Auto-Relevancy Baseline

A Hybrid System Without Human Feedback

Cody Bennett [c_bennett@tcdi.com] – TREC Legal Track (Learning; TCD1): TCDI - http://www.tcdi.com

On obtaining a Request for Production and automatically emulating a typical eDiscovery workflow[1], a simple application of the classical Bayes algorithm upon the pseudo-hybridization of Semantic[A] and Latent Semantic Indexing[BC] systems should smooth out historically high yet noisy Recall of some LSI models and their derivatives and produce a tighter linear distribution when assessing relevant documents unsupervised.

Methods

See the TREC website for details on the differences between Interactive and Learning tasks, the mock Requests for Production, and other information regarding scoring and assessing. Team TCD1's participation will be discussed without the repetition of most of that information.

Baseline Participation

TCD1's submission assumes that by building a blind baseline mechanism, the result is an automated distribution useful as a statistical snapshot, part of a knowledge and/or eDiscovery paradigm, and 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.

Therefore, TCD1's baseline submission considers a hybridization of Semantic and LSI[3] systems. The features are mostly conceptual, as strict keyword targeting was purposefully not used in order to ascertain the effectiveness of Semantic + LSI.

[STEP 0] For verbosity, the baseline was submitted to TREC Legal Learning track using:

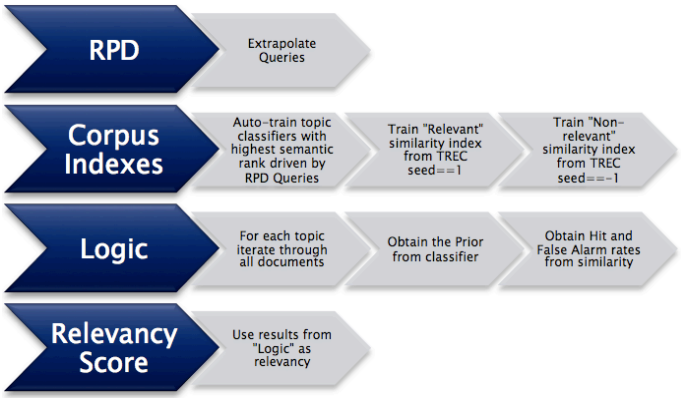
- 685,592 de-duped Enron emails and attachments Semantically indexed[4]
- A subset of 2010 TREC seed values = {1,-1} LSI conceptually indexed[5] - see *Seed Document Count*
- 8 topic iteration; 1 topic used as an attempted control
- 1 run submission; most teams submitted the maximum 3

Seed Document Count

topic	seed relevant docs	seed not relevant docs
200	230	0
201	168	40
202	994	47
203	67	82
204	59	38
205	333	122
206	19	10
207	80	11

Data inputs were two-fold – Request for Production features and seed stratum. Output was relevancy and rank among other metadata described in TREC requirements.

The run was automatic with no intervention, no feedback loop and no previous TREC seed sets. The method used a black box approach absorbing a Request for Production and mechanically determined relevancy and rank as output. As part of the relevancy assessment, the black box emulated a machine learning topic expert. Similar to some Web methods, the initial topics within the legal document were expanded upon using a mixture of Natural Language Processing, Semantic indexing and targeted contextual hit building.



High level of Algorithm

Semantic Query Expansion

[STEP 1] By using proprietary methods to locate topic request lines from the Request for Production and remove noise[D], 8 simple queries were created and used to drive query expansion[6]. As an example, initial topic truncation for topic 201 was:

[1] Essentially, Collection, Processing, Review, Analysis and near-Production were automated based on the verbiage of the Request for Production and TREC provided exemplars.

[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] The Semantic engine is proprietary and therefore will not be dissected in detail; consists of a mixture of NLP and Semantic mapping with a prebuilt verbose English training strata – see Dahlgren.

[5] The values and feature sets for indexing are proprietary and therefore will not be dissected in detail. While repeating the experiment when using historical Dumais *et al* LSI, the relevancy results are assumed to be approximately similar.

[6] This is counterintuitive to how eDiscovery typically handles keyword expansion.

Initial Search Feature – topic 201

Before	… 201. All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
After	"structured commodity transactions prepay", 201

[STEP 2] Queries from Step 1 were submitted to the Semantic index. Semantic breakdowns of term senses found at sentence level in context within the Enron corpus were returned. For each of these breakdowns, and for each of the documents found in the breakdowns, top sentence hits within highest rank documents > 80% of up to 400 bytes in length were used to populate a topic model directory (the feature counts are listed to the right). Using these extracted features, 8 topic model directories were created[7].

		P(H) Features	
topic	initial search features	ling docs	ling unique features
200	11	28	227
201	4	264	1351
202	5	4	31
203	9	23	196
204	11	98	773
205	10	113	661
206	9	8	47
207	9	101	725

Topic "Expert"

[STEP 3] A proprietary categorizer[8] ingested the 8 topic model directories as topic expert training[9].

Iteration

→ For every topic

→ For every document

*A document categorization request to the topic expert model returns the value used as **P(H)** (or .1 if none).*

*Perform a cosine based LSI similarity and return the highest relevant seed document score > .8 (or .5 if none). This value is used as **P(D|H)**.*

*Perform a cosine based LSI similarity and return the highest non-relevant seed document score > .8 (or .5 if none). This value is used as **P(D|H')**.*

Relevancy (Rl) was then found using Bayes[E]:

$$Rl = \frac{P(D\,|\,H)P(H)}{P(D\,|\,H)P(H) + P(D\,|\,H')(1-P(H))}$$

Probability Rank (Ra) was proprietary function.

→ End for

→ End for

[7] Further, this method directs the topic model build from the influence of the initial Request for Production.

[8] The categorizer can be as simple as a vector cosine comparison across a conceptual index.

[9] This type of human expert emulation simulates TREC's Legal Interactive track, except the role of the Legal expert is replaced by the Semantic knowledge of the machine.

Results

All Run Results

TCD1's baseline run averaged higher than other "baseline" submissions save one during both preliminary result assessments (Oct. 2010) and full raw result assessments (Jan. 2011):

Oct. 2010 Assessment Scores

	topic								avg		
run	200	201	202	203	204	205	206	207	actual	est	acc
otL10rvlT	39.8	85.5	100.2	100.5	85.2	84.6	98.9	86.6	85.1	98.8	86.1
xrceLogA	77.9	93.8	99.4	92.2	73.8	71.8	74.9	92.0	84.4	88.8	95.0
xrceCalA	77.9	93.8	99.4	92.2	73.8	71.8	74.9	92.0	84.4	82.7	97.8
DUTHsdtA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	90.1	92.0
DUTHsdeA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	82.4	99.3
DUTHlrgA	90.6	85.0	97.8	103.4	98.0	82.2	88.0	18.7	82.9	96.3	86.1
otL10FT	97.9	94.9	98.8	105.6	68.8	84.9	88.3	21.7	82.6	96.6	85.5
tcd1	67.2	61.6	98.2	77.9	76.2	57.5	97.5	87.4	77.9	55.8	71.6
rmitindA	72.9	85.8	96.7	102.5	79.2	87.7	78.3	19.8	77.8	53.3	68.5
otL10bT	52.4	82.1	102.7	108.3	49.5	65.2	97.6	51.1	76.1	99.2	76.1
xrceNoRA	83.2	66.1	85.5	95.9	35.2	54.4	76.5	92.0	73.6	73.2	99.4
BckExtA	78.9	75.1	90.0	38.6	67.5	72.7	85.5	80.9	73.6	49.7	67.5
BckBigA	80.7	75.1	89.9	36.1	67.4	72.7	85.5	80.9	73.5	49.5	67.3
rmitmlsT	66.6	61.6	104.3	83.8	45.6	57.8	57.9	16.3	61.7	70.7	87.3
BckLitA	44.1	76.9	88.2	77.0	42.5	74.6	57.7	11.7	59.0	49.6	88.4
rmitmlfT	68.7	59.9	90.7	52.2	47.5	56.1	58.6	15.7	56.1	67.1	83.6
ITD	-	45.6	67.5	20.7	41.6	35.2	29.7	74.7	44.9	61.8	72.6
URSK70T	51.0	17.7	18.6	23.8	62.2	22.4	87.6	22.6	38.2	91.0	42.0
URSK35T	51.3	27.7	18.4	27.6	40.5	30.3	90.3	18.3	38.0	93.2	40.8
URSLSIT	51.0	19.2	21.3	23.8	50.6	22.4	87.6	22.5	37.3	83.5	44.7

On average, with 2 topics falling below simulated control (discussed below), the baseline was essentially the median.

Jan. 2011 Assessment Scores

	topic								avg
run	200	201	202	203	204	205	206	207	actual
xrceCalA	77.9	97.1	97.8	91.3	73.8	66.7	77.8	92.0	84.3
xrceLogA	77.9	97.1	97.8	91.3	73.8	66.7	77.8	92.0	84.3
otL10FT	97.9	89.2	96.6	97.7	68.8	81.1	88.4	21.7	80.2
DUTHlrgA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6
DUTHsdeA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6
DUTHsdtA	90.6	90.9	72.1	97.5	98.0	80.9	88.2	18.7	79.6
otL10rvlT	39.8	88.3	64.5	83.4	85.2	82.9	99.6	86.6	78.8
BckExtA	78.9	74.0	75.4	71.4	67.5	75.4	85.0	80.9	76.1
BckBigA	80.7	74.1	75.4	66.4	67.4	75.4	85.0	80.9	75.7
rmitindA	72.9	92.3	72.5	98.0	79.2	85.0	80.9	19.8	75.1
tcd1	67.2	55.3	85.0	76.1	76.2	53.3	98.8	87.4	74.9
xrceNoRa	83.2	73.5	76.1	79.7	35.2	58.7	78.9	92.0	72.2
otL10bT	52.4	88.4	67.8	84.5	49.5	65.6	98.3	51.1	69.7
BckLitA	44.1	74.9	75.7	85.4	42.5	77.8	63.1	11.7	59.4
rmitmlsT	66.6	58.6	72.2	64.5	45.6	54.2	61.8	16.3	55.0
rmitmlfT	68.7	57.2	70.4	47.6	47.5	52.8	62.3	15.7	52.8
ITD		44.8	88.3		41.6	36.1	26.4	74.7	52.0
URSK70T	51.0	18.9	13.0	44.5	62.2	24.6	88.9	22.6	40.7
URSLSIT	51.0	21.0	21.1	44.5	50.6	24.6	88.9	22.5	40.5
URSK35T	51.3	25.1	15.2	45.1	40.5	27.7	91.8	18.3	39.4

TCD1 ranked top 2 in highest individual topic recall @ 200k documents (98.8%). Since TCD1 along with other teams appear to use topics 200 or 207 as controls and by removing the lowest score, a highest min-1 also shows TCD1 as median.

submission	max	topic
otL10rvlT	99.6	206
tcd1	98.8	206
otL10bT	98.3	206
DUTHlrgA	98.0	204
DUTHsdeA	98.0	204

submission	highest min-1	topic
xrceCalA	73.8	204
xrceLogA	73.8	204
rmitindA	72.5	202
DUTHlrgA	72.1	202
DUTHsdeA	72.1	202
[tcd1]	[55.3]	[201]

Control

The baseline system's simulated control – topic 200 – used 0

"non-relevant" documents; all document P(D|H') were deliberately scored as .5. The reasoning follows the possibility that Legal Learning judges/assessors were purposefully skewing topic seeds with false positives/false negatives. Topic scores below the control in a production system i.e. topics 201, 205 due to deteriorated Hit Rate and/or False alarm rates (seed topics used to define these rates) would be targets for reassessment of exemplars.

Further:

- Seed documents were occasionally not semantic representatives of the topic and conceptually ambiguous and noisy, at least noisier than other topics and therefore caused anomalies during vector comparisons[10].
- The types of features derived from the linguistic search expansion used to train the topic expert appear to be critical to the system; features extracted highlight the importance of a smart and very clear topic "expert" model[11].

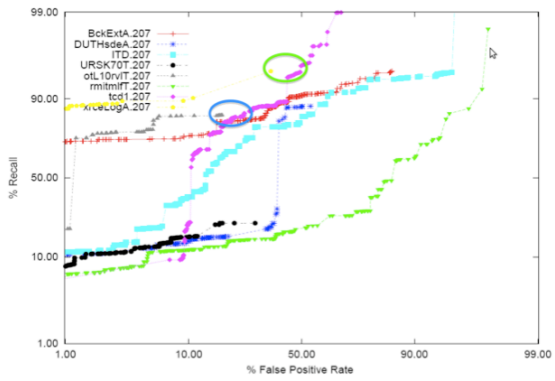
Overall, the baseline system in current form:

- Is a good QC indicator of the salience of topic exemplars when comparing to a control.
- Is dynamic; the system produced scores below the control but also top tier results on various topics. The high entropy will be a focus on future system iterations.
- Averages up to 85% returned @ 200k, closer to desired automation (ignoring the control and results below the control - 201, 205).
- The algorithm's document ranking probability estimates were considerably off on average (55.8%). In most cases, the rank algorithm needs to be adjusted upward.

Further work is needed to determine best next steps for increasing the recall at lower document cutoffs (increased precision). But it is clear from the entropy of scores delivered by the second raw assessment that topics require seed cleansing - something a statistical QC feature in future runs could determine automatically. Last, the baseline system while affected by errant seed exemplars did appear more robust in smoothing the LSI distribution due to the hybridization of a Semantic built expert in the application of Bayes.

Graphical Comparison

Below is the amount of noise and recall of relevant documents @ 200k cutoff generated by top submissions for topic 207:



Topic 207 - TREC Legal Presentation – TCD1 is Pink[12]

For topic 207, TCD1's system achieves the highest recall (99+%) but at the cost of false positives[13]. The blue circle shows the baseline approaching congruency of most other systems at recall ~88% / precision ~75%. The green circle shows the approach to highest recall capability at the cost of noise. The initial false positive rate of TCD1's submission is the likely result of the lack of use of strict keywords and phrases during topic building.

Conclusions

While the hypothesis was proven to a point based on initial tests and other teams using LSI, more noise detection and elimination is needed to achieve both high R and P > 98% @ < 200k documents returned. For automation and QA/QC purposes, 88% of non-biased topics[14] may be an acceptable threshold for use in knowledge systems[15] compared to the accuracy of human review[16]. However, it is of direct interest to judge the cost of noise as a monetary value similar to valuations performed in TREC Legal Interactive task. Significantly lowering the noise will provide a cleaner plateau to begin questioning, "what/when is the probability that the system may produce an errant document in review and what is this cost?[17]"

Moving beyond cost, further enhancements to the system will improve the precision of the topic model expert. In future iterations, work will be done to ascertain at the time of seed building the viability of the seeds[18]; checks to see if seed documents semantically overlap and cause inconsistency and/or tainting P(D|H) and P(D|H'). Also, a second and third pass before final scoring might be interesting to develop, where new (D|H) are learned from the semantic process.

Simply, the use of strict keyword features in addition to hybrid

[10] The use of seed topics is a mandatory step for this system since, in real rolling eDiscovery requests the exemplar training is iterative and dynamic.
[11] Some semantic features may be missing but could be found through multiple recursions. Also, noise introduced during topic expert feature building (P|H) appears to cause a similar dilemma as conflicting P(D|H) and P(D|H').

[12] Gordon Cormack, Maura Grossman
[13] TREC is using F1 as the official yardstick. If F2 is used, recall is weighted dominantly.
[14] The control (200) and 201, 205
[15] Grossman mentioned Xerox and TCDI topic 207 scores at TREC Legal Learning task presentation at Gaithersburg, Md.
[16] See *Inside Counsel*, Jan. 2011 "Computerized E-Discovery Review is Accurate and Defensible" - a real world test showing machine vs. human review with machine @ ~83% while human review teams @ ~76% and ~72%.
[17] The same question should be asked for human reviewers.
[18] In an active review, people and systems try to improve their models, not deliberately try to break them with false positives / false negatives – although human assessors do make mistakes. QA/QC is critical in determining these issues.

features employed on the topic expert should increase relevancy scores and decrease false positives. Online topic learning may add more precision to the topic expert. Also, Latent Dirichlet Allocation[F] would add an automated topic feature set for juxtaposition. Even further, use of rough fuzzy hybridization appears as a promising black box approach to automated IR tasks and learning[GH].

But regardless of future sophistications, TCD1's simplistic single-run hybrid baseline produced a peak topic score ~98.8% recall @ 200k directed by the Request for Production verbiage unsupervised. Next plans will reduce the document cutoff it takes to attain this recall "baseline".

[A] K. Dahlgren 1988. Naïve Semantics for Natural Language Understanding. http://www.cognitionsearch.com

[B] Deerwester, S., S. Dumais, T. Landauer, G. Furnas, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science* 41(6): 391-407

[C] ContentAnalyst – current LSI patent (and related) owner: http://www.contentanalyst.com

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

[E] Jaynes, E. T., 2003, *Probability Theory: The Logic of Science*, Cambridge University Press

[F] Blei, David M.; Lafferty, John D. (2006). "Correlated topic models". *Advances in Neural Information Processing Systems* 18.

[G] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough based approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.

[H] R. Jensen and Q. Shen, (2008) Rough and Fuzzy Hybridization, in Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches, John Wiley & Sons, Inc., Hoboken, NJ, USA.