Overview of the NTCIR-16 Lifelog-4 Task

Liting Zhou
Cathal Gurrin
Graham Healy
Dublin City University
Ireland

Hideo Joho
Tsukuba University
Ireland

Binh Nguyen
University of Science, VNU-HCM
Vietnam

Rami Albatal
Perfogram Ltd
Ireland

Frank Hopfgartner
University of Sheffield
UK

Duc-Tien Dang-Nguyen
University of Bergen & Kristiania University College
Norway

ABSTRACT

NTCIR-16 saw the fourth edition of the Lifelog task, which aimed to foster comparative benchmarking of approaches to automatic and interactive information retrieval from multimodal lifelog archives. In this paper, we describe the test collection employed, along with the tasks, the submissions and the findings from this NTCIR16 Lifelog-4 LEST sub-task. We finish by suggesting future plans for lifelog tasks.

KEYWORDS

lifelog, information retrieval, quantified self, personal data

SUBTASKS

LSAT subtask (English)

1 INTRODUCTION

NTCIR-16 [10] hosted the fourth edition of the Lifelog task. The aim of the lifelog task is to foster comparative benchmarking of approaches to automatic and interactive information retrieval from multimodal lifelog archives. In this edition of the Lifelog task, we focused on a single subtask, the Lifelog Semantic Access (sub)Task (LEST), which is a conventional ad-hoc retrieval task for lifelogs. This task had been central to the previous NTCIR Lifelog tasks at NTCIR-12 [4], NTCIR-13 [5] and NTCIR-14 [6].

Before we begin our review of the submissions for the lifelog task, we remind readers of the the concept of lifelogging by presenting the definition proposed by Dodge and Kitchin [3], who refer to lifelogging as 'a form of pervasive computing, consisting of a unified digital record of the totality of an individual's experiences, captured multimodally through digital sensors and stored permanently as a personal multimedia archive'. The lifelog task at NTCIR was proposed because the organisers identified that technological progress had resulted in the potential for lifelogging to become a commonplace activity, thereby necessitating the development of new forms of personal data analytics and retrieval tools that are designed to accommodate the particular idiosyncrasies of lifelog data. As per the definition above, the organisers consider lifelogs to be longitudinal multimodal archives of continuous capture personal data which have been integrated into one large user model, the applications of which are many and varied. A key assumption underlying this lifelog task is that retrieval from lifelogs is a fundamental task and a building block for a wide range of impactful applications.

The contribution of this paper is in terms of the task introduction and overview comparison of the performance of submitting teams. The rest of the paper describes the dataset employed, the LSAT task, the topics, the comparison between participants and finally we outline our thoughts on how lifelog benchmarking will proceed in the coming years.

2 DATASET DESCRIPTION

2.1 Overview

NTCIR-16-Lifelog-4 reuses an existing dataset, the LSC'21 dataset [8], which is a multimodal dataset that is four months in size, from one active lifelogger. The LSC (Lifelog Search Challenge) is a parallel grand-challenge activity that takes place annually at the ACM ICMR conference and attracted sixteen participants in its most recent edition [7] at ACM ICMR 2021. This dataset was chosen due to its ready availability and the fact that the dataset was already processed and anomymised for the LSC challenge in 2020 and 2021.

The dataset consists of three files, which were made available to participants who signed up for the lifelog task and who agreed to the terms of access. The data continued within these files was gathered using multiple wearable sensors, such as PoV cameras, biometric smartwatches and location and activity loggers on a smartphone. The data gathering phases occurred in 2015 (four weeks), 2016 (eight weeks) and again in 2018 (four weeks), giving a total of 4 months of lifelog data, which was gathered 24 x 7 and organised into minutes in an XML form. For dataset examples, see [7]. The UTC timestamp was the used as the alignment factor for these data sources.

The three files that comprise the dataset were:

- Metadata for the collection (2.8MB), consisting of textual XML metadata representing time, physical activities, biometrics and locations[1] of one individual for four months during 2015-2018.
- Core Image Dataset (38GB) of 183,432 wearable camera images, fully redacted and anonymised in 1024 x 768 resolution, captured using OMG Autographer and Narrative Clip devices. These images were captured during regular waking

[1]The dataset did not include biometric data for the four weeks in 2015, due to the lack of available devices at the time

hours by the same one individual. All faces and readable text have been removed, as well as certain scenes and activities manually filtered out (by the data gatherer / lifelogger) to respect privacy expectations.

- Visual Concepts (79.9MB) extracted from the non-redacted version of the visual dataset. The Visual Concepts data file includes detected scenes and concepts for each image (processed over the non-redacted version of the images). the objects detected automatically from the image. We use the object category list of 2014-2017 COCO datasets [11] with 80 labels for annotation.

This data was made available to all participants who signed up to participate in the task (eight in total) and who completed the provided data agreement forms.

3 EVALUATION TASK DETAILS

As stated, the NTCIR-16 installation of the Lifelog task focused on one sub-task, the LEST sub-task. The LEST sub-task was the most popular sub-task all previous times that the Lifelog task was run. Other tasks that examined approaches to event segmentation, multimodal annotation and insight generation were not facilitated at NTCIR-16.

3.1 Lifelog Semantic Access sub-Task (LEST)

The LEST sub-task required participants to process 48 topics using a lifelog retrieval system and return ranked results as submissions for evaluation. Participants were free to take part in an interactive or automatic manner. Automatic runs would assume that there was no involvement of a user in the search process, except perhaps in the query construction phase (non-interactive). Interactive runs assumed that a user was engaged actively in the search process, from query formation and refinement, to selection of potentially relevant images for submission.

Automatic runs assume that there was no user involvement in the search process once the query was formed. Submissions could include up to 100 images for each topic in rank order. There was no time limit on how long it can take for an automatic run. The aim of these runs were to facilitate the comparison of different back-end ranking algorithms. We had four different systems submit automatic runs.

Interactive runs assume that there is a user involved in the search process that generates a query and selects which images are considered correct for each topic. This may be a single phase, or may contain multiple phases of relevance feedback or query reformulation. In interactive runs, the maximum time allowed for any topic was set to 300 seconds. Participants were asked to include a seconds-elapsed indicator to facilitate comparison of systems at time-cutoffs between 0 and 300 seconds.

3.2 Topics

The topics (queries) for the LEST sub-task followed the typical TREC format of ID, Query, Description and Narrative, as shown in Listing 1 and Listing 2. To ensure comparability with previous NTCIR-Lifelog topics, we included the topic type, which can be ad-hoc or known-item and the user id (uid) which refers to the user who generated the query (in this case, always u1) since there was only one user in this particular dataset. The description provides useful information to the user about the subject of the topic, while the narrative helps to disambiguate what is relevant and not (for interactive runs).

ID	Title	Num-Rel
16001	Baggage Carousel	49
16002	I'm in the mirror	9
16003	Keys to my door	18
16004	My panda	52
16005	Shopping for antiques	187
16006	Eating sushi	175
16007	Cereal with milk for breakfast	1
16008	Working on the flight	224
16009	Meeting people in a room with red carpet	59
16010	Sharing breakfast in a hotel	24
16011	What's in the refrigerator?	156
16012	Recording for TV	140
16013	Lunch-time at the desk	108
16014	Money at the ATM	36
16015	Saturday morning coffee with a friend	483
16016	Eating before flying	133
16017	Drinking at home	171
16018	Buying alcohol	35
16019	Exercising in the park	99
16020	Dogs	61
16021	Eating icecream	15
16022	On public transport in Ireland	324
16023	Cooking in the kitchen	188
16024	Taking medication on the weekends	47

Table 1: Ad-Hoc Topics

In all, there were 48 topics prepared for the LEST sub-task, 24 of each type. The ad-hoc topics were similar to conventional text retrieval topics, that aimed to find as many relevant items as possible for a given query (average of 116 per topic). These relevant items may be contained in one or many different events. The list of ad-hoc Topics are shown in Table 1. The topic id starts at 16001 and the Num-Rel column refers to the number of relevant items found for each topic in the pooled relevance judgements.

The 24 known-item topics focused on solving targeted information needs that were typically solved by a single event with a small number of relevant images (average 8.25 per topic). Known-item topics are designed to simulate the human memory process of finding or remembering a specific event or activity (e.g. solving a task or attending a certain location). Table 2 shows the 24 known-item topics along with the number of relevant items in the relevance judgements. It is worth noting that the known-item topics are based on existing topics from the LSC'21 [7] benchmarking workshop and as such, would be familiar to any participant who took part in the LSC'21 exercise.

3.3 Relevance Judgements

The pooling technique was used to generate the relevance judgements. The union of all images for a topic from each submitted

Listing 1: Ad-Hoc Topic 16001

```
<topic>
<id>16011</id>
<type>adhoc</type>
<uid>u1</uid>
<title>What's in the refrigerator?
</title>
<description>Find examples of when I
    was looking inside the refrigerator
     at home
</description>
<narrative>Any moment that shows the
    lifelogger looking inside a
    refrigerator is considered relevant
     once the refrigerator is in the
    lifelogger's home. Looking inside
    refrigerators in hotels or stores
    is not considered relevant
</narrative>
</topic>
```

Figure 1: Example Relevant Image for Adhoc Topic 16011 (What's in the refrigerator?)

Listing 2: Known-item Topic 16033

```
<topic>
<id>16033</id>
<type>known-item</type>
<uid>u1</uid>
<title>Asking for Directions</title>
<description>Find the moment when I was
     lost and looking for directions on
     a street</description>
<narrative>The lifelogger was lost and
    looking for directions on a street,
     close to an Asian restaurant
    called Maple Leaf. The lifelogger
    must be clearly seen iterating with
     one or more people on a street
    with a restaurant in the background
    .</narrative>
</topic>
```

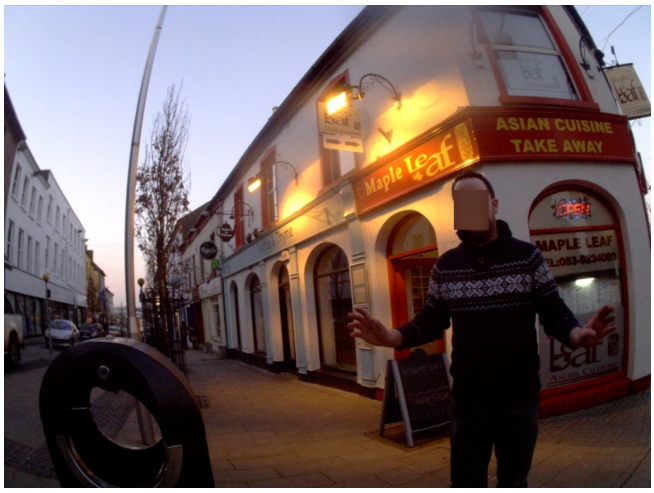

Figure 2: Example Relevant Image for Known-Item Topic 16033 (Asking for Directions)

official run was manually judged by the original lifelogger (u1) and the images that were judged relevant formed the basis of the relevance judgements. For the known-item topics, some additional relevant items were identified from the LSC'21 workshop, where appropriate, and added to the relevance judgements. For these official NTCIR16 results, only the official submissions were considered when generating the relevance judgments. Unofficial runs will we amended after the NTCIR16 workshop and revised qrels made available from the NTCIR-Lifelog website.

4 EVALUATION RESULTS

Given that there were two types of submissions to the LSAT subtask, we will examine each of them separately. Five participants

ID	Title	Num Rel
16025	Building a computer	36
16026	Northside shopping centre	1
16027	T-shirt sale	1
16028	Putting a display/monitor in my car	2
16029	Dissertation on a whiteboard	30
16030	BBQing marshmallows	11
16031	No junk mail	1
16032	Blue cups	4
16033	Asking for directions	2
16034	Coffee while waiting	25
16035	Telescope in the mirror	1
16036	Buying a blood pressure monitor	1
16037	Orange kids suitcase	2
16038	TagHeuer watch	1
16039	Boarding pass for PVG	1
16040	Colleague with heavy envelope	6
16041	Show me the hotel name	1
16042	Computer chip laboratory	16
16043	Buying fruit	1
16044	Scrambled eggs	13
16045	Reading the newspaper	33
16046	Technology photo - The lifeloggers toolkit	3
16047	Birds in a cage	2
16048	Glenisk yoghurt	4

Table 2: Known-Item Topics

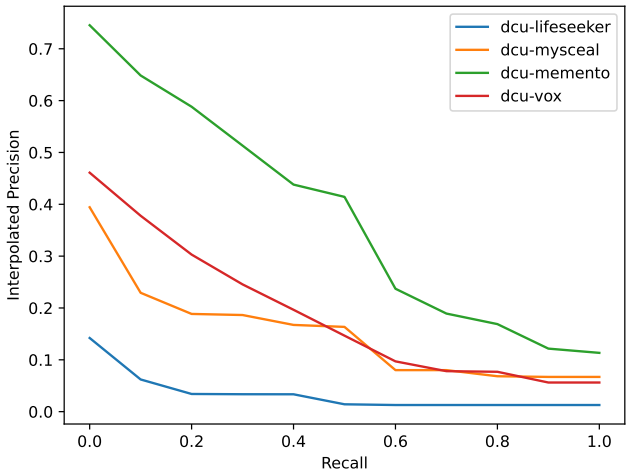

Figure 3: Comparing the best runs of the four automatic teams

submitted results to the LSAT task and we will discuss the automatic and interactive tasks separately. The evaluation results of participating teams are analysed and presented in the following subsections.

4.1 LEST-Automatic Runs

Three distinct participants submitted automatic runs and four different systems were employed:

HCMUS-DCU LifeSeeker. This team used two distinct lifelog retrieval systems and submitted six automatic submissions [12]. The systems implemented the Bottom-Up-Attention model [2] and used a mask-RCNN pre-trained model for concept extraction and integration of the metadata. A weighted Bag-of-words approach was used for generating ranked lists. From six official runs, the best approach (DCULifeSeeker08) produced a best mAP score of 0.0299 which was a run that found 320 relevant items. For this best run, P@5 was 0.0833.

HCMUS-DCU Myscéal. Similarly to LifeSeeker, Myscéal employs a visual concept-based retrieval process that utilises the visual concepts, and non-visual metadata such as GPS coordinates, semantic locations, time, and date, which are indexed using ElasticSearch. The Myscéal system was the top performing system at both the LSC'20 and LSC'21 challenges. For their automatic run (DCUMYSCEAL01), an expert user produced one set of topics and presented them to the system in order to generate 48 ranked lists. This approach got an mAP score of 0.1366 and found 366 of the relevant items. The P@5 score for this run was 0.1917, which was notably better than the LifeSeeker system which also implemented a conventional concept-based retrieval process.

DCUVOX and DCUMemento. Moving beyond the conventional concept-based retrieval process, the two interactive systems (DCUMemento and DCUVOX) took part in an automatic manner. Both systems leveraged image-text embeddings from various CLIP models to develop their respective search and ranking functionality [1]. Both systems employ different variations of the underlying CLIP models, with DCUVOX using the ViT-B/32 model and DCUMemento using a weighted ensemble of scores from larger ViT-L/14 and ResNet-50x64 models. Additionally, the DCUMemento system incorporated manually restructured queries to extract the visual concepts from metadata such as date, time, etc. to facilitate an automatic stage-wise search process, with the temporal information acting as a filter over the results. DCUVox (which was designed to support voice interaction) implemented a different query reformulation strategy that produced shorter summarised versions of the official topic descriptions.

The official score of DCUVOX group (DCUVOXLSAT01) for this task is 0.1748 and for the Memento group (DCULSAT01), it is 0.0.3605. DCUVOX found 815 and DCUMemento found 1201 relevant items. As shown in Figure 3, the memento received the highest score of all automatic runs, by a significant margin. Both of these systems (based on CLIP models) performed better than the other two automatic systems, which clearly highlights the benefits of the CLIP-type of embedding model for multimodal lifelog retrieval tasks.

4.2 LEST-Interactive Runs

Two groups submitted interactive runs to the LEST sub-task, based on the outputs of two interactive retrieval systems.

THUIR THUIR built an interactive lifelog search engine that included a multi-functional and flexible human feedback mechanism for result retrieval [9]. The feedback mechanism is used to combine

the ternary feedback and negative keywords in keywords filter fields. Additionally, result presentation is designed for interaction that can show relevant images in T-shape fixation distribution and timeline viewing is added to provide temporal information. THUIR highlight both unofficial and official runs in their paper. The unofficial results (using existing queries from a previous LSC workshop instance) highlight a best MAP score of 0.6410 for known-item tasks and 0.7160 for ad-hoc tasks. These scores are the highest of any runs at NTCIR-16-Lifelog-4, official or unofficial, but are not shown in Figure 4, due to the use of a different topic-set. The official runs, the top run (THUIRLLLSAT02) produced a MAP figure of 0.1604 over 45 topics in which items were found within timeframe for submission. Considering the number of relevant items found in the official figures, the best run found 741 items with a P@5 score of 0.3200.

HCMUS-DCU Myscéal team also submitted interactive runs, but they were unofficial runs, so they did not contribute to the relevance judgements used to calculate the scores. Myscéal was designed to facilitate novice users who are not familiar with lifelog retrieval and the indexed dataset, and supports multi-query search based on temporal relationships, which they propose to be a key feature of lifelog data. Besides, in order to address the shortcomings of using a fixed list of "keyword" concepts which are obtained from pre-trained object detectors, Myscéal applies a query expansion process to address lexical mismatches in queries. Additionally, the system supports visual similarity, mapping of results, and powerful temporal query handlers. The highest MAP score for the Myscéal team is 0.3980 over the 44 topics that they found relevant items for. This run was conducted by the system developer as an expert user, so the user will know both the data and the concepts.

Additionally, a novice run was performed over the ad-hoc topics only. In order to compare the performance of a novice and expert on these types of conventional search topics, we also ran an expert run over these ad-hoc topics and performed a fully-automatic run also. Naturally, the expert performed better with a MAP score of 0.2234, compared to 0.1302, with a MAP of 0.0674 for the fully automatic run (mentioned above). In terms of number of relevant items found, the expert could find 681, the novice could find 430 and the fully automatic system found 337 items. The expert in-the-loop located more than double the number of relevant items that were found by the fully-automatic system. Finally, considering P@5, the expert scored 0.5478, the novice 0.4909 and the automatic run 0.2916. The benefit of employing an expert user is clearly shown by these results, while the novice user performs better than the automatic system, which is to be expected.

5 DISCUSSION AND FUTURE PLANS

The NTCIR-Lifelog task attracted eight participants who sought and accessed the lifelog datasets. However, only three groups managed to submit runs by the deadline for submissions. These three groups represented five different systems / approaches to lifelog retrieval.

For automatic runs, it is becoming apparent from these NTCIR runs, but also from related activities in the Lifelog Search Challenge that CLIP-based systems preform significantly better that conventional concept-based approaches on the multimodal lifelog archives. Within the CLIP model submissions, the larger CLIP models seemed to outperform smaller models, but since there are differences in how the queries were formulated (level of human involvement), this needs more investigation.

For interactive runs, the impact of the level of expertise of the searcher will always be a key factor in the performance of a system. THUIR's unoffical run was the top performing interactive system, though it utilised older LSC queries, rather than the official NTCIR-16 topics. According to these results, the THUIR interactive lifelog system is shown to perform exceptionally well. The Myscéal interactive system, which was considered state-of-the-art at LSC, submitted (late) unofficial interactive runs which produced a good MAP score on the official topics. Due to the interactive nature of these systems, it is not possible to deeply analyse the comparative performance of both systems. For such a comparison, it is useful to examine the LSC workshop outputs which require all teams to operate under identical constraints.

The organisers intend to keep promoting lifelog search as a challenging and impactful research topic. For future challenges, we intend to focus our proposed NTCIR challenges onto a range of lifelog-related challenges that are relevant to an NTCIR audience. We will gauge the level of interest in running an automatic LSAT sub-task at NTCIR-17, which would use the new LSC'22 18 month lifelog dataset. If there is sufficient interest, then we can propose a Lifelog-5 task.

Additionally, we will explore the potential of focused-domain lifelog challenges, such as the pilot NTCIR-16 RCIR task, which explored the impact of biometric measures of reading comprehension on the text search process. We also hope to run a quantified-self pilot task in the coming years. Additionally, we will look into options of running diary search tasks or other personal data sources with more textual content. The core challenge of interactive access to multimodal lifelog challenges will continue be run through the vehicle of the LSC workshop series [7, 8].

6 CONCLUSION

In this paper, we described the data and the activities from the lifelog-4 LEST sub-task at NTCIR-16. Although it is difficult to draw many conclusions from these findings, we do note that there is still a lot of research that needs to be done to develop annotation and search tools for lifelog archives. In future years, we hope to continue this lifelog task (e.g NTCIR-17 Lifelog-5), and we will reduce both the size of the collections and the number of sub-tasks that are on offer and focus effort on the tasks that are most likely to attract interest from NTCIR participants.

REFERENCES

[1] Naushad Alam, Ahmed Alateeq, Yvette Graham, Mark Roantree, and Cathal Gurrin. 2022. DCU at the NTCIR16 Lifelog-4 Task. In *Proceedings of The 16th NTCIR Conference on Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6077–6086.

[3] Martin Dodge and Rob Kitchin. 2007. 'Outlines of a world coming into existence': pervasive computing and the ethics of forgetting. *Environment and planning B: planning and design* 34, 3 (2007), 431–445.

[4] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, and Rami Albatal. 2016. Ntcir lifelog: The first test collection for lifelog research. In *Proceedings*

of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. 705–708.

[5] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, Rashmi Gupta, Rami Albatal, and Duc Tien Dang Nguyen. 2017. Overview of NTCIR-13 Lifelog-2 Task. In *Proceedings of The 13th NTCIR Conference Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics.

[6] Cathal Gurrin, Hideo Joho, Frank Hopfgartner, Liting Zhou, V-T Ninh, T-K Le, Rami Albatal, D-T Dang-Nguyen, and Graham Healy. 2019. Overview of the NTCIR-14 Lifelog-3 task. In *Proceedings of The 14th NTCIR Conference Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics.

[7] Cathal Gurrin, Björn Þór Jónsson, Klaus Schöffmann, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Graham Healy. 2021. Introduction to the Fourth Annual Lifelog Search Challenge, LSC'21. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 690–691.

[8] Cathal Gurrin, Tu-Khiem Le, Van-Tu Ninh, Duc-Tien Dang-Nguyen, Björn Þór Jónsson, Jakub Lokoč, Wolfgang Hürst, Minh-Triet Tran, and Klaus Schöffmann. 2020. Introduction to the Third Annual Lifelog Search Challenge (LSC'20). In *Proceedings of the 2020 International Conference on Multimedia Retrieval*. Association for Computing Machinery, New York, NY, USA, 584–585. https://doi.org/10.1145/3372278.3388043

[9] Zhiyu He, Jiayu Li, Wenjing Wu, Min Zhang, Yiqun Liu, and Shaoping Ma. 2022. THUIR at the NTCIR-16 Lifelog-4 Task: Enhanced Interactive Lifelog Search Engine. In *Proceedings of The 16th NTCIR Conference on Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics.

[10] Makoto P. Kato, Hiroaki Ohshima, Ying-Hsang Liu, and Hsin-Liang Chen. 2022. Overview of the NTCIR-16 Data Search 2 Task. In *Proceedings of the NTCIR-16 Conference*.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[12] Thao-Nhu Nguyen, Tu-Khiem Le, Van-Tu Ninh Ninh, Ly-Duyen Tran, Manh-Duy Nguyen, Minh-Triet Tran, Thanh-Binh Nguyen, Annalina Caputo, Sinead Smyth, Graham Healy, and Cathal Gurrin. 2022. DCU and HCMUS at NTCIR-16 Lifelog-4 Task. In *Proceedings of The 16th NTCIR Conference on Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics.