Overview of the TREC 2015 Contextual Suggestion Track

Adriel Dean-Hall
University of Waterloo

Charles L. A. Clarke
University of Waterloo

Jaap Kamps
University of Amsterdam

Julia Kiseleva
Eindhoven University of Technology

Ellen Voorhees
NIST

1. INTRODUCTION

The TREC Contextual Suggestion Track evaluates point-of-interest (POI) recommendation systems, with the goal of creating open and reusable test collections for this purpose. The track imagines a traveler in a unknown city seeking sites to see and things to do that reflect his or her own personal interests, as inferred from their interests in their home city. Given a user's profile, consisting of a POI list and rating from a home city, participants make recommendations for attractions in a target city (i.e., a new context).

For example, imagine a group of information retrieval researchers with a November evening to spend in beautiful Gaithersburg, Maryland. A contextual suggestion system might recommend a beer at the Dogfish Head Alehouse[1], dinner at the Flaming Pit[2], or even a trip into Washington on the metro to see the National Mall [3].

This is the fourth year that the track has operated (since TREC 2012). If you are familiar with the track from previous years, here are the big changes this year:

- The track moved from the open web to a fixed set of documents.
- The track was split into two tasks:
  1. A *live* task, in which participants set up a server and were sent requests over a period of about three weeks.
  2. A *batch* task, which was similar to the task run in previous years.

The live task reflects the track's long term goal of creating a "living lab" service for POI recommendation.

[1] `www.dogfishalehouse.com`
[2] `www.flamingpitrestaurant.com`
[3] `www.nps.gov/nacc`

24th Text REtrieval Conference November 18-20, Gaithersburg, Maryland.

2. TASK OVERVIEW

For both tasks participants were asked to develop a system that is able to make suggestions for a particular person (based upon their profile) with respect to a particular geographic context (i.e., the target city). As input to the task, participating research groups were given a set of profiles, a set of example suggestions, and a set of contexts. Each profile corresponded to a single user, indicating that user's preference with respect to each example suggestion, while each context represented a target city that the user might visit. For each profile/context pairing, participating researchers were required to return a ranked list of 50 proposed suggestions. Each suggestion was expected to be be appropriate to the profile (based on the user's preferences) and the context (according to the target city).

Profiles correspond to the stated preferences of real individuals, recruited through crowdsourcing. These crowdsourced workers first judged example attractions in seed locations, representing their home cities, later returning to judge suggestions proposed by the participating research groups for various target cities. The live and batch tasks differ primarily through the way in which systems interacted with users, with live participants providing a online server to respond to new profile/contexts on demand.

Details of what the profiles and contexts contain are given on the track homepage[4]. For example, one suggestion might be to have a beer at the Dogfish Head Alehouse, and the profile might include a negative preference with respect to this suggestion. Each training suggestion includes a title, description, and an associated URL. Each context corresponds to a particular geographical location (a city). For example, the context might be Gaithersburg, Maryland. Participants returned results either by setting up a server and participating in the live experiment, or by submitting during the batch experiment.

3. COLLECTION

Early in 2015, a few volunteers crawled the web for documents that they considered to represent points-of-interest in selected target locations and submitted their set of documents to us. We merged the documents, and after some cleaning and de-duping, these POIs formed the based for the collection used in both tracks. Only documents from this collection could be returned as suggestions from either track.

[4] `sites.google.com/site/treccontext`

4. TASK DESCRIPTIONS

4.1 Live Task

For a few weeks, participants in the live task registered services with us. They were periodically sent requests for suggestions. These requests included the city the request was being made for and details on the person making the requests so that responses could be personalized. Responses from participants were made up of an ordered list of suggestions taken from the collection (and from the correct city).

The first request made to services contained no personal assessor information and only a location. This first set of suggestions was then shown to assessors. The ratings assessors gave for this first set of suggestions were then sent to services for further requests. Ratings from further requests were again sent to services during multiple rounds of suggestion requests and assessments. During this process suggestions from multiple services were combined into a single list then shown to assessors.

These assessors were recruited from Mechanical Turk (MT). For the first round of assessment, workers were recruited from the general MT pool. For additional rounds of assessment, workers were sent messages through MT asking them to return to complete additional assessment tasks.

4.2 Batch Task

For the batch task requests made during the live tasks were sampled. Any POIs that were rated for that requests were also included in the request with the ratings stripped out as a set of candidate suggestions. For participants in the batch task only these candidates were allowed to be made as suggestions (instead of all points-of-interest in the collection).

5. RESULTS

Preliminary results for the batch task are shown in Figure 1; preliminary results for the live task are shown in Figure 2. As in previous years, precision@5 provides the primary evaluation measure, and runs are sorted by the measure. The tables also show mean reciprocal rank (MRR).

6. FINAL REMARKS

This report provides an preliminary outline of track activities and results from TREC 2015. Six groups submitted a total of nine runs to the live track, demonstrating the feasibility of our online approach to evaluation. By using a fixed test collection, instead of allowing submissions from the open web, we hope to improve the reusability of the collection when compared to previous years. As we continue to analyze track results, we will examine our success with respect to this goal. The track continues for TREC 2016, where we hope to improve our methods for online evaluation, and continue to build a reusable collection.

runid	precision@5	MRR
11	0.5858	0.7404
uogTrCSFM	0.5706	0.7190
fr	0.5583	0.6815
SCIAI_runB	0.5564	0.6995
nr	0.5507	0.6921
uogTrCSLVPC	0.5498	0.6758
22	0.5450	0.6991
SCIAI_runA	0.5403	0.6983
IITBHU_2	0.5365	0.7030
IITBHU_1	0.5308	0.6760
PLM1	0.5204	0.6765
RUN1	0.5156	0.6594
BJUTb	0.5100	0.6688
PLM2	0.5024	0.6734
LavalIVA_1	0.4645	0.6102
LavalIVA_2	0.4616	0.6088
RUN2	0.4616	0.6535
TJU_CSIR_TOPIC	0.4303	0.6064
BJUTA	0.4284	0.5795
USST1	0.4047	0.5760
TJU_CSIR_VMS	0.3346	0.4755

Figure 1: Batch results

runid	precision@5	MRR
WaterlooRunA	0.4716	0.5929
WaterlooRunB	0.4695	0.5877
IRKM2	0.4079	0.5461
IRKM1	0.3953	0.5213
UDInfoCS2015	0.3411	0.4496
LavalIVA-run1	0.2611	0.3894
uogTrCsLtrUDepCat	0.2384	0.3128
uogTrCsLtrUInd	0.1605	0.2538
TJU_BASE	0.1342	0.1844

Figure 2: Live results