Multi-Faceted Question Fusion in the TREC 2022 CrisisFACTS Track

Nathaniel W. Rollings
nrolling@umd.edu

Peter A. Rankel
par@terpmail.umd.edu

Douglas W. Oard
oard@umd.edu

Abstract

To address the challenges of multi-faceted questions in rapidly evolving environments, this paper introduces a system with an architecture based on recency-weighted and score-weighted reciprocal rank fusion of per-facet ranked lists. In the absence of existing data for parameter tuning, a small test collection built to support formative evaluation was developed and employed in system refinement. Issues of duplication were addressed through pruning near duplicates and, in one variation, synthesizing rather than simply selecting responses.

1 Introduction

In the event of a crisis, responders need regular updates to multi-faceted standing questions but have limited time available to consume information, so prioritization of provided information is critical. However, as the environment is quickly changing, previously useful information may become outdated in short order. Management of crisis response requires updated information—a report on the size of a wildfire from the previous day is of little use if the fire has since doubled in size. This updated information may appear on one or many platforms and, particularly on social media sites, may appear many times on a given platform.

The 2022 Text Retrieval Conference's (TREC-2022) inaugural CrisisFACTS Track centers on providing relevant and recent answers to the questions that crisis response managers would include in daily reports. The goal is to provide this as a single ranked list each day, emphasizing coverage of each facet for which information is available for that day. Each of the three key issues—facet coverage, relevance, and recency—plays a role in the design of the system discussed in this paper.

2 The CrisisFACTS Task

The goal of a CrisisFACTS system is to select materials for inclusion in summary reports on the organized response to a natural disaster. These multifaceted reports are typically prepared daily, so the task is to organize content on which each daily report can be based.

Facets (which within the track are called "queries"[1]) were drawn from information found in Federal Emergency Management Agency (FEMA) Incident Command System (ICS) ICS-209 forms that are prepared

[1] In the CrisisFACTS track, "query" refers both to a facet and to the field in the facet description that contains a natural language question. For clarity, we refer to the first of those as a facet and the second as a question.

Event ID	Event	Date	Facets	Number of Days
CrisisFACTS-001	Lilac Wildfire	2017	52	5
CrisisFACTS-002	Cranston Wildfire	2018	52	3
CrisisFACTS-003	Holy Wildfire	2018	52	4
CrisisFACTS-004	Hurricane Florence	2018	51	6
CrisisFACTS-005	Maryland Flood	2018	48	2
CrisisFACTS-006	Saddleridge Wildfire	2019	52	2
CrisisFACTS-007	Hurricane Laura	2020	51	2
CrisisFACTS-008	Hurricane Sally	2020	51	3

Table 1: Event information.

by command center personnel each day during an event. For a given event, there are between 48 and 52 facets, as shown in Table 1. While the set of facets for each event is identical across the days of the event (although on some days there may be no relevant content to be found some facets), the facet sets do differ between event types. Each facet includes a "query" field containing a question describing the facet (e.g., "Have airports closed") and an "indicative terms" field that contains keywords similar to what might be seen in a Web query (e.g., "airport closed"). We refer to these as the "question" and the "indicative terms."

The TREC-2022 CrisisFACTS track provides a collection of short text passages extracted from social media (Twitter and Facebook), news sources, and Reddit relating to one of eight different natural disasters, referred to generically as "events". Each event extends over several days, ranging from two days for the shortest event to eighteen for the longest (although the track created relevance judgments for no more than six days for any one event). The passages are provided as JSON, with each entry including a Unix timestamp, the source, and the extracted text passage. There were 45 to 14,000 passages per event-day pair in the dataset for this year's track.

Runs are to be submitted by participating teams as a set of ranked lists, one for each event-day pair, using start and stop times for each event-day pair that are specified as Unix timestamps. The top k passages of this list are evaluated by the track organizers, with k varying by event-day pair in a manner not known in advance by the participating teams. However, the track organizers recommend submitting at least the top 100 passages per event-day pair for evaluation. Evaluation for the track involves two types of metrics: summarization metrics and fact-matching metrics. The summarization metrics include Rouge-2 F_1 and BERTScore F_1; they compare the top k passages of a run against gold-standard summaries for each event-day pair. The fact-matching metrics compare the top k passages to a set of gold-standard facts previously developed by the assessors.

As this is the first year of the track, there was no labeled data available for training or parameter tuning.

3 System Design

Our overall system is shown in Figure 1. The first step is to create one ranked list for each facet. Next, these per-facet ranked lists are fused to create a single ranked list. This fusion process is based on rank, score, and recency. Finally, duplicated content is suppressed in a result selection step using either extractive or abstractive techniques.

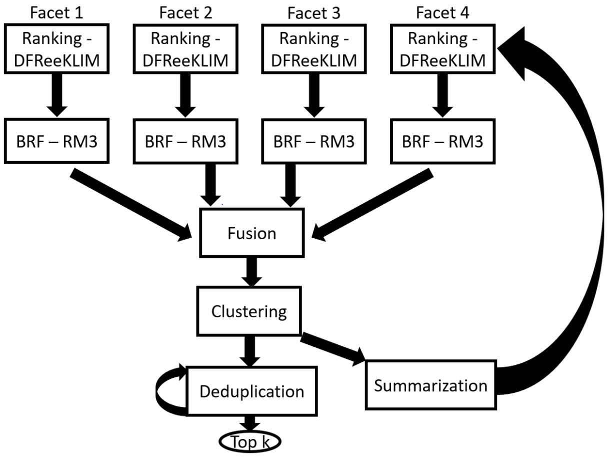

Figure 1: The primary system used in this paper.

3.1 Facet Ranking

Our system employs DFReeKLIM — a parameter-free divergence from randomness model from the TREC-2011 Microblogging Track [3] — as the initial step in generating per-facet ranked lists. This model uses a facet's indicative terms to generate a ranked list, and this initial result is then passed into an RM3 blind relevance feedback (BRF) algorithm [1]. BRF seeks to identify additional terms related to a particular facet that may not be generalizable outside of this event. For example, if a user wants to know what evacuation shelters are available (the facet) and it turns out that high schools are frequently being used as shelters, BRF may aid in automatically adding the term high schools to the query for this facet for this event.

3.2 Fusion

The next step is to combine the per-facet ranked lists for an event-day pair into a single ranked list for that event-day pair. Reciprocal rank fusion (RRF) is a well-established technique, although it is more traditionally used to combine ranked lists from different systems rather than from the same system with different queries [5]. Given the set of passages D provided by the track for an event-day pair, the set of facets Q specified for that event, and a ranking r_q for each $q \in Q$, eq. (1) shows the traditional reciprocal rank fusion algorithm. The constant 60 in the denominator is a default suggested by Cormack et al [5].

$$RRF(d \in D) = \sum_{q \in Q} \frac{1}{60 + r_q(d)} \tag{1}$$

While RRF provides an initial step toward our fusion algorithm, it does not consider two potential problems:

1. To the extent that retrieval scores are informative, we may be able to improve over the use of reciprocal rank alone. This information might, for example, be leveraged to identify scenarios where the first few ranks received much higher scores than the next lower-ranked passages.

2. Neither RRF nor DFreeKLIM are sensitive to recency. Passages marked as slightly less relevant but much more recent may be more useful than older passages that DFReeKLIM may have given a higher rank to.

The first issue was addressed by incorporating each passage's normalized score weight per query into the numerator of the RRF calculation, shown in eq. (2). This step, which Mansouri et al called "modified RRF" [7], is equivalent to performing a normalized CombSUM [6] on the relevance scores by facet. In the equation below, the w_q is the min-max normalized per-facet relevance score.

$$CS\_RRF(d \in D) = \sum_{q \in Q} \frac{w_q}{k + r_q(d)} \tag{2}$$

This approach still neglects any consideration of recency, however. To that end, a normalized recency score, t, is calculated as shown in eq. (3) with the timestamp of the passage as o, the start time for the event-day pair as s, and the event end time as n. This t is interpolated with the normalized score in the overall fusion calculation as shown in eq. (4). Tuning of the interpolation factor λ was performed using the formative evaluation process described in Section 4; a value of 0.9 was chosen. This λ puts heavy emphasis on the relevance of documents, while ensuring similar but outdated passages are pushed down the list — an important consideration for our result selection process that is discussed in the next section.

$$t = \frac{o_q - s}{n - s} \tag{3}$$

$$TW\_CS\_RRF(d \in D) = \sum_{q \in Q} \frac{\lambda w_q + (1 - \lambda)t}{k + r_q(d)} \tag{4}$$

3.3 Result Selection

We tried two ways of selecting which results to retain in our final ranked list. In our first approach, which we call extractive deduplication, we simply remove near duplicates from the result set. Our second approach, which we call abstractive deduplication, first generates a summary for each near-duplicate cluster and then reruns the entire passage ranking process.

3.3.1 Extractive Deduplication

The fusion process generates a single ranked list, as required by the track, but nothing has yet been done to address the potential for duplicate content for one facet displacing content that might be relevant to other facets. These problems occur because nearly identical passages published in different streams, and in some cases multiple times within the same stream (e.g., several accounts tweeting the same information), could all receive similar scores from the fusion techniques described in the previous section. To resolve this issue, we use the Universal Sentence Encoder [4] to embed the text of each passage in the top 100 passages of the fused list and find the cosine similarity with each other passage in the top 100.[2] Passages with a similarity above

[2] We selected 100 as a cutoff because the track guidelines suggested submitting a ranked list containing at least 100 passages per event-day pair. However, in actual practice, the k selected by the organizers for evaluation was sometimes larger than 100.

a threshold of 0.7 (selected through tuning using the formative evaluation technique discussed in Section 4) were clustered using single link clustering.

Once clusters were identified, all but the highest-ranked passage in each cluster were removed. However, if an element to be removed happened to have been the top-ranked passage in one of the facet-specific lists, it was retained in the fused list. Retaining a facet's top-ranked passage (if it made it into the top 100 passages in the fused list) was intended to minimize the risk of coverage loss from extractive deduplication. More refined approaches that consider facet relevance scores might be crafted, but we did not expect our formative evaluation technique (see Section 4) to be sufficiently discriminating to tune the parameters of such an approach.

Once clustered items are removed, the list is consolidated and passages below each of the removed elements move up to fill the gaps. Since passages outside of the original top 100 may move into the top 100, we repeat the process in order to identify any new clusters and remove duplicates. When the top 100 passages remain constant after a clustering and deduplication iteration or fewer than 100 total passages remain, the list is considered to be finalized and the final ranked list is output for submission.

3.3.2 Abstractive Deduplication

While the extractive deduplication process removes highly related passages, it will sometimes remove passages containing some unique information. Instead of simply removing the passage, an alternative approach used in one variant of our system is to summarize the information within a cluster. Summarization was performed by aggregating the text from each passage in a cluster and then constructing a summary of that text with the Hugging Face `sshleifer/distilbart-cnn-12-6` summarization model [8]. We set the desired output length from that model to be 15 words longer than the mean length of the passages being summarized. The summarized text was then temporarily added to the collection and the entire system—starting from facet ranking—was rerun. On this subsequent run, the summarization process was not repeated. Instead, extractive deduplication was performed, with the additional constraint that a summarized passage, if in the top 100, would not be removed in this process.

One additional change when using the summarization system is the removal of the recency-weighting component of the fusion calculation (by setting λ to 0 in eq. (4)). This was necessary when the full system was run for the second time since it was not clear how the timestamp of a summarized element should best be set, so our summaries lacked timestamps. To facilitate comparison with our other runs, we therefore also deleted recency weighting when the system was run for the first time; our results with summarization should therefore be compared to our run in which recency-weighting was ablated.

4 Formative Evaluation

The lack of labeled data (e.g., gold-standard facts, or gold-standard summaries) in this first year made parameter tuning and formative evaluation challenging. After checking with the organizers, we therefore chose to manually annotate a small portion of the collection for this purpose. To do this, the first author of this paper created a set of regular expressions to score 11 facets for each day in event 1 (the Lilac Wildfire). Our intent was to loosely model the Comprehensiveness fact-matching evaluation metric. The 11 facets were selected because they had easily identifiable and exhaustively enumerable answers. For example, the facet regarding the containment level of the fire is well defined for each day, but finding definitive answers for a facet on obstacles facing firefighters, an open set, would have been more difficult.

What we call the Formative Evaluation score for each day was calculated using eq. (5), with i representing a fact in the set of ground truth facts in the set I of regular expressions, S being the top 25 passages from the final ranked list, and w being the weight associated with a fact. This weight w is determined by the number of distinct elements that would be required to fully address the facet, and it sums to one for each facet of an event-day pair. For example, consider a situation where there is a facet requiring the location of all available shelters for evacuees and three shelters are open during the time window for the event-day pair undergoing evaluation. Each unique correct shelter location included in the top 25 passages would thus add $\frac{1}{3}$ to the Formative Evaluation score. Whether these locations were spread across one, two, or three passages, the full point would be awarded if all three locations appear in the top 25 passages. The final Formative Evaluation score is the mean score over the 7 days of the Lilac Wildfire.[3]

$$score(s) = \sum_{i \in I} \begin{cases} w & i \in S \\ 0 & i \notin S \end{cases} \tag{5}$$

One issue with this approach is that selection of these specific easily-answerable facets risks introducing some bias. Facets without such clear answers could have notably different properties that could impact the effectiveness of some techniques or parameter values in ways that would not be well characterized by our Formative Evaluation score.

Another unaddressed concern is the presence of inaccurate information in the top results. If a fire has spread to 100 acres by the end of the day for an event-day pair, the system's top result in the fused list could list the fire as being only 50 acres, but as long as some other lower-ranked passage still in the top 25 states that it is at 100 acres the Formative Evaluation score would award full credit for correctly addressing the facet. A real user would, however, see the incorrect answer before the correct answer, which is clearly undesirable.

Finally, the design of our Formative Evaluation metric was motivated by one specific fact-matching metric (Comprehensiveness). Tailoring formative evaluation for different metrics might have resulted in different choices of system designs, or perhaps different parameter choices.

5 Submissions

We submitted a total of six runs, as follows:

- *mrr_main*: Our primary submission implemented the system as shown in Figure 1, but without summarization. The next four runs were based off of this system, with various ablations.
- *combsum*: This version of the system removed the reciprocal rank portion (the denominator) of eq. (4) from the main system and relied on the sum of the interpolated score and recency weights.
- *mrr_no_dd*: This run removed the result selection step from the main system.
- *mrr_all*: A misnamed run, it actually set the λ in eq. (4) to one in order to ignore recency. Comparison with the *mrr_main* system thus illustrates the effectiveness of recency weighting.
- *mrr_nobrf*: This run omitted RM3 blind relevance feedback from the per-facet ranking process.

[3] Only five days of Event CrisisFACTS-001 (the Lilac Wildfire) were annotated for the official evaluation, but our Formative Evaluation score is averaged over seven days of that event.

				Component Ablation				
			BSTRD	STRD	BSTD	BTRD	BSRD	BSTR
Event	# Days	Median	*mrr_main*	*mrr_nobrf*	*combsum*	*rr_now*	*mrr_all*	*mrr_no_dd*
001	5	0.1042	0.0801	0.0986	0.1157	**0.1131**	0.0892	0.0861
002	3	0.2709	0.3404	**0.3565**	0.2981	0.2450	0.3532	0.3275
003	4	0.2957	0.3460	0.3391	**0.3557**	0.3297	0.3365	0.3346
004	6	0.0721	0.0489	0.0515	0.0513	0.0251	0.0550	**0.0760**
005	2	0.1898	0.2398	0.2488	0.1309	0.1655	**0.2809**	0.2277
006	2	0.0273	0.0182	**0.1455**	0.0273	0.0273	0.0182	0.0182
007	2	0.0412	0.0236	0.0412	0.0530	**0.0765**	0.0236	0.0236
008	3	0.2055	0.2308	0.2087	**0.2820**	0.2078	0.2035	0.2126
Microaverage		0.1512	0.1613	**0.1750**	0.1656	0.1656	0.1643	0.1624
Macroaverage		0.1508	0.1660	**0.1862**	0.1642	0.1487	0.1700	0.1633
Macroaverage excluding 001		0.1575	0.1782	**0.1987**	0.1712	0.1538	0.1816	0.1743

Table 2: Comprehensiveness at k, averaged over days per event, ablations ordered by microaverge.

- *rr_now*: This run removed consideration of the normalized score weight by setting λ to 0 in eq. (4).
- *mrr_sum*: This run included the summarization step. It was built as an addition to *mrr_all*, and thus did not use recency-weighting in its fusion step (because our summaries have no timestamp).

The first six of those runs satisfy the track's definition of an extractive run. Run *mrr_sum* is an abstractive run as defined in the track guidelines. All runs are automatic runs as defined in the track guidelines, although we note that our use of a manually constructed tuning set for eleven facets from event CrisisFACTS-001 as a basis for formative evaluation makes our results for that event less comparable to automatic runs from teams that did not make use of our tuning set (which we did share with other track participants).

6 Results

The fact-based assessment scores from the submitted runs are shown in Table 2, and the F_1 values for the BERTScore calculated using ICS-209 forms are shown in Table 3. The first row provides the official run names, while the second row contains a code indicating the features included in that run, as specified below. The final row of Table 2 shows the macroaverage excluding the first event, as some facets from that event were used for formative evaluation.

- B - includes RM3 blind relevance feedback in the facet ranking step.
- S - includes per-facet relevance score weighting in the fusion step.
- T - includes recency weighting in the fusion step.
- R - includes reciprocal rank fusion in the fusion step.
- D - includes extractive deduplication in the result selection step.
- A - includes abstractive deduplication in the result selection step.

While there are some interesting trends in the results for both tables, the differences between our six runs, and between any one run and the median across all submissions to the track, were not found to be

				Component Ablation					
			BSTRD	BSTR	BSRD	STRD	BSTD	BTRD	BSRDA
Event	# Days	Median	*mrr_main*	*mrr_no_dd*	*mrr_all*	*mrr_nobrf*	*combsum*	*rr_now*	*mrr_sum*
001	5	0.5121	0.4973	**0.5163**	0.5056	0.5113	0.4945	0.4927	0.4956
002	3	0.5300	0.5224	**0.5735**	0.5225	0.5158	0.5300	0.5209	0.5018
003	4	0.5054	0.5054	0.4977	0.5146	0.5094	0.4999	0.4878	**0.5186**
004	6	0.4774	0.4592	0.4592	0.4604	0.4696	0.4573	0.4581	**0.4799**
005	2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
006	2	0.4939	0.4940	0.4989	**0.4954**	0.4934	0.4901	0.4935	0.4870
007	2	0.5059	0.5103	0.5103	**0.5138**	0.5115	0.5101	0.5044	0.5098
008	3	0.4944	0.4958	0.4958	0.4945	0.4923	0.4827	0.4846	**0.5053**
Macroaverage		0.4399	0.4356	**0.4439**	0.4383	0.4379	0.4331	0.4303	0.4372

Table 3: ICS BERTScore F_1 values, averaged over days per event, ablations ordered by macroaverage.

statistically significant at $p < 0.05$ by two-tailed paired t-tests for either of the metrics in Tables 1 or 2, even before applying a Bonferroni correction for repeated tests.

While not significantly different, the relatively good Comprehensiveness score for *mrr_nobrf* warrants further investigation. Based on our formative evaluation, this system had been second to last of our submitted runs. The results from *mrr_nobrf* also serve to highlight one aspect of the formative evaluation process that we mentioned in Section 4. The lack of BRF in this run pushed it into nearly last place in the formative evaluation, yet it was first when evaluated by the Comprehensiveness metric, and consistently in the top three when evaluated by BERTScore. This disparity may be the result of the bias of our formative evaluation toward easily answerable questions.

Per-event macroaverage rankings for a variety of metrics are shown in Table 4. As can be seen, Formative Evaluation was not strongly predictive of Comprehensiveness system preference order, although without statistically significant differences it is hard to read too much into that observation. Also of note, deleting extractive deduplication (*mrr_no_dd*) seems to have affected BERTScore evaluation based on the ICS-209 forms somewhat differently from BERTScore evaluation based on NIST-written summaries or based on Wikipedia pages.

		Component Ablation					
	BSTRD	BSRD	BTRD	BSTD	STRD	BSTR	BSRDA
Metric	*mrr_main*	*mrr_all*	*rr_now*	*combsum*	*mrr_nobrf*	*mrr_no_dd*	*mrr_sum*
Formative Evaluation	1	2	2	4	6	7	4
Comprehensiveness	3	2	7	4	1	5	6
Redundancy/Verbosity	5	4	1	3	7	2	6
ics.bertscore.f1	5	2	7	6	3	1	4
ics.rouge2.f1	4	3	2	6	7	5	1
nist.bertscore.f1	6	5	2	4	3	7	1
nist.rouge2.f1	6	4	3	1	5	7	2
wiki.bertscore.f1	6	3	4	5	1	7	2
wiki.rouge2.f1	6	5	3	2	1	7	4

Table 4: Run rankings for different metrics, ablations ordered by formative evaluation score.

7 Conclusion and Future Work

The result of the runs we submitted to the TREC-2022 CrisisFACTS track have shown the task to be tractable, and we have seen some promising results from relatively simple fusion and deduplication techniques. We have not, however, been able to measure statistically significant differences between the techniques that we have tried, and we are looking forward to learning whether such differences can be observed between systems from different teams using this first year's test collection.

There are a number of potential directions for future work. The track casts the task as a ranking problem, and in the second year of the track it will be a ranking problem for which some training data will be available. This suggests that investigation of learning-to-rank techniques may be productive. Alaofi et al have challenged us to consider where queries come from [2], suggesting that we might also productively use manual runs to explore the effect of different queries for the same set of facets. Moreover, we might also hope to find some hierarchical structure in the facets that are of interest for an event, and even simple fusion techniques of the type we have explored this year might be adapted to exploit such structures. These and other opportunities cause us to see CrisisFACTS as a rich venue for future research, and we are glad to hear that the track will continue for a second year.

References

[1] Abdul-Jaleel, N., Allan, J., Croft, W. B., Diaz, F., Larkey, L., Li, X., Smucker, M. D., and Wade, C. UMass at TREC 2004: Novelty and HARD. In *Proceedings of TREC 2004* (Gaithersburg, MD, USA, 2004).

[2] Alaofi, M., Gallagher, L., McKay, D., Saling, L. L., Sanderson, M., Scholer, F., Spina, D., and White, R. W. Where do queries come from? In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain* (2022), pp. 2850–2862.

[3] Amati, G., Amodeo, G., Bianchi, M., Celi, A., Di Nicola, C., Flammini, M., Gaibisso, C., Gambosi, G., and Marcone, G. FUB, IASI-CNR, UNIVAQ at microblogging track of TREC 2011. In *Proceedings of TREC 2011* (Gaithersburg, MD, USA, 2011).

[4] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y., Strope, B., and Kurzweil, R. Universal sentence encoder. *CoRR abs/1803.11175* (2018).

[5] Cormack, G. V., Clarke, C. L. A., and Büttcher, S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *SIGIR '09* (Boston, MA, USA, 2009).

[6] Fox, E. A., and Shaw, J. A. Combination of multiple searches. In *The Second Text REtrieval Conference (TREC-2)* (Gaithersburg, MD, USA, 1994).

[7] Mansouri, B., Oard, D. W., and Zanibbi, R. DPRL systems in the CLEF 2022 ARQMath lab: Introducing MathAMR for math-aware search.

[8] Shleifer, S. distilbart-cnn-12-6. `https://huggingface.co/sshleifer/distilbart-cnn-12-6`. Accessed: 2022-11-05.