The Thesis Committee for Haofeng Zhou
certifies that this is the approved version of the following thesis:

Crowdsourcing Construction of Information Retrieval Test Collections for Conversational Speech

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Matthew Lease

Byron Wallace

Crowdsourcing Construction of Information Retrieval Test Collections for Conversational Speech

by

Haofeng Zhou, B.E.; M.S.; Ph.D.

Thesis

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Information Studies

The University of Texas at Austin

May 2015

To My Family!

Acknowledgments

After working for ten years, I never imagined I would be given an opportunity to broaden my knowledge. Even though it was one of the biggest challenges I ever met, this thesis is the realization of my work.

I am grateful to Dr. Matthew Lease for what I have learned from him. I am lucky to have Dr. Lease as my advisor and it has been my honor to work in his lab. His support, guidance, and motivation is what made this thesis possible. His great passion for research makes me reflect on what I have, and inspires me to take on bigger challenges.

I appreciate the valuable feedback of Dr. Byron Wallace on this thesis.

I would like to thank Edward Banner for his work to generate the best transcripts.

I would also like to thank Ivan Oropeza, Hyunjoon Jung, Aashish Sheshadri and Donna Vakharia for their support and comments as well as the joyful experience while we were together.

Finally, I thank my family and friends who have been there for me. This thesis is not only my milestone, but also evidence of your unconditional love and support.

HAOFENG ZHOU

The University of Texas at Austin
May 2015

Abstract

Crowdsourcing Construction of Information Retrieval Test Collections for Conversational Speech

Haofeng Zhou, MSINfoStds
The University of Texas at Austin, 2015

Supervisor: Matthew Lease

Building a test collection for an ad hoc information retrieval system on conversational speech raises new challenges for researchers. Traditional methods for building test collections are costly, and thus they are not feasible to apply to large scale conversational speech data. Constructing a large test collection on conversational speech with high quality at low cost is challenging. Crowdsourcing may represent a promising approach. Crowd workers tend to be less expensive than professional assessors, and crowd workers can work simultaneously to perform jobs on a large scale. However, despite the benefits of scale and cost, the quality of the results delivered by crowd workers may suffer. This thesis focuses on relevance judging, one of the key components of a test collection. We adopt two crowdsourcing platforms: oDesk and MTurk, use audio clips and various versions of transcripts, conduct multiple experiments under diverse settings, and analyze the results qualitatively and quantitatively. We delve into what factors influence the quality of relevance judgments on conversational speech. We also investigate differences between relevance judgements from experts and crowd workers. This thesis also describes best practices for the design of crowdsourcing tasks to improve crowd workers' performance. Ultimately, these may assist

researchers in building high-quality test collections on conversational speech at low cost and scale through crowdsourcing.

Table of Contents

Chapter 1: Introduction

In the past decades, information institutions such as museums, archives and libraries, as well as other organizations and individuals, have accumulated huge volumes of interviews, lectures, and speeches that were recorded on tapes, cassettes, or other formats. These information institutions maintain a large number of audio collections, with great potential for research and educational purposes. They may also intend to make it feasible for the public and professionals (users) to access these materials irrespective of where they are or when they need it. All in all, they need an information retrieval (IR) system to represent, organize and store the collections to meet users' information needs.

Building IR systems on text-based data has a long history, and researchers have made great achievements in this area. Generally, one of the key tasks is to build a test collection, including a document collection, search queries, and a set of relevance judgments, to train and evaluate the efficiency and effectiveness of an IR system. Both academic and industrial institutions have built many such test collections, e.g., TREC and CLEF, for comparing IR models, algorithms, and techniques. These test collections are of high quality, and some of them are large enough to support the requirements of modern information retrieval tasks such as web searching.

Yet, building a search engine for audio collections, as well as the corresponding test collections, remains an open challenge. Some techniques developed for traditional IR tasks are not feasible to apply to such data directly. For example, algorithms for textual data cannot deal with audio data directly. Usually, the audio data is converted into text format via transcription with additional efforts and cost. For example, among the methods of transcription, professional transcription can provide transcripts of high quality, even though the raw audio is full of "*disfluencies, heavy accents, age-related coarticulations, un-cued speaker and language switching and emotional speech*".[1]

[1] `http://catalog.ldc.upenn.edu/LDC2012S05`

However, the endeavor to recruit experts is so costly that it limits the scalability of processing a large amount of audio data to build a test collection. For example, Oard et al. [29] mentioned that, on a large collection of interviews, they spent $2,000 per interview for interviewing, digitizing and data entry, as well as another $2,000 per interview for indexing to built a test collection.

To reduce the cost of building test collections, researchers are exploring automated and human-computational methods beyond professional work in multiple areas to create large test collections. For example, researchers have developed diverse systems and algorithms for Automatic Speech Recognition (ASR) to automate the process of transcription, as well as other methods, such as detection of sentence boundary and dysfluency, to solve problems newly introduced by ASR. ASR techniques use methods from the fields of statistics and artificial intelligence. They are usually cheaper than professional transcription. However, the use of ASR raises several issues. First, the accuracy of the ASR transcripts is often lower than that of professional transcription because computers remain less accurate than people in natural language processing. Second, more transcription data of high quality is needed to build accurate models in ASR systems. It is also possible that an ASR model built on a particular dataset might be more or less effective on a new one from a different domain. The higher the accuracy and generality an ASR system requires, the greater the amount of high quality data needed to train the system increasing effort and cost. Additionally, ASR transcripts usually lack sentence boundaries and punctuation, which are expected by readers.

Emerging crowdsourcing methods provide an alternative approach for building a test collection on a large scale, especially since Amazon's Mechanical Turk (MTurk) was launched in 2005. Many people around the world are now employed or contracted in the crowdsourcing ecosystem. Their cost tends to be much lower than that of traditional professionals, and they can work simultaneously to perform jobs on a large scale. Yet the skills of crowd workers vary dramatically: some rival experts, while others perform poorly. Therefore, despite the benefits of scale and cost, the quality of the result delivered

by crowd workers, e.g., accuracy, may suffer. Those developing systems for transcription and building a test collection with crowdsourcing seek to match expert-level quality while simultaneously realizing cost savings and greater scalability.

In brief, the diversity of automated and human-computational methods mentioned above potentially introduces some new and poorly understood influences in the process of building audio test collections, thus making the process more challenging on quality than ever.

In this thesis, we will briefly introduce the crowdsourcing platforms involved in the next section, followed by a review of related work from multiple aspects, such as information retrieval and crowdsourcing. Then, we will propose the research questions as well as some hypotheses to be tested.

In Chapter 2, we introduce the preparation of data and platform we will use in this thesis, and summarize pilot experiments on collecting crowd relevance judgements. In Chapter 3, we investigate the factors that influence the quality of relevant judgements, and test the hypotheses mentioned above. Chapter 4 continues the investigation from a different perspective. Chapter 5, the last chapter, will summarize our findings, as well as best practices, and discuss the potential steps for future work.

1.1 CROWDSOURCING PLATFORMS

Two crowdsourcing platforms, MTurk and oDesk, are used in the this thesis.

Amazons Mechanical Turk (MTurk)

MTurk [4, 26] is a well-known crowdsourcing platform in both research and industry, and it remains one of the most prominent paid crowd work [20] platforms today. *Requesters* can post Human Intelligence Tasks (HITs) on MTurk for a specified payment, and the crowd workers on MTurk (turkers), can accept them and do the work. Once the turkers submit the results, the requesters can approve or reject them. The requesters only need to pay for the approved results, and they can award bonuses to turkers. MTurk charges 10% as the service

fee. MTurk is more suitable for small and simple tasks than complex or collaborative ones.

Though it has many limitations, MTurk has garnered nearly exclusive focus in research on crowd work that many researchers now struggle with those limitations in order to conduct their own research [1, 12]. One of the limitations complained about by requesters is MTurk lacks effective support for quality control. At the other end, turkers are unsatisfied with the low payment and high rejection rate from requesters [17].

MTurk exposes API interfaces based on Amazons Web Services, and all function calls are through either SOAP or RESTful services. Though it provides APIs in several programming languages on its web site, all are third-party packages instead of official ones, and Amazon is not responsible to update them.

oDesk

oDesk (www.odesk.com) provides a general purpose online labor marketplace whose focus on specialized and higher-skilled forms of labor (freelancers), distinguishes it from relatively unskilled work often posted to MTurk. It is more like a real marketplace, and requires more management effort, such as recruiting freelancers and building trust with them. Requesters can post jobs, and freelancers can bid on them. Requesters can then review the applications and negotiate with the qualified freelancers for the final contracts. Once contracts are accepted, freelancers can optionally work under a monitoring "Work Diary" system, and update the status of their work. Requesters can also set up several milestones in the whole process so that the they can control the progress and quality efficiently, and terminate the contract if necessary. When contracts end, requesters and freelancers evaluate each other with rating and feedback. oDesk collects 10% of the total payment. oDesk is more suitable for large and complex projects such as building a website, designing a mobile app, or collecting market information for business strategy purposes[2].

[2] `http://thenextweb.com/entrepreneur/2013/10/26/freelancing-new-normal-odesk/`

1.2 RELATED WORK

Having IR test collections for speech data is key to being able to evaluate alternative search algorithms, and thereby foster research in developing search algorithms in this domain. There are two types of speech data, conversational/spontaneous speech and prepared/released speech. The TREC Spoken Document Retrieval (SDR) Track is one of the well-known data set for prepared/released speech, and Garofolo et al. [13] mentioned that building test collections on such data "is a solved problem". A well-known data set for conversational/spontaneous speech is the 2005 CLEF[3] Cross-Language Speech Retrieval (CL-SR)[4] track [36].

In this thesis, we focus more on the conversational/spontaneous speech.

Many researchers are working in this area. For example, Oard et al. [29] used search-guided relevance assessment to create a test collection which they believed was the first realistic information retrieval test collection for spontaneous conversational speech. This combined both manual efforts and automatic techniques. They recruited four assessors to identify topically relevant segments for 28 topics developed from actual user requests. They found that search-guided assessment resulted in strong inter-annotator agreement to support formative evaluation during system development. Marge et al. [24] worked to process meeting speech with MTurk, and achieved high-quality results and found that MTurk was very useful for summarization systems.

Quality is one of the major concerns in crowdsourcing for audio/speech data. The main concern is whether the lower cost comes at the expense of the final quality of test collections or data sets. Many effective strategies, methods and techniques have been proposed to address this. For example, Novotney et al. [28] introduced an ASR system trained on MTurk transcriptions that achieves similar performance as one trained on expert transcrip-

[3]CLEF Initiative (Conference and Labs of the Evaluation Forum, formerly known as Cross-Language Evaluation Forum), available at `http://www.clef-initiative.eu`

[4]CL-SR is a CLEF track to evaluate search systems on noisy automatically generated transcripts of spoken documents. Speech was transcribed by ASR or manually. The transcripts were then used for relevance judgement and information retrieval experiments. More details see `http://www.clef-initiative.eu/track/clsr`

tions, but at only a fraction of the cost. They proposed a procedure to maintain the quality of the annotator pool without high quality annotation, and found that higher disagreement among turkers did not have a significant effect on the performance. They suggested that the more resources allocated to data size, the better the performance was. Marge et al. [25] described transcription of audio consisting of route instructions of robots through MTurk. They used ROVER [11] to merge multiple noisy transcriptions and achieved a good result. They also indicated that quality was less influenced by payment than expected.

Aside from common methods of quality control, such as comparing with a gold standard and being evaluated by experts, some researchers have explored alternative methods as well. Parent et al. [31] proposed "self-confidence as a quality control", where they asked the workers to check a checkbox containing the judgement "I'm not 100% confident of my answer" for any transcription they were not sure of, and found that such self-assessment could be "used as a first filter to improve the quality of the data". Gruenstein et al. [14] introduced a filtering technique their work of collecting orthographically transcribed continuous speech data from an online educational game called *Voice Scatter*. That technique automatically identified a subset of transcribed utterances obtained from MTurk, which improved the quality to a level of "human-quality". Audhkhasi et al. [3] considered a scenario with no reference transcripts as a gold standard. They proposed new metrics to evaluate the reliability of the noisy transcripts collected from MTurk when combining them and the result was also encouraging. Evanini et al. [10] adopted a merging strategy to process the results from MTurk for non-native speech, and the final quality (accuracy) was comparable to those obtained from expertise. Williams et al. [38] discussed three methods in crowd transcription: incremental redundancy, treating ASR as a transcriber, and using a regression model. These methods allow them to explicitly balance among precision, recall, and cost.

Additionally, to improve and guarantee quality, some researchers have proposed a multi-stage strategy with decomposition. Noronha et al. [27] developed *PlateMate*, a system that allows users to take photos of their meals and receive estimates of food intake

and composition. A 3-stage process of "Tag-Identify-Measure" along a series of HITs was used in *PlateMate*. The results showed that that system was nearly as accurate as a trained dietitian and more accurate than self-reporting. Lee et al. [22] proposed a two-stage model that automatically controlled the quality at each stage using different methods: a Support Vector Machine (SVM) classifier at the first stage, and word level confidence scores at the second stage. This delivered a high-quality with little-effort. Parent et al. [31] also used a two-stage approach to transcribe one year of data. ASR transcripts were assessed in regard to being understandable and correct in the first stage, and only those recordings which were completely intelligible by ASR were sent to the second stage for transcription. When "gold-standard" quality control was used at double cost, they achieved results close to NIST expert agreement. Another notable work is *Soylent* by Bernstein et al. [5]. They integrated crowdsourcing of word processing for three endeavors: writing, editing, and macro programming. They proposed the pattern of "Find-Fix-Verify" to refine the results gradually. Additionally, they pointed out that the task size and the complexity influenced the quality, and the best strategy was to make the task as small as possible. Chen et al. [7] expressed a similar point of view in their work when they collected multilingual data summarizing simple and short videos.

Another aspect of quality control is how to block cheaters results efficiently and effectively. Cheaters, spammers or malicious workers, are crowd workers primarily interested in producing quick generic answers rather than correct ones in order to optimise their time-efficiency to earn more money. Eickhoff et al. [9] investigated this issue on several crowdsourcing platforms, explored multiple methods to identify cheaters, and described patterns in how cheaters cheat. Difallah et al. [8] analyzed cheaters' behaviors and approaches, especially when they were organized, and highlighted the insufficiency of existing techniques of cheaters detection.

Beyond MTurk, our own prior work investigated other crowd transcription platforms [40]. We conducted a qualitative assessment on 1-888-Type-It-Up, Transcription Hub, CastingWords, 3Play Media, Rev, TranscribeMe, Quicktate, SpeakerText, and a quan-

titative assessment on the first three as well as oDesk. We further discussed the tradeoffs among quality, cost, risk, and effort in such alternative options.

ASR is not a perfect solution. It not only can make large numbers of errors in the recognition process, but it also delivers an output without sentence boundaries. Both of these make ASR transcripts hard to read. Byrne et al. [6] used ASR in their project to transcribe automatically the USC-SFI MALACH collection of oral history interview of Holocaust survivors. They remarked that the audio was "heavily accented, emotional and elderly spontaneous speech". From 2004 to 2006, the researchers applied several ASR algorithms, and finally achieved 25% mean WER (Word Error Rate) [30]. However, as can be seen in the MALACH dataset[5], there is no punctuation in the ASR text, which makes it difficult to read and understand. Hence, more approaches and algorithms were proposed to handle these problems. For example, to address the problem of dysfluency detection, Conditional Random Fields (CRFs), first proposed by Lafferty et al. [21], were used by Liu et al. [23] to detect edited words as well as to identify the sentence boundary. Max-Margin Markov Networks (M^3Ns), proposed by Taskar et al. [34] have been used as well. Qian and Liu [32] further combined CRF and M^3N in their new multi-step stacked learning method to detect dysfluency, and achieving F1 score of 0.841.

Additionally, understandability or readability of transcripts is also important. Jones et al. [18] proposed a framework to measure the readability of automatic speech-to-text (STT) transcripts, in which they quantified the readability of a text in terms of four factors: the accuracy of the answers to the related questions, time required to read the text and questions, time taken to answer the questions, and a subjective score of the difficulties of the text. They found that the reference text is "more readable for human readers because disfluencies that made STT hard to read have been removed".

In HCI, Kittur et al. [19] investigated the utility of MTurk for collecting user measurements, and discussed considerations in the design of user evaluation tasks. After comparing two designs and their effectiveness, the authors pointed out that "special care is

[5] https://catalog.ldc.upenn.edu/LDC2012S05

needed in formulating tasks in order to harness the capabilities of the approach". For example, explicitly verifiable questions should be included, "the on-obvious random or malicious completion" could be prevented through effort balance, and multiple methods should be used to detect suspect responses.

In the area of relevance judgement, we are interested in a user study by Huang et al. [16] about how the assessors make relevance judgements. They discussed the problem of topical relevance and identified 5 types of topical relevance, i.e. direct relevance, indirect relevance, context relevance, relevance by comparison, and relevance as a pointer. They explored the roles of these types in judgement about overall topical relevance, and also pointed out that these types may overlap, e.g., direct and indirect, indirect and context, and direct and context. They concluded "Topical relevance is situational; it is a combination of factors with different weights according to different preferences or user situations."

1.3 RESEARCH QUESTIONS

The purpose of this thesis is to study how to build a large test collection of high quality for audio data with low cost. In other words, this thesis is interested in exploring methodologies to build a test collection on conversational speech at scale using crowdsourcing. In particular, we focus on the aspect of collecting the data of good relevance judgements from crowd workers.

The following research questions will be considered in this thesis:

- RQ_1: What factors influence the quality of relevance judgments on conversational speech data?
- RQ_2: What difference in behavior of relevance judging is observed between traditional assessors and crowd workers? Why?
- RQ_3: What are best practices to design tasks in a way that promotes and enables quality work by crowd workers?

Before we start our work, we also have several hypotheses.

- HA: An audio clip may be better than the transcript for crowd workers to perform relevance judgement.
- HB: Freelancers from oDesk will outperform turkers from MTurk in relevance judging at the same cost.
- HC: The better quality of the transcripts or audio clips, the easier to understand, and better quality of relevance judgement will result.
- HD: Narratives are more informative than descriptions of topics and thus can yield better relevance judgements. (Figure 3.1 and 4.2 show 2 samples of narratives and descriptions of topics.)

This thesis has the following contributions.

- We identify the main issues that block the crowd workers delivering good relevance judgements for conversational speech.
- We describe diverse aspects of crowd worker behavior in performing transcription and relevance judging.
- We propose a practical methodology and framework to support building a high-quality test collection for conversational speech at scale at low cost.

Chapter 2: Fundamental Work

In this chapter, we will review what we have done prior to work on this thesis. We will discuss the raw data we are using and how we preprocess the data. We will summarize the experience and lessons learned in the pilot work as well. Several technical issues will also be discussed.

2.1 Data Preparation

In this section, we introduce the data used in this thesis, and how we pre-process it.

2.1.1 Data Sources

The raw data to be used derives from three sources:

- The USC-SFI MALACH Interviews and Transcripts English (LDC2012S05). This dataset contains audio files and some transcripts created manually by professional transcribers. However, it is a subset of the whole MALACH project[1]. Some audio files as well as some transcripts are missing. Furthermore, its audio files and transcripts are not a one-to-one mapping, i.e. some audio files have no corresponding transcripts, and vice-versa.

- Audio of additional interviews provided by MALACH project personnel. This makes our audio files more complete, but we still lack transcripts for many audio files.

- The CLEF 2005 Cl-SR Test Collection (CLEF2005CL-SR). It contains relevance judgments *qrels* based on segments of the transcripts in MALACH project, it also contains the information needs and queries in terms of topics on which the relevance judgment is based. The segments in CLEF2005CL-SR were derived from the

[1] `http://malach.umiacs.umd.edu/`

output of ASR instead of professional transcription, and there is no alignment information provided with the original audio files. Additionally, the audio files covered by CLEF2005CL-SR have minimal overlap with LDC2012S05, which led us to turn to the MALACH project audio mentioned above.

2.1.2 Data Selection

We mapped the segments in CLEF2005CL-SR to the interviews in LDC2012S05, then transformed the relevance judgements *qrels* in CLEF2005CL-SR from segment-topic pairs into interview-topic pairs. In other words, we obtained a complete list of relevant segments for a given topic in a particular interview, and we could further identify the interviews that provided the most relevant and non-relevant information. Finally, we chose the interviews of #00058 and #25891 for this thesis. Thus for, we have used only #00058 in our experiments, which consists of four 30-minute audio files, among which only the first one has a gold transcript provided by LDC2012S05, while the others were not transcribed yet before we started our work.

Interview #00058 was sent to freelancers recruited through oDesk for transcription. The freelancers had achieved around 10% WER (Word Error Rate), which is very close to that of the experts. Using several quality control approaches, such as a screening test, cross-checking, and gold-standard verification, we are confident about the quality of these transcripts. Thus we treated these transcripts as equivalent in quality to those in LDC2012S05. It should be noted that the transcription is file-based, i.e. if the freelancers were given a complete 30-minute audio file, what they returned would also a transcript file covering the entire 30-minute audio.

Interview #00058 covers 13 segments. We dropped the segments with less relevant and non-relevant topics, and only chose the top-2 segments with the most relevant topics, as summarized in Table 2.1.

These transcripts have a few time tags at the switches of speakers that help us to align them to the corresponding audio clips.

Segment	abbr.	Number of Relevant Topics	Number of Non-Relevant Topics
VHF00058-067118.004	V004	2	2
VHF00058-067132.005	V005	2	2
VHF00058-067153.007	V007	3	8
VHF00058-067222.012	V012	2	4
VHF00058-067232.013	V013	3	6

Table 2.1: Interview #00058 and its segments for experiments

2.1.3 Data Preprocessing

We aligned the ASR segments in CLEF2005CL-SR to the transcripts from oDesk and LDC2012S05. Though there were still some errors introduced by both ASR and manual transcripts, the results of the alignment provided rough offsets of the starting and ending points of each ASR segment in the file-based transcripts. Most ASR segments could be located within a certain transcript file, but 1 segment, as underlined in Table 2.1, crosses two consecutive files. We then identified the exact offsets for the ASR segments in oDesk's transcripts manually, and extracted the corresponding segments of transcripts from those file-based transcripts. We refer to these segments as the raw version of transcripts, or "raw transcripts" in short. The corresponding ASR segments are called "ASR transcripts".

We also performed similar work on the audio files. It took more effort because we needed to listen to the audio carefully around the cutting point. We first roughly aligned the audio files with the transcripts, then located the exact time offsets in the audio files for the segments we planned to use in our experiments of relevance judging, and finally extracted the corresponding audio segments, i.e. audio clips.

ASR transcripts are usually hard to read. We hold the hypothesis that the better quality of the transcripts or audio clips, the easier to understand, and better quality of relevance judgement will result. Thus, we believe that *the ASR transcripts from CLEF2005CL-SR will lead to low accuracy of relevance judgement because they are lower quality*. Similarly, according to the same hypothesis, we hope to make the transcript as easy to read as

Question	Segment	Topic IDs (Underlined are relevant)
Q00058-001-S	VHF00058-067153.007	2224, 1288, 3013, 3004
Q00058-002-S		1877, 1225, 1288, 3014
Q00058-003-M		2012, 1551, 2198, 1897
Q00058-004-S	VHF00058-067232.013	1508, 1551, 15602, 2012
Q00058-005-S		1508, 1225, 3027, 1288
Q00058-006-M		3027, 1429, 1311, 1225
Q00058-007-S	VHF00058-067118.004	1166, 1330, 2012, 1551
Q00058-008-S		2012, 3005, 3015, 1225
Q00058-009-M		1166, 3015, 3005, 1345
Q00058-010-S	VHF00058-067132.005	1746, 1225, 1624, 2384
Q00058-011-S		3005, 3015, 1897, 3026
Q00058-012-M		1225, 3015, 2384, 3005
Q00058-013-S	VHF00058-067222.012	2012, 1508, 1225, 1620
Q00058-014-S		3016, 1620, 3026, 1508
Q00058-015-M		1620, 2012, 1508, 3026

Table 2.2: Questions and topic distribution

possible. We have two ideas. One is to provide "true verbatim" transcripts of the audio with the expectation that crowd workers will collect everything they need from the text to make relevance judgements; the other is to rephrase the existing transcripts to make them more meaningful, e.g., removing some dysfluencies and filler words. We thought the meaningful transcript would be more easier for the crowd workers. Thus, we asked a native speaker to refine the segments of #00058 with references to the raw transcripts and the audio clips. It took about 1.55 hours on average for him to refine one 3-minute segment. We call such meaningful transcripts the best version of transcription, or the "best transcripts" in short.

Therefore, we have following data sets which are aligned with the segments in CLEF2005CL-SR: the **audio clips**, the **ASR transcripts** from CLEF2005CL-SR, the **raw transcripts** from LDC2012S05 and oDesks transcription, and the **best transcripts**.

2.2 PILOT ON MTURK

2.2.1 HIT Design

From *qrels* in CLEF2005CL-SR, we know which topics are relevant and non-relevant to a given segment. We randomly organized these topics in 3 question forms for that segment, 2 of which are in the form of single-choice (1 of 4 topics is relevant), and the remaining one is a multiple-choice (2 of 4 topics are relevant). Thus there are 15 question forms in total for all 5 segments. Table 2.2 shows the arrangement of such topics and question forms. We did not inform the crowd workers whether a given task had one or more correct answers.

A sample of the HIT design of the question form for an audio clip is shown in Figure 2.1. In this design of HIT, the task asked the turkers to judge 4 topics with 3 options: Relevant, Non-relevant and Not Sure. The option of "Not Sure" would make crowd workers feel relaxed from the situation of unable to determine the exact one. In the analysis phase, we treated "Not Sure" as the same as "Non-relevant". Additionally, we adopted a high-contrast color scheme so that it would be easier to locate the topics and options.

2.2.2 Experiments on MTurk

We organized the HITs in batches. Initially, we posted the HITs with 5 assignments per HIT. We had the intention that all assignments could be eventually accepted and correctly answered by turkers. When all 15 HITs expired and some assignments were either never taken by any turkers or their submissions were rejected, these assignments would be posted again within new HITs. These new HITs would automatically have a higher amount of reward. A batch would be terminated when all assignments were validly answered, or the amount of the reward had reached a threshold. This mechanism is called "reposting", and its implementation will be discussed in Section 2.3.

We set 3 measurements to assess the performance in these experiments: participant rate (PR), effectiveness, and accuracy. The participant rate is defined as the ratio of the number of submissions to the total number of available assignments at the beginning of

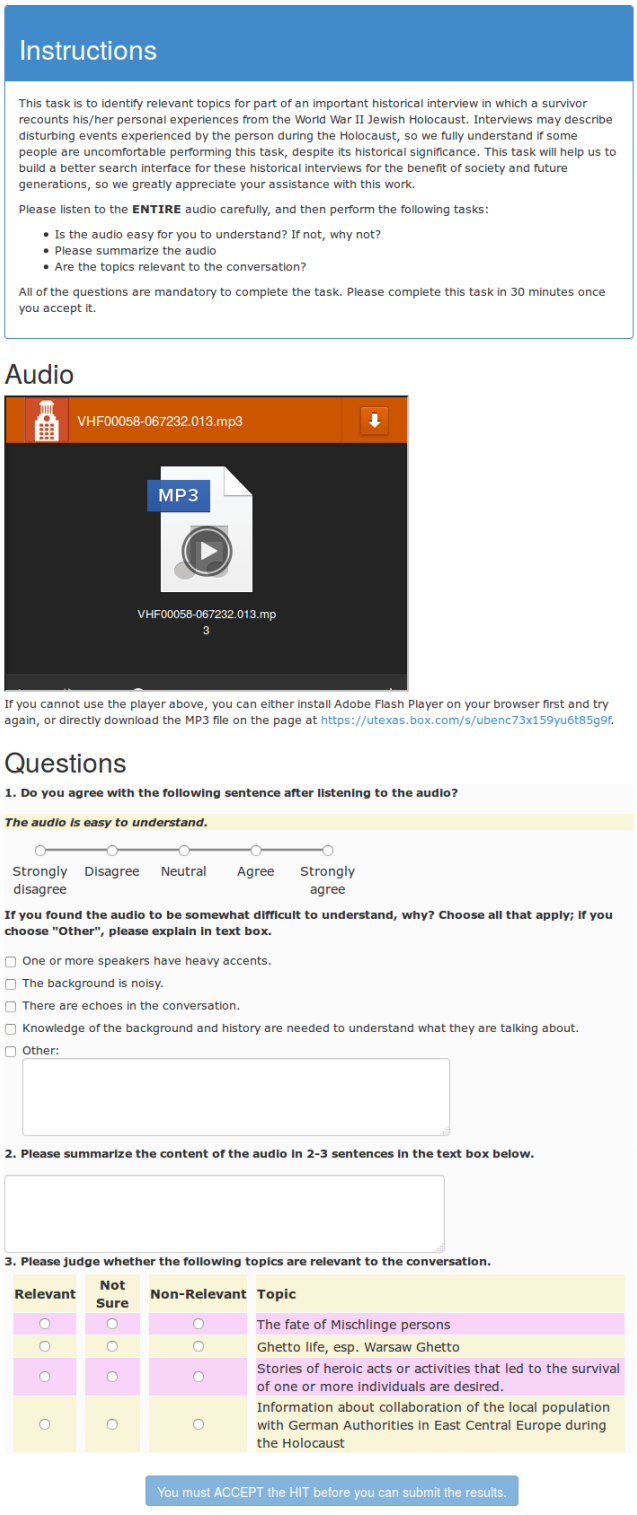


Figure 2.1: A sample of the HIT design of the question form for an audio clip

each round for a batch. Theoretically, we expect that PR in a batch can reach 100% after publishing and reposting several rounds of HITs. The effectiveness is defined as the ratio of the number of accepted submissions to the number of total ones. A submission is rejected only if it is from a possible cheater, and nothing to do with whether it is correct or not. The higher the effectiveness is, the lower number of possible cheaters involved. The accuracy is the ratio of the number of correct submissions to the number of accepted ones, and it does not take the rejected submission into account. High PR does not imply high accuracy, but high accuracy with high PR can give the requesters more confidence in the quality of the results. It should be pointed out that a submission is judged to be correct if and only if all its answers to the questions of the relevance judgment are correct.

We published 3 batches of HITs of relevance judgement for the audio clips. The batches of #1 and #2 were consecutive, and there was no verification question involved, the payment was increased from $0.01 to $0.12 per assignment eventually. A verification question was added in the batch of #3 for each HIT to prevent poor results, and its payment varied from $0.03 to $0.09 per assignment with the increment of $0.03 per round. The increase in the payment in all 3 batches was intended to attract more workers to increase the PR.

2.2.3 Lessons Learned

Table A.1 in Appendix A lists the results for reference. Briefly, we have following observations and considerations:

- It is obvious that the values of PR and effectiveness in the best transcripts are greater than those in audio clips, which implies that textual content seems to attract more workers. In other words, the best transcripts may be more popular than those audio clips for relevance judgment on MTurk.

- There is no significant difference between the values of accuracy in both tables. The accuracy is only around 20% on average, except for the batch of audio clips with ver-

ification questions that has only the accuracy of 11%. Furthermore, we did not find any evidence that audio clips perform better than transcripts in relevance judgement on MTurk or vice versa.

- In batches for audio clips, we found from round 1 to 3 in each batch, the PRs first go down and then rise up. We called it as a pattern of "high-low-higher". However, this pattern does not appear in the best transcripts. The reason is likely that we put the 2nd round of reposting on hold for 2 days after the 1st round of HITs expired. It suggests that there might be a "cool down" phase for a HIT to be reposted, and it might not be a good practice to repost a HIT without any waiting period once it expires.

- To identify cheaters accurately in the process was not an easy job. We took some heuristic criteria to reject a submission and show the turker a clear justification: 1) the response time of the submission for a turker was too short to listen to the entire audio or to read the complete transcript, 2) all questions of relevance judgement were answered with same options, or 3) some mandatory fields were skipped.

- Turkers accepted less payment to do the HITs of relevance judgement for transcripts than those for audio clips. However, we should further include the cost spent on transcription and refinement. Explicitly, we paid $25 for 1-hour audio-length in transcription on average that implies $1.25 per 3-minute segment. Each segment is associated with 3 HITs with 15 assignments available to which such transcription cost can be apportioned, i.e. each assignment additionally has an implicit cost of $0.08 or above. Thus, we actually paid more for the HITs for transcripts than those for audio clips even if we excluded the cost and effort on refinement phase. It seems the transcripts might not have separated value once produced.

Therefore, in the light of these experiments of relevance judgement, it seems that neither audio clips nor the best transcripts were feasible for relevance judging on MTurk. Yet, we cannot not simply draw the conclusion that using crowdsourcing to build a test

collection on conversational speech is impossible. The lessons we learned prompt us to reconsider the HIT design, the strategy to attract good crowd workers, as well as the potential measurement issues, which we will address in the following chapters.

2.3 TECHNICAL SUPPORT

We hope our work is reusable in the future and involves less manual management in the whole process. Thus, we developed an automation framework to support following points:

1. **Batch management**: Once we design a template of HITs, and prepare the data, this framework can automatically manage these data as well as synchronize and track the progress of the HITs on MTurk. The HITs are organized in batches. A batch can contain several HITs with different questions of relevance judgement. Due to the limitation of the service interface exposed by MTurk, there is no way to manage the HITs in batch at MTurk side, and it should be the responsibility of the framework.

2. **Automatic publishing**: The framework supports to publish HITs automatically in some scenarios with a specified strategy, e.g., the mechanism of "reposting" with increased payment when HITs expire.

3. **Automatic blocking**: Because our data covers only 5 segments of interview of #00058, which will be re-used in different batches, it is a higher probability for a turker to encounter duplicated HITs in different batches. A turker is expected to do more HITs within one batch. Once a turker completes a HIT for a certain type of question of relevance judgment, we do not want him/her to do the same type of question in other batches. One solution is *hard-block*, i.e. to block a turker for all other HITs with MTurk's API, which is not acceptable in our project because it also blocks the turker from conducting other questions in the same batch. Another solution is *soft-block*, which is implemented by a friendly reminder in the HIT interface that tells the turkers there will be no pay if they are found to work on the same type of

questions. But this option will discourage the turkers, and it is not their responsibilities to remember what they have done. One workaround is to employ external HITs, where to deploy the real tasks on requesters' servers and use HITs on MTurk as an interface to bridge both sides. Such solutions usually involve more effort to design and maintain the local web servers. In July 2014, MTurk released a new function for its qualification feature, where requesters can add a qualification as "DoesNotExist" in HIT design, which means that a turker can do the HIT if a certain qualification has not been assigned to him/her. Therefore, we monitor the new turkers who accept our assignments and submit the results on MTurk at their first time in our project, and assign corresponding qualifications to them to prevent them from taking any further assignments with same type of questions.

Additionally, because our framework is Java-based, and MTurk Java SDK has not been updated since 2013, some of the new features provided by MTurk cannot be used through APIs, e.g., the new feature of qualification mentioned above. Thus we have to modify those SDK packages to fulfill our requirements.

2.4 SUMMARY

Data is essential to the success of this project. We therefore reviewed from where we got the data and how we processed it. We also summarized what we have learned from the pilot work on relevance judgement. We also identified the technical issues we had solved. These points set up a solid starting point for us to continue the work discussed in the following chapters.

Chapter 3: Analysis of Individual Performance

In this chapter, we will touch the research problems of RQ_1 and RQ_2:

> RQ_1: What factors influence the quality of relevance judgments on conversational speech data?
>
> RQ_2: What difference in behavior of relevance judging is observed between traditional assessors and crowd workers? Why?

We will use the data introduced in Section 2.1 directly, and will not try to add any information, such as the background of the conversation, or the knowledge about some religious activities, to clarify any points in transcripts, audio clips, or questions. This implies that we will pay less to process the data. We expect that crowd workers can deliver good results without such additional information.

We will first introduce potential factors to be considered in the experiments. Next, we will discuss how we design the task interface, and how we conduct the experiments with those potential factors. Then, we will focus on the results and look for the influence of those factors as well as the correlation among them. Based on these analyses, we will examine what we have found, and possible next steps.

3.1 FACTORS

These factors are derived from an iterative process of task design and experiments. At the beginning, according to our experience, we try to test the hypotheses of HA ("An audio clip may be better than the transcript for crowd workers to perform relevance judgement.") and HB ("Freelancers from oDesk will outperform turkers from MTurk in relevance judging at the same cost."). But issues arise required changes in our experimental design, including switching platforms, updating question forms, and adopting new measurements to evaluate the quality of the results. These factors are identified and further analyzed.

Overall, the following factors are taken into account in our experiments:

- **Data format**: There are two categories of data: audio clips and transcripts. The audio clips were extracted from the raw interview tapes, as described in Section 2.1.3. We share these audio clips on Box[1] so that crowd workers can access them easily.[2] As mentioned in Section 2.1.3, there are multiple versions of transcripts, including ASR transcripts, raw transcripts, and the best transcripts. We set the best transcripts as the baseline in our experiments for the purpose of comparison.

- **Topic format**: We use the topic data in CLEF2005CL-SR directly. Specially, we only focus on the topics whose IDs are listed in Table 2.2. Furthermore, a topic contains several parts: id (num), title, desc(ription) and narr(ative). Figure 3.1 shows a sample. Either the content in description or narrative field can be used as content of the "Conversation Segment" listed in the HITs or question forms for crowd workers to conduct relevance judgement. The narrative is more informative than the description in most topics. But the description will play the role of a baseline because we think it is sufficient for crowd workers to understand what we are looking for. Another consideration is how many topics should be included for a given segment. In Section 2.2, we once used 4 topics per HIT, but it seems to be not a good practice to control the participants. Thus, we decided to include all relevant and non-relevant topics of a give topic in a question form or HIT, though the distribution of the numbers of relevant and non-relevant topics is uneven, i.e. there are fewer relevant ones.

- **Understandability**: The hypotheses of HA and HC imply that understandability has positive influence on the quality of relevance judgement. To measure the understandability for the test, we ask crowd workers to self-evaluate with a 5-grade scale whether they find the transcript or audio clip easy to understand. It is a subjective measurement. Furthermore, we also ask them to write a summary in brief for

[1] `www.box.com`

[2] Because the audio clips are confidential to some degree, we only share them with links instead of public access, and the links will expire in several days once the experiments are done.

```
<top>
<num>3005
<title>Death marches
<desc>Experiences on the death marches conducted
by the SS to evacuate the concentration camps as
the allied armies approached.
<narr>Of interest are descriptions of the
preparation for the marches by the camp
administration and by the inmates, conditions
during the march, shooting of people who stayed
behind, assistance from people living at the
route of the march, escape or hiding from the
march.
</top>
```

Figure 3.1: Topic #3005 in CLEF2005CL-SR

the segment, which provides another measure of their understanding. Meanwhile, the requirement of summarization effectively prevents poor data from tainting our results.

- **Platforms**: Switching between oDesk and MTurk is another dimension we considered in our experiment. We first conducted the experiment on oDesk because, as the hypothesis of HB implies, we believe the freelancers on oDesk, who provided more clear background, are much closer to the experts according to our experience in transcription tasks. However, we quickly found in the experiments that relevance judgements are hard to attract enough freelancers on oDesk, which leads to concerns of scalability. We turned to MTurk instead.

It is well-known that one of the advantages of crowdsourcing is it can be much cheaper than recruiting experts. However, we exclude this factor from our considerations in the experimental design because, according to the low performance shown in Section 2.2, our first priority now is the quality of relevance judgements though cost and scalability are also taken into account. We will keep the amount of reward around $2.00 per task to

reduce its effect on the quality of relevance judgement.

We had several considerations when we finalized such amount of reward per task. First, we prefer to pay a fair reward to crowd workers. Because the minimum wage rate in the US is $7.25 per hour[3], and a crowd worker usually spends 10 minutes working on it, to pay $2.00 per task may be high enough. Second, we wanted to fix cost in comparing oDesk and MTurk, i.e. keeping the amount of reward as similar as possible on both platforms. We started our experiment on oDesk with $2.00 per task, i.e. question form, though we know a freelancer on oDesk usually earn tens of dollars in a project. We worried about how such low pay may affect the popularity of the task. On the other hand, we were aware that the most common reward on MTurk is less than $1.00, just several cents, so the $2.00 should be high enough to attract crowd workers and encourage them to do the tasks seriously. Additionally, in later iterations, we rewarded some crowd workers with bonus for their exceptional behavior. However, we were aware of the risk that paying more to turkers may attract more cheaters and decrease quality. We put this off to future work .

A reasonable measurement is essential to evaluate the quality of relevance judgements, i.e. the performance of the crowd workers. At the beginning, we adopted simple accuracy as the measurement, similar to what we did in Section 2.2. However, due to imbalanced distribution of the relevant and non-relevant topics, that accuracy did not work well. Thus, we used F1 score instead. However, F1 score is rather sophisticated for the crowd workers to understand, so we had to roughly explain what it is in the instructions.

In total, we conducted 9 iterations of experiments. Table 3.1 shows the settings of the factors mentioned above in each iteration.

Iterations #2 to #9 have a bar of payment, i.e. a worker can get the payment only if his/her result achieves the minimum F1 score bar. At the beginning, the bar was arbitrarily set as 0.6, and later it was changed to 0.5 due to the feedback from crowdworkers.

This chapter covers the first 7 iterations, and Chapter 4 covers iterations #8 and #9.

Iteration #1 was special. It proved our concern that the reward of $2.00 per task

[3] `http://www.dol.gov/whd/minwage/chart.htm`

#Iteration	Description
1 (1.A, 1.T)	oDesk, Topic Desc, Accuracy, each freelance will be assigned one Audio clip and one Best transcript on different segments, $2.00 per task ($4/freelancer), dynamic bonus for accuracy
2	Audio clips, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.6
3	Best transcript, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.6
4	Audio clips, MTurk, Topic Narrative, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary
5	Best transcript, MTurk, Topic Narrative, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary
6	Raw transcript, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary
7	ASR transcript, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary
8	Best transcript, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary, $1 bonus of good justification
9	Best transcript, MTurk, Topic Desc, F1 score, only pay for F1$\geq$0.5, $1 bonus of good summary, $0.5 bonus of direct and concise justification

Table 3.1: Iterations to collect data

oDesk did not attract a sufficient number of workers. Originally, we expected a freelancer only took one question form with either an audio clip or a transcript. When we found it was difficult to recruit enough freelancers for our task, we decided each freelancer should complete two question forms in different data formats on different segments. However, we still failed to meet the goal of each question form being completed by 3 freelancers. Four question forms only had 2 freelancers per task, and this sparsity introduces some issues in the analysis discussed later.

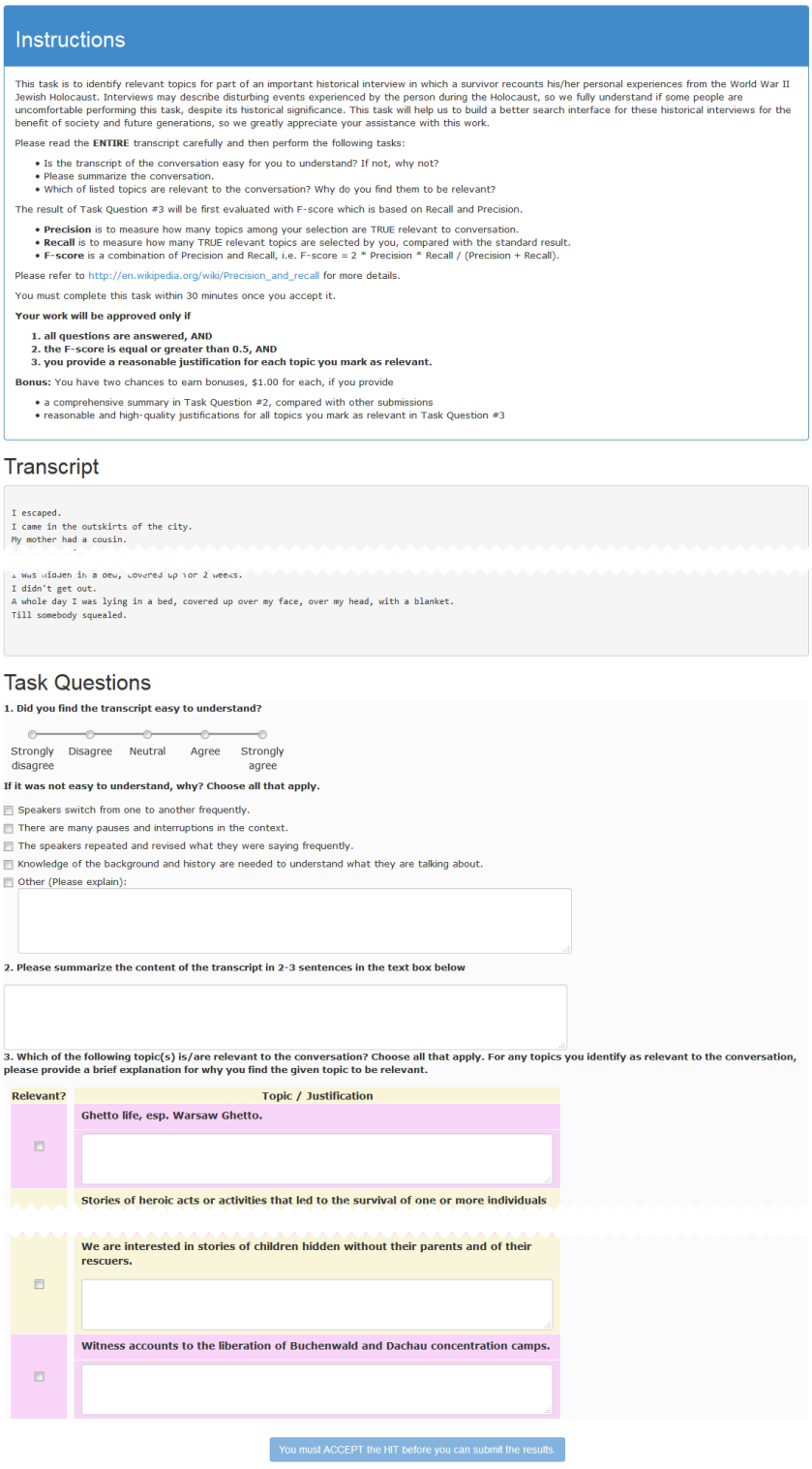


Figure 3.2: A sample HIT for a best transcript in iteration #8

3.2 TASK DESIGN

Figure 3.2 is a sample page in iteration #8 posted on MTurk for a "best transcript". A detail version of text is available in Appendix B. The freelancers from oDesk read a similar version in Microsoft Word format instead of the web page. Generally, our task interface consists of 3 sections:

- **Instructions**: Compared with the design in Figure 2.1, the instructions here are more detailed. They contain a brief introduction of the background of the project and a list of requirements of the HIT. The introduction of background gives the workers an overall picture of what they will work for and what knowledge they may need to complete the task. We stress the requirements, though they may be duplicated with the detailed instructions in the *Task Questions* section, with the intention to draw the workers' attention once they decide to accept the HIT. For example, we reject any submission that skips one or more questions because we emphasize "All questions are mandatory to complete the task." We also clarify how submitted work will be evaluated, i.e. F1 score.

- **Conversation Segment**: This section contains either an audio clip or a transcript. For an audio clip, workers can either play it online or download it, as shown in Figure 2.1. For a transcript, we organize the entire segment of transcript sentence by sentence, line by line, to ease readability. Specially, for ASR transcripts, we use the filler words such as "uh", and the annotations such as "<Breath>" to break the lines because they do not contain any punctuation,. This process was performed manually, but could be automated.

- **Task Questions**: In this section, 3 types of questions are presented to crowd workers.

  - *Quality Questions*: The workers are asked to provide subjective feedback on the quality of the segment. For example, if the segment is an audio clip, the workers can evaluate whether it is easy for them to understand its content when they lis-

ten to it. In the case of a transcript, workers comment on whether they have any difficulties reading the text. Their self-evaluation will be elicited on a 5-grade scale ranging from "Strongly Disagree" to "Strongly Agree". Aside from the option of "Strongly Agree", workers are expected to further clarify the problems encountered when listening to the audio clip or reading the transcript. We suggest some predefined options based on our perception of difficulties for audio clips and transcripts, but we also provide an open text box for them to identify additional types of difficulties encountered. Using Javascript, the clarification request will prompt automatically responding to how the workers choose on the 5-grade bar. If the workers select the option of "Strong Agree", this part will be hidden; otherwise, it will appear.

- *Verification Question*: We ask workers to summarize what they hear or read in the segment section. The original purpose of this question is to filter out poor results, but the answer also provides another perspective for ensuring and gauging worker comprehension.
- *Relevance Question*: Relevance judgement is the core task to support. Compared with the design in Section 2.2, where we treated the option of "Not Sure" as non-relevant (with some complexity of payment of reward and bonus for that option), here we simplify the relevance question and only ask the workers to identify relevant topics. Additionally, the justification text box in Figure 3.2 only appears in the last 2 iterations. It helps us to obtain a clearer picture of what the workers' thought processes and whether they were confident in their judgements.

We revised the design iteratively to fix some problems in light of feedback from previous iterations, clarifying points which confused workers, or changing settings. For example, we changed the bonus policy and updated the the section of instruction accordingly after receiving some negative feedback about the none-or-all payment we originally offered.

3.3 ANALYSIS

In this section, we analyze the experiments. We test the hypotheses in Section 1.3, and investigate the root causes.

3.3.1 Audio Clips vs. Best Transcripts

At the beginning of our work, we planned to focus on different versions of transcripts because we knew an ASR transcript is difficult to read and therefore should be less amenable to elicit reliable relevance judgements. However, when we take the audio clips into account, it is difficult to compare the audio clip to the transcript from understandability perspective.

Meanwhile, we hypothesized that students in a class may learn more quickly from teacher's lecture than from self-learning, i.e. having to read the textbooks by themselves. So we have the hypothesis HA. We also thought that perhaps the crowd workers might prefer the audio clips to the transcripts, like students preferring to listening to their teachers rather than reading the materials by themselves. Thus, we focus on testing hypothesis H_A with the null hypothesis HA_0:

> An audio clip and its best transcript are the same for crowd workers to perform relevance judgement. (The difference between the means of F1 score of audio clips and the best transcripts is 0.)

We used the best transcripts and audio clips on both oDesk and MTurk, assessing whether we would observe any significant difference in results between those two data formats on either platform. Table 3.2 shows the results.

As mentioned in Section 3.1, because we relied upon fewer freelancers to work on the task on oDesk, its results may be more biased to some extent. For example, the results for audio clip of V012 were extremely low, i.e. $mean \pm std = 0.250 \pm 0.250$. Thus, we decided to exclude the results of V012 on oDesk for both audio clip and transcript. For the purpose of fairness, we excluded the same items from MTurk as well. Therefore, we excluded V012 from analysis.

Criteria		Raw		≥ 0.5		≥ 0.6	
Data Type		Mean±Std	Cnt	Mean±Std	Cnt	Mean±Std	Cnt
Overall	A	0.506±0.262	13	0.655±0.114	9	0.721±0.078	6
	T	0.540±0.250	13	0.662±0.078	10	0.694±0.049	8
All w/o V012	A	0.553±0.235	11	0.675±0.106	8	0.721±0.078	6
	T	0.506±0.268	10	0.665±0.048	7	0.681±0.031	6
V004	A	0.317±0.273	3	0.667±0.000	1	0.667±0.000	1
	T	0.444±0.314	3	0.667±0.000	2	0.667±0.000	2
V005	A	0.600±0.200	2	0.800±0.000	1	0.800±0.000	1
	T	0.333±0.333	2	0.667±0.000	1	0.667±0.000	1
V007	A	0.635±0.045	3	0.635±0.045	3	0.667±0.000	2
	T	0.635±0.045	3	0.635±0.045	3	0.667±0.000	2
V012	A	0.250±0.250	2	0.500±0.000	1	−−−−	0
	T	0.656±0.123	3	0.656±0.123	3	0.733±0.067	2
V013	A	0.675±0.146	3	0.675±0.146	3	0.762±0.095	2
	T	0.575±0.175	2	0.750±0.000	1	0.750±0.000	1

(a) Results of iteration #1 (oDesk / Audio Clip & Best Transcript)

Criteria		Raw		≥ 0.5		≥ 0.6	
Data Type		Mean±Std	Cnt	Mean±Std	Cnt	Mean±Std	Cnt
Overall	A	0.582±0.287	25	0.707±0.133	20	0.750±0.112	16
	T	0.548±0.283	25	0.668±0.153	20	0.753±0.123	13
All w/o V012	A	0.583±0.294	19	0.712±0.150	15	0.776±0.122	11
	T	0.527±0.268	19	0.645±0.138	15	0.733±0.109	9
V004	A	0.500±0.258	5	0.625±0.072	4	0.667±0.000	3
	T	0.533±0.067	5	0.533±0.067	5	0.667±0.000	1
V005	A	0.413±0.387	5	0.833±0.167	2	0.833±0.167	2
	T	0.481±0.249	5	0.601±0.070	4	0.667±0.000	2
V007	A	0.728±0.094	5	0.728±0.094	5	0.767±0.058	4
	T	0.760±0.131	5	0.760±0.131	5	0.760±0.131	5
V012	A	0.560±0.285	5	0.700±0.058	4	0.700±0.058	4
	T	0.567±0.327	5	0.708±0.182	4	0.778±0.157	3
V013	A	0.708±0.177	5	0.708±0.177	5	0.822±0.137	3
	T	0.398±0.372	5	0.829±0.029	2	0.829±0.029	2

(b) Results of iteration #2 (MTurk/Audio Clip) and #3 (MTurk/Best Transcript)

Table 3.2: Results of iteration #1, #2, and #3

Table 3.2a shows, on oDesk, though there was some noise among the segments, overall the audio clips seem somewhat better than transcripts for maximizing relevance judgement agreement with the gold standard, achieving 0.553±0.235 and 0.506±0.268, respectively. Further considering that any poor results may ruin the quality, we filtered out workers with F1 score lower than 0.5. We found that the audio clips still outperform transcripts, with 0.675±0.106 compared to 0.665±0.048. If we raise the filtering bar up to 0.6, the advantage of audio clips is more obvious than transcript, i.e. 0.721±0.078 to 0.681±0.031.

As to the results from MTurk in Table 3.2b, we find the results for audio clips are still better than those for the best transcripts in most cases, though the values are very close, whether or not we exclude the results of V012. The only exception is when we raise the bar of F1 score to 0.6 and include the results of V012. In that case, the F1 score on audio clips is 0.750±0.112, just a bit lower than that on transcript, i.e, 0.753±0.123. But the standard deviation of transcripts is a bit higher than that of audio clips.

It seems that turkers may have less preference regarding audio clips vs. transcripts, but freelancers on oDesk appear to have a stronger preference for audio clips.

Diving into the results of 5 segments on both oDesk and MTurk, we cannot perceive an obvious pattern like "transcript is better" or "audio clip is better". For example, for segment of V013, both oDesk and MTurk show that the best transcript outperform the audio clip.

Statistically, paired-sample t-tests for the two data formats on oDesk (#1.A vs. #1.T) and MTurk (#2 vs. #3)[4] show failures to reject the null hypothesis HA_0 at $\alpha = 0.05$ with p-value of 0.7849 and 0.6996, respectively.

If we further replace the data of the best transcript with the data of the raw and ASR transcripts om MTurk in the paired-sample t-tests, we will have the same conclusion with p-value of 0.6178 and 0.6453, respectively.

Therefore, though results for the audio clips seems somewhat better than those for

[4]The V012 is included in both tests.

transcripts, we fail to reject the null hypothesis HA_0. Thus, our hypothesis of HA does not seem to hold. The workers have no preference between audio clips and the best transcripts in performing the relevance judgement task.

3.3.2 oDesk vs. MTurk

Another dimension we investigate is the platform. Freelancers on oDesk usually have clear background information, as well as the feedback history of the tasks they had done. It is thus easier for us to recruit experts there, similar to what we had done for transcription, vs. MTurk, where the workers are largely anonymous, except for their statistics information of their performance. Thus, we will test the hypothesis of HB with the null hypothesis HB_0:

> The freelancers from oDesk perform the same as the turkers in relevance judging. (The difference between the means of F1 score of oDesk and MTurk is 0.)

As shown in the Table 3.2, we found the F1 score from both oDesk and MTurk are very close. For example, when using audio clips, the overall F1 score on oDesk is 0.553$\pm$0.235, close to 0.583$\pm$0.294 on MTurk. Similarly, workers for the best transcripts on oDesk and MTurk are competitive with 0.506$\pm$0.268 and 0.527$\pm$0.268, respectively. Further analysis was done to target the freelancers on oDesk who had related background to our tasks. One freelancer is a daughter of Holocaust survivors, her performance was 0.667 for audio clips and 0.572 for the best transcripts. Another freelancer claimed to be a Jewish, so we expected she might know more about the history of Holocaust. However, her performance was 0.667 for both audio clips and the best transcript. Based on this small sample, it seems that the related background had no significant influence on the quality of relevance judgement.

Because there are fewer submissions on oDesk than those on MTurk, we randomly select submissions from MTurk per segment to pair the data on oDesk so that we can apply paired-sample t-test. Therefore, a paired-sample t-tests for audio clips between oDesk and

MTurk shows a failure to reject the null hypothesis HB_0 at $\alpha = 0.05$ with p-value of 0.3505, and for the best transcript with p-value of 0.3025.

Therefore, we fail to reject the null hypothesis HB_0. Our hypothesis of HB does not seem to hold, which is surprising. However, it also shows a solid evidence to support use of MTurk for scalability.

3.3.3 Understandability of Audio Clips and Transcripts

The hypothesis HC involves understandability. Before we test that hypothesis, we will first analyze the understandability itself. We summarize the input of understandability from all iterations, regardless of whether they are from MTurk or oDesk. Table 3.3 illustrates the distribution of understandability on different data formats. We treat both U4 ("Agree") and U5 ("Strongly Agree") as no difficulty in understanding the audio clips or transcripts.

Table 3.3 shows that about 77.8% of workers working on audio clips reported no difficulties in understanding the conversations; the corresponding number for the best transcripts is 60.2%. It is also consistent with our perception that the understandability of ASR transcirpts is the worst, lower than the best and raw transcripts.

Further analysis reveals that pain points of the workers are also centralized. Audio clips exhibit the main problem (67.6%) of heavy accents, which follows our expectation. Yet, transcripts show some different result (47.5%) than our assumptions. It is interesting that difficulties for the raw transcripts have 55% concerns about pauses and interruptions in the context, more than those for the best and ASR transcripts, as well as the overall numbers. We are not sure why the turkers had less concerns on the ASR transcripts since, compared to the raw transcripts, they should have more pauses and interruptions in the context.

Looking into the feedback on the option of "Others", we find more details about what the workers were concerned about beyond our perceptions. For example, one of the workers for audio clips complained:

- Didn't know the context of the story

		Cnt	U1	U2	U3	U4	U5	AVG
Audio Clip	V004	13	0.077	0.000	0.231	0.538	0.154	3.692
	V005	12	0.000	0.167	0.333	0.333	0.167	3.500
	V007	13	0.000	0.077	0.077	0.462	0.385	4.154
	V012	12	0.000	0.000	0.167	0.583	0.250	4.083
	V013	13	0.000	0.000	0.000	0.692	0.308	4.308
	Subtotal	63	0.016	0.048	0.159	0.524	0.254	3.952
Best Transcript	V004	18	0.056	0.278	0.167	0.333	0.167	3.278
	V005	18	0.111	0.056	0.222	0.444	0.167	3.500
	V007	17	0.059	0.176	0.176	0.412	0.176	3.471
	V012	18	0.000	0.111	0.167	0.556	0.167	3.778
	V013	17	0.000	0.176	0.235	0.471	0.118	3.529
	Subtotal	88	0.045	0.159	0.193	0.443	0.159	3.511
Raw Transcript	V004	5	0.200	0.200	0.400	0.200	0.000	2.600
	V005	5	0.000	0.200	0.200	0.400	0.200	3.600
	V007	5	0.000	0.200	0.800	0.000	0.000	2.800
	V012	5	0.000	0.400	0.000	0.600	0.000	3.200
	V013	5	0.000	0.400	0.200	0.200	0.200	3.200
	Subtotal	25	0.040	0.280	0.320	0.280	0.080	3.080
ASR Transcript	V004	5	0.400	0.600	0.000	0.000	0.000	1.600
	V005	5	0.600	0.200	0.200	0.000	0.000	1.600
	V007	5	0.200	0.600	0.200	0.000	0.000	2.000
	V012	5	0.600	0.400	0.000	0.000	0.000	1.400
	V013	5	0.800	0.000	0.200	0.000	0.000	1.400
	Subtotal	25	0.520	0.360	0.120	0.000	0.000	1.600

Table 3.3: Distribution of understandability

- Using jargon which I was unfamiliar with
- Using KM instead of miles

Similarly, another crowd worker for best transcripts said:

> The transcript appears to cut off in the middle and move onto a completely new person. It seems that the first story is left unfinished or at the very least completely interrupted.

These complaints point to the context issue we did not address in the task design. Implicitly, they are also consistent with what we will find in Section 4.2.1.

Media Type	Options	Frequency
Audio Clip	One or more speakers have heavy accents.	67.65
	Others	44.12
	Knowledge of the background and history are needed to understand what they are talking about.	23.53
Best Transcript	Others	42.86
	Knowledge of the background and history are needed to understand what they are talking about.	8.93
	The speakers repeated and revised what they were saying frequently.	5.36
	Speakers switch from one to another frequently.	5.36
Raw Transcript	There are many pauses and interruptions in the context.	55
	The speakers repeated and revised what they were saying frequently.	40
	Others	40
ASR Transcript	Others	64
	There are many pauses and interruptions in the context.	52
	The speakers repeated and revised what they were saying frequently.	40
Overall Transcript	Others	47.52
	There are many pauses and interruptions in the context.	25.74
	The speakers repeated and revised what they were saying frequently.	20.79

Table 3.4: Top-3 issues in understandability

```
1. Removing Understandability:DataFormat:Platform, FStat = 0.005847, pValue = 0.93913
2. Removing DataFormat:Platform, FStat = 0.58968, pValue = 0.4435
3. Removing Understandability:DataFormat, FStat = 1.2389, pValue = 0.29688
4. Removing DataFormat, FStat = 0.51378, pValue = 0.67326
5. Removing Understandability:Platform, FStat = 0.93013, pValue = 0.33601
6. Removing Platform, FStat = 0.25212, pValue = 0.61614

ans =


Linear regression model:
    F1 ~ 1 + Understandability

Estimated Coefficients:
                         Estimate        SE        tStat       pValue
                         ---------     --------     ------     ----------

    (Intercept)            0.62485     0.057993     10.775     1.2788e-21
    Understandability     -0.036307     0.01626     -2.233       0.026665


Number of observations: 201, Error degrees of freedom: 199
Root Mean Squared Error: 0.277
R-squared: 0.0244,  Adjusted R-Squared 0.0195
F-statistic vs. constant model: 4.99, p-value = 0.0267
```

Figure 3.3: Generate regression model with stepwise

3.3.4 Correlation between Understandability and Quality

Now, we test the hypothesis of HC. We want to know whether there is any correlation between F1 score and understandability. Furthermore, we want to identify any interactions between understandability and platform or data format if they exist.

We adopt a stepwise method to generate a regression model, starting with model in Rogers-Wilkinson notation[37]:

$$F1 \sim Understandability * Platform * DataFormat \tag{3.1}$$

which means this model has all interactions among $Understandability$, $Platform$ and $DataFormat$. $Platform$ and $DataFormat$ are dummy variables.

Figure 3.3 shows the process of stepwise method in MATLAB. In the process, $Platform$ and $DataFormat$, as well as related interactions are finally removed, and the

final regression model is:

$$F1 = 0.62485 - 0.036307 * Understandability \tag{3.2}$$

Therefor, we know that the correlation between the understandability and the F1 score cannot not be positive. Thus, our hypothesis of HC, which expects a positive correlation between the understandability and the quality of relevance judgement, does not hold. [5]

A potential explanation for this failed hypothesis is that when the workers are reading the transcript or listening to the audio clip, what they are paying attention to is what the queries focused on. We listed all relevant and non-relevant topics in the question form, which might overwhelmed and thus distracted the workers' focus.

The model in Equation 3.2 also indicates that F1 score has no correlation with $Platform$ and $DataFormat$, which are consistent with our analysis in previous sections.

3.3.5 Narratives vs. Descriptions of Topics

In iteration #1, #2, and #3, all question forms used the content in the description field of the topics for workers to make judgements. However, as a sample topic shown in Figure 3.1, a topic in CLEF2005CL-SR contains both description and narrative fields, and the narrative has more detailed information than the description. Thus, we wonder whether the quality of relevance judgement is diminished by insufficient information in the topic lists in the relevance questions. If we adopt narratives to replace descriptions, might the workers performed better? Thus, we will test the hypothesis of HD with the null hypothesis HD_0:

> Though narratives are more informative than descriptions of topics, they yield equivalent relevance judgements. (Given audio clips or the best transcripts, the difference between means of F1 score of narrative and description of topics is 0.)

[5] There is a potential defect in the analysis above. Except for the models related with MTurk and the best transcripts, the data samples in other models are so small that they may lead to such biased results.

Criteria		Raw		≥ 0.5		≥ 0.6	
Data Type		Mean±Std	Cnt	Mean±Std	Cnt	Mean±Std	Cnt
Overall	A	0.428±0.224	25	0.590±0.080	14	0.658±0.022	8
	T	0.394±0.331	25	0.694±0.145	12	0.758±0.107	9
All w/o V012	A	0.485±0.177	20	0.606±0.077	12	0.658±0.022	8
	T	0.468±0.315	20	0.711±0.139	11	0.758±0.107	9
V004	A	0.480±0.113	5	0.556±0.079	3	0.6670.000	1
	T	0.180±0.223	5	0.500±0.000	1	$---$	0
V005	A	0.390±0.253	5	0.667±0.000	2	0.667±0.000	2
	T	0.560±0.248	5	0.833±0.167	2	0.833±0.167	2
V007	A	0.590±0.152	5	0.667±0.000	4	0.667±0.000	4
	T	0.688±0.117	5	0.688±0.117	5	0.735±0.078	4
V012	A	0.200±0.245	5	0.500±0.000	2	$---$	0
	T	0.100±0.200	5	0.500±0.000	1	$---$	0
V013	A	0.480±0.075	5	0.533±0.047	3	0.600±0.000	1
	T	0.443±0.364	5	0.739±0.055	3	0.739±0.055	3

Table 3.5: Results with narrative fields as the options

We conducted iteration #4 and #5 for audio clips and the best transcripts, respectively. We kept the same configuration as those in iteration #2 and #3, and replaced the descriptions in the topics with the corresponding narratives. Table 3.5 shows the results of iteration #4 and #5.

We compared the values between Table 3.5 and Table 3.2b, with the segment of V012 included. Statistically, a paired-sample t-test for audio clips between narrative and description of the topics (iteration #2 vs. #4) rejects the null hypothesis at $\alpha = 0.05$ with p-value of 0.0488, and thus cannot reject the alternative hypothesis that there is a significant difference between the narrative and description of topics, i.e. the F1 score using topic narratives is significant lower than that with descriptions, which implies the turkers prefer the descriptions.

However, similar t-tests for the best transcripts between narrative and description of the topics (iteration #3 vs. #5) failed to reject the null hypothesis at $\alpha = 0.05$ with p-value of 0.0945, which means there is no significant difference in the quality of the relevance

judgements when we replace the descriptions of the topics with the narratives when judging written transcripts.

Why the difference? The workers, though motivated by the amount of reward, might feel bored completing a long task. They might also prefer to maximize their benefit as quick as possible. A narrative is usually longer than a description, and the workers may read it less carefully. If so, they may not fully understand what the topic is about, and so, the final judgement may be not sound.

Thus, overall, we can say our hypothesis of HD does not seem to hold. The narrative does not boost the quality of relevance judgement. On MTurk, description seems to be preferable in task design. Alternatively, we may have to design other mechanisms to encourage or require workers to read the narrative more carefully.

3.3.6 Comparison among Transcripts

Before iteration #6, we only used the best transcripts in our experiments. In light of the hypothesis of HC ("The better quality of the transcripts or audio clips, the easier to understand, and better quality of relevance judgement will result."), garbage in, garbage out. We expected the unreadable ASR transcripts would prevent the workers from providing good results. To test this hypothesis, we will compare the results for the ASR transcripts and the raw transcripts with the best transcripts. We used the same question forms, kept the same amount of reward and bonus and other settings, and replaced the segment section with corresponding transcripts. The null hypothesis H_0 here is:

> Different versions of transcripts yield the same quality of relevance judgements. (The difference between the means of F1 score for each paired versions of transcripts is 0.)

Table 3.6 shows the means and their deviations of the F1 score over 3 types of transcripts after we include data from iteration #3, #6 and #7. The means are very close, and the deviations are comparable. For example, if we filter out the results whose F1 score

Criteria		Raw		≥ 0.5		≥ 0.6	
Data Type		Mean±Std	Cnt	Mean±Std	Cnt	Mean±Std	Cnt
All	A	0.544±0.283	25	0.677±0.148	19	0.763±0.120	12
	R	0.548±0.206	25	0.665±0.137	16	0.746±0.110	10
	B	0.548±0.283	25	0.668±0.153	20	0.753±0.123	13
All w/o V012	A	0.640±0.169	20	0.670±0.150	18	0.759±0.125	11
	R	0.573±0.186	20	0.678±0.144	13	0.765±0.114	8
	B	0.527±0.268	19	0.645±0.138	15	0.733±0.109	9
V004	A	0.513±0.136	5	0.611±0.079	3	0.667±0.000	2
	R	0.447±0.126	5	0.583±0.083	2	0.667±0.000	1
	B	0.533±0.067	5	0.533±0.067	5	0.667±0.000	1
V005	A	0.733±0.226	5	0.733±0.226	5	0.889±0.157	3
	R	0.629±0.232	5	0.767±0.205	3	0.900±0.100	2
	B	0.481±0.249	5	0.601±0.070	4	0.667±0.000	2
V007	A	0.682±0.103	5	0.682±0.103	5	0.756±0.063	3
	R	0.650±0.090	5	0.650±0.090	5	0.711±0.063	3
	B	0.760±0.131	5	0.760±0.131	5	0.760±0.131	5
V012	A	0.160±0.320	5	0.800±0.000	1	0.800±0.000	1
	R	0.447±0.245	5	0.611±0.079	3	0.667±0.000	2
	B	0.567±0.327	5	0.708±0.182	4	0.778±0.157	3
V013	A	0.631±0.087	5	0.631±0.087	5	0.694±0.039	3
	R	0.566±0.188	5	0.698±0.119	3	0.762±0.095	2
	B	0.398±0.372	5	0.829±0.029	2	0.829±0.029	2

Table 3.6: Comparison among transcripts (B=Best, R=Raw, A=ASR)

is less than 0.5, the ASR, raw and the best transcripts have the values of 0.677±0.148, 0.665±0.137, and 0.668±0.153, respectively.

Table 3.7 show the p-values of the paired-sample t-test between pairs of those three types of transcripts under the null hypothesis H_0. The results fail to reject the null hypothesis of no significant differences among these three versions of transcripts.

Why could workers deliver equal quality results for ASR transcritps with best transcripts? One potential explanation is the payment and bonus we paid. On MTurk, $2.00 is a high reward for HITs. The additional bonus could raise the total payment for a worker to $4.00. Most workers seemed to take the question forms seriously and try their best to overcome the difficulties in reading the ASR transcripts. If we reduced the payment, there

	Best	Raw
Raw	0.9975	
ASR	0.9529	0.9581

Table 3.7: The p-values for comparisons between the pair of platform/data formats

might appear some large difference.

We originally planned to introduce more variant versions of ASR transcripts such as automatic annotation, dysfluency and sentence boundary detection. We believed that such versions could result in a median F1 score between the best and ASR transcripts. Gloom with current results, however, we decided not to pursue this further.

3.4 SUMMARY

In this chapter, we developed several iterations of experiments to address the research questions "What factors influence the quality of relevance judgments on conversational speech data?", and "What difference in behavior of relevance judging is observed between traditional assessors and crowd workers? Why?"

We tested several hypotheses listed in Section 1.3. We found the hypotheses of HA, HB, HC and HD do not seem to hold. In the analysis of understandability, user patterns and justification, we found that workers, motivated by the amount of award and forced by the verification question in the question forms, paid more attention to the audio clips and transcripts than to the topics. The crowd workers reported challenges such as lacking context or background when they were doing the jobs.

What do these points imply? First of all, we should consider the quality of both segments and topics. Crowd workers may fully understand what the segments are talking about though they may skip some unfamiliar words. But if they do not read the topics carefully, they may also make wrong decisions. Secondly, we should provide support to help the workers to fully understand both topic and document(segment) before they make relevance judgements. Additionally, we should enable these ideas to be implemented at

scale, e.g., through automation or crowdsourcing.

Chapter 4: Further Analysis

In Chapter 3, what we mainly found was that the raw data, regardless of audio clips, transcripts, or topics, is insufficient for crowd workers to deliver competitive results with the experts. The raw data, when put into the question forms, lacks contextual information such as time, background, and explanation of specific terms. However, instead of providing more contextual information in the question forms, in this chapter, we will start from a comparison on patterns between workers' judgements and the gold standard, then we will conduct iterations #8 and #9 to collect workers' justifications on their judgements. Then we will analyze the collected data with majority vote and discuss issues of agreement among crowd workers and experts.

We will continue to address the research questions of RQ_1 and RQ_2, and try to answer RQ_3 in this chapter.

> RQ_3: What are best practices to design tasks in a way that promotes and enables quality work by crowd workers?

4.1 Comparison between User Patterns and gold standard

Despite changing various settings, as mentioned in previous sections, we failed to achieve the goal of enabling crowd workers to provide competitive quality of relevance judgements as experts. However, we also noticed that some topics defined in the gold standard were never marked as relevant by crowd workers before iteration #9. Thus, we wondered whether problems in the gold standard may be partially to blame.

The gold standard in CLEF2005CL-SR is based on a raw judgement file, where each topic is judged from 5 aspects: context, point, comparison, indirect and direct, as well as overall assessment [16]. The assessment is made on a 5-grade scale. The gold standard was constructed from that raw judgements by treating as relevant any values of direct or

Segment	Topic	Context	Comparison	Pointer	Indirect	Direct	Overall	Support≥40%	Segment	Topic	Context	Comparison	Pointer	Indirect	Direct	Overall	Support≥40%
V004	1166	3	0	0	4	3	3		V007	2012	0	0	0	0	0	0	
	1225	0	0	0	0	0	0			2198	0	0	0	0	0	0	
	1330	-1	-1	-1	-1	-1	0			2224	3	0	0	0	0	2	
	1345	0	0	0	0	0	0			3004	0	3	0	0	0	2	45.95
	1551	0	0	0	0	0	0			3013	0	0	0	0	0	0	
	2012	0	0	0	0	0	0			3014	0	0	0	0	0	0	
	3005	0	3	0	0	0	3	50	V012	1225	0	0	0	0	0	0	41.67
	3015	0	0	0	2	2	2	83.33		1508	0	0	0	0	0	0	
V005	1225	0	0	0	0	0	0			1620	0	0	0	0	0	0	
	1624	0	0	0	0	0	0			2012	0	0	0	0	3	3	
	1746	0	0	0	0	0	0			3016	0	0	0	0	0	0	
	1897	0	0	0	0	0	0			3026	3	0	0	0	3	3	47.22
	2384	4	0	0	0	3	3	51.43	V013	1225	0	0	0	0	0	0	
	3005	0	0	0	0	4	3	68.57		1288	0	0	0	0	0	0	
	3015	0	0	0	0	1	1	65.71		1311	0	0	0	0	3	3	
	3026	0	0	0	0	0	0			1429	0	0	0	0	0	0	
V007	1225	0	0	0	0	0	0			1508	0	0	0	0	0	0	
	1288	3	0	0	2	0	3			1551	4	0	0	0	0	2	41.67
	1551	4	0	0	0	4	4	94.59		15602	0	0	0	2	0	2	50
	1877	0	0	0	0	0	0			2012	0	2	0	0	0	1	
	1897	3	0	0	2	2	2	97.3		3027	1	0	0	0	2	2	83.33

Table 4.1: Frequent users' patterns vs. CLEF2005CL-SR gold standard

indirect relevance no less than 2. The raw judgement data distributed in the test collection also includes only the combined, post-adjudicated judgements made by 2 experts. We could not ascertain whether there was any disagreement between the experts, which might help us to analyze the disagreement between the gold standard and crowd workers' judgements.

After referring to the method of how the gold standard was generated from the raw judgement, we are interested in whether the crowd workers utilize alternative relevance criteria or thresholds. For example, they may tend to make judgements emphasizing context or comparison. Thus, we consolidate all iterations before iteration #8 and applied the

technique of frequent pattern mining, which was first introduced by by Agrawal et al. [2]. We treated each worker's submission as a "transaction", and attempted to find subsets of topics (itemsets) which are common to at least a minimum number of the submissions. The ratio of the number of the common subsets of topics to the total number of submissions is called as "support". Subsets of topics with support higher than a threshold, a.k.a. minimum support, are considered as frequent. We used FP-Growth algorithm [15] to extract such frequent topic sets with the minimum support of 0.4. Table 4.1 shows the frequent topics and their values of support.

For these frequent topics, we tried to figure out whether there is a single rule to follow, like the "direct$\geq$2 & indirect$\geq$2" in gold standard. However, we did not find such a rule. It seems there is no single preference for crowd workers to make judgements. It should be reasonable because, though we assume that the crowd workers in MTurk are even from statistical perspective, actually each worker has his/her own idea of what the relevance means. The diversity of workers' preferences may result in multiple overlapped or conflicting rules. Thus, it is not easy to generate an alternative gold standard like what CLEF2005CL-SR did.

4.2 JUSTIFICATION ANALYSIS

4.2.1 Justification with Workers' Own Words

In iteration #8, we asked the crowd workers to justify their judgements. Such requirement indeed makes the tasks more effortful, but it may further discourage cheating and encourage more understanding to the conversations. More importantly, when the crowd workers disagree with the gold standard, we can look at their rationales (justifications) to better understand why they made the judgments they made, which will offer us an insight into how to improve the task design, and determine whether their disagreement represents a valid subjective different interpretation or is a bona fide error.

Appendix C shows the data we obtained. Overall, the workers prefer direct infor-

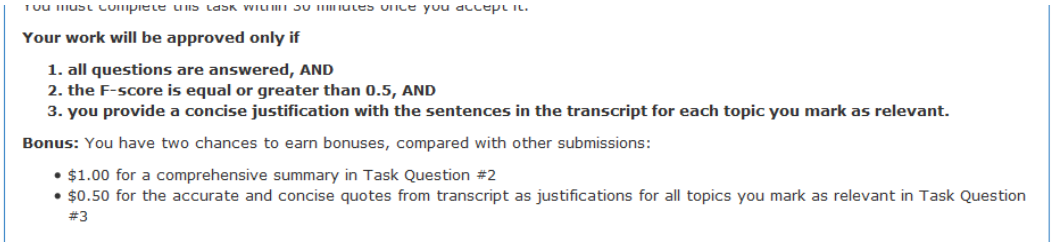
mation, i.e. they use exact sentences to support their judgement, such as #W1-005 for segment of V005, and #W1-007 for V007. But some workers (#W3-004 and #W3-012) are likely cheaters because their justifications seems to be less to do with the transcripts but more like to imagine what should happen based on the descriptions of the topics. For example, #W3-012 justified the topic #1508 (The fate of Mischlinge persons.):

> "Learning the fate of people who survived the concentration camps would be extremely relevant. It would be interesting to know the impact the war had on the group as a whole, since many were killed in such a short time span."

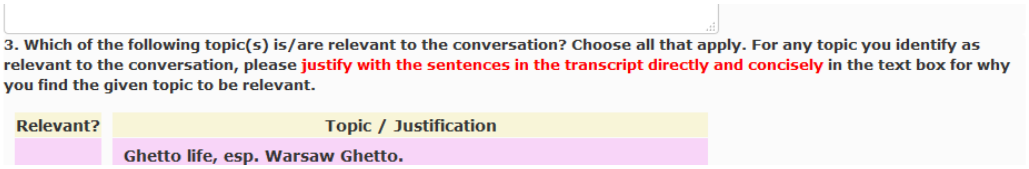
It is interesting that, according to the justification, it seems that workers ignore some specific key words they are not familiar with in making their judgements. For example, the segment of V007 has a term of "Kosher", which refers to dietary practices of Judaism. It seems #W1-007 had caught that word, but misunderstood the meaning and wrongly associated it with "ceremonies and rituals related to death and burials". The term "Mischlinge" appears in the topic of 1508 instead of the transcript of V012, which is a German term used during the Third Reich to denote persons deemed to have both Aryan and Jewish ancestry. It seems #W3-012 and #W4-012 did not understand the word.

Crowd workers also suffered from the lack of background information or context of the conversation. For example, #W3-013 identified topic of 1429 as relevant for the transcript of V013. But the story of V013 happened in 1939, and the topic has a very clear clue of "1942-1945" so that it should not be identified. The reason seems to be that the information of "1939" was talked in the beginning of the interview, but that segment of V013 started from the middle of the interview, which missed such time information.

In summary, it seems many errors were introduced because of lack knowledge and information about the specific terms, history background, and the story context which is in other segments.



(a) Updates in instructions



(b) Updates in questions

Figure 4.1: Updates in iteration #9 (Justification with exact words)

4.2.2 Justification with Exact Words

We conducted a new experiment iteration (#9) on MTurk, in which we asked turkers to support their relevance judgements by quoting evidence in the transcript instead of writing justifications in their own words. Our intentions is to identify whether workers' relevance judgements are based on "direct relevance" or "context relevance" as defined by Huang et al. [16], understanding their mistakes, and validating their answers even if not matching the gold standard.

We updated the instructions and the question of relevance judgement in the HIT used iteration #8, as shown Figure 4.1, in which the Figure 4.1a and 4.1b highlight the changes, respectively.

The justifications we collected are listed in Appendix D. Most of the turkers *used* the sentences in transcripts, either in the manner of direct or indirect speech[1]. Some workers also presented their logic behind the justifications. For example, #W4-005 provided the

[1] *Direct speech* is "the reporting of what someone has said or written by quoting his or her exact words", as defined at `http://www.collinsdictionary.com/dictionary/english/direct-speech`. *Indirect speech* is "the reporting of something said or written by conveying what was meant rather than repeating the exact words", as defined at `http://www.collinsdictionary.com/dictionary/english/indirect-speech`

following justification when making the judgement for topic #3005 in segment V005:

> "And they marched us the same distance probably from Biala to Rawa." Was the first sentence the speaker tells us, which indicates they were marching. Then shortly after, him saying that they lost people along the way, he states "Next morning, they marched us out of Rawa" and eventually comes to tell us "This is the closest city to Poland. In BRASLAU, we found out that the Warsaw fell. We saw the Red Cross. They were wearing arm bands from the Red Cross. Women came and brought some fresh bread and cheese and coffee" this shows that the allied armies approached. Another clue about the allied armies were that he later states "Because the roads were crowded with army units, going back and forth."

However, in this iteration, 6 of 25 workers (#W4-004, #W2-005, #W5-005, #W4-007, #W3-013 and #W4-013) still used their own words to express their justifications as those did in iteration #8. We paid each of them $1.00 as bonus for their efforts though we rejected their work. It seems that our instructions were not clear enough for the turkers to follow. Perhaps, the phrase of "copy and paste" would be better for them to understand what we had intended by "direct sentences".

We further compared the F1 score of this iteration with iteration #8 and #3 in Table 4.2, where the symbol "D" in the column of "Data Type" stands for this "Direct Justification" iteration, "J" is for iteration #8 where turkers provided justifications in their own words, and "T" refers to iteration #3 as the baseline for comparison because all these three iterations employed the best transcripts.

Statistically, this result in iteration #9 is the same as those of iteration #8, but different with those of iteration #3. With null hypothesis that the difference between 2 mean values of the paired iterations is 0, the paired-sample t-test at $\alpha = 0.05$ for iteration #9 against iteration #3 fails to reject the null hypothesis with p-values of 0.2824, accepts the null hypothesis with p-values of 0.0276 for iteration #9 against iteration #8.

However, we observe that iteration #9 is more efficient to prevent poor results than

Criteria		Raw		≥ 0.5		≥ 0.6	
Data Type		Mean±Std	Cnt	Mean±Std	Cnt	Mean±Std	Cnt
Overall	D	0.625±0.194	25	0.674±0.133	22	0.746±0.097	15
	T	0.548±0.283	25	0.668±0.153	20	0.753±0.123	13
	J	0.456±0.285	25	0.649±0.097	16	0.703±0.059	11
All w/o V012	D	0.623±0.214	20	0.686±0.145	17	0.775±0.099	11
	T	0.527±0.268	19	0.645±0.138	15	0.733±0.109	9
	J	0.478±0.276	20	0.657±0.098	13	0.711±0.063	9
V004	D	0.493±0.271	5	0.617±0.126	4	0.733±0.067	2
	T	0.533±0.067	5	0.533±0.067	5	0.667±0.000	1
	J	0.280±0.254	5	0.667±0.000	1	0.667±0.000	1
V005	D	0.660±0.174	5	0.725±0.130	4	0.800±0.000	3
	T	0.481±0.249	5	0.601±0.070	4	0.667±0.000	2
	J	0.608±0.114	5	0.608±0.114	5	0.733±0.067	2
V007	D	0.622±0.104	5	0.622±0.104	5	0.733±0.067	2
	T	0.760±0.131	5	0.760±0.131	5	0.760±0.131	5
	J	0.693±0.053	5	0.693±0.053	5	0.693±0.053	5
V012	D	0.633±0.067	5	0.633±0.067	5	0.667±0.000	4
	T	0.567±0.327	5	0.708±0.182	4	0.778±0.157	3
	J	0.367±0.306	5	0.611±0.079	3	0.667±0.000	2
V013	D	0.718±0.203	5	0.798±0.140	4	0.798±0.140	4
	T	0.398±0.372	5	0.829±0.029	2	0.829±0.029	2
	J	0.331±0.316	5	0.686±0.114	2	0.800±0.000	1

Table 4.2: Results of iteration #9 (Justifications with exact words)

iteration #3 and #8. Only 3 submissions were lower than the bar of F1$\geq$0.5, compared to the 5 and 9 submissions in iteration #3 and #8, respectively. The means of overall F1 score in iteration #9 is also a bit higher than those in other 2 iterations, though they are statistically the same. One of the possible reasons is that the requirement of quoting sentences from the transcripts might have forced the turkers to concentrate on the transcripts more seriously, so that they understood the transcripts better than those in other iterations. Additionally, the higher means in this iteration vs. iteration #8 implies that the "direct" factor is likely to improve relevance judgements to some extent.

However, the workers might pay less attention to the contents of the topics, compared with their efforts on the transcripts, as we pointed out in other iterations. It is a

```
<top>
<num>3015
<title>Mass shootings
<desc>Accounts by Jewish survivors from the sites
 of mass shootings, and by those who personally
 witnessed the Jews being marched to the sites of
 mass shooting.
<narr>Of particular interest are accounts of
 people who survived a mass shooting, also of
 interest are accounts from those who personally
 witnessed victims being selected and being
 marched to a mass shooting site, particularly
 accounts of resistance or other unusual events.
</top>
```

Figure 4.2: Topic #3015 in CLEF2005CL-SR

challenge for the design of the task to ask crowd workers to concentrate on both transcripts and topics simultaneously.

Alternatively, the "description" field of the topics might be insufficient for the turkers to make accurate judgements. For example, topic #3005 and #3015, as shown in Figure 3.1 and 4.2, both have the common word "match". But they can be distinguished by the narrative information. Iterations #4 and #5 had demonstrated that using only "narrative" field did not lead to better results as we expected. But what would happen if we were to present the topic in a mixed way, where some topics use the "description" while some use the "narrative", especially for those topics with similar description information? Furthermore, how might we identify the topics that may be closer? These issues can be further explored in the future.

4.3 Analysis with Majority Vote

In all 9 iterations, we have focused on evaluating the average individual performance of the crowd workers. While, we once expected some workers could perform as good as those professionals. As we illustrated in Chapter 3 and Section 4.2.2, their performance failed

Iteration	Mean±Std		
	Raw	F1≥0.5 for peer	F1≥0.6 for peer
1.A	0.506±0.262	0.655±0.114	0.721±0.078
1.T	0.540±0.250	0.662±0.078	0.694±0.049
2	0.582±0.287	0.707±0.133	0.750±0.112
3	0.548±0.283	0.668±0.153	0.753±0.123
4	0.394±0.331	0.694±0.145	0.758±0.107
5	0.428±0.224	0.590±0.080	0.658±0.022
6	0.548±0.206	0.665±0.137	0.746±0.110
7	0.544±0.283	0.677±0.148	0.763±0.120
8	0.456±0.285	0.649±0.097	0.703±0.059
9	0.625±0.194	0.674±0.133	0.746±0.097

Table 4.3: Summary of the individuals' performance

to meet our expectations. For convenience in comparing with the analysis in the following sections, Table 4.3 summarizes their overall performance in all 9 iterations.

Let us next look into the data in another way where we analyze the crowd workers' consensus response instead of average individual behavior.. We employed the method of majority vote for this purpose. A topic for a given segment (document) will be considered relevant here if the majority of workers who were working on that segment agree on that topic.

More precisely, we relax majority vote to vary the minimum agreement threshold to be 0.4, 0.5 and 0.6, and we applied the filter of peers' performance as we did before. Table 4.4 shows the results.

Compared to Table 4.3, where the higher filter of peers' performance boosts the F1 scores, the effect of the filter is more complex in majority vote. The F1 filter of 0.5 pushes up the values, but the filter of 0.6 then draws them down, and sometimes it makes the values worse than those without any filters. It seems that filter of 0.5 is a sufficient bar to screen out poor data.

In the columns of "Raw" in Table 4.3 and 4.4, except for iteration #4 and #5, both of which use the narrative as topic information, majority vote achieves better F1 score.

	Iteration	Mean±Std		
		Raw	F1≥0.5 for peer	F1≥0.6 for peer
Majority=0.4	1.A	0.651±0.163	0.698±0.139	0.579±0.342
	1.T	0.714±0.165	0.731±0.163	0.691±0.088
	2	0.648±0.209	0.733±0.149	0.720±0.073
	3	0.600±0.091	0.671±0.126	0.705±0.085
	4	0.413±0.276	0.705±0.221	0.521±0.484
	5	0.397±0.076	0.567±0.091	0.520±0.292
	6	0.605±0.174	0.685±0.145	0.777±0.137
	7	0.523±0.331	0.737±0.067	0.787±0.137
	8	0.581±0.084	0.648±0.043	0.720±0.073
	9	0.698±0.139	0.731±0.091	0.758±0.087
Majority=0.5	1.A	0.651±0.163	0.698±0.139	0.579±0.342
	1.T	0.714±0.165	0.731±0.163	0.691±0.088
	2	0.667±0.163	0.720±0.073	0.720±0.073
	3	0.667±0.163	0.731±0.091	0.731±0.091
	4	0.390±0.388	0.721±0.221	0.521±0.484
	5	0.433±0.253	0.567±0.091	0.520±0.292
	6	0.573±0.159	0.698±0.139	0.777±0.137
	7	0.617±0.371	0.777±0.137	0.787±0.137
	8	0.667±0.163	0.701±0.098	0.747±0.073
	9	0.720±0.073	0.758±0.087	0.758±0.087
Majority=0.6	1.A	0.538±0.326	0.698±0.139	0.587±0.335
	1.T	0.548±0.361	0.731±0.163	0.710±0.062
	2	0.667±0.163	0.787±0.137	0.787±0.137
	3	0.667±0.163	0.747±0.073	0.720±0.073
	4	0.390±0.388	0.655±0.157	0.505±0.476
	5	0.433±0.253	0.567±0.091	0.520±0.292
	6	0.573±0.159	0.731±0.091	0.627±0.376
	7	0.617±0.371	0.777±0.137	0.787±0.137
	8	0.667±0.163	0.747±0.073	0.720±0.073
	9	0.720±0.073	0.747±0.073	0.747±0.073

Table 4.4: F1 score based on majority vote

However, such pattern does not apply to the filter columns.

The highest score (0.787 ± 0.137) in Table 4.4 appears at several cells, e.g., at iteration #7, which used the ASR transcripts with the filter of F1$\geq$0.6. However, Iteration #9 has a smaller standard deviation with a remarkably high score.

Table 4.5 shows the p-values of a couple of paired-sample t-tests on the mean values between Table 4.3 and the majority sections in Table 4.4 at the significant level of $\alpha = 0.05$. We first test the null hypothesis that the difference between the mean values of majority vote and individual performance for each setting is 0. The underlined p-values shown in the first section of Table 4.5 suggest the corresponding null hypotheses have to be rejected. For example, when majority=0.5, regardless the F1 filter, there are significant differences between the corresponding items in Table 4.3 and Table 4.4. Next, we want to know what such differences imply, i.e. whether the mean value in Table 4.3 is less than or greater than the corresponding one in Table 4.4. We apply left-tailed and right-tailed paired-sample tests, respectively. The left-tailed paired-sample t-test assesses the null hypothesis that the mean value in Table 4.3 minus the corresponding one in Table 4.4 is no less than 0, while the right-tailed paired-sample t-test assesses the null hypothesis that the mean value in Table 4.3 minus the corresponding one in Table 4.4 is no greater than 0. As shown in the second and third sections of Table 4.5, the underlined values refer to rejecting the null hypotheses.

For Table 4.5, we can figure out that the mean values at majority of 0.4, 0.5 and 0.6 without filter are the best because they are neither the same as the mean values of corresponding individuals' performance nor less than them. Particularly, the value for iteration #9 has a lower standard deviation when majority is 0.5 and 0.6, which implies the workers could deliver more consistent results in such settings. This suggests it is a good practice for further work.

Majority	Raw	F1≥0.5	F1≥0.6
H_0: individual-majority=0			
H_A: individual-majority≠0			
0.4	0.0128	0.4071	0.0052
0.5	0.0027	0.0202	0.0003
0.6	0.0130	0.0024	0.0000
H_0: individual-majority≥0			
H_A: individual-majority<0			
0.4	0.0064	0.7965	0.9974
0.5	0.0013	0.9899	0.9999
0.6	0.0065	0.9988	1.0000
H_0: individual-majority≤0			
H_A: individual-majority>0			
0.4	0.9936	0.2035	0.0026
0.5	0.9987	0.0101	0.0001
0.6	0.9935	0.0012	0.0000

Table 4.5: p-values of paired-sample t-test (individual vs. majority)

4.4 AGREEMENT AMONG CROWD WORKERS

Until now, we only compared the results from crowd workers to the gold standard to calculate the values of measurements, i.e. the F1 score on individual performance or majority vote. However, we may have noticed that crowd workers sometimes had high consistency with each other to mark some topics as relevant or non-relevant, regardless of the gold standard. We are interested in such agreement among the workers.

For each pair of segment and topic marked as relevant by crowd workers, we calculate its agreement among the workers. For example, if that segment has 5 workers, and three of them marked the topic as relevant, we have agreement of $3/5 = 0.6$. More accurately, it is a positive agreement. Meanwhile, for instance, if only 1 of 5 workers marked that topic as relevant, we calculate its negative agreement as $1 - (1/5) = 0.8$, which means 80% of the workers have a agreement that the topic for the given segment is non-relevant. Then we summarize the agreement by iteration and segment in Table 4.6.

Roughly, the workers have agreement around 0.7 regardless of positive and nega-

Mean±Std		Raw	F1≥0.5 for peer	F1≥0.6 for peer
Iteration	1.A	0.699±0.122	0.916±0.116	0.906±0.120
	1.T	0.675±0.095	0.900±0.149	0.883±0.162
	2	0.751±0.069	0.815±0.054	0.867±0.095
	3	0.757±0.063	0.785±0.059	0.810±0.137
	4	0.710±0.050	0.852±0.135	0.754±0.007
	5	0.755±0.051	0.874±0.124	0.955±0.089
	6	0.745±0.030	0.730±0.041	0.762±0.183
	7	0.773±0.038	0.835±0.102	0.900±0.099
	8	0.731±0.068	0.798±0.127	0.901±0.148
	9	0.794±0.037	0.813±0.062	0.800±0.126
Segment	V004	0.743±0.047	0.889±0.100	0.963±0.111
	V005	0.722±0.067	0.833±0.131	0.846±0.147
	V007	0.806±0.042	0.811±0.040	0.817±0.059
	V012	0.701±0.081	0.852±0.140	0.798±0.179
	V013	0.723±0.068	0.774±0.082	0.845±0.117
Overall		0.739±0.070	0.832±0.108	0.855±0.132

Table 4.6: Agreement among workers

tive, which means approximately 70% of the workers for a given task of relevance judgement have a common sense to agree or disagree on whether a topic is relevant or not. In other words, for a given pair of segment and topic, 70% workers will agree or disagree on its relevance.

4.5 Agreement between Workers and Experts

Now, we further consider the behavior differences between the workers and experts (gold standard). We want to diagnose following potential factors that could contribute to observing different workers' results vs. the gold standard:

A) Errors in the gold standard for a given segment, e.g. a topic is said to be relevant to a segment but not contained in the gold standard (false positive), or a topic is said to be non-relevant but appearing in the gold standard as relevant (false negative).

B) The topic is hard for crowd workers to judge whether it is relevant or non-relevant to

a given segment. (It may also be difficult for the experts.)

C) The crowd workers and experts have different points of views on the same topic for a given segment.

D) The instructions of the task for a given segment are so unclear or ambiguous on the criterion of relevance that the crowd workers have no idea how to do the work.

E) The instructions are clear on the criterion of relevance, but it misleads the crowd workers to provide consistent but different responses than what is in the gold standard.

We will consider two aspects in the diagnosis: 1) Do the workers' results agree with the gold standard for a given segment? and 2) Do the workers agree with each other on such segments?

The combinations of these two aspects are:

- Workers agree with each other and their results also agree with the gold standard (AA): This is a "perfect" combination because it is what we most expect, i.e. we wish the crowd workers can perform as well as those experts.

- Workers agree with each other but their results disagree with the gold standard (AD): Workers' results have a high consistency, but they disagree with the gold standard. For example, majority of crowd workers mark a topic as relevant for a given segment, but such topic is non-relevant according to the gold standard. This combination may reveal some potential problems such as factors A, C and E mentioned above.

- Workers disagree with each other and the result also disagree with the gold standard (ND): Workers' results has lower consistency, i.e. workers disagree with each other on the relevant or non-relevant topics for a given segment. Their results also disagree with the gold standard. For example, the minority of crowd workers agree on a non-relevant topic for a given segment, but the gold standard says that the topic is

relevant. It may be caused by the factor D because the workers had no clear definition of relevance as we require, or factor B, i.e. the topic is hard to judge.

- Workers disagree with each other but their results agree with the gold standard (NA): This seems unlikely. If workers' results agree with the gold standard, it should imply the results are consistent with each other, i.e. workers agree with each other.

In combinations of AD and ND, we will not further diagnose, e.g., to investigate the variables to differentiate the factors of A, C and E in AD. We leave it for future work.

We preprocess the data in all iterations. For each segment, we calculate a topic list with majority vote, i.e. choosing the topics which the majority of workers mark as relevant. Such a topic list will be referred to as the "silver standard". The majority of workers are defined by a parameter of majority rate (MR). Then, for each submission, we calculate F1 scores with the corresponding gold standard and silver standard, respectively. Thus, these 2 F1 scores can be presented on a scatter chart, as shown in Figure 4.3 when MR=0.5 without any F1 filter.

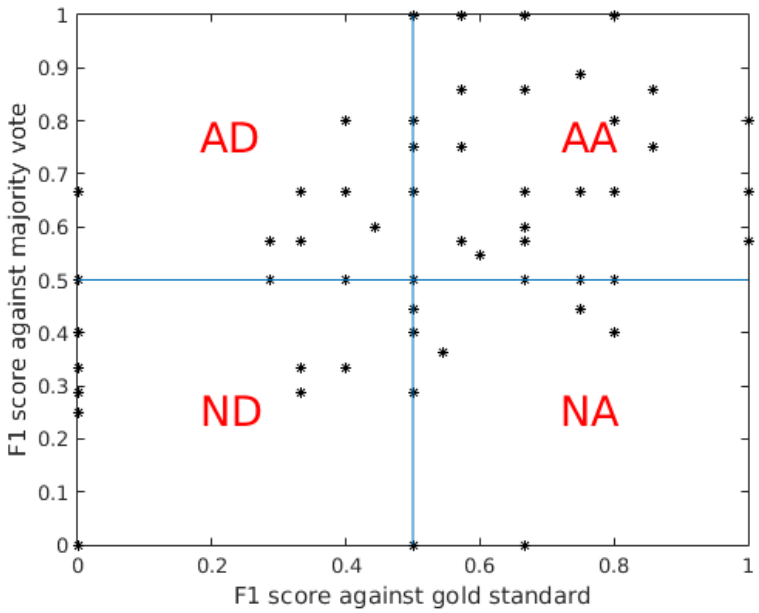


Figure 4.3: Combination of F1 scores against gold standard and majority vote

For all the points in the chart, suppose we set the threshold for agreement with the gold standard and threshold for agreement among workers as 0.5. The chart can be divided into 4 areas, each of which is a combination mentioned above. Table 4.7 shows the ratios of the workers belonging to the 4 combinations with both thresholds as 0.5. According to Table 4.7, most workers belong to AA, and in most settings, workers in NA are fewer than those in other 3 combinations. Furthermore, workers in AD are fewer than those in ND, which means factors D and B may play more important roles in relevance judgements.

MR	AD / ND	AA / NA				
	Raw		F1≥0.5		F1≥0.6	
0.4	0.155	0.637	0.097	0.646	0.066	0.438
	0.146	0.062	0.204	0.053	0.235	0.261
0.5	0.106	0.633	0.084	0.659	0	0.460
	0.195	0.066	0.217	0.040	0.301	0.239
0.6	0.093	0.482	0.049	0.642	0.022	0.434
	0.208	0.217	0.252	0.058	0.279	0.266

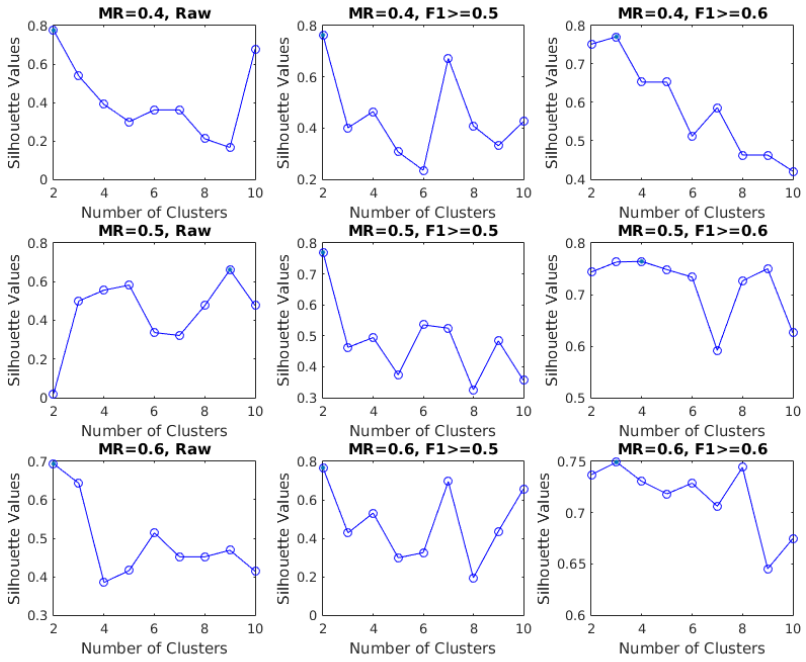
Table 4.7: Combination across MRs and F1 filters

However, to divide the space in the chart into 4 areas with predefined thresholds is arbitrary. We have no idea what the best values of the thresholds should be. Thus, we take another approach to cluster the pairs of workers' performance of gold standard vs. majority vote.

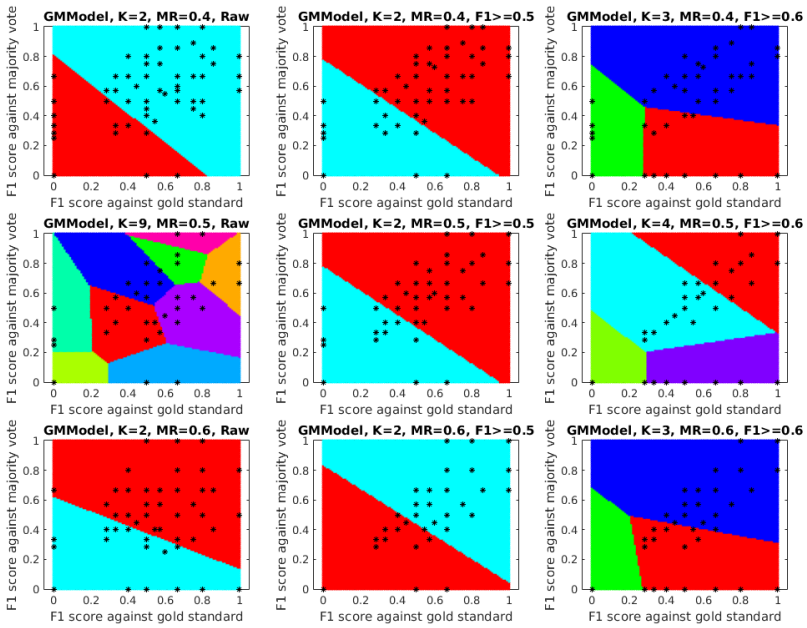
For each setting (MR and F1 filter), we use the *evalclusters* tool[2] in MATLAB to explore the best number of clusters (K) between 1 to 10 with Silhouette values [33] and Gaussian mixture distributions[3].

Figure 4.4 shows the how the K is determined and how the data is clustered in each setting. Figure 4.4a illustrates that, in each setting, the higher the Silhouette value is, the better the value of K to be used for clustering.

[2] http://www.mathworks.com/help/stats/evalclusters.html
[3] http://www.mathworks.com/help/stats/gmdistribution.html



(a) Selection of K



(b) Clustering with the best K

Figure 4.4: Clustering with Gaussian mixture distributions

Figure 4.4b uses different colors to represent the regions of the cluster in the scatter chart. For example, when MR=0.5, the raw data can be clustered into 9 groups. Theoretically, each group should be interpreted more or less with the factors mentioned above, but we leave this analysis for future work. Additionally, 5 of 9 settings generate 2-cluster results, which can be roughly categorized into AA and ND. But the points in the clusters are consistent with what we demonstrated in Table 4.7, i.e. most workers belong to AA.

4.6 SUMMARY

In this chapter, we further developed an experiment which required the turkers to justify their judgements using quotations drawn from the given transcripts. We conducted analysis with majority vote, and found this experiment was a best practice because it improves the F1 score significantly.

We also analyzed the consistency issue among crowd workers. The workers provided approximately 70% agreement on their submissions. As to the performance between crowd workers and professionals (represented by the gold standard), the analysis was based on combinations of agreement among workers as agreement with the gold standard or not. It suggests further analysis is still needed.

Chapter 5: Conclusion

The work of this thesis is to propose and assess methods to build a test collection for conversational speech at scale with crowdsourcing. In previous work, we had successfully solved the problem of retrieving transcripts of high quality from audio files through crowdsourcing. In this thesis, we are focusing on how to recruit crowd workers to deliver relevance judgements of high quality by investigating the following research questions.

- RQ_1**: What factors influence the quality of relevance judgments on conversational speech data?**

  First, we identify 4 factors, including data format, topic format, understandability, and platforms. We design the first round of experiments on the "raw data" with the intention to assess their influence on the quality of relevance judgements.

  We found that there is no significant correlation between understandability and quality of relevance judgement, across varying scenarios of audio clips, topic descriptions vs. narratives, and different platforms (MTurk vs. oDesk), and different versions of transcripts, such as ASR, raw and the best manual transcripts. In other words, the quality of the transcript has (remarkably!) no (observed) influence on the quality of relevance judgement. In the case of variant topic format, we even found the more verbose "narrative" to have negative effect on the result in some cases, even though it is more informative than the "description".

  However, when we asked crowd workers to justify their judgements, especially our requirement to quote the sentences in the transcripts instead of explaining in their own words, we obtained the best result in this thesis, i.e. majority vote agreement with gold standard assessments. Hence, we think use of such justifications can be an important factor to improve the quality of relevance judgements.

- RQ_2**: What difference in behavior of relevance judging is observed between**

traditional assessors and crowd workers? Why?

Most of our hypotheses are based on comparing crowd vs. experts-produced relevance judgements. However, the first round of experiments largely fails to reject most of our null hypotheses and thus find different behaviors of relevance judgement.

We also noticed that better summary/understandability does not imply better relevance judgement. Workers' justifications for their judgement illustrated some points that are assumed to be common knowledge for experts, but the crowd workers were required to present such information explicitly in the tasks.

We also analyzed the agreement among crowd workers as well as their agreement with the experts represented by the gold standard. The crowd workers have an agreement of approximately 70%. However, their agreement with gold standard is a more complex topic. Most workers' judgements agree with those of experts. But it is still unclear why the disagreement exists. We think further investigation is warranted.

- RQ_3**: What are best practices to design tasks in a way that promotes and enables quality work by crowd workers?**

  Among all 9 iterations of experiments, we identified several best practices:

  - Encourage crowd workers to pay more attention to the content of the conversational speech. One of the best ways we adopted is to require them to justify their judgement with the sentences in transcripts, as we did in iteration #9. This method can efficiently remove poor data and improve the final quality.
  - Decompose payments for different sub-tasks. A worker's performance can be evaluated from multiple aspects, e.g., writing a comprehensive summary or concise justification in our experiments. Beyond their answers to the questions of relevance judgement, their efforts on these aspects should also be rewarded because such efforts provide the potential pathways to the high-quality results.

5.1 FUTURE WORK

After the work in this thesis, we have multiple points to explore in the future:

- Provide more information about the domain-specific vocabulary. Where are the workers most impacted by unknown terms, the query or the segment? How do we identify such words? Is it possible to use crowdsourcing to do the work? Which information and in what format should be added? For example, shall we just provide a link to Wikipedia for a specific location or to provide term definitions directly in the context in which the terms occur?

- Because current relevance judgement is segment-based, knowledge of previous segments could be especially important to make correct judgements. For example, speakers may talk about their ages, families, relationships with others in previous segments. Similarly, it seems general background, e.g. familiarity with the timeline of WWII, could help the workers to make better judgement. How to provide such information in the task efficiently? How to extract such information from the context? Additionally, if we provide a longer conversation (audio clip or transcript), and ask workers to work on just a specified portion, how do we best design the task and incentive the workers?

- Instead of judging relevance overall, we can have workers judge multiple aspects of relevance (and inducing relevance from these aspects), and asking repeated questions [39]. The point of this design is that even if a worker disagrees with the gold standard, if the worker is self-consistent (same answer to repeated questions), we can trust their responses (intra-judge self-consistency vs. inter-judge consistency)

- We can try to ask crowd workers to do some precursor tasks to support the data preparation, e.g., to ask workers to define unfamiliar words, rewrite topic description in their own words, or summarize relevant articles on Wikipedia or other background resources for the segments. .

- Workers seem to prefer direct evidence to make decisions, such as matching the exact query words in the transcript, which will accelerate their completing the task. However, searching such words (like using Ctrl-F in browser) could be automated without any need for human assessors, and such word-spotting is not relevance assessment. We hope to encourage workers to read the transcript thoroughly. We can highlight important query words, instruct the workers to pay more attention to these words. The key point here is educating workers that judging relevance does not equal word-spotting, and that we already know where query words occur in the document (transcript). Alternatively, we can either convert the transcription into image format or use Javascript to block the searching function

- We can make a concerted effort to recruit experts who are familiar with the history of Holocaust and WWII, as well as the religious traditions of Judaism. We assume they would be more familiar with most of the specific terms.

- Rather than measure crowd judgement accuracy vs. gold, instead measure change in rank correlation of systems who participated in the CLEF evaluation. Ultimately it doesn't matter if judges disagree with one another or gold; what matters is whether their judgements yield a stable ranking of systems. We can refer to Voorhees's work [35]. Roughly, we could download past system runs from CLEF on our topics as well as the *trec_eval* tool[1] from NIST, then we put crowd judgements in simple text format and evaluate IR runs on crowd judgements vs. gold judgements. Once get the results, we could measure rank correlation of systems under crowd vs. gold evaluation.

[1] `http://trec.nist.gov/trec_eval/`

Appendix A: Pilot Relevance Judgement

Data Format	Batch	Verification	Round	Reward (cents)	#Assignment (A)	#Response: Accept (B) Reject (C) Total (D)	#Corr. Resp. (E)	Participant Rate (D/A)	Effectiveness (B/D)	Accuracy (E/B)
Audio Clips	1	N	1	1	75	13 / 8 / 21	1	0.28	0.619	0.077
			2	2	62	11 / 7 / 18	5	0.08	0.611	0.455
			3	3	51	21 / 5 / 26	4	0.51	0.808	0.190
			subtotal		188	45 / 20 / 65	10	0.35	0.692	0.22
	2	N	1	6	75	11 / 17 / 28	0	0.37	0.393	0
			2	9	64	6 / 10 / 16	1	0.25	0.375	0.167
			3	12	152 (58)	79 / 5 / 84	22	0.55	0.940	0.278
			subtotal		291	96 / 32 / 128	23	0.44	0.75	0.24
	3	Y	1	3	75	3 / 0 / 3	0	0.04	1.0	0
			2	6	72	0 / 0 / 0	0	0	0	0
			3	9	90 (72)	15 / 0 / 15	2	0.17	1.0	0.13
			subtotal		237	18 / 0 / 18	2	0.07	1.0	0.11
Best Transcripts	1	N	1	3	75	34 / 37 / 71	5	0.946	0.479	0.147
			2	6	41	30 / 11 /41	8	1.0	0.732	0.267
			3	9	11	2 / 9 / 11	0	1.0	0.182	0
			subtotal		127	66 / 57 / 123	13	0.968	0.537	0.197
	2	Y	1	3	75	22 / 26 / 48	7	0.64	0.458	0.318
			2	6	53	32 / 4 / 36	6	0.679	0.889	0.188
			3	9	21	17 / 0 / 17	8	0.810	1.0	0.471
			subtotal		149	71 / 30 / 101	21	0.678	0.703	0.296

Table A.1: Results of pilot relevance judgement

Appendix B: HIT Interface Template

This appendix illustrates the template of HITs in the experiments. As mentioned in Chapter 3, some wording and instructions may be changed iteration to iteration, but the main frame is stable.

NOTES:

- The template demonstrated below is used for iteration #8 in, and the conversational segment uses the best transcripts.
- Tags with "< ... >" stand for variables to be replaced by the content.

INSTRUCTIONS

This task is to identify relevant topics for part of an important historical interview in which a survivor recounts his/her personal experiences from the World War II Jewish Holocaust. Interviews may describe disturbing events experienced by the person during the Holocaust, so we fully understand if some people are uncomfortable performing this task, despite its historical significance. This task will help us to build a better search interface for these historical interviews for the benefit of society and future generations, so we greatly appreciate your assistance with this work.

Please read the **ENTIRE** transcript carefully and then perform the following tasks:

- Is the transcript of the conversation easy for you to understand? If not, why not?
- Please summarize the conversation.
- Which of listed topics are relevant to the conversation? Why do you find them to be relevant?

The result of Task Question #3 will be first evaluated with F-score which is based on Recall and Precision.

- Precision is to measure how many topics among your selection are TRUE relevant to conversation.
- Recall is to measure how many TRUE relevant topics are selected by you, compared with the standard result.
- F-score is a combination of Precision and Recall, i.e. $F - score = 2 * Precision * Recall/(Precision + Recall)$.

Please refer to `http://en.wikipedia.org/wiki/Precision_and_recall` for more details.

You must complete this task within 30 minutes once you accept it.

Your work will be approved only if

1. all questions are answered, AND
2. the F-score is equal or greater than 0.5, AND
3. you provide a reasonable justification for each topic you mark as relevant.

Bonus: You have two chances to earn bonuses, $1.00 for each, if you provide

- a comprehensive summary in Task Question #2, compared with other submissions
- reasonable and high-quality justifications for all topics you mark as relevant in Task Question #3

TRANSCRIPT

$< Transcript\ Segment >$
...
...

TASK QUESTIONS

1. **Did you find the transcript easy to understand?**

   ○ ○ ○ ○ ○

   Strongly disagree Disagree Neutral Agree Strongly agree

   If it was not easy to understand, why? Choose all that apply.

   ☐ Speakers switch from one to another frequently.

   ☐ There are many pauses and interruptions in the context.

   ☐ The speakers repeated and revised what they were saying frequently.

   ☐ Knowledge of the background and history are needed to understand what they are talking about.

   ☐ Other (Please explain): [ ]

2. **Please summarize the content of the transcript in 2-3 sentences in the text box below**

   [ ]

3. **Which of the following topic(s) is/are relevant to the conversation? Choose all that apply. For any topics you identify as relevant to the conversation, please provide a brief explanation for why you find the given topic to be relevant.**

Relevant?	Topic / Justification
	$< TOPIC_1 >$
☐	
	$< TOPIC_2 >$
☐	
	...
	$< TOPIC_n >$
☐	

Appendix C: Justifications with Workers' Own Words

In this appendix, the justifications provided by the turkers in the iteration #8 are listed by topics. The relevant topic in the gold standard is underlined on its ID. The justifications of the topics that were marked as relevant by turkers are in *italic*, otherwise they are in plain text.

C.1 JUSTIFICATIONS FOR VHF00058-067118.004 (V004)

1166 : Hasidim and their unquestioning faith.

W3-004: Faith was not really discussed.

1330 : Eyewitness accounts that describe the personalities and actions of Raoul Wallenberg and Adolf Eichmann.

W3-004 : Neither of those names were mentioned.

W4-004 : *It is important for people to understand how these men were.*

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W3-004 : *Since the discussion was from the point of view of someone from the local population, this would be relevant.*

1551 : Ghetto life, esp. Warsaw Ghetto.

W3-004 : There was no mention of Ghetto life.

W4-004 : *Important from a historical perspective.*

3015 : Accounts by Jewish survivors from the sites of mass shootings, and by those who personally witnessed the Jews being marched to the sites of mass shooting.

W1-004 : *This is a first hand account of someone who was a participant of a forced march of Poles and Jews during which there were numerous killings.*

W3-004 : *The shooting of several captives was discussed so this would be relevant.*

W4-004 : *So no one forgets the brutality of the Holocaust.*

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

W3-004 : Buchenwald and Dachau were not mentioned.

W4-004 : *People need to understand the realization of hope that these people felt.*

3005 : Experiences on the death marches conducted by the SS to evacuate the concentration camps as the allied armies approached.

W2-004 : *This is the only option for things I have enough information in this account to respond to. The other options mention specific names or events that may or may not apply to this particular story. But in this transcript, the poignant experience of a Death March is certainly detailed.*

W3-004 : *The interviewee was marched 23 kilometers without food or water and many with him on the march died.*

1345 : Attitudes and feelings of inmates of Birkenau and Buchenwald towards the possibility of Allied air strikes. How affected by the bombing of the factory adjacent to the camp August 24,1944

W3-004 : The names Buchenwald and Dachau were not mentioned.

C.2 JUSTIFICATIONS FOR VHF00058-067132.005 (V005)

1746 : Life in the concentration camp, specifically Auschwitz and Birkenau.

1624 : Jewish Women Couriers in Poland during the Holocaust.

W4-005 : *they fed and cared for people*

3005 : Experiences on the death marches conducted by the SS to evacuate the concentration camps as the allied armies approached.

W1-005 : *"And they marched us the same distance probably from Biala to Rawa."*

W2-005 : *The interviewee was being marched/transported around Poland before escaping.*

W4-005 : *i imagine the groups mentioned were part of the evacuation*

3015 : Accounts by Jewish survivors from the sites of mass shootings, and by those who personally witnessed the Jews being marched to the sites of mass shooting.

W1-005 : *"They counted off twenty people, said run, and opened fire."*

W2-005 : *This person tells the story of being told to run by the SS, who then shot at them.*

W4-005 : *the story is about a survivor who witnessed shootings and father was shot*

1897 : Issues of the overcrowded accommodation in the ghettos during the Holocaust.

W3-005 : *they were stationed once in a church for 500 people and there was over 5000 in that church*

W4-005 : *thousands were forced into a building to accommodate 500*

3026 : Resistance other than organized resistance groups or partisans, both pre-war and during the war, in or outside the camps.

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

2384 : Reports on material assistance provided by the Red Cross to Holocaust survivors (food, shelter, transportation, emigration / immigration)

W2-005 : *When they got off the train at Braslau, they were given bread, cheese, and coffee by the Red Cross.*

W3-005 : *women with red crosses gave out food to the jews in this time*

W4-005 : *this survivor was provided food and assistance*

W5-005 : *In this interview, the individual gives an account about being able to see the Red Cross and having the women who were involved in the organization provide them with real food after having very limited access to it or having only water. They were given bread, cheese, and coffee by the Red Cross.*

C.3 JUSTIFICATIONS FOR VHF00058-067153.007 (V007)

2224 : Religious observances in the ghettos and in DP camps with particular emphasis on ceremonies and rituals related to death and burials.

W1-007 : *"When we came in to the ghetto, there was no apartments. We moved in with a couple that had a 4 or 5 year old child. My mother was still Kosher."*

1288 : Strengthening religious faith as a result of Holocaust experience

3013 : Accounts regarding Poland's Pre-war Yeshiva and its influence on its graduates and their descendants.

3004 : Experiences in locations and buildings were Jews were assembled for transportation to concentration and death camps (assembly points).

W2-007 : *The speaker mentioned being beaten and having their possessions taken before being taken to the ghetto itself.*

W3-007 : *The warsaw ghettos eventually fed into a concentration camp.*

1887 : Stories of people who were interned in one of the Volkswagen armaments labor camps.

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

3014 : Discussions pertaining to Eastern European Zionist movements, especially Hashomer Hatzair.

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W5-007 : *"Judenrats" - individuals secretly working for the Nazis, are described as tricking the selected population into believing they would be allowed to take their possessions, less furniture, with them into the ghetto. In fact all possessions were confiscated upon arrival.*

1551 : Ghetto life, esp. Warsaw Ghetto.

W1-007 : *The whole transcript is about this topic.*

W2-007 : *He speaks of the harsh conditions, cramped living quarters, lack of food, and people dying regularly, even in the streets.*

W3-007 : *This was all about the Warsaw ghetto and how they were told to move there and what they could carry and a bit about their day to day life.*

W4-007 : *yes the ghetto life was filled with people that were dieing and the ghetto did not have enough food*

W5-007 : *Upon enter the ghetto, dead people are seen lying in the streets. Aside from laborers, people were forced to try to survive on the foods supplied by their captors, which were insufficient to support life. People coming into the ghetto were forced to live with families already there.*

2198 : Interested in descriptions of the daily horror witnessed by these Sonderkommando units. Also want information about the events that culminated in the blowing up of Crematorium III in Birkenau on Octorber 7, 1944 (Sonderkommando Uprising).

1897 : Issues of the overcrowded accommodation in the ghettos during the Holocaust.

W1-007 : *"When we came in to the ghetto, there was no apartments. We moved in with a couple that had a 4 or 5 year old child."*

W2-007 : *He talks about their being no apartments, and having to live with a family and their child. He doesn't speak of how many people were in his family specifically, but mentions his mother being in that apartment as well.*

W3-007 : *The interview at the end explained the shared accomodations in the warsaw ghetto.*

W4-007 : *again there was overcrowding and people could not cook food and after a while it was so full it was not livable*

W5-007 : *Brief mention is made of having to move in with people already there due to a lack of enough apartments.*

C.4 JUSTIFICATIONS FOR VHF00058-067222.012 (V012)

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

W3-012 : *This topic is relevant because the transcript above recounts a survivor's experience while in a camp. It would be nice to have an account of the camps liberation, to see "how the story ends" so to speak.*

1620 : Nursing and patient care provided by Jewish nurses in concentration camps during the Holocaust.

W3-012 : *This information would be relevant, as it pertains to conditions within the concentration camps. I'm certain the nurses themselves have plenty of interesting (and relevant) stories to tell.*

3016 : Forced labor making bricks.

W3-012 : *Though people being forced to make bricks is much less severe than what people in concentration camps had to deal with, their stories would be relevant,*

as they too are being forced to do things against their will. Their plight is at least somewhat similar to the plight of those who were in concentration camps (though on a very different scale).

3026 : Resistance other than organized resistance groups or partisans, both pre-war and during the war, in or outside the camps.

W1-012 : *The man resisted by fighting back against the man who beat him.*

W2-012 : *This answer seems to be the only and most accurate one for this short tale in that it describes the "one-man" resistance put up by the captive and his escape. [As a side note, these transcripts are hard to read in the sense that it's emotional and sad, these stories are, but this one is uplifting at least in the short term since it ends with his escape]*

W3-012 : *Stories such as these would be relevant to show how individuals tried to make a difference. We are frequently told of group efforts, but rarely hear about individual efforts. A story telling of an individual's efforts would be much more interesting, and would probably be easier for most people to relate to.*

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W3-012 : *This information would be relevant, because it would explain how authorities were handling matters during the war, and whether they were helping or hurting the situation. It would also be relevant to discuss how their lack of action allowed the concentration camps to be open for so long.*

1508 : The fate of Mischlinge persons.

W3-012 : *Learning the fate of people who survived the concentration camps would be extremely relevant. It would be interesting to know the impact the war had on the group as a whole, since many were killed in such a short time span.*

W4-012 : *One man decided to resist while working at the camp he was afraid of getting killed if he didn't do a good job but he still got beat for doing what he was supposed to. Eventually they were going to kill the man but he escaped the death camp and that saved his life.*

C.5 JUSTIFICATIONS FOR VHF00058-067232.013 (V013)

1551 : Ghetto life, esp. Warsaw Ghetto.

W2-013 : *This might be relevant but this part of the story doesn't address the ghetto specifically. It might be useful background knowledge to build a better understanding of where he is running to.*

W3-013 : Not relevant to the conversation. No mention of specific living conditions besides hiding away in a bed for 2 weeks.

W4-013 : *Warsaw is mentioned in the story*

15602 : Stories of heroic acts or activities that led to the survival of one or more individuals are desired.

W1-013 : *The man who survives to tell the story talks about what he has done in order to do so.*

W2-013 : *This is an action packed story, the narrator is definitely heroic.*

W3-013 : Not justified by this story. His trek did not save anybody else and he left his own mother behind.

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W2-013 : *The narrator is clearly distrustful of some of the locals he encounters on his journey, he understands that they have a relationship with the German Authorities and might give him away.*

W3-013 : *The farmer told him that if the other people staying at the farm knew he was there then they would inform the German authorities. Indicates that there were certain residents in East Central Europe who would inform the Germans of where Jews were staying.*

W4-013 : *people were working together to hide jewish identity*

1508 : The fate of Mischlinge persons.

W3-013 : No indication from the transcript that this man is a Mischlinge.

3027 : Separation from loved ones in the Holocaust.

W1-013 : *The man must leave some of his family behind.*

W2-013 : *Having read a number of horrible WWII stories I have a feeling that the narrator might not see his parents again, but I am just drawing a conclusion.*

W3-013 : *Had to leave his mother behind while he traveled to Warsaw because she would just slow him down. This was probably a common theme between the younger and elder people in the family*

1288 : Strengthening religious faith as a result of Holocaust experience.

W3-013 : *Constantly reminding himself to praise God in polish. Indicates that these peoples have faith in their God. God will guide them to the other side of this terrible time of treachery.*

1429 : Describe impressions, drama, and personal experiences of German and Austrian Jews who escaped the Final Solution and returned as American soldiers to fight Hitler in Europe and North Africa 1942-1945

W3-013 : *This transcript describes the personal experience of one man traveling to somewhere he would find safer than where he was.*

1311 : We are interested in stories of children hidden without their parents and of their rescuers.

W3-013 : Not relevant here.

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

W3-013 : Not relevant here.

Appendix D: Justifications with Quotations in Transcripts

In this appendix, the justifications provided by the turkers in the iteration #9 are listed by topics. The relevant topic in the gold standard is underlined on its ID. The justifications of the topics that were marked as relevant by turkers are in *italic*, otherwise they are in plain text.

D.1 Direct Justifications for VHF00058-067118.004 (V004)

1166 : Hasidim and their unquestioning faith.

W1-004 : *Many of the Jewish people inflicted by the German brutality were unquestioning in their faith, and they remained devout through times of torture. One example given by our narrator is, "He was a heavy Hasidic Jew. And he couldn't keep up. They tied a chain around his waist. And tied him to the half track. He should walk but he couldn't. He fell.That's how he died."*

1330 : Eyewitness accounts that describe the personalities and actions of Raoul Wallenberg and Adolf Eichmann.

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W1-004 : *The narrator gives insight toward the collaboration of the local population with the German Authorities through a governmental takeover. He writes, "The first thing they did is take out twelve Jews that belonged to the city government in the the market place."*

1551 : Ghetto life, esp. Warsaw Ghetto.

W4-004 : *The speaker seems to be talking about Ghetto life, specifically in Warsaw when he mentions that they were still fighting for Warsaw. He describes how*

they were brought to large ranch and given small amounts of food. This is relevant because it is the makings of ghetto life, the cramped quarters, lack of food and soldiers constant watch. This where the speaker ends up, so he is describing it.

W5-004 : *The references to Poland, where Warsaw is located.*

3015 : Accounts by Jewish survivors from the sites of mass shootings, and by those who personally witnessed the Jews being marched to the sites of mass shooting.

W1-004 : *Many Jewish people were shot without reason. They merely could not continue marching under such horrid conditions. These conditions are brought to light when the narrator writes, "They marched us on a dirt road to a city Biala without water, without food. If you're on for a little water people used to hand out, you were shot."*

W2-004 : This boy has witnessed death and violence on a forced walk by the Nazis. He even says, "Nobody was used to walk twenty three kilometers in one day."

W3-004 : *They picked up men from fifteen to fifty and me and my father. Next morning they marched us out, Poles and Jews. Four people they shot. Four were brought back. They put up a gallow. And we had to see them hang all of them four.*

W4-004 : *This is an eye witness account because he talks about hearing shots fired, realizing that soldiers shot people who tried to escape. They were mostly likely marched out, shot and then hung in gallows to instill fear. This is also relevant because it shows the harsh realities of the war and how these people lived in constant fear. It sets the tone for the entire transcript.*

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

3005 : Experiences on the death marches conducted by the SS to evacuate the concentration camps as the allied armies approached.

W2-004 : *This boy has witnessed death and violence on a forced walk by the Nazis. He even says, "Nobody was used to walk twenty three kilometers in one day."*

1345 : Attitudes and feelings of inmates of Birkenau and Buchenwald towards the possibility of Allied air strikes. How affected by the bombing of the factory adjacent to the camp August 24,1944

D.2 DIRECT JUSTIFICATIONS FOR VHF00058-067132.005 (V005)

1746 : Life in the concentration camp, specifically Auschwitz and Birkenau.

W5-005 : *They were only given water to drink and no food to eat.*

1624 : Jewish Women Couriers in Poland during the Holocaust.

3005 : Experiences on the death marches conducted by the SS to evacuate the concentration camps as the allied armies approached.

W1-005 : *No matter how fast you walk, if you're behind, you'll always fall behind. That's what they said. That's what they did. They counted off twenty people, said run, and opened fire.*

W2-005 : *The Nazis marched Jews from town to town in this narrative, and shot those too slow to escape, but it is not entirely clear that this was after allied armies began to approach.*

W3-005 : *Marching is referenced a few times. Two examples: "And they marched us the same distance probably from Biala to Rawa." "I was always marching."*

W4-005 : *"And they marched us the same distance probably from Biala to Rawa." Was the first sentence the speaker tells us, which indicates they were marching. Then shortly after, him saying that they lost people along the way, he states "Next morning, they marched us out of Rawa" and eventually comes to tell us "This is the closest city to Poland. In BRASLAU, we found out that the Warsaw*

fell. We saw the Red Cross. They were wearing arm bands from the Red Cross. Women came and brought some fresh bread and cheese and coffee" this shows that the allied armies approached. Another clue about the allied armies were that he later states " Because the roads were crowded with army units, going back and forth."

3015 : Accounts by Jewish survivors from the sites of mass shootings, and by those who personally witnessed the Jews being marched to the sites of mass shooting.

W1-005 : *No matter how fast you walk, if you're behind, you'll always fall behind. That's what they said. That's what they did. They counted off twenty people, said run, and opened fire.*

W2-005 : *The speaker describes being shot at as the Nazis sent the Jews running, and his actions helping tend his father's gunshot wound.*

W3-005 : *The person saw his father and others shot: "They marched us out on a road. They told us that they're gonna count off twenty people at a time. And tell us to run. And they're gonna shoot." "That's what they said. That's what they did. They counted off twenty people, said run, and opened fire. A few fell. My father was hit in the thigh."*

W4-005 : *This man was clearly a survivor of the mass shootings that the Jews were marched to during the holocaust. The man begins his story with him in a group that were marching, resting, marching, etc. etc. The man tells us as one point "We traveled the whole night. We came back to SHERRATZ. They marched us out on a road. They told us that they're gonna count off twenty people at a time. And tell us to run. And they're gonna shoot. If somebody doesn't run fast, he will be shot. I was always marching." He goes on to say "That's what they did. They counted off twenty people, said run, and opened fire. A few fell." He claimed his father was shot in the thigh, but with his help, they both made it out alive and were able to walk to the nearest city to properly dress his father's*

gunshot wound before the walk home.

W5-005 : *They were told to march as they were being shot at.*

1897 : Issues of the overcrowded accommodation in the ghettos during the Holocaust.

3026 : Resistance other than organized resistance groups or partisans, both pre-war and during the war, in or outside the camps.

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

2384 : Reports on material assistance provided by the Red Cross to Holocaust survivors (food, shelter, transportation, emigration / immigration)

W2-005 : *The Red Cross provided bread, cheese, and coffee to the Jews during their forced march.*

W3-005 : *The person recounted being given bread and coffee by the red cross while being transported by the SS: "We saw the Red Cross. They were wearing arm bands from the Red Cross. Women came and brought some fresh bread and cheese and coffee."*

W4-005 : *At one point in the man's recollection of his horrific experience in the holocaust, he states "We saw the Red Cross. They were wearing arm bands from the Red Cross. Women came and brought some fresh bread and cheese and coffee.". That shows me that the red cross was providing help to those in the holocaust.*

W5-005 : *People had red crosses on their arms and gave them coffee and bread.*

D.3 DIRECT JUSTIFICATIONS FOR VHF00058-067153.007 (V007)

2224 : Religious observances in the ghettos and in DP camps with particular emphasis on ceremonies and rituals related to death and burials.

W2-007 : *My mother was still Kosher. Most people were a long time in the ghetto. They were cooking whatever they could get. So my mother used to put in a BOAT. It shouldn't splash from the other side to her side of the stove. After a few months, it didn't work.*

1288 : Strengthening religious faith as a result of Holocaust experience

3013 : Accounts regarding Poland's Pre-war Yeshiva and its influence on its graduates and their descendants.

3004 : Experiences in locations and buildings were Jews were assembled for transportation to concentration and death camps (assembly points).

W3-007 : *they took us into where the SS was staying. They took everything away, beat the hell out of us, and told us to go*

W4-007 : *Jews were moved en masse to the Warsaw ghetto, where it would have been easy to send them elsewhere.*

1887 : Stories of people who were interned in one of the Volkswagen armaments labor camps.

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

3014 : Discussions pertaining to Eastern European Zionist movements, especially Hashomer Hatzair.

W2-007 : *n meantime the Warsaw ghetto was formed. And they start putting people from all over. And we were told we're gonna move to the Warsaw ghetto. This was the beginning of nineteen forty. And they told us to prepare. The Judenrat told us to prepare, putting everything in suitcases. You couldn't take any furniture. You could only take what you could carry. And they tricked us to put everything together. When we had to go to the ghetto, they took us into where*

the SS was staying. They took everything away, beat the hell out of us, and told us to go.

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W4-007 : *The Judenrat collaborated with the Nazis by telling Jews to pack their belongings in suitcases, though the Nazis conveniently stole everything that had been packed together like this.*

W5-007 :

1551 : Ghetto life, esp. Warsaw Ghetto.

W1-007 : *In meantime the Warsaw ghetto was formed. When we had to go to the ghetto, they took us into where the SS was staying. They took everything away, beat the hell out of us, and told us to go. When we came into the Warsaw ghetto, we were already beaten, downtrodded.*

W2-007 : *They took two hundred people. They marched them out, so at least they got food where they worked building air fields and unloading under train loads with ammunition, all kind of things. And there's no work, only rations they gave you. Even a child couldn't survive.*

W3-007 : *And we came into the ghetto. You saw people dying in the street They were cooking whatever they could get. So my mother used to put in a BOAT. It shouldn't splash from the other side to her side of the stove. After a few months, it didn't work. We were all on the same boat.*

W4-007 : *Ghetto life involved hard work and little food, and though the speaker's mother had kept Kosher, many Jews were unable to do so and survive.*

W5-007 : *He states that their were no rations, that they would eat whatever the could get, that you saw people dying in the streets, that he was forced to live with another family, and that people in the Ghetto were forced to work.*

2198 : Interested in descriptions of the daily horror witnessed by these Sonderkommando units. Also want information about the events that culminated in the blowing up of Crematorium III in Birkenau on Octorber 7, 1944 (Sonderkommando Uprising).

1897 : Issues of the overcrowded accommodation in the ghettos during the Holocaust.

W1-007 : *When we came in to the ghetto, there was no apartments. We moved in with a couple that had a 4 or 5 year old child.*

W2-007 : *When we came in to the ghetto, there was no apartments. We moved in with a couple that had a 4 or 5 year old child.*

W3-007 : *When we came in to the ghetto, there was no apartments. We moved in with a couple that had a 4 or 5 year old child*

W4-007 : *The speaker and his family moved into an apartment that already housed a family with a child.*

D.4 DIRECT JUSTIFICATIONS FOR VHF00058-067222.012 (V012)

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

1620 : Nursing and patient care provided by Jewish nurses in concentration camps during the Holocaust.

3016 : Forced labor making bricks.

3026 : Resistance other than organized resistance groups or partisans, both pre-war and during the war, in or outside the camps.

W1-012 : *And I took the pitchfork And I put it in his belly. And I left it there.*

W2-012 : L*et's say you got up 5:00. You took him for an inspection. They were wearing white gloves. Go against the grain. Every time I got a beating. They're not clean enough. And this was going on for weeks.*

W3-012 : *I was restless again. Because if you stay in a situation like this, you know that the end will come. They'll beat you to death. he tried to hit me again. And I took the pitchfork And I put it in his belly. I escaped.*

W4-012 : *He resisted on his own by stabbing the abusive guard and fleeing captivity.*

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W5-012 : *References to the bodies and the SS.*

1508 : The fate of Mischlinge persons.

W2-012 : *Not till I came to, he tried to hit me again. And I took the pitchfork And I put it in his belly. And I left it there. I opened the door to the back field. And I escaped.*

D.5 Direct Justifications for VHF00058-067232.013 (V013)

1551 : Ghetto life, esp. Warsaw Ghetto.

W4-013 : *His sisters worked in Warsaw and he hid in the ghetto after escaping.*

15602 : Stories of heroic acts or activities that led to the survival of one or more individuals are desired.

W1-013 : *At 4:00 he woke me up in the morning. He brought me a half of bread. And then they said you could stay overnight. And they gave me some food. And I helped them, to finish off their chores in the evening. They took me in. They gave me some food and asked me questions. I was hidden in a bed, covered up for 2 weeks. I didn't get out. A whole day I was lying in a bed, covered up over my face, over my head, with a blanket.*

W2-013 : *And I say, "Shlomo, give me an arm band. Because without that arm band you're dead." He says, "You want to tell me you are here? Every SS is*

looking for you. The man died. The Folge Deutsch died". I said, "Don't give me questions and answers". If they catch me, it's too bad. Give me an arm band". They gave me an arm band.

W3-013 : *The woman's tale is heroic, as were the actions of people who helped her, and the actions of the man who gave her an arm band. This woman sought to survive and had to be very brave to set out on her journey, giving the risks of getting caught and the price that would be paid if she were caught.*

W4-013 : *People risked their lives to help the escapee and hide him when it was dangerous to harbor escaping Jews.*

2012 : Information about collaboration of the local population with German Authorities in East Central Europe during the Holocaust.

W1-013 : *"But you have to sleep on the other side from those people. Because if they know you're Jewish, they'll take you to the German like it's night now". Germans were marching recruits. And I encounter an assessment in a Folge Deutsch, the biggest bastard you ever saw, to tell it on the tape. And they say that the foreman is a German. And he doesn't let one more Jew come in here.*

W2-013 : *It's in the outskirts of Warsaw. Germans were over the Polish girls. And it's like a laughing, joking.*

W5-013 : *Till somebody squealed. This is an example of someone collaborating with the German authorities by turning someone in.*

1508 : The fate of Mischlinge persons.

3027 : Separation from loved ones in the Holocaust.

W1-013 : *"I'm going back to Warsaw where my sisters are working". My mother says she wants to go with me. I say the same thing like you went to the village. "You're just gonna hold me back because it's 77 kilometers. It's not easy for a middle-aged woman to walk". I talked her out of it.*

W2-013 : *My mother says she wants to go with me. I say the same thing like you went to the village. "You're just gonna hold me back because it's 77 kilometers. It's not easy for a middle-aged woman to walk". I talked her out of it.*

W3-013 : *This story speaks of separation from parents and siblings and really drives home how families were separated due to fear, the need to hide, etc.*

W4-013 : *He had been separated from his family, and though he briefly met his mother, he had to leave her behind on his journey to Warsaw.*

W5-013 : *My mother says she wants to go with me. This is an example of the protagonist having to leave his mother.*

1288 : Strengthening religious faith as a result of Holocaust experience.

1429 : Describe impressions, drama, and personal experiences of German and Austrian Jews who escaped the Final Solution and returned as American soldiers to fight Hitler in Europe and North Africa 1942-1945

1311 : We are interested in stories of children hidden without their parents and of their rescuers.

W1-013 : *He says, "I know you're SONA's son. I recognize you. I have 6 people working, just picking the grain from the fields. I could keep you only over the night". And I said the same thing, "Praise God" in Polish. And then they said you could stay overnight. I waited till nightfall. They took me in.*

W3-013 :*This person clearly had to hide on her own. The narrative indicates her mother may have been too slow for them to safely escape.*

1225 : Witness accounts to the liberation of Buchenwald and Dachau concentration camps.

Bibliography

[1] Eytan Adar. Why I Hate Mechanical Turk Research (and Workshops). In *CHI'11 Workshop on Crowdsourcing and Human Computation*, May 2011.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] Kartik Audhkhasi, Panayiotis Georgiou, and Shrikanth S. Narayanan. Accurate transcription of broadcast news speech using multiple noisy transcribers and unsupervised reliability metrics. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 4980–4983, May 2011.

[4] Jeff Barr and Luis Felipe Cabrera. AI gets a brain. *Queue*, 4(4):24–29, May 2006.

[5] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. Soylent: A word processor with a crowd inside. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 313–322, New York, NY, USA, 2010. ACM.

[6] William Byrne, David Doermann, Martin Franz, Samuel Gustman, Jan Hajic, Douglas Oard, Michael Picheny, Josef Psutka, Bhuvana Ramabhadran, Dagobert Soergel, Todd Ward, and Wei-Jing Zhu. Automatic recognition of spontaneous speech for access to multilingual oral history archives. *Speech and Audio Processing, IEEE Transactions on*, 12(4):420–435, July 2004.

[7] David L. Chen and William B. Dolan. Building a persistent workforce on mechanical turk for multilingual data collection. In *Proceedings of The 3rd Human Computation Workshop (HCOMP 2011)*, August 2011.

[8] Djellel Eddine Difallah, Gianluca Demartini, and Philippe Cudré-Mauroux. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In Ricardo A. Baeza-Yates, Stefano Ceri, Piero Fraternali, and Fausto Giunchiglia, editors, *Proceedings of the 1st International Workshop on Crowdsourcing Web Search*, volume 842 of *CEUR Workshop Proceedings*, pages 26–30. CEUR-WS.org, April 2012.

[9] Carsten Eickhoff and ArjenP. de Vries. Increasing cheat robustness of crowdsourcing tasks. *Information Retrieval*, 16(2):121–137, 2013.

[10] Keelan Evanini, Derrick Higgins, and Klaus Zechner. Using amazon mechanical turk for transcription of non-native speech. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 53–56, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] Jonathan G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, pages 347–354, Dec 1997.

[12] Karn Fort, Gilles Adda, Benot Sagot, Joseph Mariani, and Alain Couillault. Crowdsourcing for language resource development: Criticisms about amazon mechanical turk overpowering use. In Zygmunt Vetulani and Joseph Mariani, editors, *Human Language Technology Challenges for Computer Science and Linguistics*, volume 8387 of *Lecture Notes in Computer Science*, pages 303–314. Springer International Publishing, 2014.

[13] John S Garofolo, Cedric GP Auzanne, and Ellen M Voorhees. The trec spoken document retrieval track: A success story. *NIST Special Publication: The 8th Text REtrieval Conference (TREC 8)*, 500(246):107–130, 2000.

[14] Alexander Gruenstein, Ian McGraw, and Andrew M. Sutherl. A self-transcribing speech corpus: collecting continuous speech with an online educational game. In *Proceedings of ISCA International Workshop on Speech and Language Technology in Education (SLaTE)*, 2009.

[15] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA, 2000. ACM.

[16] Xiaoli Huang and Dagobert Soergel. Relevance judges' understanding of topical relevance types: An explication of an enriched concept of topical relevance. *Proceedings of the American Society for Information Science and Technology*, 41(1):156–167, 2004.

[17] Lilly C. Irani and M. Six Silberman. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 611–620, New York, NY, USA, 2013. ACM.

[18] Douglas A. Jones, Florian Wolf, Edward Gibson, Elliott Williams, Evelina Fedorenko, Douglas A. Reynolds, and Marc A. Zissman. Measuring the readability of automatic speech-to-text transcripts. In *8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, 2003*, pages 1585–1588, 2003.

[19] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[20] Aniket Kittur, Jeffrey V. Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, pages 1301–1318, New York, NY, USA, 2013. ACM.

[21] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors, *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*, pages 282–289. Morgan Kaufmann, 2001.

[22] Chia-ying Lee and James Glass. A transcription task for crowdsourcing with automatic quality control. In *12th Annual Conference of the International Speech Communication Association (Interspeech)*, pages 3041–3044, 2011.

[23] Yang Liu, Elizabeth Shriberg, Andreas Stolcke, Dustin Hillard, Mari Ostendorf, and Mary Harper. Enriching speech recognition with automatic detection of sentence boundaries and disfluencies. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(5):1526–1540, Sept 2006.

[24] Matthew Marge, Satanjeev Banerjee, and Alexander Rudnicky. Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization. In *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 99–107, 2010.

[25] Matthew Marge, Satanjeev Banerjee, and Alexander I Rudnicky. Using the amazon mechanical turk for transcription of spoken language. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5270–5273, March 2010.

[26] Katharine Mieszkowski. "I make $1.45 a week and I love it". *Salon*, July 24 2006. `http://www.salon.com/2006/07/24/turks_3/`.

[27] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z. Gajos. Platemate: Crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 1–12, New York, NY, USA, 2011. ACM.

[28] Scott Novotney and Chris Callison-Burch. Cheap, fast and good enough: Automatic speech recognition with non-expert transcription. In *Human Language Technologies:*

The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, HLT '10, pages 207–215, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[29] Douglas W. Oard, Dagobert Soergel, David Doermann, Xiaoli Huang, G. Craig Murray, Jianqiang Wang, Bhuvana Ramabhadran, Martin Franz, Samuel Gustman, James Mayfield, Liliya Kharevych, and Stephanie Strassel. Building an information retrieval test collection for spontaneous conversational speech. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 41–48, New York, NY, USA, 2004. ACM.

[30] DouglasW. Oard, Jianqiang Wang, GarethJ.F. Jones, RyenW. White, Pavel Pecina, Dagobert Soergel, Xiaoli Huang, and Izhak Shafran. Overview of the clef-2006 cross-language speech retrieval track. In Carol Peters, Paul Clough, FredricC. Gey, Jussi Karlgren, Bernardo Magnini, DouglasW. Oard, Maarten Rijke, and Maximilian Stempfhuber, editors, *Evaluation of Multilingual and Multi-modal Information Retrieval*, volume 4730 of *Lecture Notes in Computer Science*, pages 744–758. Springer Berlin Heidelberg, 2007.

[31] Gabriel Parent and Maxine Eskenazi. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 312–317. IEEE, Dec 2010.

[32] Xian Qian and Yang Liu. Disfluency detection using multi-step stacked learning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 820–825. Association for Computational Linguistics, 2013.

[33] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65, November 1987.

[34] Benjamin Taskar, Carlos Guestrin, and Daphne Koller. Max-margin markov networks. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 25–32. MIT Press, 2004.

[35] Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 315–323, New York, NY, USA, 1998. ACM.

[36] Ryen W. White, Douglas W. Oard, Gareth J.F. Jones, Dagobert Soergel, and Xiaoli Huang. Overview of the clef-2005 cross-language speech retrieval track. In Carol Peters, FredricC. Gey, Julio Gonzalo, Henning Mller, GarethJ.F. Jones, Michael Kluck,

Bernardo Magnini, and Maarten de Rijke, editors, *Accessing Multilingual Information Repositories*, volume 4022 of *Lecture Notes in Computer Science*, pages 744–759. Springer Berlin Heidelberg, 2006.

[37] G. N. Wilkinson and C. E. Rogers. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 22(3):pp. 392–399, 1973.

[38] Jason D. Williams, I. Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. Crowd-sourcing for difficult transcription of speech. In *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pages 535–540, Dec 2011.

[39] Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. Multidimensional relevance modeling via psychometrics and crowdsourcing. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '14, pages 435–444, New York, NY, USA, 2014. ACM.

[40] Haofeng Zhou, Denys Baskov, and Matthew Lease. Crowdsourcing transcription beyond mechanical turk. In *Scaling Speech, Language Understanding and Dialogue through Crowdsourcing Workshop, Workshop in the First AAAI Conference on Human Computation and Crowdsourcing (HCOMP-13), AAAI Technical Report WS-13-25*, 2013.