics, CD008803, similarly has no M. This also happens to be the topic with the second largest number of Y.

As a general tendency, we can observe that the relative number of Y / M / N in the CLEF dataset varies dramatically across topics. At the one end we have one topic consisting of 14.06% Y (CD008760), and one topic consisting of 15.79% Y (CD010705). At the other end we have five topics with less than 0.1% Y (CD011548, CD011549, CD012019, CD011515, and CD009263). The number of N also varies wildly, from 52 up to 78,679.

	Y||MN					YM||N				
		RF					RF			
Topic	static	unigram	bigram	combined	baseline	static	unigram	bigram	combined	baseline
ALL	0.169	0.176	0.124	0.203	0.014 ± 0	0.313	0.314	0.218	0.337	0.053 ± 0
CD008122	0.331	0.274	0.327	0.344	0.042 ± 0.013	0.744	0.706	0.652	0.748	0.146 ± 0.001
CD008587	0.045	0.033	0.043	0.094	0.004 ± 0	0.076	0.063	0.062	0.109	0.009 ± 0.001
CD008759	0.477	0.543	0.283	0.549	0.047 ± 0.001	0.562	0.620	0.326	0.609	0.101 ± 0.010
CD008892	0.278	0.342	0.329	0.511	0.022 ± 0.001	0.323	0.376	0.361	0.462	0.043 ± 0.002
CD009175	0.085	0.095	0.003	0.059	0.002 ± 0.001	0.206	0.156	0.025	0.130	0.013 ± 0.002
CD009263	0.060	0.022	0.000	0.103	0.000 ± 0.000	0.116	0.104	0.003	0.038	0.002 ± 0.001
CD009694	0.435	0.447	0.494	0.843	0.084 ± 0.014	0.734	0.774	0.411	0.694	0.102 ± 0.018
CD010213	0.040	0.061	0.018	0.053	0.002 ± 0.000	0.260	0.250	0.195	0.226	0.042 ± 0.003
CD010296	0.450	0.535	0.074	0.541	0.011 ± 0.002	0.512	0.563	0.082	0.568	0.017 ± 0.005
CD010502	0.209	0.254	0.186	0.334	0.028 ± 0.003	0.339	0.409	0.323	0.467	0.080 ± 0.007
CD010657	0.176	0.206	0.070	0.196	0.028 ± 0.001	0.386	0.406	0.213	0.421	0.079 ± 0.003
CD010864	0.079	0.054	0.013	0.020	0.002 ± 0.001	0.084	0.082	0.113	0.133	0.023 ± 0.000
CD011053	0.065	0.063	0.019	0.048	0.007 ± 0.005	0.105	0.105	0.035	0.080	0.011 ± 0.005
CD011126	0.111	0.107	0.018	0.042	0.003 ± 0.001	0.145	0.141	0.027	0.070	0.003 ± 0.001
CD011420	0.062	0.056	0.263	0.215	0.021 ± 0.000	0.341	0.336	0.644	0.742	0.178 ± 0.000
CD011431	0.216	0.166	0.167	0.231	0.026 ± 0.004	0.649	0.626	0.662	0.669	0.262 ± 0.018
CD011515	0.050	0.028	0.071	0.042	0.001 ± 0.001	0.298	0.369	0.302	0.360	0.017 ± 0.001
CD011602	0.002	0.002	0.002	0.003	0.001 ± 0.000	0.018	0.014	0.021	0.037	0.004 ± 0.002
CD011686	0.015	0.012	0.005	0.047	0.002 ± 0.001	0.289	0.201	0.111	0.162	0.005 ± 0.001
CD011912	0.212	0.195	0.453	0.266	0.013 ± 0.001	0.374	0.365	0.447	0.481	0.031 ± 0.007
CD011926	0.428	0.540	0.028	0.129	0.008 ± 0.000	0.479	0.569	0.037	0.165	0.013 ± 0.002
CD012009	0.051	0.149	0.027	0.041	0.009 ± 0.002	0.387	0.317	0.192	0.455	0.085 ± 0.010
CD012010	0.090	0.125	0.102	0.106	0.002 ± 0.001	0.253	0.295	0.272	0.354	0.050 ± 0.001
CD012083	0.612	0.436	0.335	0.602	0.022 ± 0.003	0.373	0.313	0.243	0.378	0.040 ± 0.004
CD012165	0.072	0.075	0.013	0.073	0.005 ± 0.001	0.347	0.348	0.046	0.291	0.031 ± 0.002
CD012179	0.183	0.193	0.075	0.201	0.015 ± 0.002	0.374	0.343	0.123	0.356	0.033 ± 0.002
CD012216	0.016	0.016	0.013	0.012	0.014 ± 0.008	0.268	0.246	0.222	0.285	0.089 ± 0.023
CD012281	0.012	0.024	0.091	0.155	0.001 ± 0.000	0.026	0.027	0.080	0.210	0.003 ± 0.001
CD012599	0.054	0.059	0.080	0.042	0.002 ± 0.000	0.266	0.266	0.260	0.253	0.074 ± 0.004

Table 2: Average precision score for each topic, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The combined model uses the static and RF bigram as subcomponents.

	Y||MN					YM||N				
		RF					RF			
Topic	static	unigram	bigram	combined	baseline	static	unigram	bigram	combined	baseline
ALL	0.741	0.815	0.668	0.824	0.104 ± 0.024	0.513	0.617	0.519	0.657	0.028 ± 0.009
CD008122	0.800	0.794	0.772	0.788	0.018 ± 0.033	0.403	0.455	0.415	0.453	0.005 ± 0.013
CD008587	0.839	0.838	0.836	0.896	0.034 ± 0.047	0.772	0.746	0.696	0.759	0.012 ± 0.026
CD008759	0.746	0.764	0.612	0.736	0.019 ± 0.037	0.685	0.703	0.612	0.668	0.015 ± 0.030
CD008892	0.891	0.884	0.788	0.883	0.048 ± 0.052	0.040	0.534	0.694	0.486	0.006 ± 0.027
CD009175	0.936	0.916	0.546	0.915	0.073 ± 0.111	0.027	0.532	0.285	0.532	0.011 ± 0.029
CD009263	0.465	0.920	0.117	0.861	0.041 ± 0.084	0.418	0.408	0.122	0.557	0.006 ± 0.020
CD009694	0.826	0.832	0.678	0.813	0.045 ± 0.091	0.521	0.795	0.320	0.683	0.061 ± 0.073
CD010213	0.278	0.834	0.647	0.825	0.038 ± 0.049	0.065	0.590	0.556	0.341	0.002 ± 0.009
CD010296	0.928	0.924	0.723	0.924	0.028 ± 0.042	0.906	0.909	0.588	0.918	0.022 ± 0.034
CD010502	0.346	0.617	0.757	0.646	0.019 ± 0.030	0.298	0.587	0.405	0.609	0.002 ± 0.014
CD010657	0.739	0.741	0.345	0.757	0.034 ± 0.044	0.473	0.453	0.404	0.503	0.006 ± 0.018
CD010864	0.914	0.885	0.837	0.854	0.197 ± 0.193	0.215	0.506	0.571	0.619	0.017 ± 0.036
CD011053	0.909	0.913	0.537	0.903	0.076 ± 0.112	0.913	0.913	0.766	0.906	0.105 ± 0.095
CD011126	0.921	0.929	0.819	0.910	0.048 ± 0.091	0.933	0.935	0.860	0.917	0.096 ± 0.094
CD011420	0.719	0.715	0.831	0.823	0.114 ± 0.144	0.572	0.575	0.585	0.627	0.015 ± 0.034
CD011431	0.763	0.733	0.696	0.703	0.025 ± 0.048	0.017	0.162	0.275	0.173	0.003 ± 0.011
CD011515	0.947	0.945	0.948	0.947	0.459 ± 0.290	0.398	0.178	0.679	0.721	0.005 ± 0.020
CD011602	0.879	0.864	0.877	0.890	0.448 ± 0.283	0.750	0.786	0.806	0.870	0.059 ± 0.098
CD011686	0.937	0.910	0.844	0.875	0.284 ± 0.231	0.584	0.285	0.457	0.811	0.022 ± 0.034
CD011912	0.871	0.874	0.854	0.883	0.053 ± 0.067	0.843	0.850	0.654	0.841	0.032 ± 0.044
CD011926	0.933	0.933	0.483	0.916	0.017 ± 0.047	0.928	0.926	0.483	0.909	0.024 ± 0.041
CD012009	0.713	0.734	0.362	0.476	0.150 ± 0.158	0.584	0.592	0.362	0.476	0.026 ± 0.042
CD012010	0.020	0.671	0.579	0.744	0.064 ± 0.105	0.004	0.534	0.261	0.581	0.001 ± 0.013
CD012083	0.925	0.900	0.835	0.897	0.122 ± 0.144	0.180	0.512	0.727	0.605	0.117 ± 0.102
CD012165	0.818	0.824	0.308	0.828	0.013 ± 0.035	0.779	0.769	0.234	0.774	0.002 ± 0.012
CD012179	0.804	0.790	0.403	0.819	0.010 ± 0.022	0.750	0.723	0.363	0.769	0.002 ± 0.012
CD012216	0.669	0.655	0.597	0.577	0.444 ± 0.289	0.669	0.655	0.583	0.581	0.112 ± 0.103
CD012281	0.880	0.886	0.931	0.923	0.054 ± 0.095	0.716	0.745	0.622	0.762	0.031 ± 0.054
CD012599	0.080	0.413	0.807	0.877	0.053 ± 0.067	0.154	0.384	0.422	0.476	0.001 ± 0.009

Table 3: WSS@95 score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The combined model uses the static and RF bigram as subcomponents.

	Y||MN					YM||N				
		RF					RF			
Topic	static	unigram	bigram	combined	baseline	static	unigram	bigram	combined	baseline
ALL	0.640	0.762	0.633	0.779	0.130 ± 0.024	0.349	0.460	0.339	0.510	0.027 ± 0.007
CD008122	0.459	0.496	0.378	0.481	0.016 ± 0.015	0.289	0.320	0.040	0.332	0.003 ± 0.003
CD008587	0.782	0.848	0.769	0.845	0.029 ± 0.028	0.419	0.475	0.393	0.412	0.012 ± 0.012
CD008759	0.031	0.276	0.325	0.368	0.021 ± 0.020	0.031	0.276	0.325	0.368	0.016 ± 0.016
CD008892	0.828	0.887	0.576	0.875	0.031 ± 0.031	0.072	0.358	0.576	0.390	0.014 ± 0.014
CD009175	0.986	0.966	0.596	0.965	0.123 ± 0.111	0.010	0.381	0.264	0.269	0.015 ± 0.015
CD009263	0.515	0.970	0.167	0.911	0.091 ± 0.084	0.018	0.061	0.047	0.218	0.008 ± 0.008
CD009694	0.876	0.882	0.728	0.863	0.095 ± 0.091	0.565	0.720	0.228	0.708	0.051 ± 0.051
CD010213	0.019	0.520	0.582	0.727	0.029 ± 0.029	0.001	0.043	0.274	0.061	0.001 ± 0.002
CD010296	0.918	0.914	0.638	0.917	0.026 ± 0.026	0.918	0.914	0.418	0.917	0.019 ± 0.018
CD010502	0.335	0.629	0.626	0.684	0.014 ± 0.014	0.324	0.581	0.163	0.585	0.004 ± 0.004
CD010657	0.550	0.526	0.331	0.553	0.028 ± 0.028	0.057	0.058	0.103	0.047	0.007 ± 0.007
CD010864	0.964	0.935	0.887	0.904	0.247 ± 0.193	0.254	0.423	0.383	0.351	0.021 ± 0.022
CD011053	0.959	0.963	0.587	0.953	0.126 ± 0.112	0.959	0.957	0.587	0.953	0.078 ± 0.070
CD011126	0.971	0.979	0.869	0.960	0.098 ± 0.091	0.971	0.979	0.869	0.960	0.073 ± 0.070
CD011420	0.769	0.765	0.881	0.873	0.164 ± 0.144	0.343	0.530	0.575	0.534	0.020 ± 0.020
CD011431	0.707	0.665	0.724	0.695	0.036 ± 0.034	0.019	0.029	0.033	0.064	0.003 ± 0.003
CD011515	0.997	0.995	0.998	0.997	0.509 ± 0.290	0.171	0.012	0.386	0.575	0.007 ± 0.008
CD011602	0.929	0.914	0.927	0.940	0.498 ± 0.283	0.800	0.836	0.856	0.920	0.109 ± 0.098
CD011686	0.987	0.960	0.894	0.925	0.334 ± 0.231	0.069	0.051	0.198	0.798	0.018 ± 0.017
CD011912	0.886	0.902	0.704	0.897	0.051 ± 0.048	0.877	0.866	0.460	0.877	0.027 ± 0.026
CD011926	0.302	0.871	0.383	0.867	0.033 ± 0.034	0.302	0.871	0.383	0.867	0.025 ± 0.024
CD012009	0.763	0.784	0.412	0.526	0.200 ± 0.158	0.437	0.457	0.270	0.285	0.024 ± 0.024
CD012010	0.070	0.721	0.629	0.794	0.114 ± 0.105	0.027	0.180	0.067	0.226	0.003 ± 0.003
CD012083	0.975	0.950	0.885	0.947	0.172 ± 0.144	0.168	0.540	0.294	0.618	0.084 ± 0.075
CD012165	0.072	0.362	0.179	0.442	0.020 ± 0.019	0.039	0.347	0.087	0.367	0.003 ± 0.003
CD012179	0.141	0.482	0.205	0.464	0.008 ± 0.008	0.141	0.367	0.200	0.401	0.003 ± 0.003
CD012216	0.719	0.705	0.647	0.627	0.494 ± 0.289	0.576	0.599	0.303	0.627	0.078 ± 0.075
CD012281	0.930	0.936	0.981	0.973	0.104 ± 0.095	0.724	0.726	0.453	0.730	0.040 ± 0.038
CD012599	0.129	0.301	0.857	0.619	0.051 ± 0.048	0.092	0.089	0.109	0.067	0.001 ± 0.002

Table 4: WSS@100 score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The combined model uses the static and RF bigram as subcomponents.

	Y||MN					YM||N				
		RF					RF			
Topic	static	unigram	bigram	combined	baseline	static	unigram	bigram	combined	baseline
ALL	3349.448	1305.034	3798.000	1224.655	6405.696 ± 272.238	5708.400	5173.467	5500.600	4378.900	7131.769 ± 36.629
CD008122	1034	964	1189	991	1880.775 ± 29.665	1358	1300	1835	1276	1905.126 ± 6.638
CD008587	1998	1390	2113	1418	8890.107 ± 252.363	5317	4803	5559	5378	9042.512 ± 105.947
CD008759	903	675	630	589	912.361 ± 18.900	903	675	630	589	917.406 ± 15.133
CD008892	258	170	636	187	1452.336 ± 46.459	1391	962	636	914	1478.265 ± 21.061
CD009175	80	190	2282	195	4947.315 ± 626.643	5586	3492	4156	4125	5558.439 ± 85.764
CD009263	38214	2340	65642	6984	71659.995 ± 6650.289	77389	73961	75061	61604	78178.984 ± 632.362
CD009694	20	19	44	22	145.670 ± 14.589	70	45	125	47	152.735 ± 8.235
CD010213	14915	7297	6348	4144	14753.984 ± 445.766	15185	14543	11039	14269	15174.940 ± 22.935
CD010296	379	394	1665	382	4481.412 ± 121.595	379	394	2677	382	4516.729 ± 84.967
CD010502	1986	1108	1116	944	2942.072 ± 42.574	2018	1252	2500	1238	2973.515 ± 12.346
CD010657	836	882	1244	831	1806.271 ± 52.839	1753	1752	1668	1772	1846.911 ± 12.177
CD010864	90	164	283	240	1886.456 ± 482.878	1869	1445	1546	1625	2451.308 ± 54.825
CD011053	92	83	923	106	1952.562 ± 250.296	92	97	923	106	2060.096 ± 157.085
CD011126	174	128	784	238	5414.703 ± 543.169	174	128	784	238	5564.254 ± 418.618
CD011420	58	59	30	32	209.813 ± 36.150	165	118	107	117	246.108 ± 5.095
CD011431	346	396	326	361	1139.302 ± 40.689	1160	1148	1144	1106	1178.709 ± 3.773
CD011515	20	36	14	24	3553.976 ± 2097.615	6003	7160	4452	3079	7190.031 ± 54.788
CD011602	435	529	448	370	3088.216 ± 1740.224	1229	1011	886	495	5485.916 ± 605.398
CD011686	123	382	997	710	6291.365 ± 2182.568	8787	8965	7573	1903	9270.875 ± 161.630
CD011912	160	138	417	145	1334.026 ± 67.988	173	188	760	173	1368.069 ± 35.908
CD011926	2827	524	2501	537	3915.805 ± 135.943	2827	524	2501	537	3948.887 ± 97.154
CD012009	127	116	316	254	428.898 ± 84.684	302	291	392	383	523.100 ± 12.992
CD012010	6352	1907	2537	1405	6049.525 ± 719.498	6645	5601	6374	5284	6807.614 ± 22.027
CD012083	8	16	37	17	266.458 ± 46.415	268	148	228	123	294.965 ± 24.196
CD012165	9488	6521	8394	5706	10013.510 ± 193.570	9824	6673	9337	6468	10189.351 ± 31.388
CD012179	8446	5097	7813	5269	9750.778 ± 80.874	8446	6225	7863	5893	9800.761 ± 31.725
CD012216	61	64	77	81	109.840 ± 62.640	92	87	152	81	200.049 ± 16.240
CD012281	695	631	183	263	8851.610 ± 939.890	2728	2706	5400	2669	9479.651 ± 374.432
CD012599	7009	5626	1153	3070	7636.046 ± 387.458	7308	7328	7171	7512	8035.456 ± 13.165

Table 5: Last rel score for all topics in the CLEF dataset, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The combined model uses the static and RF bigram as subcomponents.

	Y||MN					YM||N				
		RF					RF			
Metric	static	unigram	bigram	combined	baseline	static	unigram	bigram	combined	baseline
AP	0.169	0.176	0.124	0.203	0.014 ± 0.000	0.313	0.314	0.218	0.337	0.053 ± 0.000
WSS@95	0.741	0.815	0.668	0.824	0.104 ± 0.024	0.513	0.617	0.519	0.657	0.028 ± 0.009
WSS@100	0.640	0.762	0.633	0.779	0.130 ± 0.024	0.349	0.460	0.339	0.510	0.027 ± 0.007
Last Rel	3349.448	1305.034	3798.000	1224.655	6405.696 ± 272.238	5708.400	5173.467	5500.600	4378.900	7131.769 ± 36.629

Table 6: Aggregate scores, evaluated using either inclusion decisions based on full text (Y||MN), or based on abstract and title (YM||N). The combined model uses the static and RF bigram as subcomponents.

5.2 Performance

No single model performs best on all topics. Generally however, **RF unigram** consistently outperforms the static model, and the **combined** model (**static** + **RF bigram**) outperforms the other three models.

Surprisingly, the **RF unigram** model consistently outperforms the **RF bigram** model, despite using a subset of the features of the **RF bigram** model. For this reason it seems likely that a stacked model consisting of the **static** model and the **RF unigram** model would have achieved better performance than the stacked model submitted as our official run.

The **RF unigram** model is particularly adept at finding all relevant articles, resulting in better last rel score than the **static** model for 19 topics out of 29, and a better last rel score than the **RF bigram** model for 24 out of 29. This also results in a WSS@100 score of 76.2% for the **RF unigram**, versus 64.0% for the **static** model, and 63.3% for **RF bigram**.

Note however that last rel generates scores of wildly varying scale, and the large last rel scores for **static** and **RF bigram** are therefore almost entirely due to a few large outliers. In particular, 59% of the information contained in the last rel score for **RF bigram** is due to a single topic with a large number of candidate articles (CD009263). The metric may thus be useful when interpreted on individual topics, but not when averaged. The WSS@100 metric, which is equivalent to last rel on individual topics, produces scores on the same scale and therefore makes sense also when averaged.

6 Conclusions

Our best system combines a static model and a relevance feedback model using stacking. The workload reduction to retrieve 95% of relevant articles is estimated at 82.4% on average, with a minimum workload reduction of 47.6%, and a maximum workload reduction of 94.7%. The workload reduction is consistent across topics, and we note a workload reduction less than 70% in only two topics. Due to the highly variable number of candidate articles in different topics, however, we

may still need to screen several thousands of articles to find all relevant articles in any given systematic review.

Our remarks on the implementation of the shared task model and task organization from last year [5] remain valid for this edition of the TAR task.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 676207.

Bibliography

[1] Cormack, G.V., Grossman, M.R.: Autonomy and reliability of continuous active learning for technology-assisted review. arXiv preprint arXiv:1504.06868 (2015)

[2] Cormack, G.V., Grossman, M.R.: Technology-assisted review in empirical medicine: Waterloo participation in clef ehealth 2017. Working Notes of CLEF pp. 11–14 (2017)

[3] Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Overview of the CLEF technologically assisted reviews in empirical medicine. In: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation forum, Dublin, Ireland, September 11-14, 2017. CEUR Workshop Proceedings, CEUR-WS.org (2017)

[4] Kanoulas, E., Li, D., Azzopardi, L., Spijker, R.: Overview of the CLEF technologically assisted reviews in empirical medicine 2018. In: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings (2018)

[5] Norman, C., Leeflang, M., Névéol, A.: Limsi@CLEF eHealth 2017 task 2: Logistic regression for automatic article ranking (2017)

[6] O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., Ananiadou, S.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic reviews 4(1), 5 (2015)

[7] Petersen, H., Poon, J., Poon, S.K., Loy, C.: Increased workload for systematic review literature searches of diagnostic tests compared with treatments: Challenges and opportunities. JMIR medical informatics 2(1), e11 (2014)

[8] Suominen, H., Kelly, L., Goeuriot, L., Kanoulas, E., Azzopardi, L., Spijker, R., Li, D., Névéol, A., Ramadier, L., Robert, A., Palotti, J., Jimmy, Zuccon, G.: Overview of the clef ehealth evaluation lab 2018. clef 2018. In: 8th Conference and Labs of the Evaluation Forum. Lecture Notes in Computer Science, Springer (2018)