A Logistic Regression Approach for NTCIR-11 Temporalia

Ray R. Larson

University of California, Berkeley

School of Information

For the Temporalia TIR task Berkeley used two retrieval methods for different submitted runs, including

1. Logistic Regression with Blind Feedback
2. Logistic Regression without Feedback

This poster presentation summarizes our approach to retrieval and the results when these methods were applied to the Temporalia TIR task.

Note that in this is preliminary approach using only the raw text content of the documents and topics, we did not attempt to identify or calculate actual dates, or parse the temporal phrasing in finding and ranking results.

Logistic Regression Ranking

Probability of relevance is based on Logistic regression from a sample set of documents to determine values of the coefficients. The sample used for this task was trained on data and queries from the 2nd TREC evaluation, hence we refer to it as TREC2.

At retrieval the probability estimate is obtained by:

$$P(R \mid Q,D) = \frac{e^{\log O(R \mid Q,C)}}{1 + e^{\log O(R \mid Q,C)}} = b_0 + \sum_{i=1}^{m} b_i X_i$$

For some set of *m* statistical measures, X_i, derived from the collections and queries

TREC2 Ranking Elements

Term Freq for:

$$\log O(R\,|\,C,Q) = c_o + c_1 \frac{1}{\sqrt{|Q_c|+1}} \sum_{i=1}^{|Q_c|} \frac{qtf_i}{ql+35}$$

Query

$$+\, c_2 \frac{1}{\sqrt{|Q_c|+1}} \sum_{i=1}^{|Q_c|} \log \frac{tf_i}{cl+80}$$

Document

$$+\, c_3 \frac{1}{\sqrt{|Q_c|+1}} \sum_{i=1}^{|Q_c|} \log \frac{ctf_i}{N_t}$$

Collection

$$+\, c_4 |Q_c|$$

# Matching Terms

Where *C* denotes a document component (i.e., an indexed part of a document which may be the entire document) and *Q* a query, *R* is a relevance variable, log *O(R|C,Q)* is the log odds that document component *C* is relevant

to query *Q*,

$|Q_c|$ is the number of matching terms between a document component and a query,

qtf_i is the within-query frequency of the *i*th matching term,

tf_i is the within-document frequency of the *i*th matching term,

ctf_i is the occurrence frequency in a collection of the *i*th matching term,

ql is query length (i.e., number of terms in a query like $|Q|$ for non-feedback situations),

cl is component length (i.e., number of terms in a component), and

N_t is collection length (i.e., number of terms in a test collection).

c_k are the *k* coefficients obtained though the regression analysis.

Pseudo-Relevance Feedback

Term selection from top-ranked documents in the initial TREC2 search is based on the classic Robertson/Sparck Jones probabilistic model:

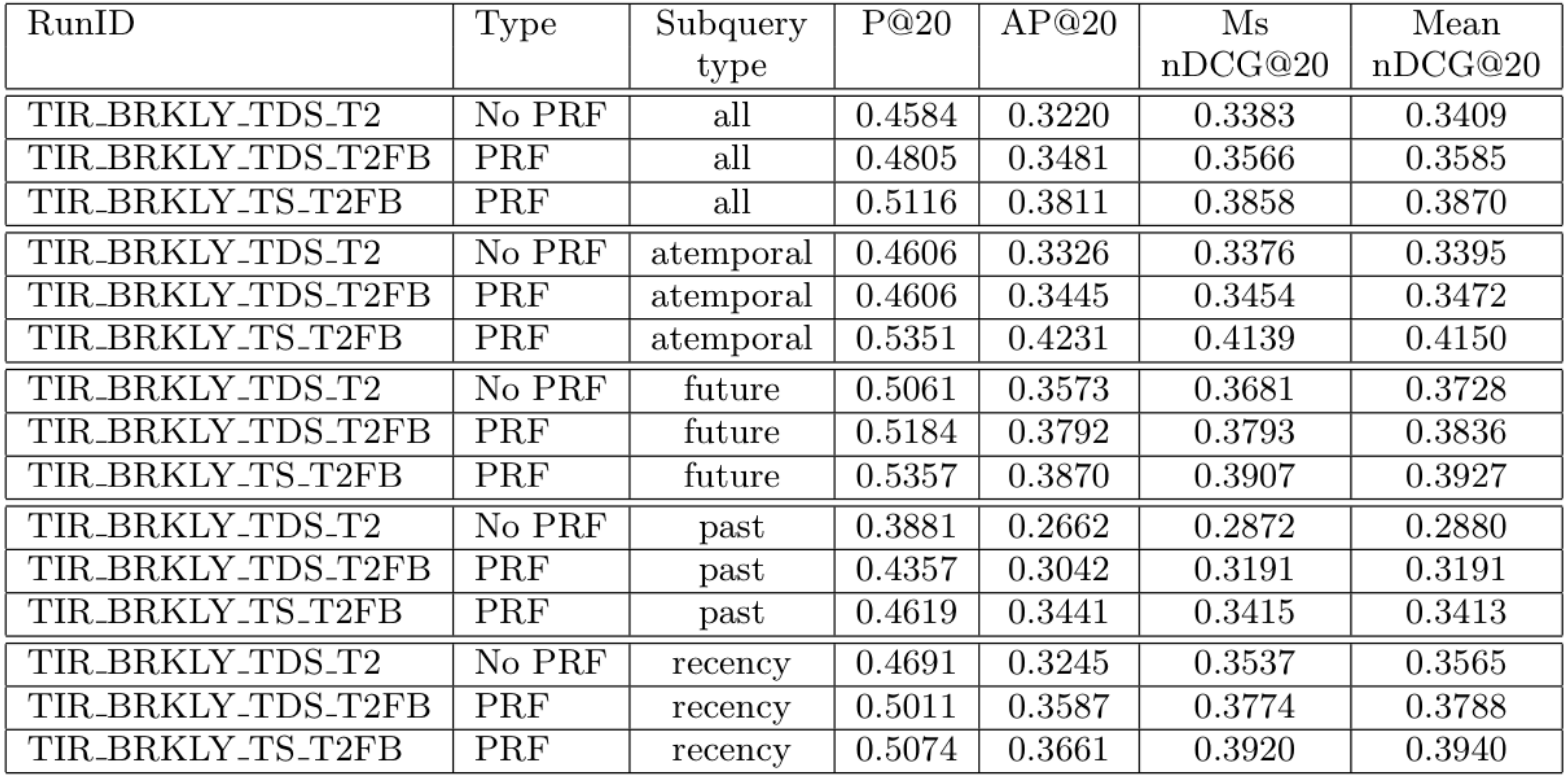
		Document Relevance		
For each term *t*		**+**	**-**	
Document indexing	**+**	R_t	$N_t - R_t$	N_t
	-	$R - R_t$	$N - N_t - R + R$	$N - N_t$
		R	$N - R$	N

Top *x* new terms are taken from top *y* documents
For each term in the top y assumed relevant set…

$$termwt = \log \frac{\left(\dfrac{R_t}{R - R_t}\right)}{\left(\dfrac{N_t - R_t}{N - N_t - R + R_t}\right)}$$

Terms are ranked by *termwt* and the top *x* selected for inclusion in the revised query.

Results for each of the 3 runs for each type of Temporalia TIR topic



RunID	Type	Subquery type	P@20	AP@20	Ms nDCG@20	Mean nDCG@20
TIR_BRKLY_TDS_T2	No PRF	all	0.4584	0.3220	0.3383	0.3409
TIR_BRKLY_TDS_T2FB	PRF	all	0.4805	0.3481	0.3566	0.3585
TIR_BRKLY_TS_T2FB	PRF	all	0.5116	0.3811	0.3858	0.3870
TIR_BRKLY_TDS_T2	No PRF	atemporal	0.4606	0.3326	0.3376	0.3395
TIR_BRKLY_TDS_T2FB	PRF	atemporal	0.4606	0.3445	0.3454	0.3472
TIR_BRKLY_TS_T2FB	PRF	atemporal	0.5351	0.4231	0.4139	0.4150
TIR_BRKLY_TDS_T2	No PRF	future	0.5061	0.3573	0.3681	0.3728
TIR_BRKLY_TDS_T2FB	PRF	future	0.5184	0.3792	0.3793	0.3836
TIR_BRKLY_TS_T2FB	PRF	future	0.5357	0.3870	0.3907	0.3927
TIR_BRKLY_TDS_T2	No PRF	past	0.3881	0.2662	0.2872	0.2880
TIR_BRKLY_TDS_T2FB	PRF	past	0.4357	0.3042	0.3191	0.3191
TIR_BRKLY_TS_T2FB	PRF	past	0.4619	0.3441	0.3415	0.3413
TIR_BRKLY_TDS_T2	No PRF	recency	0.4691	0.3245	0.3537	0.3565
TIR_BRKLY_TDS_T2FB	PRF	recency	0.5011	0.3587	0.3774	0.3788
TIR_BRKLY_TS_T2FB	PRF	recency	0.5074	0.3661	0.3920	0.3940

All of our submitted runs for the GeoTime track used probabilistic retrieval using TREC2 logistic regression algorithm described in detail above. Two of our submitted runs also used pseudo or blind relevance feedback along with the TREC2 algorithm, indicated by "PRF" in the Feedback Type column. For each RunID in the table, those with "TDS" in the RunID name used the Title, Description and Subquery elements of the topics, and those with "TS" did not use the Description. As the scores in this table show, using both the title and subquery elements along with blind feedback gives the best results for this collection and approach, while including the description leads to slightly worse results.

Overall, this approach was considerably less effective than the approaches taken by other participants in this task, undoubtedly due to our pure text-only approach with no attempt to explicitly leverage temporal cues and information.