Search Success Evaluation with Translation Model

Cheng Luo*, Yiqun Liu, Min Zhang, and Shaoping Ma

State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Department of Computer Science and Technology, Tsinghua University
Beijing 100084, China
yiqunliu@tsinghua.edu.cn
http://www.thuir.cn

Abstract. Evaluation plays an essential way in Information Retrieval (IR) researches. Existing Web search evaluation methodologies usually come in two ways: offline and online methods. The benchmarks generated by offline methods (e.g. Cranfield-like ones) could be easily reused. However, the evaluation metrics in these methods are usually based on various user behavior assumptions (e.g. Cascade assumption) and may not well accord with actual user behaviors. Online methods, in contrast, can well capture users' actual preferences while the results are not usually reusable. In this paper, we focus on the evaluation problem where users are using search engines to finish complex tasks. These tasks usually involve multiple queries in a single search session and propose challenges to both offline and online evaluation methodologies. To tackle this problem, we propose a search success evaluation framework based on machine translation model. In this framework, we formulate the search success evaluation problem as a machine translation evaluation problem: the ideal search outcome (i.e. necessary information to finish the task) is considered as the reference while search outcome from individual users (i.e. content that are perceived by users) as the translation. Thus, we adopt BLEU, a long standing machine translation evaluation metric, to evaluate the success of searchers. This framework avoids the introduction of possibly unreliable behavior assumptions and is reusable as well. We also tried a number of automatic methods which aim to minimize assessors' efforts based on search interaction behavior such as eye-tracking and click-through. Experimental results indicate that the proposed evaluation method well correlates with explicit feedback on search satisfaction from search users. It is also suitable for search success evaluation when there is need for quick or frequent evaluations.

Keywords: search engine evaluation, search success evaluation, user behavior

1 INTRODUCTION

Evaluation plays a critical role in IR research as objective functions for system effectiveness optimization. Traditional evaluation paradigm focused on assessing system performance on serving "best" results for single queries. The Cranfield method proposed by Cleverdon [4] evaluates performance with a fixed document collection, a query set, and relevance judgments. The relevance judgments of the documents are used to calculate various metrics which are proposed based on different understanding of users'

* This work was supported by Natural Science Foundation (61622208, 61532011, 61472206) of China and National Key Basic Research Program (2015CB358700).

behavior. We refer this type of evaluation paradigm as *offline* evaluation, which is still predominant form of evaluation.

To line up the evaluation and the real user experience, *online* evaluation tries to infer users' preference from implicit feedback (A/B test [17], interleaving [14]), or explicit feedback (satisfaction [11]). Online methods naturally take user-based factors into account, but the evaluation results can hardly be reused.

Offline and Online methods have already achieved great success in promoting the development of search engine. However, offline evaluation metrics do not always reflect real users' experience [26]. The fixed user behavior assumptions (e.g. Cascade assumption) behind offline metrics may lead to failures on individual users. Consider an example in our experiment (depicted in Figure 1), user A and B worked on the same task in one search engine and behaved in similar ways, the offline measurements should also be similar. However, the actual normalized scores given by external assessors showed that there was a relatively great difference in their success degrees.

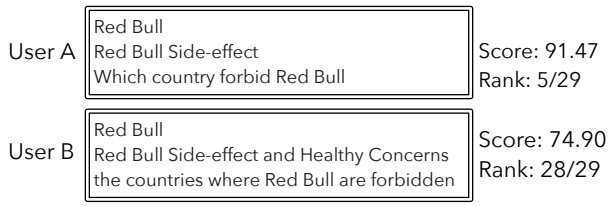

Fig. 1. An Example of Two User Sessions with Similar Offline Evaluation Results but Completely different Online Feedback (Scores by assessors and Rank among all 29 participants)

In a typical search, according to the Interpretive Theory of Translation (ITT) [15], the search process can be modelled as three interrelated phases: (1) reading the content, (2) knowledge construction and (3) answer presentation. Inspired by this idea, we formalize the search success evaluation as a machine translation evaluation problem and propose a Search Success Evaluation framework based on Translation model (SSET). The ideal search outcome, which can be constructed manually, is considered as the *"reference"*. Meanwhile, the individual search outcome collected from a user is regarded as a *"translation"*. In this way, we can evaluate *"what degree of success the user has achieved"* by evaluating the correspondence between the ideal search outcome and individual search outcome. We investigate a number of machine translation (MT) evaluation metrics and choose BLEU (Bilingual Evaluation Understudy) for its simpleness and robustness.

To reduce the effort of manually construction of ideal search outcome, we also propose an automatic extraction method with various users' behavior data. Experiments indicate that evaluation with automated extracted outcome performs comparatively as well as with manually organized outcomes. Thus, it is possible to perform automated online evaluation including relatively large scale of users. In summary, our contribution includes: (1) A search evaluation framework based on machine translation model. To the best of our knowledge, our study is among the first to evaluate success with machine translation models. (2) An extraction method for the automatic generation of references with the help of multiple users' search interaction behavior (e.g. eye-tracking) is proposed and enables quick or frequent evaluations. (3) Experiment framework and data shared with the research community.

2 Related Work

Online/Offline Search Evaluation. Cranfield-like approaches [4] introduced a way to evaluate ranking systems with a document collection, a fixed set of queries, and rel-

evance assessments from professional assessors. Ranking systems are evaluated with metrics, such as Precision, Recall, nDCG etc. The Cranfield framework has the advantage that relevance annotations on query-document pairs can be reused.

Beyond Cranfield framework, IR community strives to make evaluation more centred on real users' experience. The *online* evaluation methods, observing user behavior in their natural task procedures offer great promise in this regard. The *satisfaction* [11] method will ask the users to feedback their satisfaction during the search process explicitly, while the *interleaving* [14], *A/B testing* [17] methods try to infer user preference depending on implicit feedbacks, such as click-through etc. The evaluation results can hardly be reused for other systems which are not involved in the online test.

Session Search Evaluation. Beyond serving "best" results for single queries, for search sessions with multiple queries, several metrics are proposed by extending the single query metrics, i.e. the *nsDCG* based on *nDCG* [12] and *instance recall* based on *recall* [23]. Yang and Lad [31] proposed a measure of expected utility for all possible browsing paths that end in the *k*th reformulation. Kanoulas et al. [16] proposed two families of measures: one model-free family (for example, *session Average Precision*) that makes no assumption about the user's behavior and the other family with a simple model of user interactions over the session (*expected session Measures*).

Search Success Prediction. Previous researchers intuitively defined *search success* as the *information need fulfilled* during interactions with search engines. Hassan et al. [10] argued that relevance of Web pages for individual queries only represented a piece of the user's information need, users may have different information needs underlying the same queries. Ageev et al. [1] proposed a principled formalization of different types of "success" for informational tasks. The success model consists of four stages: *query formulation*, *result identification*, *answer extraction* and *verification of the answer*. They also presented a scalable game-like prediction framework. However, only binary classification labels are generated in their approach.

What sets our work apart from previous approaches is the emphasis on the outcomes the users gained through multiple queries. Our framework evaluates the success based on the information gained by users rather than implicit behavior signals. Ageev et al.'s definition about "success" was designed to analyze the whole process of their designed informational tasks. In our work, we simplify this definition and mainly focus on in what degree the user has gained enough information for certain search tasks.

Machine Translation Evaluation. Machine Translation models have been explored in Information Retrieval research for a long time [3, 7]. However, machine translation evaluation methods have not been explored in search success evaluation problem. Several automatic metrics were accomplished by comparing the translations to references, which were expected to be efficient and correlate with human judgments. BLEU was proposed by Papineni et al. [24] to evaluate the effectiveness of machine translation systems. The scores are calculated for individual language segments (e.g. sentences) combining modified *n-gram* precision and brevity penalty. Several metrics were proposed later by extending BLEU [5, 28].

Based on our definition of success, we mainly focus on whether a user has found the key information to solve the task. BLEU offers a simple but robust way to evaluate how good the users' outcome are comparing to pre-organized ideal search outcome on *n-gram* level. Other MT evaluation metrics could be adopted in this framework in a similar way as BLEU and we would like to leave them to our future work.

3 METHODOLOGY

3.1 Search Success Evaluation with Translation Model (SSET)

During a typical Web search, the user's information gathering actions can be regarded as "distilling" information gained into an organized answer to fulfill his/her information need [3]. We take the view that this distillation is a form of translation from one language to another: from documents, generated by Web page authors, to search outcome, with which the user seeks to complete the search task. Different from the standard three step translation process (understanding, deverbalization and re-expression [27, 19]), "re-expression" is not always necessary in search tasks. In our framework, we retain the re-expression step by asking the participants to summarize their outcomes with the help of a predefined question so that we can measure the success of search process. The details of the framework will be represented in Section 4.

Comparing to previous success evaluation methodologies, we put more emphasis on user perceived information corresponding to the search task. We at first define some terminologies to introduce our framework:

Individual Search Outcome: for a specific user engaged in a search task, *search outcome* is the information gained by the user from interactions to fulfill the task's information need.

Ideal Search Outcome: for a certain topic, *ideal search outcome* refers to all possible information that can be found (by oracle) through reading the relevant documents provided by the search engine to fulfill the task's information need.

Search Success: *Search Success* is the situation that the user has collected enough information to satisfy his/her information need.

For a particular search task, a user read a sequence of words R^U. The search outcome of the user can be described by another sequence of words as S^J, while the ideal search outcome can be represented by a sequence of words as T^K. We assume users' search outcomes and the ideal search outcomes have identical vocabularies due to both of them come from the retrieved documents.

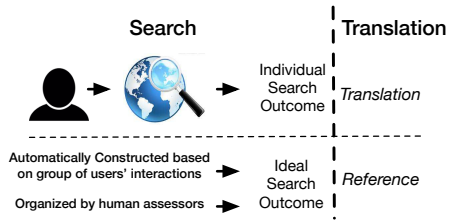

Fig. 2. Overview of Search Success Evaluation Framework with Translation Model

In this work, we propose a search success evaluation framework with translation model (SSET), which is presented in Figure 2. Suppose the user's individual search outcome is a *"translation"* from examined documents in the search session, we can treat the ideal search outcome as a *"reference"*, which can be constructed manually by human assessors, or automatically based on group of users' interaction behaviors.

When evaluating a translation, the central idea behind a lot of metrics is that "the closer a machine translation is to a professional human translation, the better it is" [24]. We assume that *"the closer a user's individual search outcome is to an ideal search outcome, the more successful the search is"*. The SSET model is proposed to evaluate the search success by estimating the closeness between the individual search outcome and the ideal search outcome.

In SSET, we first compute a modified *n-gram* precision for the individual search outcome S^J, according to the ideal search outcome T^K:

$$p_n = \frac{\sum_{n\text{-}gram \in S^J} min(c(n\text{-}gram, T^K), c(n\text{-}gram, S^J))}{\sum_{n\text{-}gram\prime \in T^K} c(n\text{-}gram\prime)} \tag{1}$$

where $c(n\text{-}gram, S)$ indicates the times of appearances of the *n-gram* in S. In other words, one truncates each *n-gram*'s count, if necessary, to not exceed the largest count observed in the ideal search outcome. Then the brevity penalty BP is calculated by considering the length of the individual search outcome (c), and the length of the ideal search outcome (r):

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \leq r \end{cases} \tag{2}$$

The SSET score combine both the modified precision of *n-gram* in different lengths and the brevity penalty, where w_n is the weight of the modified precision of *n-gram*, we use the typical value $N = 4$ in our experiment.

$$score_{SSET} = BP \cdot exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{3}$$

In this way, we can measure how close the individual search outcome is to ideal search outcome. The ideal search outcome organized by human assessors could be reused to evaluate other retrieval systems.

However, the generation process of ideal search outcome is still expensive and time-consuming. The individual search outcome generated by explicit feedback would also bring unnecessary effort to the users. We further explore the automation generation of *ideal search outcome* and *individual search outcome* based on multiple search interaction behaviors.

3.2 Automated Search Outcome Generation

We employ a bag-of-words approach, which has proven to be effective in many retrieval settings [18, 9], to generate pseudo documents as the users' individual search outcome and ideal search outcome.

Consider a user u involving in certain task t, we can calculate a modified TF-IDF score for each *n-gram* in the snippets and titles read by the user, the IDFs of terms are calculated on the titles and snippets of all the tasks' SERPs:

$$s_{n\text{-}gram} = \sum_{r \in ViewedBy\_u_t} (c\,(n\text{-}gram, r) \cdot w_r) \cdot IDF_{n\text{-}gram} \tag{4}$$

where w_r denotes the weight of documents. We can estimate w_r with the user's clicks or eye-fixations. For the score based on user clicks, $w_r = \#clicks\_on\_r$. For the score based on fixations, $w_r = \sum_{f \in fixations\_on\_r} \log\left(duration_f\right)$.

Thus, we can construct the user's individual search outcome *indiv_so* by joining the top-k *n-grams* with greatest scores in different lengths. Note that all the *n-grams* appearing in the task description are removed because we want to capture what extended information the user have learned from the search process. We could calculate $s'_{n\text{-}gram}$ for the *n*grams with group of users' clicks or eye-fixations in a similar way. The ideal search outcome could be organized by utilizing group of users' interactions, which is assumed to be a kind of "wisdom of crowds".

More fine-grained user behaviors (e.g. fixations on the words) are promising to help the generation of search outcome extraction. Due to the limit of experimental settings in our system, we would leave them to our future work.

Based on different ideal/individual search outcome extraction methods, we get several SSET models which are shown in Table 1. The performance of these models will be discussed in Section 5. In the experiments we find that the SSET models with eye-tracking data always outperform the models with clickthrough information, thus, we only report the performance of models with eye-tracking in the remainder of this paper.

Table 1. SSET with Different Outcome Measurements

	Individual Search Outcome	**Ideal Search Outcome**
SSET1MM	Summarized by participants	Organized by assessors
SSET2MA	Summarized by participants	Generated based on Eye-fixation
SSET3AM	Generated based on Eye-fixation	Organized by assessors
SSET4AA	Generated based on Eye-fixation	Generated based on Eye-fixation

4 EXPERIMENT SETUPS

We conducted an experiment to collect user behaviors and search outcomes for completing complex tasks. During the whole process, users' queries, eye fixation behaviors on Search Engine Result Pages (SERPs), clicks and mouse movements are collected.

Search Task. We selected 12 informational search tasks for the experiment. 9 of them were picked out from recent years' TREC Session Track topics. According to the TREC Session Track style, we organized 3 tasks based on the participants' culture background and the environment they live in. The criteria is that the tasks should be clearly stated and the solutions of them cannot be retrieved simply by submitting one query and clicking the top results. Each task contains three parts: an information need description, an initial query and a question for search outcome extraction. The description briefly explains the background and the information need. To compare user behavior on query level, the first query in each task was fixed. We summarized the information needs and extracted key words as relatively broad queries. People may argue that the fixed initial queries might be useless for searcher. Statistics shows that there are average 2.33 results clicked on the SERPs of initial queries. At the end of the task, the question is showed to the participants which requires them summarize the information gained in the searching process and the answers were recorded by voice.

Experimental System. We built an experimental search system to provide modified search results from a famous commercial search engine in China. First, all ads and sponsors' links were removed. Second, we removed vertical results to reduce possible behavior biases during searching process [30]. Third, we remove all the query suggestions because we suppose that query reformulation might reflect potential interests of users. Besides these changes, the search system looks like a traditional commercial search engine. The users could issue a query, click results, switch to the landing pages and modify their queries in a usual way. All the interactions were logged by our background database, including clicks, mouse movement and eye-tracking data.

Eye-tracking. In the experiment, we recorded all participants' eye movements with a Tobii X2-30 eye-tracker. With this tracking device, we are able to record various ocular behaviors: fixations, saccades and scan paths. We focus on eye fixations since fixations and durations indicate the users' attention and reading behavior [25].

Participants We recruited 29 undergraduate students (15 females and 14 males) from a University located in China via email, online forums, and social networks. 17 of 29 participants have perfect eyesight. For the others, we calibrated the eye-tracker carefully

to make sure the tracking error was acceptable. All the participants were aged between 18 and 26. Ten students are major in human sciences, fifteen are major in engineering while the others' majors range from arts and science. All of them reported that they are familiar with basic usage of search engines.

Procedure. The experiment proceeded in following steps, as shown in Figure 3. First the participants were instructed to read the task description carefully and they were asked to retell the information need to make sure that they had understood the purpose of the search tasks. Then the participants could perform searches in our experimental system as if they were using an ordinary search engine. We did not limit their time of searching. The participants could finish searching when they felt satisfied or desperate. After searching for information, the participants were asked to judge and rate queries/results regarding their contribution to the search task. More specifically, they were instructed to make the following three kinds of judgments in a 5-points Likert scale, from strong disagreement to strong agreement:

- For each **clicked result**, how useful it is to solve the task?
- For each **query**, how useful it is to solve the task?
- Through the search **session**, how satisfied did the participant feel?

At last, the system would present a question about the description, which usually encourage the participant to summarize their searches and extract search outcome. The answers from users would be recorded by voice. We notice that answering the question by voice-recording could not only reduce participants' effort but also give them a hint that they should be more serious about the search tasks.

Fig. 3. Experimental Procedure (Translated from original Chinese system)

5 EXPERIMENTAL RESULTS AND DISSCUSSIONS

This section will lead to answers to 3 research questions:

RQ1: How well do the result of SSET correlate with human assessments? Can we use it as an understudy of human assessments?

RQ2: What's the relationship between SSET, and offline/online metrics?

RQ3: Does automatic methods work for SSET? Can we extract the ideal/individual search outcomes based on a single user's or group of users' behavior automatically?

5.1 Data and Assessments

In our experiment, we collected search behavior and success behavior from 29 participants on 12 unique tasks. To evaluate search success, we recruited 3 annotators to assess the degree of success based on the users' answer after each task. The assessors were instructed to make judgments with magnitude estimation (ME) [29] methods, rather than ordinal Likert scale. ME could be more precise than traditional multi-level categorical judgments and ME results were less influenced by ordering effects than multi-points scale [6]. For each task, before assessments, the assessors were represented with the task description and the question. The records of 29 participants are randomly listed on

a webpage, each assessor make judgments sequentially. For each record, the assessor can listen to the record one or more times and then assign a score between 0 and 100 to the record in such a way that the score represents how successful the record is. The score was normalized according to McGee et al.'s method [22]. In this paper, we use the mean of normalized scores from three assessors as the Ground Truth of search success evaluation.

While assessing the participants' answers, we find that the question of Task 10 ("Please tell the support conditions of the Hong Kong version iphone to domestic network operators.") fails to help the participants to summarize their search outcome depending on the task description ("You want to buy a iphone6 in Hong Kong. Please find the domestic and Hong Kong price of iphone6, how to purchase iphone in Hong Kong, whether it is necessary to pay customs on bringing iphone home, whether the Hong Kong version of iphone would support domestic network operators, etc."), because the question just focuses on a detailed fact about the task. Thus, in the reminder analysis of this paper, Task 10 and corresponding data is removed and we have 319 sessions (11 tasks with 29 participants) in total.

After assessments, we asked the assessors to organized *standard answers* for the 12 tasks. More specifically, the three assessors were instructed to search information about the tasks with the retrieval system that were used by the participants. Note that the assessors did not perform any search before assessments for individual search outcome to avoid potential biases,e.g., they may prefer the individual outcomes similar to the documents examined by them. Then, the assessors organized their own answers and summarized the *ideal search outcomes* based on both their own answers and the 29 participants'. In addition, all the recorded voices were converted to text, with discourse markers removed, which were regarded as users' individual search outcomes.

5.2 SSET v.s. Human Assessment.

With our proposed evaluation framework, Search Success Evaluation with Translation model (SSET), we attempt to evaluate what degree of success a searcher has achieved in a certain task. The normalized scores for three assessors are regarded as *Ground Truth* of the performance evaluation of search success evaluation model.

For each session (a user in a certain task), the input of SSET includes a *"reference"*, the ideal search outcome, and a *"translation"*, the individual search outcomes from each participants and the SSET outputs the degree of success in a value range.

We calculate the correlation of SSET1MM model and the Ground Truth. The SSET1MM model uses the *ideal search outcomes* organized by external assessors as *"references"* and use the *answers of questions* (individual search outcomes) as *"translations"*. The correlation on each task is shown in Table 2.

The results show that SSET1MM correlates with the human judgments on most of tasks. The Pearson's r is significant at 0.01 for 10 of 11 tasks, which makes this method as an automated understudy for search success evaluation when there is need for quick or frequent evaluations.

We notice that the performance of SSET1MM varies with the tasks. It may suggest that the SSET is task-sensitive, in other words, SSET1MM is not appropriate for all kinds of tasks. From the facet of search goal identified by Li et al. [20], we can classify the tasks into 2 categories: *specific* (well-defined and fully developed) and *amorphous* (ill-defined or unclear goals that may evolve along with the user's exploration). Thus, we find SSET performs better on the specific task (Task 5,9,4,3,6,8,7) rather than on the amorphous tasks (Task 2,1,12,11,7). For an amorphous task (e.g. find a ice breaker

Table 2. Correlation between SSET1MM and the Ground Truth (*, **: correlation significant at 0.01, 0.001 level)

Tasks	Correlation with Ground Truth	
	Pearson's r	Kentall's τ-b
5	0.879**	0.363*
9	0.822**	0.600**
4	0.789**	0.670**
3	0.774**	0.551**
6	0.719**	0.524**
2	0.706**	0.378*
1	0.631**	0.295
12	0.630**	0.533**
8	0.629**	0.546**
11	0.552*	0.406*
7	0.537*	0.315

game), it is very difficult to construct a "perfect" search outcome including all possible answers. Therefore, SSET is more appropriate to evaluate the tasks which are well-defined and have restrained answers.

5.3 SSET v.s. Offline/Online Metrics

SSET attempt to combine the advantages of offline and online evaluation methods. The tasks and ideal search outcomes organized by human experts offline can be reused easily and the individuals' search outcomes can be collected online efficiently and effectively. In this section, we investigate the relationship between SSET and offline/online metrics.

Previous work [13] reported that session cumulated gain (sCG) [12] correlated well with user satisfaction. We use sCG as a offline measure of the search outcome, which is the sum of each query's information gain. For each query, its gain is originally calculated by summing the gains across its results. In this work, we use the participants' subjective annotation ("how useful it is to solve the task?") as a proxy of the query's gain, e.g. $SearchOutcome = sCG = \sum_{i=1}^{n} gain(q_i)$.

The correlation between SSET1MM and sCG are shown in Table 3. There is weak correlation between SSET1MM and sCG. It is partly due to the difference in cognitive abilities between users. Consider the example in Section 1, two users search for "the side effects of red bulls", they issued similar queries, viewed similar SERPs and got quite close sCG scores. However, the information they gained for completing the task differed at quality and quantity. In other words, it means that the offline metric may lead to failure to evaluate in what degree the user has achieved success in complex tasks.

User satisfaction is a session/task level online evaluation metrics. In our experiment, we asked the users to rate their satisfaction for each task. The correlation of SSET1MM and user satisfaction is shown in Table 3.

Experiment shows less of a relationship between SSET1MM and user satisfaction. Jiang et al. [13] reported that the satisfaction was mainly affected by two factors, search outcome and effort. However, the search success evaluation mainly focuses on the search outcome of users. No matter the degree of success is assessed by external assessors or the SSET systems, they are not aware of the effort that the user has made to achieve the search outcome. This could be a plausible explanation for the closeness to uncorrelated between SSET and satisfaction. Jiang et al. proposed an assumption that the satisfaction is the value of search outcome compared with search effort. As our proposed SSET is also a measurement of search outcome, we investigate the correlation between SSET/Search Effort.

Table 3. Correlation Comparison between SSET and Offline/Online Metrics (*, **: correlation significant at 0.01, 0.001 level)

Tasks	Correlation (Pearson's r)		
	SSET1MM v.s. SAT	SSET1MM/#Queries v.s. SAT	SSET1MM v.s. sCG
7	−0.144	0.762**	0.186
2	0.027	0.574*	0.218
4	0.131	0.574*	0.208
6	0.232	0.568*	0.159
5	−0.001	0.554*	0.252
1	0.125	0.536*	0.285
3	−0.212	0.527*	0.140
8	0.257	0.432	0.196
9	−0.037	0.329	0.127
11	0.087	0.252	0.063
12	0.094	0.227	0.080

Search effort is the cost of collecting information with the search engine, e.g., formulating queries, examining snippets on SERPs, reading results, etc. We follow the economic model of search interaction proposed in [2]. For a particular search session, we can use Q (number of queries) as a proxy of search effort. Table 4 shows that there is strong correlation between SSET1MM/#queries and user's satisfaction for most of the tasks and our proposed SSET is able to act as an indicator of search outcome.

5.4 Performance of Automated Outcome Extraction

Development of search engine is based on ongoing updates. In order to validate the effect of a change to prevent its negative consequences, the developers compare various versions of the search engines frequently. This motivate us to improve SSET with automated methods for the organization of ideal/individual search outcomes.

In Table 4, we compared the correlation between 4 different SSET models and the Ground Truth (external assessments). SSET1MM is the model which use manually organized as ideal search outcome and users' answer for questions as individual search outcome. We use SSET1MM as a baseline to evaluate other SSET models.

Table 4. Correlation (Pearson's r) Comparison between Different SSET Models and the Ground Truth (*, **: correlation significant at 0.01, 0.001 level)

Tasks	Correlation (Pearson's r)			
	SSET1MM	SSET2MA	SSET3AM	SSET4AA
5	0.879**	0.907**	−0.063	−0.263
9	0.822**	0.808**	−0.193	−0.131
4	0.789**	0.724**	−0.243	−0.108
3	0.774**	0.769**	−0.107	−0.143
6	0.719**	0.625**	−0.165	−0.006
2	0.706**	0.691**	0.143	−0.222
1	0.631**	0.685**	−0.779	−0.032
12	0.630**	0.412	−0.035	0.313
8	0.629**	0.652**	−0.080	0.354
11	0.552*	0.385	0.144	0.268
7	0.537*	0.565*	0.132	0.146

SSET2MA performs almost as well as SSET1MM. It uses the same way to collect individual search outcomes (e.g. summarized by users) but constructs the ideal search outcomes automatically based on users' eye fixations on snippets. Thus, in practical environment, we can generate ideal search outcome based on group of users' behavior.

SSET3AM and SSET4AA correlates poorly with the Ground Truth. In these two models, we adopt the individual search outcome extraction method based on the user's eye fixations on SERPs. The individual search outcomes generated automatically differs a lot from their answers. The potential two reasons are: 1) the sparsity of user behavior makes it difficult to extract search outcome. 2) what the user has read is not equal to what he/she has perceived. Similar phenomenon has been observed by previous researches [21].

We also investigate the performance of SSET2MA based on different size of users' behaviors. We randomly split the all the participants into five groups, four groups has six participants while the remaining one has five. Then we construct multiple ideal search outcomes by sequentially adding group of users' fixations into the SSET2MA model. Then we compare the correlations between SSET2MA models and the Ground Truth.

Table 5. Correlation (Pearson's *r*) Comparison between SSET2MA Models based on Different Size of Users' Behaviors and the Ground Truth (*, **: correlation significant at 0.01, 0.001 level)

Tasks	Correlation (Pearson's r) with the Ground Truth				
	$SSET2^1$	$SSET2^2$	$SSET2^3$	$SSET2^4$	$SSET2^5$
5	0.620**	0.792**	0.901**	0.907**	0.907**
9	0.508*	0.738**	0.800**	0.807**	0.808**
3	0.439	0.696**	0.769**	0.760**	0.769**
4	0.411	0.589*	0.701**	0.714**	0.724**
2	0.382	0.541*	0.660**	0.687**	0.691**
1	0.356	0.495*	0.629**	0.657**	0.685**
8	0.347	0.477	0.652**	0.654**	0.652**
6	0.298	0.433	0.589*	0.626**	0.625**
7	0.287	0.365	0.501*	0.561**	0.565*
12	0.101	0.276	0.327	0.414	0.412
11	0.220	0.287	0.342	0.383	0.385

The results are shown in Table 5, where $SSET2^k$ denotes the SSET2MA model based on the first k groups of users. As the size of users grows, the correlation between SSET2MA and the Ground Truth becomes stronger. The $SSET2MA^3$ almost performs as well as $SSET3AM^5$. In other words, in practice, we need about behavior data from about 15 people to construct a reliable ideal search outcome for SSET.

6 CONCLUSION AND FUTUREWORK

Although previous offline/online evaluation frameworks have achieved significant success in the development of search engines, they are not necessarily effective in evaluate in what degree of success the search users have achieved. In this work, we put emphasis on the outcomes the users gained through multiple queries. We propose a Search Success Evaluation framework with Translation model (SSET). The search success evaluation is formalized as a machine translation evaluation problem. A MT evaluation algorithm called BLEU is adopted to evaluate the success of searchers. Experiments shows that evaluation methods based on our proposed framework correlates highly with human assessments for complex search tasks. We also propose a method for automatic generation of ideal search outcomes with the help of multiple users' search interaction behaviors. It proves effective compared with manually constructed ideal search outcomes. Our work can help to evaluate search success as an understudy of human assessments when there is need for quick or frequent evaluation. In the future work, we plan to adopt more MT evaluation methods in this framework and compare the performance in evaluate different types of tasks. Experiments with a relatively large scale of participants will be conducted based on crowdsourcing platforms.

References

1. M. Ageev, Q. Guo, D. Lagun, and E. Agichtein. Find it if you can: a game for modeling different types of web search success using interaction data.
2. L. Azzopardi. Modelling interaction with economic models of search. In *SIGIR'14*.
3. A. Berger and J. Lafferty. Information retrieval as statistical translation. In *SIGIR'99*.
4. J. M. C.W. Cleverdon and M. Keen. Factors determiningthe performance of indexing systems. *Readings in Information Retrieval*, 1966.
5. G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *HLT'02*.
6. M. B. Eisenberg. Measuring relevance judgments. *IPM*, 1988.
7. J. Gao, X. He, and J.-Y. Nie. Clickthrough-based translation models for web search: from word models to phrase models. In *CIKM'10*.
8. G. A. Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013.
9. A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *NAACL'09*.
10. A. Hassan, R. Jones, and K. L. Klinkner. Beyond dcg: user behavior as a predictor of a successful search. In *WSDM'10*.
11. S. B. Huffman and M. Hochster. How well does result relevance predict session satisfaction? In *SIGIR '07*.
12. K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen. Discounted cumulated gain based evaluation of multiple-query ir sessions. In *Advances in Information Retrieval*. 2008.
13. J. Jiang, A. Hassan Awadallah, X. Shi, and R. W. White. Understanding and predicting graded search satisfaction. In *WSDM '15*.
14. T. Joachims. Optimizing search engines using clickthrough data. In *KDD'02*.
15. C. Jungwha. The interpretive theory of translation and its current applications. 2003.
16. E. Kanoulas, B. Carterette, P. D. Clough, and M. Sanderson. Evaluating multi-query sessions. In *SIGIR '11*.
17. R. Kohavi, R. Longbotham, D. Sommerfield, and R. M. Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 2009.
18. M. Lease, J. Allan, and W. B. Croft. Regression rank: Learning to meet the opportunity of descriptive queries. In *ECIR'09*.
19. M. Lederer. *La traduction simultanée: expérience et théorie*, volume 3. Lettres modernes, 1981.
20. Y. Li and N. J. Belkin. A faceted approach to conceptualizing tasks in information seeking. In*IPM*, 2008.
21. Y. Liu, C. Wang, K. Zhou, J. Nie, M. Zhang, and S. Ma. From skimming to reading: A two-stage examination model for web search. In *CIKM'14*.
22. M. McGee. Usability magnitude estimation. In *HFES'03*.
23. P. Over. Trec-7 interactive track report. *NIST SPECIAL PUBLICATION SP*, 1999.
24. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL'02*.
25. K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 1998.
26. M. Sanderson, M. L. Paramita, P. Clough, and E. Kanoulas. Do user preferences and evaluation measures line up? In *SIGIR'10*.
27. D. Seleskovitch. *L'interprète dans les conférences internationales: problèmes de langage et de communication*, volume 1. Lettres modernes, 1968.
28. M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *AMTA'06*.
29. S. S. Stevens. The direct estimation of sensory magnitudes: Loudness. *The American journal of psychology*, pages 1–25, 1956.
30. C. Wang, Y. Liu, M. Zhang, S. Ma, M. Zheng, J. Qian, and K. Zhang. Incorporating vertical results into search click models. In *SIGIR'13*.
31. Y. Yang and A. Lad. Modeling expected utility of multi-session information distillation. In *ICTIR'09*