Auto-Relevancy and Responsiveness Baseline II

Improving Concept Search to Establish a Subset with Maximized Recall for Automated First Pass and Early Assessment Using Latent Semantic Indexing [LSI], Bigrams and WordNet 3.0 Seeding

Cody Bennett [c_bennett@tcdi.com] – TREC Legal Track (Automatic; TCDI): TCDI - http://www.tcdi.com

Abstract

We experiment with manipulating the features at build time by indexing bigrams created from EDRM data and seeding the LSI index with thesaurus-like WordNet 3.0 strata. From experimentation, this produces fewer false positives and a smaller, more focused relevant set. The method allows concept searching using bigrams and WordNet senses in addition to singular terms increasing polysemous value and precision; steps towards a unification of Semantic and Statistical. Also, because of LSI and WordNet senses, WSD appears enhanced. We then apply an automated method for selecting search criteria, query expansion and concept searching from Reviewer Guidelines and the original Request for Production thereby returning a search result with scores across the Enron corpus for each topic. The result of the normalized cosine distance score for each document in each topic is then shifted based on the foundation of primes, golden standard, and golden ratio. This results in 'best cutoff' using naturally occurring patterns in probability of expected relevancy with limit approaching .5. Submissions A1, A2, A3, and AF include similar combinations of the above. Although we did not submit a mopup run, we analyzed the mopups for post assessment. For each of the three topics, there were documents which TAs selected as relevant in contention with their other personal assessments. The defect percentage and potential impact to a semi/automated system will also be examined. Overall the influence of humans involved (TAs) was very minimal, as their assessments were not allowed to modify any rank or probability of documents. However, the identification of relevant documents by TAs at low LSI thresholds provided a feedback loop to affect the natural cutoff. Cutoffs for A1, A2, A3 were nearly -.04 (Landau) against the Golden and Poisson means and F was nearly +.04 (Apéry). Since more work is required to decrease false positives, it is encouraging to find a natural relevancy cutoff that maximizes probable Recall of Responsiveness across differing topics.

Automated concept search using a mechanically generated semantically derived feature set upon indexed bigram and WordNet sense terms in an LSI framework reduces false positives and produces a tighter cluster of potentially responsive documents. Further, since legal Productions are essentially binary (R/NR), work was done to argue for scoring supporting this view. Obtaining Recall =>90% and Precision =>90% with a high degree of success is a two step process[1], of which we test and discuss the first (maximization of Recall) for this study. Therefore, our focus will be heavily skewed on the probability of attaining high Recall for the creation of a subset of the corpus.

Main Experiment Methods

See the TREC website for details on the mock Requests for Production, Reviewer Guidelines per topic and other information regarding scoring and assessing. Team TCDI's participation will be discussed without the repetition of most of that information.

Baseline Participation

TCDI's baseline submissions assume that by building a blind automated mechanism, the result is a distribution useful as a statistical snapshot, part of a knowledge and/or eDiscovery paradigm, and/or ongoing quality assurance and control within large datasets and topic training strata. Further, corporations' Information Management architectures currently deployed can offer hidden insights of relevancy when historically divergent systems[2] are hybridized. For TREC Legal Track 2011, TCDI's baseline submission considers a hybridization of NLP, Semantic and LSI[3] systems. 4 runs were submitted of 5 - we did not submit a "mopup" run. For runs A1, A2 and A3, some keyword filtering was tested. The Final run, AF used no keyword filtering. Multiple side experiments were performed, some discussed further. Steps for running the main experiment are listed below.

Feature Build for Indexing

[STEP 0] Baselines were submitted to TREC Legal using:

- 685,592 de-duped Enron emails and attachments conceptually indexed[4]
- Additional features per document beyond unigrams:
  - bigrams produced by a simple algorithm
  - small set of randomly selected WordNet 3.0 senses
- 3 Topics

Data inputs were the mock Requests for Production, Reviewer Guidelines, and phone conversations. Similar to some Web methods, the verbiage within the legal documents and discussions were expanded upon using a mixture of Natural Language Processing, WordNet sense non-linear distance, LSI

[1] During initial data assessment, automated maximization of Recall should be of highest value, since the Recall will carry over to human assisted systems such as Technology Assisted Review, and/or other search methodologies whose focus is to maximize Precision. In tandem, the approach will give a higher probability of attaining max P/R, and use hybridization techniques allowing for semi- / automated capabilities.

[2] Keyword vs. concept, concept vs. probabilistic, concept vs. semantic, etc. Esp. with IR systems, hybridization offers revitalization and ROI longevity.

[3] The semantic and conceptual systems could be considered plug and play for different approaches. The approach is considered modular as long as a topic model is available and exemplar data is available specifying relevant and non-relevant information.

[4] ContentAnalyst

and term and document frequency. Outputs were relevancy and rank among other metadata described in TREC Legal Track requirements.

Runs A1 and AF were automatic with no intervention, no feedback loop and no previous TREC seed sets. Runs 2 and 3 were used as subtle tests to gauge human / machine learning, with focus on how the document movement based on 40 human generated responsiveness calls per topic, out of the 1000 per topic allowable.

This automation provides a repeatable system, typical of black box approaches. This year's black box used "concept search" to obtain high recall in comparison to 2010's categorization approach. The addition of bigram and WordNet senses[A] should add focus to the standard concept search, attempting to give the syntax of LSI more semantic value.

Query Expansion

[STEP 1] By expanding on last year's methods, multiple inputs[5] were parsed and applied to the query expansion algorithm[B], creating 3 simple queries[6]. Topics 401, 402, 403:

- 401 - enrononline financial.instruments derivative.instruments commodities.enrononline enrononline.swaps enrononline.transactions enrononline.trades enrononline.commodity trading.enrononline enrononline.training
  (10 = 1 uni, 9 bi) [.11]
- 402 - otc.derivatives derivative.regulation regulate.otc botched.deregulation legal.instruments regulatory.instruments derivative
  (7 = 1 uni, 6 bi) [.17]
- 403 - environment environmental disaster oil.spill epa emissions enron.strategies habitats environmental.pollution noise.pollution oil.leaking environmental.policy environmental.training
  (13 = 6 uni, 7 bi) [.86]

[STEP 2] Queries from Step 1 were submitted to the concept index.

Of Natural Cutoffs, Mathematical Constants, and Golden Ratios[C]

We attempt to smooth quantify potentially relevant documents by applying approximations to the Golden Mean - "The desirable middle between two extremes". Further, the nature of LSI is reminiscent of fringe ideas similar to ideas from theories of Biolinguistics[D].

[STEP 3] Result scores were modified as below:

As θ = cos_sim() = $\frac{d \bullet q}{\|d\|\|q\|}$

- Probability is shifted based on ratio influence @ θ=0. Lower numbers have higher influence causing a stricter threshold:
  - ~.33 = (θ + .5) / (1 + L) - *Landau* [~.5] (**MIN**)[7]
    ~.37 = e^{-1} - *Poisson, Euler [~2.71828]* (MEAN)
    ~.38 = 1 - 1 / ϕ - Golden Mean [~1.61803] (MEAN)
    ~.42 = (θ + .5) / A - *Apéry* [~1.20205] (**MAX**)
- Conversion of 0:1 distribution to one approaching a lim of .5 based on loose rational approximations to the Golden Ratio using *Landau* for runs A1, A2, A3 and Apéry for AF
  - A1, A2, A3 = tcdicskwA1 = [**MIN**]
    AF = tcdinokaAF = [**MAX**]
- The change from [**MIN**] to [**MAX**] was based in part on runs A2 and A3, and the amount of resulting responsive documents determined by TA.

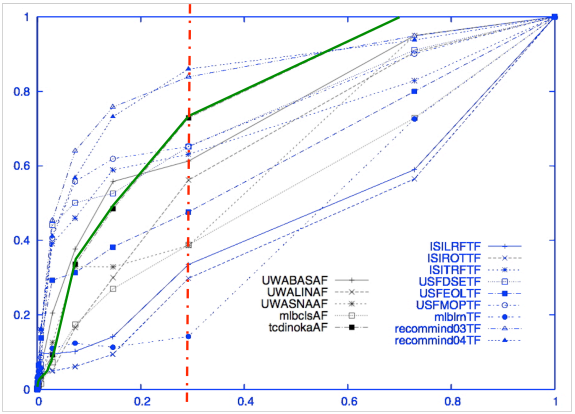
By applying this threshold conversion, a binary classification recalculation of 0:1 to <>.5 is possible. Probability of returning Responsiveness / Relevancy is mandated by values greater than .5[8].

Results

TCDI's runs without TA influence (AF) had preliminary avg. ROC AUC and Recall @ 200k scores at or above last year's highest averages[9].

To graphically set the stage, superimposed Gain graphs from the Legal Track assessors[10] of Automated and Technology Assisted (green is tcdinokaAF run, and red dashed is roughly the 30% corpora returned threshold) along with general comments are shown below:

Topic 401



- all runs appear to miss an unstated goal of >~90% Recall <=30% documents returned
- the topic and underlying relationships may be semantically heterogeneous, ambiguous

[5] Using verbiage from Mock Request for Production and portions of the Reviewer Guidelines from available at http://trec-legal.umiacs.umd.edu

[6] This is counterintuitive to how eDiscovery typically handles keyword expansion, human based analytics or other team efforts. However, it does not preclude these actions from improving the automated method's capability.

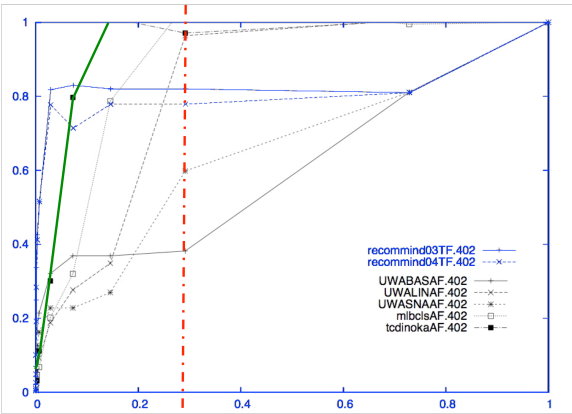
[7] Similarly approached by Robertson and Spark Jones, 1976 although for weight normalization.

[8] During litigation productions, if the document is leaving the door, it is considered Responsive / Relevant to the request - a very binary situation.

[9] Using the top average scores from tables at http://plg.uwaterloo.ca/~gvcormac/legal10/legal10a.pdf - page 3.

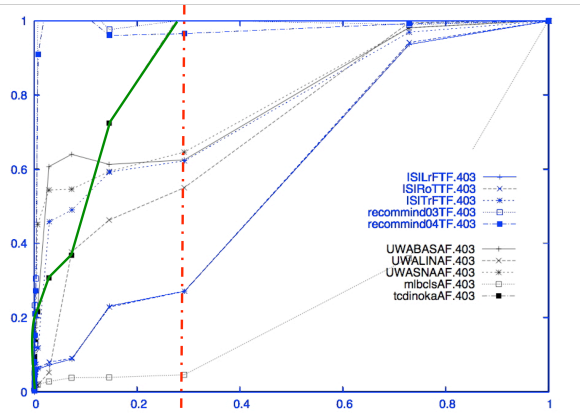
[10] Gordon Cormack, See http://trec-legal.umiacs.umd.edu

Topic 402



- automated runs appear to do well

Topic 403



- best illustration of the power of Technology Assisted Review (for this study) with automated systems following closely

Scoring for 2011 (as with 2010)

Scores for all TCDI submitted runs are listed [Figure 1]:

R@30	A1	A2	A3	AF	Avg	AF Actual
401	70.5	71.0	72.6	72.9	71.8	~91
402	94.2	93.8	92.2	97.1	94.3	~99
403	99.0	99.0	66.9	100.3	91.3	~99
Avg.	87.9	87.9	77.2	90.1	85.8	~96

ROC AUC	A1	A2	A3	AF	Avg
401	80.0	80.1	80.5	80.1	80.2
402	91.8	91.9	91.4	95.4	92.6
403	87.7	87.8	66.0	91.7	83.3
Avg.	86.5	86.6	79.3	89.1	85.4

Figure 1 - Submitted Runs

"AF Actual Recall" scores are those which were determined by documents humans have actually assessed. Other scores are obtained from the algorithmic probability[E] that documents will be Responsive / Non-responsive, but not returned by a human.

Arguably, however, document similarity and semantic degrees of separation[FG] based on "likeness" push potentially responsive outliers from the initial query direction. So, even if false negatives fall just below threshold, additional misses are less likely than higher degrees of similarity. But, "there is always one more[11]" document which may be relevant and nowhere near similar due to semantic ambiguity. The most important documents to a case arguably may be those which are in this outlier area, and more expensive to obtain.

Using this as a bookend as well as the notion of "Recall at 30% documents returned", we sought to refit a result set to naturally "break" at the center threshold of .5, so that statistical methods could be later employed to obtain outlying data.

If this natural threshold is used as described previously in the algorithm [Step 3], the combinatorial linguistic features should expose highest probability of responsiveness / non-responsiveness numerically / visually [Figure 2] with the noise falling to the left of .5 and the likely Responsive falling to the right of .5.

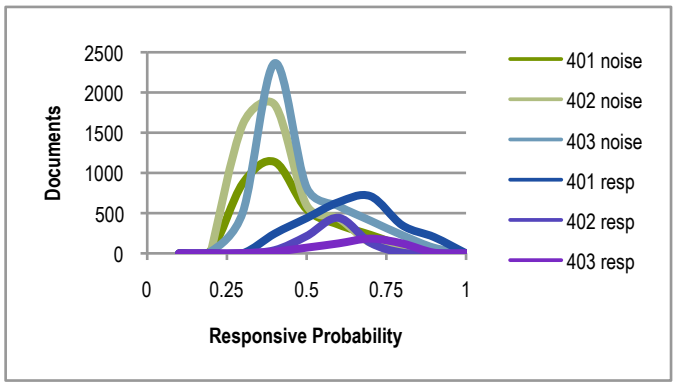


Figure 2 - Natural Cutoff (.5)

Other scores produced by the official TREC algorithm deal more with Precision. As an automated system, our Precision is baseline middle or lower as suggested by Hypothetical F1:

- Topic 401: 28.6%
- Topic 402: 8.7%
- Topic 403: 10.7%

Stepping away from probabilistically / hypothetically Responsive to those which humans actually assessed, scores are shown below: [Figure 3]

T	TP	FP	TN	FN	F1	F2	MCC
401	**2337**	**1238**	**2048**	**248**	**.76**	**.80**	**.54**
402	**800**	**1265**	**3475**	**43**	**.55**	**.64**	**.51**
403	**509**	**2109**	**2902**	**25**	**.32**	**.41**	**.32**

Figure 3 - Actuals

~98% of the False Negatives (2% FN in total for all topics) were =>40% and <50% - statistically examining the 40% area with

[11] Antithesis to the *Highlander*

sampling seems appropriate to further maximize possible Recall. The manual effort spent ascertaining initial sets with high Recall could arguably be spent on finding critical outliers.

More Work

Our Hypothetical F1 and other self-estimations need further work to automatically game TREC Legal Track's scoring algorithm. However, the automated runs' Recall appears to be successful with comparatively excellent AUC ROC and "Recall at 30% documents retrieved" [Figure 4]. Again, our goal is not to create an all encompassing document set for production, but to establish the best case subset with maximized Recall to pass to TAR.

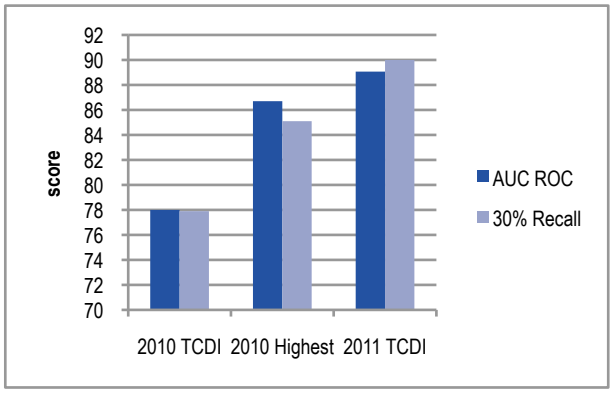


Figure 4 - 2010, 2011 Averages

A Posteriori Control

The control was the Final run since no keyword filters were used and a less restrictive (*Apéry)* threshold was employed. The control AF outperformed A1, A2 and A3. A1 assumed that by adding complex concepts into the index, a higher order of coupling would occur semantically, and therefore the application of a strict threshold. A2 and A3 measured high noise during the 40 human assessments, but there were intermittent hits of responsiveness that caused AF to apply a less intense threshold.

A1, A2 and A3 Topics 401 and 402 had "AND NOT" keyword filters as:

- dinner, lunch*, interiew*, drug.test*, gllery.openings, internetshortcut, job.application, trading.meeting, promotion

A1, A2 and A3 Topic 403 had "AND NOT" keyword filter as:

- dinner, lunch*, interiew*, drug.test*, gllery.openings, internetshortcut, job.application, trading.meeting, promotion, air.condition*, cont.air, us.air, air.force, dry.clean

AF had no keyword filters, solely using "concepts" (unigram and bigram).

Since AF was the superior run, the effects of the keyword filter appear negligible, although extensive analysis has yet to be performed.

Secondary Experiments

LSI Indexing Comparisons

One negative effect bigrams add to the LSI model, is an overall lowering of document scores. However, the removal of noise in 2 out of 3 topics may give weight to the usage of complex concepts as useful features.

For Topic 401 [Figure 5], LSI indexing "out of the box" [BoW+R] had less noise than an index using WordNet features and Bigrams [RWN+B+R] complex feature building.

For Topics 402 [Figure 6] and 403 [Figure 7 and 8], noise was reduced using complex feature building [RWN+B+R]. And in the case of Topic 403, dramatically reduced.

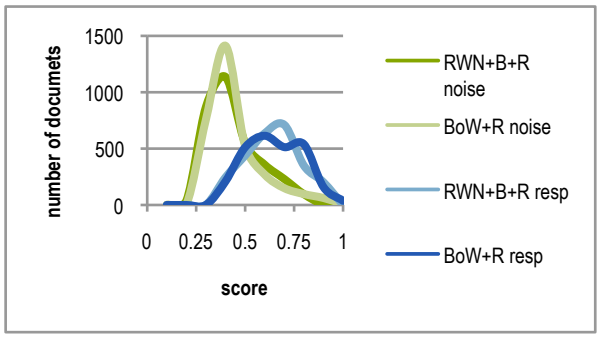


Figure 5 - Topic 401

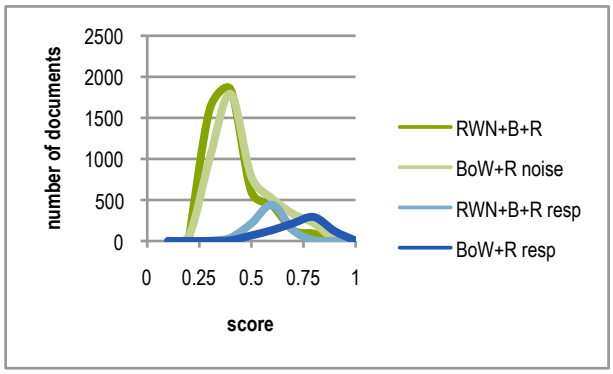


Figure 6 - Topic 402

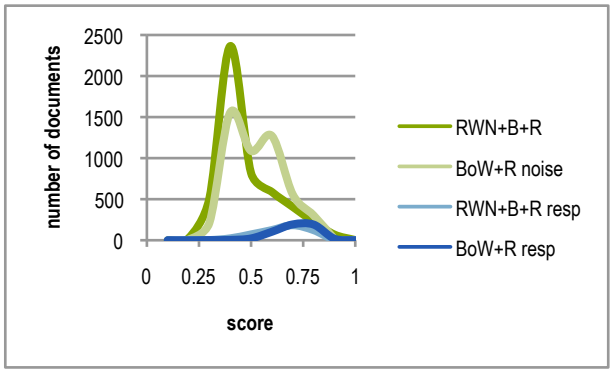


Figure 7 - Topic 403

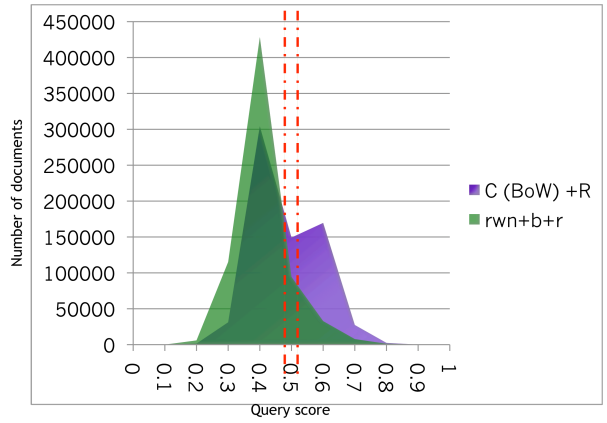


Figure 8 - Closer Examination of Topic 403

Semantic Exploration

WordNet[HI] 3.0 was used to add Ontology-like features to the statistical LSI index. While most of this work is arguably proprietary, there appears to be statistically valid Word Sense Disambiguation capabilities when the two are combined. Further, word senses seem to offer interesting context when requesting term->document relationships and overall topic modeling[12].

Conclusions

The application of a hybrid feature approach / complex concepts to Latent Semantic Indexing using very simple automated parsing and query construction appears promising in generating a high Recall set based solely on initial topic modeling (Request for Production). By reinterpreting a well known concept search method's (LSI) scoring and applying smoothing and best fit techniques found in many disciplines (besides IR)[13], automated runs across diverging topics can attain Actual Recall of ~90% with maximum documents returned at 30% of the corpus. Using this probabilistically predetermined rate of success, the subset of automatically accrued data can be sent downstream for further analysis, applied to a feedback system to further improve Recall at an attempt to completely maximize full potential, and/or to a Technology Assisted Review workflow. In any of these cases, the target should be maximizing Precision while allowing for best of breed statistical sampling / QC to assure max P / R. This automated study is not about replacing the human intelligence required to successfully complete an end-to-end review. It is one part of a display of how automated and human assisted workflows can in tandem guide a historically expensive process into a realm of data proportionality and expectation.

[12] Not quite Blei et. al with LDA, but closer to relationships manually attained with Upper Ontology.

[13] We affectionately name this method "*Blatant* Semantic Indexing".

[A] Fausto Giunchiglia, Uladzimir Kharkevich, Ilya Zaihrayeu, Concept Search: Semantics Enabled Information Retrieval, University of Trento, Italy http://www.ulakha.com/pubs/concept-search-tech-report-2010.pdf, 2010.

[B] Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[C] Dr. Bekir Taner Dincer, Mugla University for thoughts on Fibonacci, 2010.

[D] Dr. Juan Uriagereka

[E] Gordon Cormack and his extensive work with Spam detection - http://plg.uwaterloo.ca/~gvcormac/jig/

[F] Advanced Data Mining and Applications: 6th International Conference, ADMA, Finding Potential Research Collaborators in Four Degrees of Separation - By Longbing Cao, Yong Feng, Jiang Zhong.

[G] A. Budanitsky. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures, 2001.

[H] George A. Miller (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41.

[I] Christiane Fellbaum (1998, ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.