# USING CROWDSOURCING TO COMPARE DOCUMENT RECOMMENDATION STRATEGIES FOR CONVERSATIONS

Maryam Habibi Andrei Popescu-Belis

Idiap-RR-14-2012

JUNE 2012

**Centre du Parc**, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

# Using Crowdsourcing to Compare Document Recommendation Strategies for Conversations

Maryam Habibi
Idiap Research Institute and EPFL
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
maryam.habibi@idiap.ch

Andrei Popescu-Belis
Idiap Research Institute
Rue Marconi 19, CP 592
1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

## ABSTRACT

This paper explores a crowdsourcing approach to the evaluation of a document recommender system intended for use in meetings. The system uses words from the conversation to perform just-in-time document retrieval. We compare several versions of the system, including the use of keywords, retrieval using semantic similarity, and the possibility for user initiative. The system's results are submitted for comparative evaluations to workers recruited via a crowdsourcing platform, Amazon's Mechanical Turk. We introduce a new method, Pearson Correlation Coefficient-Information Entropy (PCC-H), to abstract over the quality of the workers' judgments and produce system-level scores. We measure the workers' reliability by the inter-rater agreement of each of them against the others, and use entropy to weight the difficulty of each comparison task. The proposed evaluation method is shown to be reliable, and demonstrates that adding user initiative improves the relevance of recommendations.

## Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Query formulation, Retrieval models*; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation*

## General Terms

Evaluation, Uncertainty, Reliability, Metric

## Keywords

Document recommender system, user initiative, crowdsourcing, Amazon Mechanical Turk, comparative evaluation

## 1. INTRODUCTION

A document recommender system for conversations provides suggestions of potentially relevant documents within a conversation between individuals, such as a business meetings. The system builds upon previous approaches known as implicit queries or just-in-time retrieval. Used as a virtual secretary, the system constantly retrieves documents that are related to the words of the conversation, using automatic speech recognition, but users are also allowed to make explicit queries.

Evaluating the relevance of recommendations produced by such a system is a challenging task. Evaluation in use requires the full deployment of the system and the setup of numerous evaluation sessions with realistic meetings. That is why alternative solutions based on simulations are important to find. In this paper, we propose to run the document recommender system over a conversational corpus, and to use crowdsourcing to compare the relevance of results in various configurations of the system.

Using a crowdsourcing platform, namely Amazon's Mechanical Turk, is helpful for several reasons. First, we can evaluate a large amount of data in a fast and inexpensive manner. Second, workers are sampled from the general public, which might represent a more realistic user model than the system developers, and have no contact with each other. However, in order to use AMT workers' judgments for relevance evaluation, we have to circumvent the difficulties of measuring the quality of the workers evaluations, and factor out the biases of individual contributions.

In this paper, we introduce an evaluation protocol using crowdsourcing, which estimates the quality of the workers' judgments by predicting task difficulty and workers' reliability, even if no ground truth to validate the judgments is available. This approach, named Pearson Correlation Coefficient-Information Entropy (PCC-H), is inspired by previous studies of inter-rater agreement as well as by information theory.

This paper is organized as follows. Section 2 describes the document recommender system and the different versions compared in the paper. Section 3 describes previous research on measuring the quality of workers' judgments for relevance evaluation and labeling tasks using crowdsourcing. Section 4 explains the design of the evaluation micro-tasks – "Human Intelligence Tasks" for the Amazon's Mechanical Turk. In Section 5, our proposed PCC-H method for measuring judgments qualities is explained. Section 6 presents the results of our evaluation experiments, which on the one hand validate the proposed method, and on the other hand illustrate the comparative relevance of the different versions of the recommender system.

## 2. OUTLINE OF THE DOCUMENT RECOMMENDER SYSTEM

The document recommender system under study is the Automatic Content Linking Device (ACLD, see [13]), which uses real-time automatic speech recognition [7] to extract words from a conversation, typically in a business meeting. The ACLD then filters and aggregates the words to prepare queries at regular time intervals. The queries can be addressed to a local database of meeting-related documents, including also transcripts of past meetings if available, but also to a web search engine. The results are then displayed in an unobtrusive manner to the meeting participants, which can consult them if they find them relevant and purposeful.

While it is difficult to assess the utility of the recommended documents from an absolute perspective, we rather aim at comparing several variants of the ACLD, in order to assess the improvement (or lack thereof) due to various designs. Here, we will compare four different approaches to the recommendation problem – which is in all cases a cold-start problem, as we don't assume knowledge about participants. Rather, in a pure content-based manner, the ACLD simply aims to find the closest documents to a given stretch of conversation.

The four compared versions are the following ones. Two "standard" versions as in [13] differ by the filtering procedure for the conversation words: one of them (noted AW) uses all the words (except stop words) spoken by users during a specific period (typically, 15 s) to retrieve related documents; the other one (noted KW) filters the words, keeping only keywords from a pre-defined list related to the topic of the meeting.

We implement three different algorithms for query aggregator module. The first one is ACLD using all words (AW) in which the system uses all the words spoken by users during a specific period to retrieve related documents. The second one is the ACLD using keywords only (KW), where the system applies important keywords selected by the system designer so as to search through the Internet. The third one is the ACLD using semantic search (noted SS, see [14]), which uses a graph-based semantic relatedness measure to perform retrieval. Finally, we include the most recent version which allows user initiative (noted UI), that is, explicit queries addressed by the users to the system. These are processed by the same ASR component, with participants using a specific name for the system ("John") to solve addressing problems.

In the evaluation experiments presented here, we only use human transcriptions of meetings, to focus on the evaluation of the retrieval strategy itself. We use one meeting (ES2008b) from the AMI Meeting Corpus [5] in which the design of a new remote control for a TV set is discussed. The users' requests for the UI version are simulated by modifying the transcript at 24 different locations where we think that users may ask explicit queries (a more principled approach for this simulation is currently under study). We restrict the search to the Wikipedia website, as it is one of the most popular general reference works on the Internet, and also because the semantic search system is adapted to this data (using a local copy of it, WEX, for semantic indexing). The 24 fragments of the meeting containing the explicit queries are then selected for comparison: that is, we want to know which of the results displayed by the various versions at the moment following the explicit query are considered most relevant by external judges. As the method allows only binary comparisons, as we will now describe, we will compare UI with the AW and KW versions, and then SS with KW.

## 3. RELATED WORK

Relevance evaluation is a hard task because it is subjective and expensive to be performed. Click data corpus and hiring experts are traditional ways of relevance evaluation. However, in our case click data would be expensive to produce, as would be hiring professional workers for relevance evaluation. One approach is using crowdsourcing, or peer collaborative annotation, which is relatively easy to prototype and to test experimentally. Moreover, it is a cheap and fast approach to explicit evaluation. However, it is necessary to consider some problems which are associated to this phenomenon such as spammers, reliability of workers judgment, and intrinsic knowledge of the workers [2].

Recently, many studies have considered the effect of the task design on relevance evaluation, and offer significant hints for task design in order to decrease time and cost of evaluation and increase the accuracy of results. In [8], several human factors are considered: query design, terminology, pay and its impact on cost, time and accuracy of annotations. To collect proper results, the effect of user interface guidelines, inter-agreement metrics and justification analysis were examined [1]. The study showed that asking workers to write a short explanation in exchange of bonus is an efficient method for detecting spammers. In addition, in [10], different batches of tasks were designed to measure the effect of pay, required effort and worker qualifications on the accuracy of resulting labels. Another contribution [12] has studied how the distribution of correct answers in the training data affects worker responses, and suggested to use a uniform distribution to avoid biases from unethical workers.

The Technique for Evaluating Relevance by Crowdsourcing (TERC, see [3]) emphasizes the importance of qualification control, e.g. by creating qualification tests that must be passed before performing the actual task. However, another study [1] showed that workers may still perform tasks randomly even after passing qualification tests. Therefore, it is important to perform partial validation of each worker's tasks, and weight the judgments of several workers to produce aggregate scores [3].

Several other studies have focused on Amazon's Mechanical Turk crowdsourcing platform and have proposed techniques to measure the quality of workers' judgments when there is no ground truth to verify them directly [15, 16, 6, 9, 11]. For instance, in [4], the quality of judgments for a labeling task is measured using the inter-rater agreement and majority voting. Expectation maximization (EM) has sometimes been used to estimate true labels in the absence of ground truth, e.g. in [15] for an image labeling task. In order to improve EM-based estimation of reliability of workers, the confidence of workers in each of their judgments has been used in [6] as an additional feature – the task being dominance level estimation for participants in a conversation. As the performance of the EM algorithm is not guaranteed, a new method [9] was introduced to estimate reliability based on low-rank matrix approximation.

All of the above-mentioned studies assume that tasks share the same level of difficulty. To model both task difficulty and user reliability, an EM-based method named GLAD was proposed by [16] for an image labeling task. However, this method is sensitive to the initialization value, hence a good estimation of labels requires a small amount of data with ground truth annotation [11].

# 4. SETUP OF THE EXPERIMENT

Amazon's Mechanical Turk (AMT) is a crowdsourcing platform which gives access to a vast pool of online workers paid by "requesters" to complete tasks, which are called Human Intelligence Tasks (HITs). Once designed and published, registered workers that fulfill the requesters' selection criteria are invited by AMT service to work on HITs in exchange for a (typically small) amount of money [2].

As it is difficult to find the absolute relevance evaluation for each version of the ACLD recommender system, we only focus on comparative relevance evaluation between versions. For each pair of versions, a batch of HITs was designed. Each HIT (see example in Figure 1) contains a fragment of conversation transcript with the two lists of document recommendations to be compared. Only the first six recommendations are kept for each system. The lists from the two compared versions are placed in random positions (first vs. second) across HITs, to avoid biases from a constant position.

We experimented with two different HIT designs. In the first one, with binary choice, workers are shown two choices for relevance: either the first list is considered more relevant than the second, or the other way round. In other words, they are obliged to express an explicit preference for one recommendation set. This encourages decisions, but of course may prove inappropriate when the two answers are of comparable quality – though this may be evened out when averaging over workers.

In the second HIT design, workers have four choices (as in Figure 1): in addition to the previous two options, they can indicate that both lists seem equally relevant, or equally irrelevant. In both designs, workers must select exactly one option.

To assign a value to each worker's judgment, a binary coding scheme will be used in the computations below, assigning a value of 1 to the selected option and 0 to all others. The relevance value $RV$ of each list for a given meeting fragment is computed by giving a weight to each worker judgment and averaging them. The percentage of relevance value $PRV$ is computed by assigning a weight to each part of the meeting and averaging the $RV$ for all meeting fragments.

There are 24 meeting fragments, hence 24 HITs in each batch for comparing pairs of systems, for UI vs. AW and UI vs. KW. As queries are not needed for SS vs. KW, we designed 36 HITs, with 30-second fragments for each. There are 10 workers per HIT, so there are 240 total assignments for UI-vs-KW and for UI-vs-AW (with a 2-choice and 4-choice design for each), and 360 for SS-KW. As workers are paid 0.02 USD per HIT, the cost for the five separate experiments was 33 USD, with an apparent average hourly rate of 1.60 USD. The average time per assignment is almost 50 seconds. All five tasks took only 17 hours to be performed by workers via AMT. For qualification control we allow workers with greater than 95% approval rate or with more than 1000 approved HITs.

# 5. THE PCC-H METHOD

Majority voting is the usual technique to aggregate multiple sources of comparative relevance evaluation. However, this assumes that all HITs share the same difficulty and all the workers are equally reliable. We will take here into account the task difficulty $W_q$ and the workers' reliability $r_w$, as it was shown that they have a significant impact on the quality of the aggregated judgments. We thus introduce a new computation method called PCC-H, for *Pearson Correlation Coefficient-Information Entropy.*

## 5.1 Estimating Worker Reliability

The PCC-H method computes the $W_q$ and $r_w$ values in two steps. In a first step, PCC-H estimates the reliability of each worker $r_w$ based on the Pearson correlation of each worker's judgment with the average of all the other workers judgments (see Eq. 1).

$$r_w = \frac{\sum_{a=1}^{A}\sum_{q=1}^{Q}(X_{qwa} - \bar{X_{wa}})(Y_{qa} - \bar{Y_a})}{(Q-1)S_{X_{wa}}S_{Y_a}} \qquad (1)$$

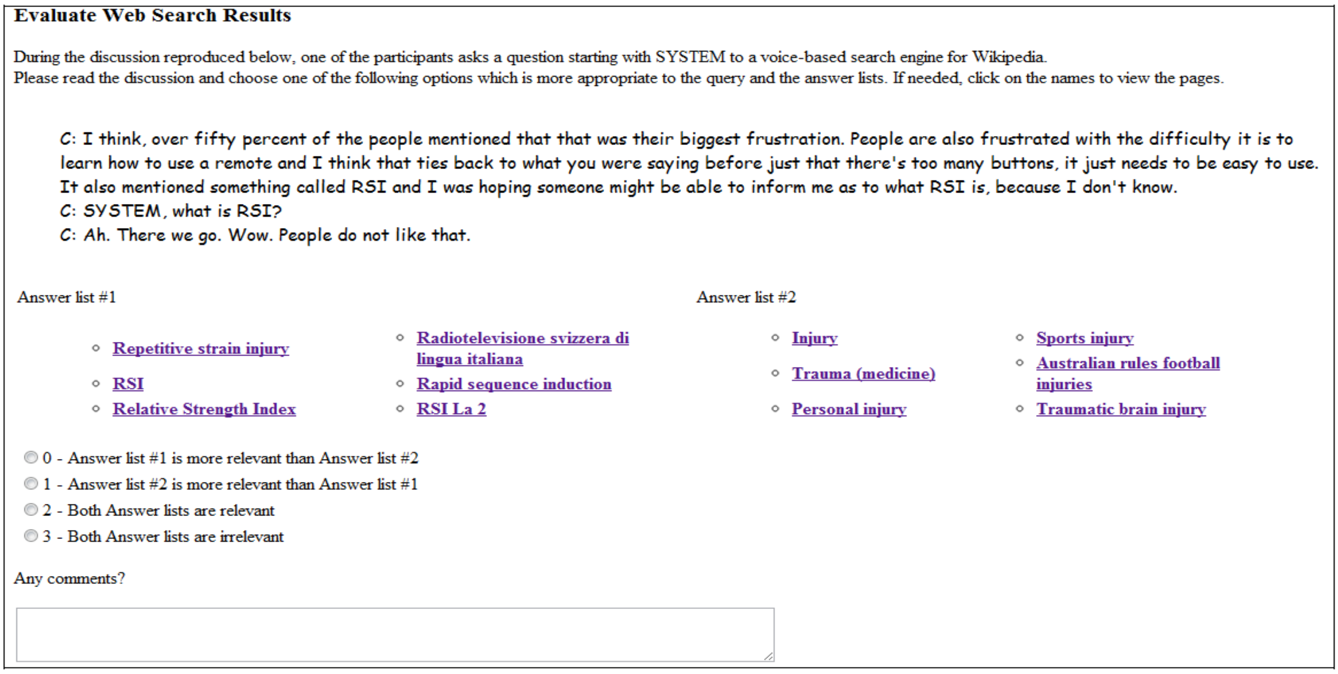
In Equation 1, $Q$ is number of meeting fragments, $X_{wqa}$ is the value that worker $w$ assigned to option $a$ of fragment $q$, $X_{wqa}$ has value 1 if that option $a$ is selected by worker $w$, otherwise it is 0. $\bar{X}_{wa}$ and $S_{X_{wa}}$ are the expected value and standard deviation of variable $X_{wqa}$ respectively. $Y_{qa}$ is the average value which all other workers assign to the option $a$ of fragment $q$. $\bar{Y_a}$ and $S_{Y_a}$ are the expected value and standard deviation of variable $Y_{qa}$.

The value of $r_w$ computed above is used as a weight for computing $RV_{qa}$ option $a$ of each fragment $q$ according to Eq. 2 below:

$$RV_{qa} = \frac{\sum_{w=1}^{W} r_w X_{wqa}}{\sum_{w=1}^{W} r_w} \qquad (2)$$

For HIT designs with two options, $RV_{qa}$ shows the relevance value of each answer list $a$. However, for the four option HIT designs, $RV_{ql}$ for each answer list $l$ is formulated as Eq. 3 below:

$$RV_{ql} = RV_{ql} + \frac{RV_{qb}}{2} - \frac{RV_{qn}}{2} \qquad (3)$$



**Evaluate Web Search Results**

During the discussion reproduced below, one of the participants asks a question starting with SYSTEM to a voice-based search engine for Wikipedia.
Please read the discussion and choose one of the following options which is more appropriate to the query and the answer lists. If needed, click on the names to view the pages.

C: I think, over fifty percent of the people mentioned that that was their biggest frustration. People are also frustrated with the difficulty it is to learn how to use a remote and I think that ties back to what you were saying before just that there's too many buttons, it just needs to be easy to use. It also mentioned something called RSI and I was hoping someone might be able to inform me as to what RSI is, because I don't know.
C: SYSTEM, what is RSI?
C: Ah. There we go. Wow. People do not like that.

Answer list #1

- Repetitive strain injury
- RSI
- Relative Strength Index
- Radiotelevisione svizzera di lingua italiana
- Rapid sequence induction
- RSI La 2

Answer list #2

- Injury
- Trauma (medicine)
- Personal injury
- Sports injury
- Australian rules football injuries
- Traumatic brain injury

○ 0 - Answer list #1 is more relevant than Answer list #2
○ 1 - Answer list #2 is more relevant than Answer list #1
○ 2 - Both Answer lists are relevant
○ 3 - Both Answer lists are irrelevant

Any comments?

**Figure 1: Snapshot of one HIT: workers read the conversation transcript, examine the two answer lists (with recommended documents for the respective conversation fragment) and select one of the four comparative choices (#1 better than #2, #2 better than #1, both equally good, both equally poor). A short comment can be added.**

In this equation, half of the relevance value of option in which both answer lists are relevant $RV_{qb}$ is added as a reward, and half of the relevance value of option in which both answer lists are irrelevant $RV_{qn}$ is subtracted as a penalty from the relevance value of each answer list $RV_{ql}$.

## 5.2 Estimating Task Difficulty

In a second step, PCC-H considers the task difficulty for each fragment of the meeting. The goal is to reduce the effect of some fragments of the meeting, in which there is an uncertainty in the workers judgments, for comparative relevance evaluation, for instance because there are no relevant search results in Wikipedia about the current fragment. To lessen the effect of uncertainty in our judgments, the entropy of answers for each fragment of the meeting is computed and a function of it is used as a weight for each fragment. This weight is used for computing $PRV$. Entropy, weight and $PRV$ are formulated in Eqs. 4, 5 and 6 below.

$$H_q = -\sum_{a=1}^{A} RV_{qa} log(RV_{qa}) \tag{4}$$

$$W_q = 1 - H_q \tag{5}$$

$$PRV_a = \frac{\sum_{q=1}^{Q} W_q RV_{qa}}{\sum_{q=1}^{Q} W_q} \tag{6}$$

In these equations, $A$ is the number of options, while $H_q$ and $W_q$ are the entropy and the weight of fragment $q$ respectively.

# 6. RESULTS OF THE EXPERIMENTS

The experiments were performed in two directions. At first, we attempt to validate the PCC-H method. Then, we apply the PCC-H method to compute $PRV$ for each answer list to conclude which version of the system outperforms the other.

In order to make an initial validation of the workers judgments, we compare the judgments of individual workers with a single expert. For each worker, the number of fragments for which the answer is equal to the expert's answer is counted and divided by the number of fragments, to compute accuracy. Then we compare this value with the $r_w$ which is estimated as the reliability measurement for each worker judgment. The percentage of agreement between each worker vs. the expert $e_w$ and the $r_w$ for each worker for one of the batches is shown in Table 1, showing that there is an agreement between these two values for each worker. In other words, workers who have more similarity with our expert also have more inter-rater agreement with other workers. Since in the general case there is no ground truth (expert) to verify workers judgments, we rely on the inter-rater agreement for the other experiments.

Firstly, equal weights for all the user evaluations and fragments are assigned to compute $PRV$s for two answer lists of our experiments, which are shown in Table 2.

In this approach, it is assumed that all the workers are reliable and all the fragments share the same difficulty. To handle workers' reliability, we suppose that workers with lower $r_w$ are outliers. One approach is to remove all the outliers. For instance, the first four workers with lower $r_w$

**Table 1: Percentage of agreement between a single worker and the expert, and a single worker and the other workers, for the KW system and 4-choice HITs**

| Worker # | $e_w$ | $r_w$ |
|---|---|---|
| 1 | 0.66 | 0.81 |
| 2 | 0.54 | 0.65 |
| 3 | 0.54 | 0.64 |
| 4 | 0.50 | 0.71 |
| 5 | 0.50 | 0.60 |
| 6 | 0.50 | 0.35 |
| 7 | 0.41 | 0.24 |
| 8 | 0.39 | 0.33 |
| 9 | 0.36 | 0.34 |
| 10 | 0.31 | 0.12 |

**Table 2: $PRV$s for AW-vs-UI and KW-vs-UI pairs**

| All workers and fragments with equal weights | | 2-choice HITs | 4-choice HITs |
|---|---|---|---|
| AW-vs-UI | $PRV_{AW}$ | 30% | 26% |
| | $PRV_{UI}$ | 70% | 74% |
| KW-vs-UI | $PRV_{KW}$ | 45% | 35% |
| | $PRV_{UI}$ | 55% | 65% |

are considered outliers and deleted and the same weight is given to the remaining 6 workers. The result of comparative evaluation based on removing outliers is shown in Table 3.

In the computation above, an arbitrary border was defined between outliers and other workers as a decision boundary for removing outliers. However, instead of deleting workers with lower $r_w$, which might still have potentially useful insights or relevance, it is rational to give a weight to all workers' judgments based on a confidence value. The $PRV$ for each answer list of four experiments based on assigning weight $r_w$ to each worker's evaluation, and equal weights to all meeting fragments are shown in Table 4.

In order to show that our method is stable on different HIT designs, we used two different HIT designs for each pair as mentioned in Section 4. We show that $PRV$ converges to the same value for each pair with different HIT designs. As observed in Table 4, $PRV$s of AW-vs-UI pair are not similar for two different HIT designs, although the answer lists are the same. Moreover, it is observed that, in several cases, there is no strong agreement among workers to decide which answer list is more relevant to that meeting fragment. Since the source of uncertainty is undefined, the effect of that fragment is reduced giving a weight to each fragment. This weight represents the difficulty level of assigning $RV_{ql}$. The $PRV$ for all experiments are represented in Table 5. As shown in Table 5 $PRV$s of AW-vs-UI pair are the same after considering task difficulty weights for computing $PRV$.

**Table 3: $PRV$s for AW-vs-UI and KW-vs-UI pairs**

| Six workers and fragments with equal weights | | 2-choice HITs | 4-choice HITs |
|---|---|---|---|
| AW-vs-UI | $PRV_{AW}$ | 24% | 13% |
| | $PRV_{UI}$ | 76% | 86% |
| KW-vs-UI | $PRV_{KW}$ | 46% | 33% |
| | $PRV_{UI}$ | 54% | 67% |

**Table 4: $PRVs$ for AW-vs-UI and KW-vs-UI pairs**

| All workers with different weights and parts with equal weights | | 2 choices HIT design | 4 choices HIT design |
|---|---|---|---|
| AW-vs-UI | $PRV_{AW}$ | 24% | 18% |
| | $PRV_{UI}$ | 76% | 82% |
| KW-vs-UI | $PRV_{KW}$ | 33% | 34% |
| | $PRV_{UI}$ | 67% | 66% |

**Table 5: $PRVs$ for AW-vs-UI and KW-vs-UI pairs**

| All workers with different weights and fragments with different weights (PCC-H method) | | 2-choice HITs | 4-choice HITs |
|---|---|---|---|
| AW-vs-UI | $PRV_{AW}$ | 19% | 15% |
| | $PRV_{UI}$ | 81% | 85% |
| KW-vs-UI | $PRV_{KW}$ | 23% | 26% |
| | $PRV_{UI}$ | 77% | 74% |

Moreover, we compare PCC-H method with the GLAD (Generative model of Labels, Abilities, and Difficulties) method [16] for estimating comparative relevance value through considering task difficulty and worker reliability parameters. We run GLAD algorithm with the same initial values for all four experiments. The $PRV$s which are computed by the GLAD method and the PCC-H method are shown in Table 6.

Based on Table 6, $PRV$s which are computed by the PCC-H method for both HIT designs are very close to those of GLAD for the 4-choice HIT design. This means that PCC-H method is able to calculate the $PRV$s independent of the HIT design.

The proposed method is also applied for comparative evaluation of SS-vs-KW search results. The $PRVs$ are calculated by three different methods as shown in Table 7. The first method considers all the workers and fragments with the same weight. The second method assigns weights computed by PCC-H method to measure $PRVs$, the third one is the GLAD method. Therefore the SS version outperforms the

**Table 6: $PRV$s computed by GLAD and PCC-H methods**

| Methods pairs | | (GLAD,PCC-H) | |
|---|---|---|---|
| | | 2-choice HIT design | 4-choice HIT design |
| AW-vs-UI | $PRV_{AW}$ | (23%,19%) | (13%,15%) |
| | $PRV_{UI}$ | (77%,81%) | (87%,85%) |
| KW-vs-UI | $PRV_{KW}$ | (47%,23%) | (23%,26%) |
| | $PRV_{UI}$ | (53%,77%) | (77%,74%) |

**Table 7: $PRV$s for SS-vs-KW**

| Methods pairs | | (Equal Weight, PCC-H,GLAD) |
|---|---|---|
| | | 4-choice HIT design |
| SS-vs-KW | $PRV_{SS}$ | (0.88,0.93,0.88) |
| | $PRV_{KW}$ | (0.12,0.07,0.12) |

KW version according to all three scores.

## 7. CONCLUSION AND PERSPECTIVES

In all the evaluation steps, the UI system appeared to produce more relevant recommendations than AW or KW. Using KW instead of AW improved $PRV$ by 10 percent. This means that using UI, i.e. when users ask explicit queries in conversation, improves over AW or KW versions, i.e. with spontaneous recommendations. Nevertheless, KW can be used as an assistant which suggests documents based on the context of the meeting along with the UI version, that is, both spontaneous and user-initiated recommendations can be made. Moreover, SS version works better than KW version, which shows an advantage of semantic search.

As for the evaluation method, PCC-H outperformed the GLAD method (which had been earlier suggested to estimate task difficulty and reliability of workers for labeling task in the lack of ground truth). Based on the evaluation results, the PCC-H method provided more stable $PRV$ with different HIT designs.

There are some instances in which the search results of both versions are irrelevant. The goal of future work will be to reduce the number of such uncertain instances, to deal with ambiguous questions, and to improve the processing of user-directed queries by recognizing the context of the conversation. Another experiment should improve the design of simulated user queries, in order to make them more realistic.

## 8. REFERENCES

[1] O. Alonso and R. A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *ECIR*, volume 6611, pages 153–164, 2011.

[2] O. Alonso and M. Lease. Crowdsourcing 101: Putting the "wisdom of the crowd" to work for you. WSDM Tutorial, 2011.

[3] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42:9–15, 2008.

[4] J. Carletta. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254, 1996.

[5] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[6] G. Chittaranjan, O. Aran, and D. Gatica-Perez. Exploiting observers' judgements for nonverbal group interaction analysis. In *IEEE Conference on Automatic Face and Gesture Recognition*, 2011.

[7] P. N. Garner, J. Dines, T. Hain, A. El Hannani, M. Karafiat, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang. Real-time ASR from meetings. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, pages 2119–2122, Brighton, UK, 2009.

[8] C. Grady and M. Lease. Crowdsourcing document relevance assessment with mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 172–179, 2010.

[9] D. R. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using lowrank matrix approximations. In *Proceeding of the Allerton Conf. on Commun., Control and Computing*, 2011.

[10] G. Kazai. In search of quality in crowdsourcing for search engine evaluation. In *ECIR*, volume 6611, pages 165–176, 2011.

[11] F. K. Khattak and A. Salleb-Aouissi. Quality control of crowd labeling through expert evaluation. In *Second Workshop on Computational Social Science and the Wisdom of Crowds (NIPS)*, 2011.

[12] J. Le, A. Edmonds, V. Hester, L. Biewald, V. Street, and S. Francisco. Ensuring quality in crowdsourced search relevance evaluation : The effects of training question distribution. In *Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation*, pages 17–20, 2010.

[13] A. Popescu-Belis, E. Boertjes, J. Kilgour, P. Poller, S. Castronovo, T. Wilson, A. Jaimes, and J. Carletta. The AMIDA automatic content linking device: Just-in-time document retrieval in meetings. In *Machine Learning for Multimodal Interaction V (Revised selected papers from MLMI 2007, Brno)*, pages 272–283, Berlin/Heidelberg, 2008. LNCS 5237, Springer-Verlag.

[14] A. Popescu-Belis, M. Yazdani, A. Nanchen, and P. Garner. A speech-based just-in-time retrieval system using semantic search. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT 2011), Systems Demonstrations*, pages 80–85, Portland, OR, 2011.

[15] P. Smyth, U. M. Fayyad, M. C. Burl, P. Perona, and P. Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems*, pages 1085–1092, 1994.

[16] J. Whitehill, P. Ruvolo, T.-F. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, pages 2035–2043. 2009.