Text Retrieval Baseline for GeoTime

Ray R. Larson
School of Information
University of California, Berkeley

Overview

- For GeoTime Berkeley participated all of the tasks, both English and Japanese Monolingual and Cross-Language tracks
- We used a number of retrieval methods for different runs, including
  - Logistic Regression with Blind Feedback
  - Logistic Regression without Feedback
  - Okapi BM-25 without feedback

Logistic Regression Ranking

Probability of relevance is based on Logistic regression from a sample set of documents to determine values of the coefficients.

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 collection and query

TREC2 Algorithm

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

Term Freq for: 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

Blind Feedback

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

For each term *t*

Document indexing	Document Relevance: +	Document Relevance: -	
+	R_t	$N_t - R_t$	N_t
-	$R - R_t$	$N - N_t - R + R$	$N - N_t$
	R	$N - R$	N

Blind Feedback

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

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

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

Okapi BM25

$$\sum_{T \in Q} w^{(1)} \frac{(k_1+1)tf}{K+tf} \frac{(k_3+1)qtf}{k_3+qtf}$$

- *Where:*
- *Q* is a query containing terms *T*
- *K* is $k_1((1-b) + b.dl/avdl)$
- k_1, *b* and k_3 are parameters , usually 1.2, 0.75 and 7-1000
- *tf* is the frequency of the term in a specific document
- *qtf* is the frequency of the term in a topic from which *Q* was derived
- *dl* and *avdl* are the document length and the average document length measured in some convenient unit
- $w^{(1)}$ is the Robertson-Sparck Jones weight:

$$w^{(1)} = \log \frac{\left(\frac{r+0.5}{R-r+0.5}\right)}{\left(\frac{n-r+0.5}{N-n-R+r+0.5}\right)}$$

GeoTime Submitted Runs

runID	Lang	mean AP	mean Q	mean nDCG	system description
BRKLY-EN-EN-02-DN	EN	0.4	0.42	0.61	Probabilistic retrieval based on logistic regression with blind feedback using DESCRIPTION and NARRATIVE text from topics.
BRKLY-EN-EN-03-D	EN	0.36	0.38	0.56	Probabilistic retrieval based on logistic regression with blind feedback using DESCRIPTION only text from topics.
BRKLY-EN-EN-04-DN	EN	0.34	0.36	0.58	Probabilistic retrieval based on logistic regression using DESCRIPTION and NARRATIVE text from topics.
BRKLY-EN-JA-01-DN	EN > JA	0.36	0.38	0.59	Probabilistic retrieval based on logistic regression with blind feedback using QUESTION and NARRATIVE text. EN->JA translation with GOOGLE Translate
BRKLY-EN-JA-02-T	EN > JA	0.35	0.36	0.54	Probabilistic retrieval based on logistic regression with blind feedback using QUESTION text only. EN->JA translation with GOOGLE Translate
BRKLY-JA-JA-01-DN	JA	0.43	0.45	0.65	Probabilistic retrieval based on logistic regression with blind feedback on QUESTION and NARRATIVE text.
BRKLY-JA-JA-02-T	JA	0.41	0.43	0.63	Probabilistic retrieval based on logistic regression with blind feedback on QUESTION text only.
BRKLY-JA-JA-03-DN	JA	0.16	0.16	0.31	Probabilistic retrieval based on OKAPI weighting on QUESTION and NARRATIVE text.
BRKLY-JA-JA-04-DN	JA	0.32	0.34	0.58	Probabilistic retrieval based on logistic regression using both QUESTION and NARRATIVE text.
BRKLY-JA-JA-05-T	JA	0.3	0.32	0.54	Probabilistic retrieval based on logistic regression using QUESTION text only.
BRKLY-JA-EN-01-DN	JA > EN	0.42	0.43	0.62	Probabilistic retrieval based on logistic regression with blind feedback using DESCRIPTION and NARRATIVE text from JA topics after Google Translate to EN topics.
BRKLY-JA-EN-02-D	JA > EN	0.38	0.39	0.56	Probabilistic retrieval based on logistic regression with blind feedback using DESCRIPTION only text from JA topics after Google Translate to EN topics.