Evaluating Stream Filtering for Entity Profile Updates in TREC 2012, 2013, and 2014 (KBA Track Overview, Notebook Paper)

John R. Frank[1], Max Kleiman-Weiner[1], Daniel A. Roberts[1]
Ellen Voorhees[2], Ian Soboroff[2]
1 KBA Organizers, Massachusetts Institute of Technology, jrf@mit.edu
2 National Institute of Standards and Technology Gaithersburg, MD ian.soboroff@nist.gov

Abstract

The Knowledge Base Acceleration (KBA) track ran in TREC 2012, 2013, and 2014 as an entity-centric filtering evaluation. This track evaluates systems that filter a time-ordered corpus for documents and slot fills that would change an entity profile in a predefined list of entities. Compared with the 2012 and 2013 evaluations, the 2014 evaluation introduced several refinements, including high-quality community metadata from running Raytheon/BBN's Serif named entity recognizer, sentence parser, and relation extractor on **579,838,246** English documents in the corpus. We also expanded the query entities to be primarily long-tail entities that lacked Wikipedia profiles. We simplified the SSF scoring, and also added a third task component for highlighting creative systems that used the KBA data. A successful KBA system must do more than resolve the meaning of entity mentions by linking documents to the KB: it must also distinguish novel **"vitally" relevant** documents and **slot fills** that would change a target entity's profile. This combines thinking from natural language understanding (NLU) and information retrieval (IR). Filtering tracks in TREC have typically used queries based on topics described by a set of keyword queries or short descriptions, and annotators have generated relevance judgments based on their personal interpretation of the topic. For TREC 2014, we selected a set of filter topics based on people, organizations, and facilities in the region between Seattle, Washington, and Vancouver, British Columbia: 86 people, 16 organizations, and 7 facilities. Assessors judged ~30k documents, which included ***most*** documents that mention a name from a handcrafted list of surface form names of the 109 target entities. TREC teams were provided with all of the ground truth data divided into training and evaluation data. We present peak macro-averaged F_1 scores for all run submissions. High scoring systems used a variety of approaches, including feature engineering around linguistic structures, names of related entities, and various types of classifiers. Top scoring systems achieved F_1 scores in the high-50s. We present results for a baseline system that performs in the low-40s. We discuss key lessons learned that motivate future tracks at the end of the paper.

Categories & Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – Information Filtering; H.3.m [Information Storage and Retrieval]: Miscellaneous – Test Collections; I.2.7 [Natural Language Processing] Text analysis – Language parsing and understanding

General Terms: Experimentation, Measurement

Introduction

This overview paper describes the progression of the KBA evaluation over the three years from 2012, 2013, and 2014.

TREC KBA is a stream filtering task focused on entity-level events in large volumes of data. Many large knowledge bases, such as Wikipedia, are maintained by small workforces of humans who cannot manually monitor all relevant content streams. As a result, most entity profiles lag far behind current events. KBA aims to help these scarce human resources by driving research on automatic systems for filtering streams of text for ***new*** information about **entities**. We refer to

such novel information as “vital,” which has a technical meaning as the highest relevance level in the assessor-generated graded relevance data in KBA.

KBA focuses on the boundary between IR and natural language understanding (NLU). Entities are widely used concept in NLU, which focuses on fully automatic algorithms. KBA transports the concept of entities into IR and uses entities as queries for an end user task. Entities are a special subset of general topics. Entities have strongly typed attributes and relationships, such as a name, birth date, father, hometown, employer, profession, which information retrieval (IR) systems can exploit to surface novel information. We refer readers to the TREC KBA 2013 Overview Paper for a more in-depth discussion of the goals of the TREC KBA track and its relation to other evaluations the data generated by and for the track, and the evaluation metrics.

This paper is organized as follows: Section 1 describes data assets generated by the evaluation. Section 2 discusses the scores from TREC KBA 2014. Section 3 concludes with three lessons learned from KBA, which motivate possible future directions that build on the KBA experience.

Data Assets

In addition to the three hundred run submissions from diverse systems, KBA produced a unique corpus and three sets of ground truth from human assessors on portions of that corpus. This data is available through NIST and also at trec-kba.org

The KBA corpus is the largest *stream* corpus ever released for open evaluations. It consists of 1.2 billion documents from a contiguous span of 19 months (13,663 hours) of news, blog, and Web content with several special substreams, including the Internet arXiv pre-print server. In 2014, we deployed BBN’s Serif NLP system on the English and likely-English documents, which make up more than half the corpus. For 2014, we expanded the corpus with two more months of Spinn3r data. Figure 1 illustrates the large scale of the corpus relative to smaller set filtered out for humans working on the particular query entities used in the evaluation.

Figure 1 depicts the corpus statistics over time, and Table 1 describes the truth data. See also powerpoint slides from trec-kba.org

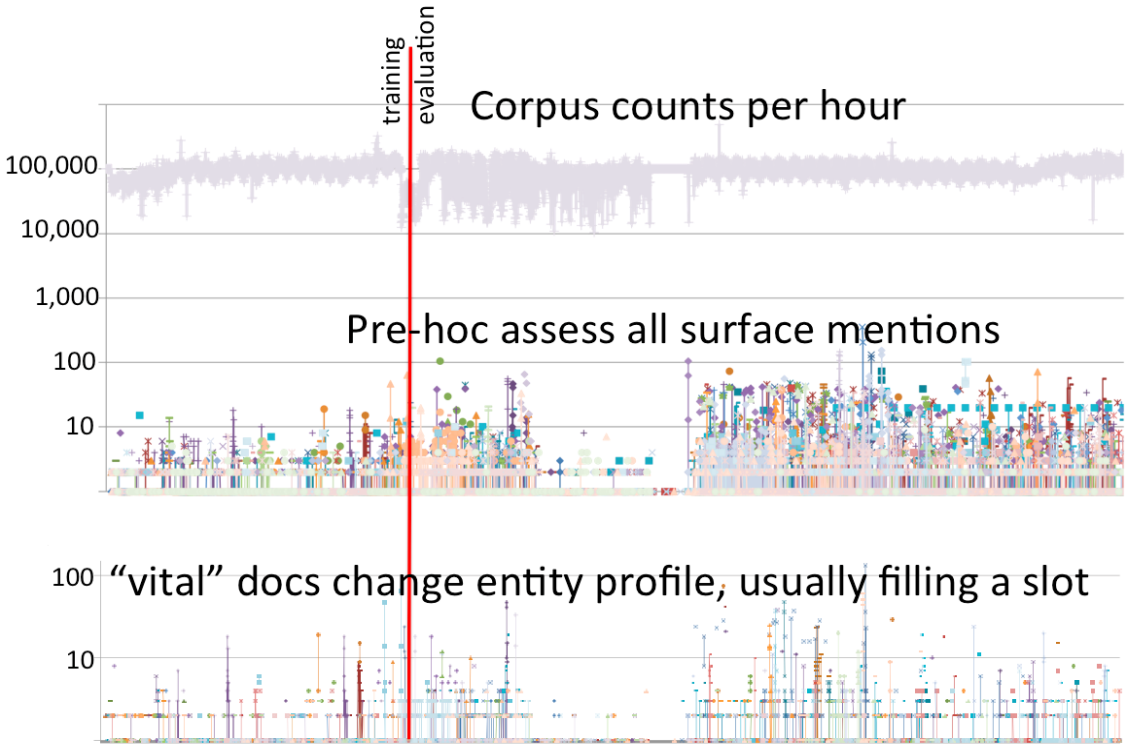

Figure 1: KBA's StreamCorpus and assessing data. (depicted using the 2013 data)

	2012	2013	2014
Corpus [1]	7 months (4,973 hours) >400M documents 40% English Oc 2011 to Apr 2012.	17 months (11,948 hours) >1B documents 60% English or unknown Oct 2011 to Feb 2013	19 months (13,663 hours) 1.2B documents 60% English or unknown Oct 2011 to April 2013
Queries (entities)	27 people, 2 organizations, all from Wikipedia	98 people, 19 organizations, and 24 facilities. Fourteen inter-related communities of entities, such small towns like Danville, KY, and Fargo, ND, and academic communities like Turing award winners.	86 people, 16 organizations, and 7 facilities all from the geographic region between Seattle and Vancouver.
Assessing	70% agreement on "central"	3198 hours have >0 vitals, 76% agreement on "vital" (replaced "central")	2503 hours have >0 vitals, 68% agreement on "vital"
Submissions	11 teams, 40 runs	13 teams, 140 runs	11 teams, 118 runs
Metrics	F_1, Scaled Utility	F_1, Scaled Utility	F_1, Scaled Utility [2]

CCR Assessing

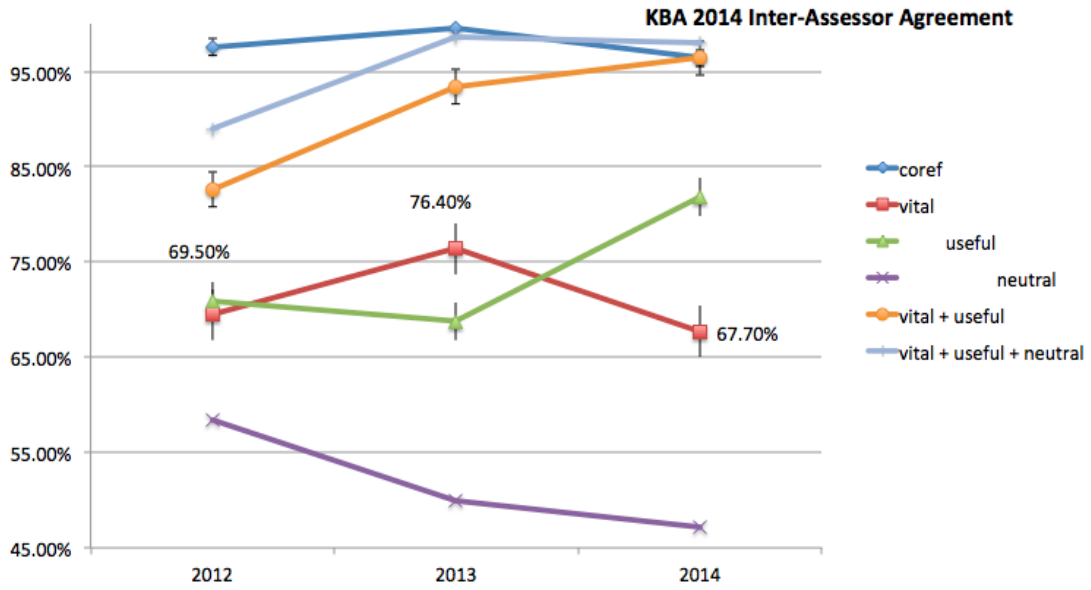

Figure 2: Inter-Assessor Agreement Scores across all three years.

The primary task in KBA is identifying documents that contain vital new information that would not have be in an up-to-profile at the time the document entered the stream. The difference between a vital update and background biographical info can be subjective in several ways. One particular aspect of subjectivity involves judging whether an event that changed the entity is being mentioned way after the fact. For example, a text might explain when an entity was born -- fifty years after the fact. Such reporting is obviously biographical (therefore "useful") and not sufficiently timely to be vital. Borderline examples exist. For example, should we consider this passage timely? "Sara helped start NXIVM in 2007." (reporting date 2011) No, that's not timely enough. That's useful, not vital. We refer the reader to assessor guidelines included with the truth data, which can be obtained through NIST.

Assessors must also decide what to include in an entity's profile. For example, for an entity with a Wikipedia article, a profile might include why the person is noteworthy. If the person is less noteworthy, the profile might simply describe how they spend their time.

Figure 2 illustrates interassessor agreement. The most important line is "Vital" colored red. This is the highest relevance level. Human agreement on the notion of vital varies across entities and domains. TREC KBA 2014 focused on less noteworthy entities than the many moderately famous people in TREC KBA 2013, and this increased subjectivity contributed to the lower inter-assessor agreement.

Top scoring systems in KBA 2014 captured linguistic signals in the sentence or short passage surrounding entity mentions. The precision gains found via these structures suggest that assessors' judgments of vitality are correlated with particular linguistic patterns, such as verbs, see MSR_KMG slides about action patterns.

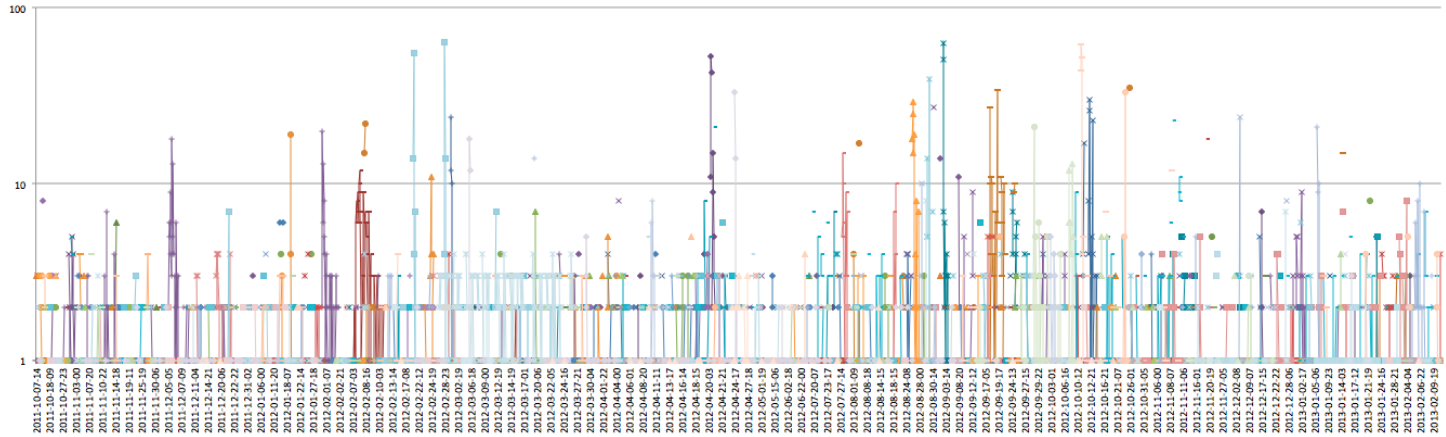

Figure 3: KBA 2013's 141 query entities and their vital document counts per hour. Several spikes are visible. Most spikes are echos in the blogosphere that reverberate after a single event, which is often characterized by changes to a few strongly typed entity attributes or relationships, such as death or the breakdown of a corporate or spousal relationship.

CCR Queries and Training Data

For 2014, an initial round of assessing was conducted in June using name match heuristics to find candidate documents for assessors to judge for each entity. Efforts were made to have one assessor complete the judging for one entity, although this was not always practical. The assessors are identified in the truth data by unique strings.

In July, track participants were provided with all the truth data divided into two sets: ~20% for training, and ~80% for evaluation. The boundary between training and evaluation was set separately for each entity at the hour by which 20% of the true positives had appeared in the stream. The remainder was used to measure the F_1 accuracy and scaled utility of these systems.

After run submissions were sent to NIST in September, we conducted an additional week of assessing to boost recall. The final scores are computed using this expanded set of truth data.

CCR Metrics (see SSF Metrics below)

The metrics for CCR are F_1 and Scaled Utility(SU) and are shown in Figure 4. Most systems had an SU below 0.333, which corresponds to a run with no output. The F_1 score is the harmonic mean of the macro-averaged precision and macro-averaged recall. In this context, macro-averaging means using the confidence cutoff for which the F_1 is highest for the system under study, and summing the precision (or recall) scores from all of the query entities and dividing by the number of entities. Some of track participants have invented a time-sensitive metric [5].

CCR Results

Figure 4 ranks the teams by their highest scoring system using the maximum F_1 or maximum Scaled Utility using two different retrieval objectives. The primary ranking is F_1 on the vital-only retrieval objective. This is a very hard task, as illustrated by the plateau of high ranking systems with scores similar to the baseline.

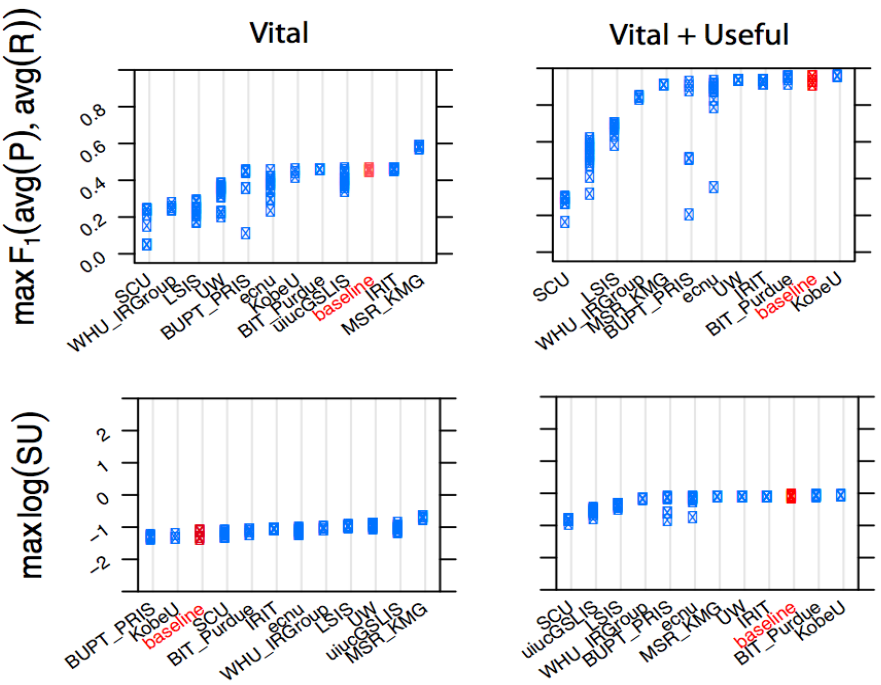

Figure 4: Official scores using "vital" and "vital+useful" as the classification objective. The red points represent the baseline system. In both 2013 and 2014, teams that built SSF systems consistently performed below the median in CCR.

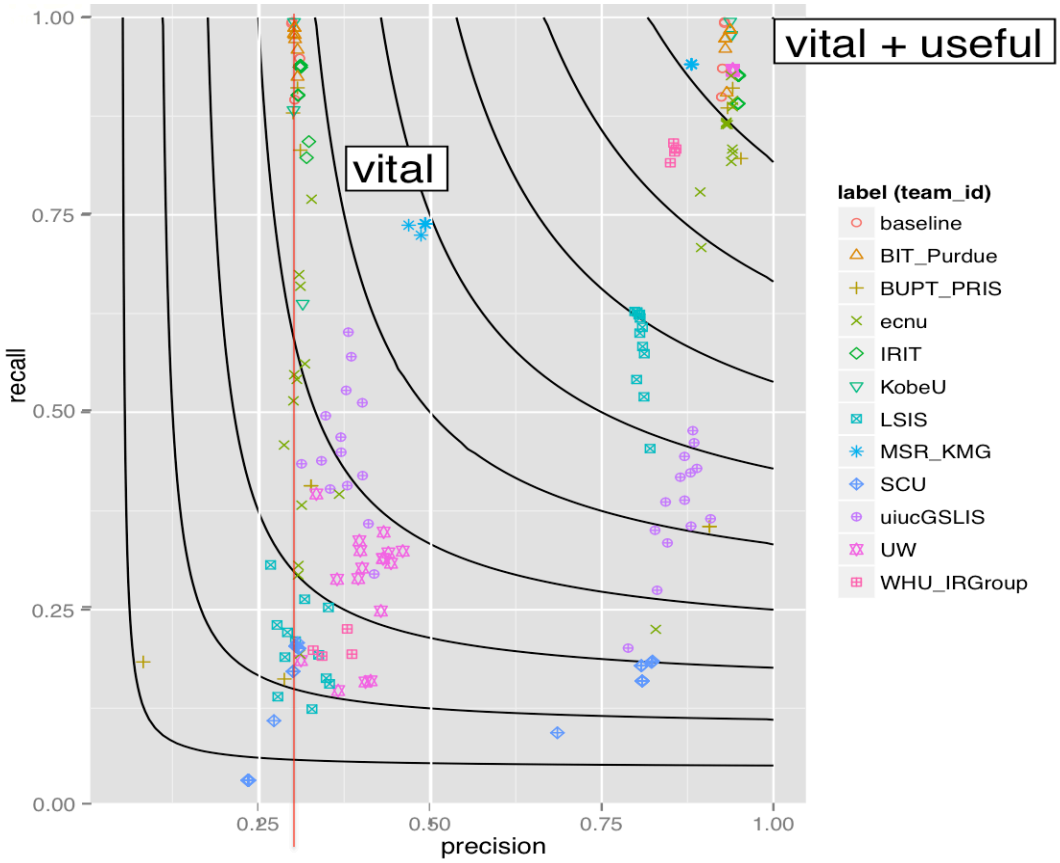

Figure 5: macro-averaged recall versus precision with curves of constant F_1 for "vital" CCR and also "vital+useful," which is equivalent to document-level to coreference resolution. The precision and recall values correspond to the maximum F_1 as a function of confidence cutoff.

CCR Baseline Systems

After considering several baseline systems to characterize the task, we decided to keep the official baseline that assigns a "vital" rating to every document that matches a surface form name of an entity and assigning a confidence score based on the number of matches of tokens in the name. See code in github [6]. macro-P=0.316, macro-R=0.520, macro-F=0.393, SU=0.3334 This is shown in the official score plots below as "baseline." In contrast to the so-called "oracle baseline" used in 2013, the baseline used in 2014 uses only the entity's canonical name instead of hand-picked variants. We also tested a similar oracle baseline that included string matches for all of the slot fills strings gathered by assessors had minimal impact on the baseline's score, and it scored almost the same as the non-oracle baseline.

As shown in Figure 5, systems had lower precision than recall. The baseline system's precision of 0.316 sets a threshold (vertical red line). Systems with precision above this line found signals that correlated with novelty. Both LSIS and UW reported improved runs shortly after the evaluation deadline, which are described in their TREC papers.

Streaming Slot Filling (SSF):

Compared with 2013, we significantly simplified the SSF evaluation process in 2014. Instead of detecting ***changes*** to particular slots, we simply asked systems to fill as many slots as possible. A summary of the most common slot types is displayed in Table 2.

		2013 (restricted to changes)				2014 (unrestricted)
num eval fills	num entities	slot type		num eval fills	num entities	slot type
232	50	Contact_Meet_PlaceTime		465	75	TITLE
96	41	Affiliate		258	24	MET_WITH
70	10	Contact_Meet_Entity		210	61	EMPLOYEE_OR_MEMBER_OF
27	10	AssociateOf		192	43	VISITED
19	12	AwardsWon		185	40	ATTENDED
10	8	Titles		145	6	VISITED_BY
6	3	TopMembers		131	64	GENDER
4	2	FoundedBy		56	30	NAME
2	2	DateOfDeath		54	12	TOP_MEMBERS_EMPLOYEES
2	1	EmployeeOf		44	19	WON_AWARD
1	1	SignificantOther		34	31	CITIES_OF_RESIDENCE
1	1	CauseOfDeath		32	7	MEMBERS
						(32 more slots, see truth data)

Table 2

To evaluate SSF systems, we considered the token overlap between the slot fills found by human assessors and those found by a system. To make compute this overlap comparison, we constructed bags of words from each systems' output and also from the truth data. Each bag of

words is a count vector in word space. Four comparison functions for bags of words are:

- dot product = <a, b> = sum_i a_i b_i
- cosine(a,b) = <a, b> / |a| / |b|
- c_TT(a,b) = sum_i 1 if a_i>0 and b_i > 0, which is similar to dot product with all counts set to one; often called "boolean overlap"
- sokalsneath(a, b) = c_TT / (c_TT + 2 (c_TF + c_FT)), which is form of normalization for c_TT. c_FT means the count of items that are in the right-side vector but not the left, and c_TF is the converse.

The dot product and c_TT metrics do not penalize systems for precision errors, so a naive system that outputs all of the text in the corpus will get a high score. The normalizations in cosine and sokalsneath penalize for precision errors, and makes them better metrics. By ignoring counts, the sokalsneath metric has the largest spread between systems. However, the familiar cosine metric properly penalizes the low-precision baseline system, which outputs entire sentences surrounding name-matching tokens. For this evaluation, cosine is the better metric.

See Table 3 in the Appendix for SSF scores.

Lessons Learned and Future Directions

KBA sits at the boundary between two different paradigms for exploiting textual content:

Natural Language Understanding (NLU)	**Information Retrieval (IR)**
• Prioritizes linguistics over user tasks	• Prioritizes user tasks over linguistics
• Universal annotation	• Relevance rating is subjective
• Seeks high interannotator agreement	• Expects a diversity of judgments
• Inference & probabilities	• Heuristics & scores to sort lists
• Reductionist, first principles	• Constructionist, emergence

Table 3

TREC KBA has established a foundation for temporally driven IR tasks at the entity level. This foundation opens up many new avenues of research. When we first proposed the track in the fall of 2011, we cited large knowledge bases and large data streams as key motivators. We also defined the track in contrast to the Knowledge Base *Population* track in the Text Analytics Conference (TAC KBP). Both evaluations aim to create knowledge bases from text mining. KBP evaluates fully automatic approaches and has consistently shown that key aspects of natural language understanding are beyond the state of the art. In contrast, KBA evaluates machine-assisted, human-driven knowledge base curation. In addition to the many lessons learned about different technical approaches, the KBA experience also taught us about how to structure evaluations for systems on the boundary between IR and NLU. We summarize three salient aspects of this intersection:

1. **While top performing KBA systems often use ranking signals generated by NLU algorithms that depend on fixed ontologies, such rigid schemas capture only a fraction of vitally relevant events.** NLU systems often generate entity attributes and relations from fixed ontologies of strongly typed properties. This approach facilitates machine learning and enables sophisticated logical inference in automatically constructed knowledge bases. However, we observed that only half of the vitally relevant documents in KBA 2013 were associated with a change to a value in one of the schematized slots. Even when such a change occurred, it often touched only tangentially on the salient information in the event. Despite the subtlety and complexity of strongly typed ontologies from evaluations such as ACE and TAC KBP, these knowledge representations evidently lack expressive power to fully capture events like this example: "M.I.A.," the rapper, tweeted that her daughter's father is using New York's child custody laws to perpetrate a

human rights violation. This information is not easily schematized into slots.

2. **When vital events trigger updates to a KB profile, they can also drive deeper research by the KB curator into the entity's past. In real KBs, this background research also updates the profile.** KBA's emphasis on vital events is motivated by real phenomena in the stream of content, see spikes in Figure 3. Streams of content have a periodic rhythm illustrated in Figure 1. These temporal structures deeply affect humans curating knowledge bases and systems designed to support them. KBA implemented a key aspect of this by instructing assessors to *define* vital documents relative to an "already up-to-date profile." This combines with a short, and subjective, time window of one to three days of vitality after a sudden event. A more natural user pattern expands the search task beyond that time window to find other information that is also missing from the profile. Updating entity profiles requires human editors to *both* monitor the stream and explore the past. For example, current events in the Kalispel Tribe build upon a long history that also needs to get into the KB.
3. **The KBA StreamCorpus captures the head of a long-tailed distribution of websites that change with varying frequency.** The KBA StreamCorpus contains many salient events for noteworthy people. Most web domains are less focused on current events, and therefore typically not crawled in the feed aggregations used to make the KBA StreamCorpus. As such, the KBA corpus skims the surface of many complex networks of related entities. Exploring these complex networks will require deeper, more targeted content harvesting to expand the corpus. All document collections are gathered at a particular time, and thus are inherently "stream corpora." The StreamCorpus processing tools [7] developed while assembling the KBA corpus are in active use building more data sets and provide a foundation for this further expansion.

We hope to address these issues in a successor track to KBA, called TREC Dynamic Domain. TREC DD starts in 2015 and builds on the foundation of TREC KBA. See trec-dd.org

Acknowledgements: The KBA organizers are grateful to all of the excellent teams who participated in making KBA such a great learning experience (see appendix). We also thank Boyan Onyshkevych and Alan Goldschen for ideas and helpful discussions. JRF, MKW, DAR thank the Fannie and John Hertz Foundation for support. Ian Soboroff and Ellen Voorhees at TREC have been instrumental in designing and organizing KBA. We also wish to thank Diffeo, MIT, University of Wisconsin, the Open Science Grid, Amazon, Spinn3r, and the arXiv for their generous support of TREC KBA.

1: http://s3.amazonaws.com/aws-publicdatasets/trec/kba/index.html
http://s3.amazonaws.com/aws-publicdatasets/trec/kba/kba-streamcorpus-2013-v0_2_0-chunk-path-to-stream-id-time.txt.xz

2: https://github.com/trec-kba/kba-scorer/

3: http://www.nist.gov/tac/2012/KBP/task_guidelines/TAC_KBP_Slots_V2.4.pdf

4: http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf

5: http://ciir.cs.umass.edu/~dietz/streameval/taia2013-camerayready.pdf

6: https://github.com/trec-kba/kba-tools/

7: https://github.com/trec-kba/streamcorpus-pipeline/

Appendix: SSF Scores

	unnormalized			normalized	
uses occurrence counts	**team_system**	**dot product**		**team_system**	**cosine**
	baseline-ssf	5107		BUPT_PRIS-ssf2	61.12
	baseline-ssf_oracle	4724		SCU-ssf_9	54.56
	ecnu-ssf_run	1061		SCU-ssf_12	54.56
	BUPT_PRIS-ssf2	782		SCU-ssf_8	54.56
	BUPT_PRIS-ssf1	601		SCU-ssf_7	54.56
	SCU-ssf_6	423		SCU-ssf_14	54.56
	SCU-ssf_11	423		SCU-ssf_13	54.56
	SCU-ssf_2	423		SCU-ssf_3	52.61
	SCU-ssf_1	422		SCU-ssf_10	52.61
	SCU-ssf_3	421		SCU-ssf_5	52.61
	SCU-ssf_10	421		SCU-ssf_6	49.89
	SCU-ssf_5	421		SCU-ssf_11	49.89
	SCU-ssf_9	247		SCU-ssf_2	49.89
	SCU-ssf_12	247		SCU-ssf_1	44.59
	SCU-ssf_8	247		ecnu-ssf_run	42.68
	SCU-ssf_7	247		BUPT_PRIS-ssf1	41.72
	SCU-ssf_14	247		SCU-ssf_4	37.03
	SCU-ssf_13	247		baseline-ssf	29.06
	SCU-ssf_4	127		baseline-ssf_oracle	26.71
ignores occurrence counts	**team_system**	**c_TT**		**team_system**	**sokalsneath**
	baseline-ssf	3560		baseline-ssf	310.22
	baseline-ssf_oracle	3227		baseline-ssf_oracle	281.14
	ecnu-ssf_run	483		BUPT_PRIS-ssf2	91.51
	BUPT_PRIS-ssf2	481		BUPT_PRIS-ssf1	90.32
	BUPT_PRIS-ssf1	380		ecnu-ssf_run	55.45
	SCU-ssf_6	269		SCU-ssf_6	36.58
	SCU-ssf_11	269		SCU-ssf_11	36.58
	SCU-ssf_2	269		SCU-ssf_2	36.58
	SCU-ssf_3	267		SCU-ssf_3	36.10
	SCU-ssf_10	267		SCU-ssf_10	36.10
	SCU-ssf_5	267		SCU-ssf_5	36.10
	SCU-ssf_1	262		SCU-ssf_1	35.79
	SCU-ssf_9	129		SCU-ssf_9	26.95
	SCU-ssf_12	129		SCU-ssf_12	26.95
	SCU-ssf_8	129		SCU-ssf_8	26.95
	SCU-ssf_7	129		SCU-ssf_7	26.95
	SCU-ssf_14	129		SCU-ssf_14	26.95
	SCU-ssf_13	129		SCU-ssf_13	26.95
	SCU-ssf_4	64		SCU-ssf_4	18.64

Participants:

2012	2013	2014
Centrum Wiskunde & Informatica (CWI)	Beijing Institute of Tech (BIT)	BIT_Purdue
HLT COE	CWI	BUPT_PRIS
IGPI	Institut de Recherche en Inform. de Toulouse (IRIT)	East China Normal University (ECNU)
LSIS	LSIS/LIA:	Kobe University
Pattern Recognition and Intelligent System (PRIS)	PRIS	IRIT
SCIA	Santa Clara University (SCU)	LSIS
U. of Amsterdam (UvA)	Stanford University	Microsoft Research, KMG
U. Delaware (UDel)	UvA	SCU
U. Helsinki	UDel	UIUC
U. Illinois, Urbana-Champaign (UIUC)	U. of Florida	U. Washington
UMass Amherst	UIUC	WHU
	UMass Amherst	
	U. Wisconsin	