An Initial Test Collection for Ranked Retrieval of SMS Conversations

Rashmi Sankepally, Douglas W. Oard
iSchool and UMIACS
University of Maryland
College Park, MD USA 20742
rashmi@umd.edu, oard@umd.edu

Abstract

This paper describes a test collection for evaluating systems that search English SMS (Short Message Service) conversations. The collection is built from about 120,000 text messages. Topic development involved identifying typical types of information needs, then generating topics of each type for which relevant content might be found in the collection. Relevance judgments were then made for groups of messages that were most highly ranked by one or more of several ranked retrieval systems. The resulting TREC style test collection can be used to compare some alternative retrieval system designs.

Keywords: ranked retrieval, test collection, relevance assessment, retrieval effectiveness

1. Introduction

With the ubiquity of smartphone usage, many people have turned to using short message services (SMS) to send quick messages that don't require an immediate response. As with other types of "conversational text" (e.g., email, discussion forums, online reviews), personal SMS text message archives can be valuable information sources in their own right, both for the people who contributed to their creation and for others (e.g., historians or researchers) who may subsequently gain authorized access to such collections.

Test collections have been built for a number of information retrieval tasks, both in shared-task venues such as the Text Retrieval Conference (TREC) and for more focused development. We are, however, not aware of any prior work on characterizing the effectiveness of searching SMS content. As part of the DARPA Broad Operational Language Translation (BOLT) program, the Linguistic Data Consortium (LDC) released some SMS corpora with the principal goal of supporting research on machine translation for informally written content (Song and others, 2014). In this paper, we describe our development of an information retrieval test collection based on those LDC SMS corpora.

The paper is organized as follows: the next section reviews some related work, the data collection is described in section 3, section 4 describes topics used in the test collection, section 5 gives the relevance assessment procedure and we conclude in section 6.

2. Related Work

Although formal dissemination-oriented content such as news stories and scientific papers have been the focus of much information retrieval research, there has also been some work on development of test collections for more informal user-generated content in which there is potentially some interaction among the content creators. Among these are the CLEF Cross-Language Speech Retrieval Track focused on interviews (in which interviewer and interviewee co-construct the interview, passages from which are to be found) (Pecina and others, 2007), the TREC Microblog track (focused on Twitter content, some of which is directed between specific users) (Lin and others, 2014), the TREC Legal Track (focused on email content, some of which is easily threaded into conversations) (Grossman and others, 2011), and the the TREC Blog Track (some of which involves comments on other blog posts) (Macdonald et al., 2009). Although differing in some details, each proved amenable to a fairly conventional approach to test collection construction involving topic design, runs from a diverse set of systems, some way of sampling documents for relevance judgments, and some rank-based or set-based evaluation measures. We are, however, aware of only one shared task evaluation involving SMS messages: The FIRE SMS-Based FAQ Retrieval task. In that task, queries were posed using SMS, and a preexisting set of answers to Frequently Asked Questions (FAQ) provided the "document" set to be searched (Contractor and others, 2013). Our goal in this paper is to switch the focus from searching using SMS queries to searching SMS content itself. We studied the process adopted by LDC to create open-domain queries as part of the same DARPA BOLT program (Griffitt and Strassel, 2016). Those queries were developed for discussion forum posts and intended to be cross-lingual to some extent. We used a modified version of that process to create our queries to search the SMS content as described in section 4..

3. Selecting the Messages

The SMS collection that we have used was assembled by the LDC for the DARPA BOLT program and released in three phases (Song and others, 2014)[1]. As released, the LDC corpora contained both English SMS messages and English text chat logs that were contributed for research use (in exchange for compensation) by individuals. Contributors were offered the opportunity to redact content that they did not wish to have distributed, and LDC reviewers also examined each message for content for which distribution would not be appropriate. Redactions are marked with sequences of hash characters ("#"). The vast majority of the messages are from SMS, so we used only the SMS mes-

[1] LDC catalog: LDC2013E49, LDC2013E63, LDC2013E84

sages as the basis for our test collection. Because the machine translation systems for which the corpora were originally designed are sometimes designed to process entire documents, the LDC grouped the SMS messages between each pair of participants into time-ordered sets that can be thought of as "conversations" (although of course in practice some such sets actually include discussions of multiple topics). Redactions proved to be rare, affecting only 202 conversations, and being limited to part or all of a single message in 151 of those conversations, so we retained conversations that included redacted messages.

This process resulted in 8,282 conversations. From these, we removed the 55 that contained only a single message and the 40 with the greatest number of messages. This resulted in 8,187 conversations for our test collection, each containing between 2 and 303 consecutive SMS messages between a pair of correspondents. For most (6,439) of these conversations, the time span between the first and last messages is no more than 24 hours. Together, the 8,187 conversations contain 121,114 messages (for a mean conversation length of 14.8 messages); only 184 of the 8,187 conversations contain 100 or more SMS messages.

For convenience, we used the LDC conversations as the unit of annotation for relevance judgments. In other words, our evaluation asks whether we are able to find a conversation that the searcher might wish to see. We also used the conversations as retrieval units when pooling system results for relevance judgment, although we additionally conducted post-hoc experiments with smaller indexing units, thus preferring conversations that exhibit a temporal concentration of content on the topic being searched for.

```
<top lang="en" type="opinion">
<num> 034 </num>
<title> farmers markets </title>
<desc>
What do people think about farmers'
markets?
</desc>
<narr>
Farmers' markets feature a retail
market where food items  are sold
directly by farmers to consumers.
To be relevant, conversations would
contain people expressing their
opinions on farmers' markets.
</narr>
</top>
```

Figure 1: An example topic.

Table 1: Topics types across the two assessment phases along with example titles for each type.

Type	Phase1	Phase2	Total	Example Title
opinion	3	4	7	new Xbox release
behavior	4	3	7	disobeying rules
experience	7	9	16	living with parents
knowledge	0	1	1	Philly bars
Total	14	17	31	

4. Topics and Queries

The most challenging aspect of creating our test collection was to develop topics that we believed reasonably represent what real users might actually search for in an SMS message archive. As a starting point, we looked to two prior observational studies for other types of conversational content. Looking first to the nature of the content, Naaman et al. report that the most common types of content users share on Twitter are opinions or complaints, self reports, random thoughts and facts (Naaman et al., 2010). Looking next to what people might ask about, Oard clustered questions from an existing question answering test collection for searching Web discussion forums, identifying four categories: "open" opinion oriented questions that do not suggest a perspective, focused opinion-oriented questions that ask for opinions on a specific aspect of a topic, experiential questions and knowledge-oriented questions (Oard, 2012).

To see which categories might be useful for an SMS collection, we examined part of our collection (about 100 conversations). After this we settled on four topic types: *opinion* (seeking personal opinions), *behavior* (seeking to learn what people do), *experience* (seeking insights from someone with experience) and *knowledge* (seeking to learn something that someone knows). We then described these question types in general terms to a colleague who had not seen the collection and asked them to craft some plausible topics of each type. We additionally looked for inspiration to topic descriptions from TREC Microblog tracks (2011, 2012, 2013) and TREC Robust tracks (2004, 2005).

We formalized each topic in a TREC-like format, as illustrated in Figure 1. Title (T) contains a few words that we expect a user might type as an initial search; Description (D) is a fully formed question styled after a question answering task; and Narrative (N) is intended to guide relevance assessment. A total of 62 topics were developed in this way (see `https://github.com/rashmisankepally/SMSTestCollection/blob/master/topics.txt`). We indexed the collection using Indri [2] with conversations as the unit of retrieval, and then checked to see if we could find at least some relevant content for each topic. This triage process yielded 36 potentially useful topics, of which we actually used the 31 for which our assessors ultimately found one or more relevant conversations. Table 1 shows the final number of topics by type for the two phases of relevance assessment as described in section 5. below.

5. Relevance Assessment

We hired 3 assessors to perform relevance judgments. Because none had prior relevance judgment experience, we divided the task into two phases. After some initial training, each assessor judged 7 topics for relevance, 3 of which

[2] https://www.lemurproject.org/indri/

Table 2: Inter-assessor agreement (kappa).

Assessor Pair	Phase 1	Phase 2
A:B	0.090	0.328
B:C	0.100	0.203
C:A	0.184	0.498

were common to all assessors. Phase 1 thus produced judgments for 15 different topics.

Pools of conversations to be judged for each topic were created by selecting the top 50 conversations from three diverse ranked retrieval systems: (1) language model (LM): Indri's (Strohman and others, 2005) default Language Modeling with $\mu = 1000$, (2) query expansion (QE): the pseudo relevance feedback model in Indri with 20 documents and 30 terms used for query expansion, and (3) BM25: Indri's implementation of BM25 term weights, with default parameters. Each of these was run three times (once each with T, TD, TDN queries, each created by concatenating all words in the indicated fields).

Assessors were asked to give a score for each topic-conversation pair that they were presented. Conversations were formatted for display as shown in Figure 2. Assessors were asked to base their decision regarding relevance on whether the information need was addressed by any part of the conversation. They had four options for assessment: HREL (worthy of being a top result); REL (somewhat relevant content); NON (no useful information about the topic); and JUNK (no useful information for any purpose). For scoring, we collapsed the NON and JUNK categories into a single "non-relevant" category. This resulted in ternary graded judgments that (for computing measures such as kappa and Average Precision) we further binarized by additionally collapsing HREL and REL into a single "relevant" category.

We evaluated assessor agreement for the 3 common topics (described below) in phase 1 and used those results to guide a conversation among the assessors with the goal of achieving a greater agreement in phase 2. In the second phase, each assessor judged relevance for 9 additional topics, 3 of which were again common. This produced judgments for the remaining 21 topics. Depth 50 pooling was again used, this time with 12 runs (the same 9 runs as phase 1, plus 3 runs (T, TD, TDN) for a system built using word2vec (Mikolov and others, 2013)). Our word2vec system was a query expansion model using 100-dimension word embedding vectors that was added to increase the diversity of the pools. Word2vec's clustered bag of words (CBAG) model was used, with context set to 10 and number of iterations set to 5.

Over the total of 31 topics that each had at least one relevant document, there are a total of 214 relevant conversations in the pools (119 for 14 topics in phase 1, 95 for 17 topics in phase 2) for an average of 6.9 relevant conversations per topic.

5.1. Assessor Agreement and Comparing System Rankings

In phase 1, the pools for topic numbers 015, 017 and 020 were judged by all three assessors; together those judgment pools contained 438 conversations on which we could compute Cohen's kappa, a chance-corrected agreement measure. In phase 2, the pools for topics 023, 024 and 032 were judged by all three assessors; in this case kappa was computed over 421 conversations. As Table 2 shows, annotator agreement improved markedly in phase 2. Some possible reasons for the lower agreement in phase 1 were: long conversations, abrupt topic shifts in a conversation, and differing interpretations of a topic. That last factor was perhaps exacerbated by cultural factors, as all the three assessors were from India or China, while the SMS messages had been collected in the USA. Post hoc analysis found that length did not measurably affect inter-annotator agreement.

```
<conversation id="SMS_ENG_20110xx.x">

 m0000 - A:[2011-02-09 19:03:03]
         Where you at.what you doing?
 m0001 - B:[2011-02-09 19:04:04]
         I'm in NYC to see a show for
         the evening!
 m0002 - B:[2011-02-09 19:19:15]
         I have an extra ticket if
         you wanna join :)
```

Figure 2: Conversation format shown to assessors.

The kappa values in phase 2 are more typical of agreement statistics for relevance judgment results in other settings (Voorhees, 2000). Ultimately however, what we care about is whether the resulting judgments can be used to determine which retrieval systems are better. To explore this, we examined the degree to which judgments from different assessors during the second phase would produce the same preference order among systems when those systems were evaluated by Average Precision, a widely used ranked retrieval measure. Of the three topics for which judgments were obtained for all assessors, only topic 024 proved to be suitable for this analysis. Assessor B judged none of the sampled conversations as relevant to Topic 032, thus precluding computation of any system rankings for that topic using Assessor B's judgments. For Topic 023, each assessor found only two relevant conversations. While it is numerically possible to compute Average Precision with only two relevant items, the use of the inverse rank in the Average Precision computation introduces substantial quantization noise when so few relevant items are known to exist. For topic 024, by contrast, no assessor judged fewer than 5 conversations to be relevant (specifically, Assessor A found 12, Assessor B found 5, and Assessor C found 16 relevant conversations). Five relevant items suffice to compute Average Precision with only minimal quantization noise effects.

Figure 3 plots a comparison for topic 024, showing Average Precision (AP) scores for 15 systems (the 12 that contributed to the pools, plus 3 passage retrieval runs—for the same three query lengths—constructed using 60-

Assessor pair	τ
A-B	0.45
B-C	0.30
C-A	0.16

Table 3: Rank correlation for Topic 024.

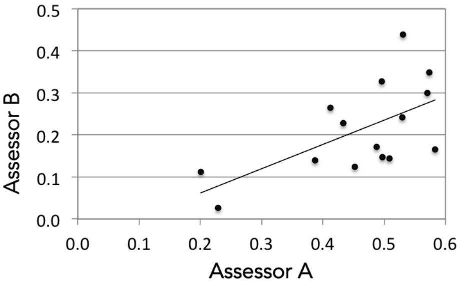

Figure 3: Average Precision Correlation for Topic 024.

word sliding windows with 45-word overlap and Indri's language model, with each document assigned the score of its highest-scoring passage) computed using relevance judgments from different assessors. Although Assessor B's scores tend to be somewhat lower than Assessor A's (perhaps because Assessor A judged more conversations to be relevant, thus setting up an easier search problem), higher AP scores computed with Assessor B's judgments are clearly predictive of higher AP scores computed with Assessor A's judgments.

We can summarize the degree of consistency of system rankings using Kendall's τ, a rank correlation number whose values range between -1 and 1. (with 1 indicating identical rankings, 0 indicating completely random swaps that would be characteristic of unrelated rankings, and -1 indicating complete reversal). As Table 3 shows, assessors A and B show reasonable consistency in system ranking. Another thing that we can observe from Table 3 is that systems ranked using Assessor B's judgments agree more strongly with the judgments of Assessors A and C than do Assessors A and C agree with each other. Thus we have some indication that choosing Assessor B's judgments would be reasonable for the six topics that had been judged by all three assessors. That's what we have done to create the final relevance judgment file in the released test collection.

6. Using the Collection

Table 4 shows phase 2 results for the systems that contributed to the pools, and for T, TD, and TDN runs with a fifth system (LM-60-45), the best of a set of passage retrieval systems we built for post hoc experiments in which our goal was to assess the effectiveness of more focused retrieval techniques. For LM-60-45 we use a sliding window to form 60-word passages, with 45-word overlap between adjacent passages. We score each passage using the Indri LM, and each conversation is assigned the maximum LM score across its passages. The best of the systems that we tried (LM-60-45 with TDN queries) achieved MAP above 0.4 (indicating that near the top of the ranked list more than 40% of the conversations were relevant, on average) and nDCG near 0.4 (indicating that on average the best system ranks documents better than half as well as the best possible ranking ,which would rank all the HREL ahead of all the REL, which are in turn ahead of all the other conversations). These results suggest that a retrieval system built with the best of these methods should be usable for typical interactive search tasks. The correlation between these 15 systems across the two measures is nearly perfect; the only reversal is between the two lowest-ranked systems. Results for phase 1 topics were broadly comparable, ranking the smaller number of systems that we tested in phase 1 consistently with those in phase 2, with MAP ranging between 0.268 and 0.463 and nDCG range in between 0.505 and 0.659.

Table 4: Phase 2 results.

Query	System	MAP	nDCG
TDN	LM-60-45	0.421	0.594
TD	LM-60-45	0.414	0.587
T	LM-60-45	0.412	0.575
T	QE	0.379	0.550
T	BM25	0.367	0.534
T	LM	0.352	0.523
TDN	LM	0.350	0.516
T	word2vec	0.343	0.506
TD	LM	0.338	0.500
TD	QE	0.325	0.499
TD	word2vec	0.304	0.460
TD	BM25	0.239	0.442
TDN	QE	0.219	0.405
TDN	word2vec	0.191	0.305
TDN	BM25	0.130	0.307

Although our principal focus has been on development of the test collection, two results seem worthy of note. First, our best passage retrieval model did substantially better than any other approach. Pooled assessment can create some degree of bias against post hoc assessment of new systems, so this result suggests that passage retrieval may be a useful approach for this task. Second, longer (TDN) queries yielded better results with passage retrieval, whereas shorter (T) queries yielded better results for every other approach. This comports with our intuition that additional context can be provided by longer queries or by longer "documents," thus suggesting that when short queries are all that the user provides we might prefer longer passage lengths (up to entire conversations).

7. Conclusion

We now have a test collection for evaluating information retrieval systems designed for SMS conversations. Topics, relevance judgments and the list of SMS conversation IDs used in the collection can be obtained from `https://github.com/rashmisankepally/SMSTestCollection`. Although the collection contains only 31 topics, a number generally considered too few for reliable statistical significance tests (Sanderson and Zobel, 2005), we have already been able to make some useful observations. Mean Average Precision results for phase 2 topics ranged between 0.130 and 0.421, values consistent with those observed on typical TREC collections, showing that the systems that contributed to the pools performed well enough to find a substantial number of relevant documents. Since we completed this work, the LDC has released larger SMS collections, so leveraging what we have learned to create larger test collections, both with more messages and with more topics, would be a natural next step. In addition to judgments on conversations, we also collected (but have not yet analyzed) message-level relevance judgments. Analyzing agreement on those message-level judgments, and using those judgments for more fine-grained evaluation, would be worth doing.

8. Bibliographical References

Contractor, D. et al. (2013). Text retrieval using SMS queries: Datasets and overview of fire 2011 track on SMS-based FAQ retrieval. In *Multilingual Information Access in South Asian Languages*, pages 86–99. Springer.

Griffitt, K. and Strassel, S. (2016). The query of everything: Developing open-domain, natural-language queries for BOLT information retrieval. In *LREC*.

Grossman, M. R. et al. (2011). Overview of the TREC 2011 legal track. In *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, USA, November 15-18, 2011*.

Lin, J. et al. (2014). Overview of the TREC-2014 microblog track (notebook draft). In *Proceedings of TREC*.

Macdonald, C., Ounis, I., and Soboroff, I. (2009). Overview of the TREC 2009 blog track. In *Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009*.

Mikolov, T. et al. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.

Oard, D. (2012). Answering questions from conversations. In *COLING Question Answering for Complex Domains Workshop*. Powerpoint slides.

Pecina, P. et al. (2007). Overview of the CLEF-2007 cross-language speech retrieval track. In *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers*, pages 674–686.

Sanderson, M. and Zobel, J. (2005). Information retrieval system evaluation: Effort, sensitivity, and reliability. In *SIGIR*, pages 162–169.

Song, Zhiyi and others. (2014). *Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus.*

Strohman, T. et al. (2005). Indri: A language model-based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, volume 2, pages 2–6.

Voorhees, E. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M*, 36(5):697–716.