Supplementary Material: Effectively Crowdsourcing Radiology Report Annotations

Anne Cocos[1,2], Aaron J. Masino[1], Ting Qian[1], Ellie Pavlick[2], and Chris Callison-Burch[2]
Department of Biomedical and Health Informatics, The Children's Hospital of Philadelphia[1]
Computer and Information Sciences Department, University of Pennsylvania[2]

1 Overview

This supplementary material provides additional detail on the content of radiology reports in the Audiological and Genetic Database (AudGenDB)[1], our method for collecting crowdsourced sentence annotations, and our experiments that examined the relationship between predicted sentiment and sentence class.

2 AudGenDB Radiology Report Data

2.1 Data Extraction

AudGenDB houses over 16,000 radiology images and associated text reports. In this work we focused on computed tomography (CT) scans of the temporal bones. For the purposes of classifying each report as *normal* or *abnormal* with respect to the thirteen anatomical structures listed in table 4, we performed a simple keyword search for each particular structure and returned sentences containing one or more matches. The resulting extracted dataset contained 10,880 unlabeled sentences from 2503 CT reports of the temporal bones.

2.2 Gold Standard Dataset

In table 1 we provide a randomly chosen sample of 40 expert-annotated sentences from our gold standard dataset.

3 Crowdsourcing Method

3.1 Crowdsourcing User Interface

We used the Amazon Mechanical Turk (MTurk) platform to actively crowdsource sentence annotations. Through the interface shown in figure 1 we asked each anonymous worker to label three sentences per Human Intelligence Task (HIT).

Workers were shown an excerpt from the original radiology report that contained the target sentence to be labeled plus two sentences before and after for context. We highlighted the target sentence and asked the worker:

> Does the highlighted sentence describe a normal or abnormal structure?

where the appropriate target structure was prepopulated for the target sentence. We also asked the workers

> How confident are you that your answer is correct?

and gave them three options: *Very confident*, *Somewhat confident*, and *Not at all confident*.

Since our workers were unscreened and assumed to have no expertise in medicine or radiology, we provided brief instructions and tooltip examples of *normal* and *abnormal* sentences tailored to the target structure in question. A full list of the example sentences provided is in 4.

For each unlabeled sentence in our dataset, we requested labels from at least two unique workers. In cases where the workers disagreed, we continued to collect annotations for that sentence until reaching at least 75% absolute agreement. In total we were able to crowdsource annotations for 717 sentences satisfying these criteria.

3.2 Quality Control

In each HIT we included at least one control sentence from the gold standard dataset. We rejected HITs in which the worker incorrectly labeled the control sentence or left one or more sentences unannotated. We also rejected the annotations of four workers whose overall accuracy on control sentences fell below 75%.

3.3 Worker Statistics

Our crowdsourced labor pool included 56 unique workers. The average number of sentences classified per worker was 99.875, with a handful of

[1] http://audgendb.chop.edu

Table 1: Randomly Selected Gold Standard Sentences

Gold Class	Sentence
Normal	The stapes is present, inserting into the oval window. The malleus and incus are intact. The malleus appears to be slightly elevated, however, the ossicular relationship appears intact. The inner ear structures, including the semicircular canals and vestibule are unremarkable. The cochlear aperture is not seen. The inner ear structures, including semicircular canal, vestibule and cochlea are unremarkable. The left mastoid air cells are clear. The middle ear is clear and the ossicles are intact. Normal vestibular aqueducts. There is a normal number of turns within the cochlea. The middle ear cavities, mastoid air cells, and paranasal sinuses are clear. The inner ear structures, including the cochlea, semicircular canals and internal auditory canals are unremarkable. On the right side, the external auditory canal appears patent. The mastoid air cells are well-pneumatized. The semicircular canal, vestibule, and internal auditory canal on the left are unremarkable. The vestibulo-cochlear nerves appear to be present. The incus and malleus are grossly intact. Otherwise, the middle ear cavities, mastoid air cells, and paranasal sinuses are clear. The inner ear structures, including the semicircular canals and vestibule are unremarkable.
Abnormal	Bilateral dysplastic cochlea. The tympanic membrane is slightly thickened relative to the left. As previously seen the right mastoid air cells are under pneumatized. No right cochlear aperture seen. There is slight dysmorphic appearance of the cochlea with a slightly prominent vestibule and proximal semicircular canal. That lateral semicircular canals are maldeveloped on both sides. The ossicles are abnormal on the right with abnormal globular shape and positioning. Marked narrowing of the right internal auditory canal. The bony bar in the right internal auditory canal is larger than normal, as on the contralateral side. Bilateral enlarged vestibules. Atresia of bony external auditory canals. Bilateral enlargement of vestibular aqueducts. The malleus and incus are fused and malformed with abnormal positioning. The left cochlea is also hypoplastic. Linear appearing density anteriorly along the right tympanic membrane of uncertain etiology. The lateral semicircular canal is not formed and the proximal portion is bulbous. The vestibular aqueducts of the temporal bones, bilaterally, appear enlarged on the axial T2 images. The distance between the ossicles and scutum is widened. The right incus and malleus are dysplastic and probably fused. The internal auditory canal is very small, compared to the left. The middle ear cavity is clear, however the study is notable for abnormal malleus, as the head is smaller, and the long process appears to be absent.

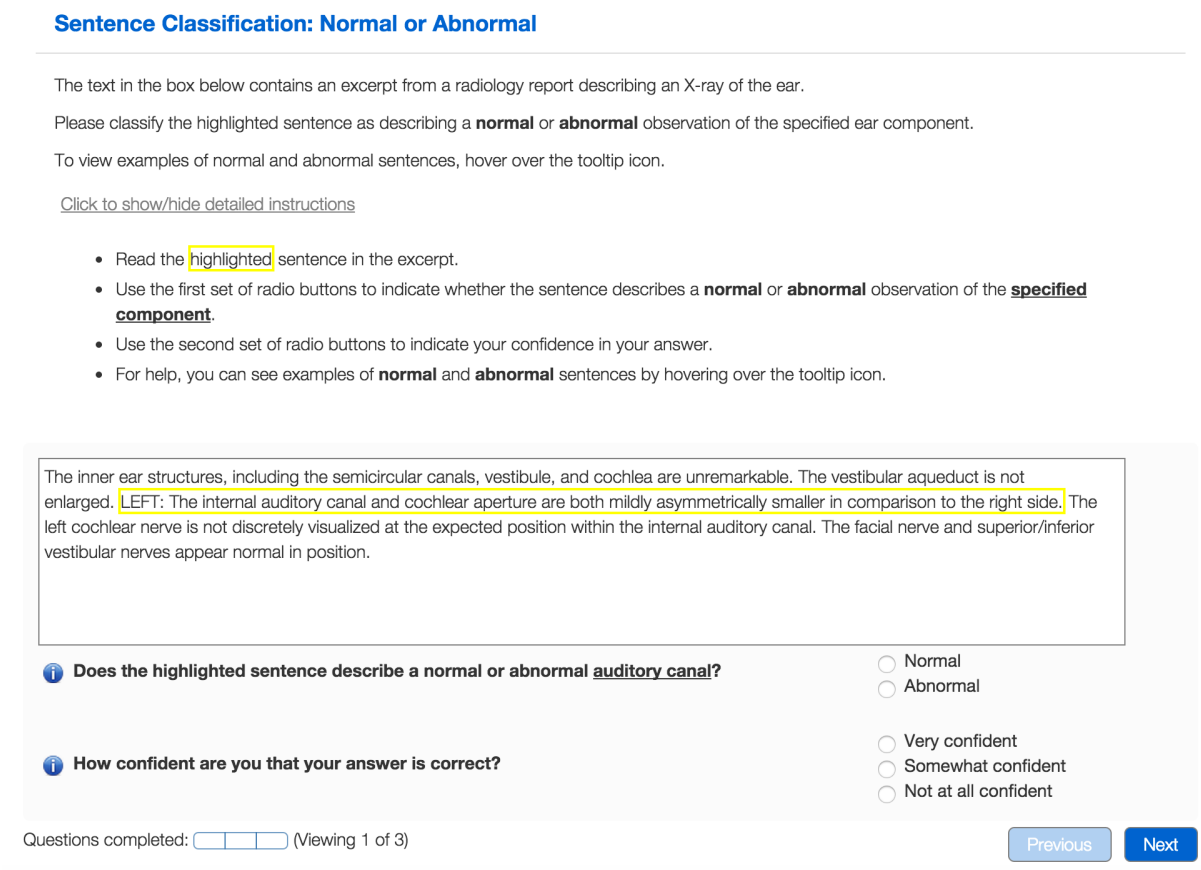


Figure 1: Annotation User Interface.

workers labeling over 200 sentences (range [3, 462]) (figure 2a). All but four workers accurately classified 80% or more of the control sentences they were presented (range [50%, 100%]) (figure 2b). The average time taken to label three sentences in each HIT was 277 seconds, or roughly 92 seconds per sentence (figure 2c).

4 Sentiment Prediction and Class Labels

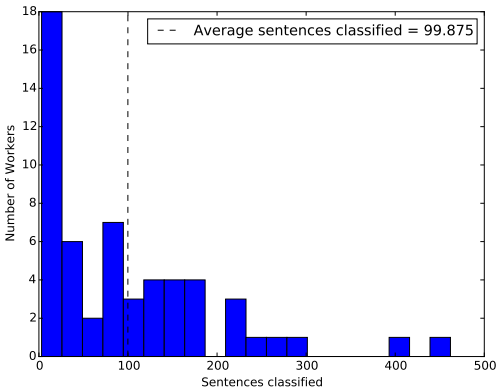
In order to investigate the extent to which non-expert MTurk workers might be able to predict *normal* or *abnormal* sentence labels based on shallow lexical or syntactic information, without domain knowledge, we ran the Stanford CoreNLP off-the-shelf sentiment prediction model on our gold standard dataset (Socher et al., 2013). Our reasoning was that if sentence classes aligned with predicted sentiment, then *normal* sentences tended to resemble typically *positive* sentences in the common language data (movie reviews) on which the model was trained.

The problem, of course, is that the sentiment lexicon for movie reviews is vastly different from that of radiology reports. For example, the word *unremarkable* is indicative of a *normal* sentence in a report, but would signify a *negative* sentiment in a movie review. For this reason, we made word substitutions for high-frequency adjectives in our gold standard data as detailed in table 2 before running sentiment prediction on it. The substitutions were chosen to more closely mirror the movie review sentiment lexicon.

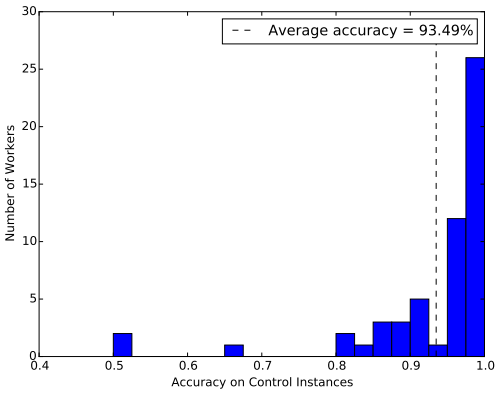
Original word	Substitution
patent unremarkable clear normal	good
abnormal	bad

Table 2: Word substitutions made in gold data prior to sentiment prediction.

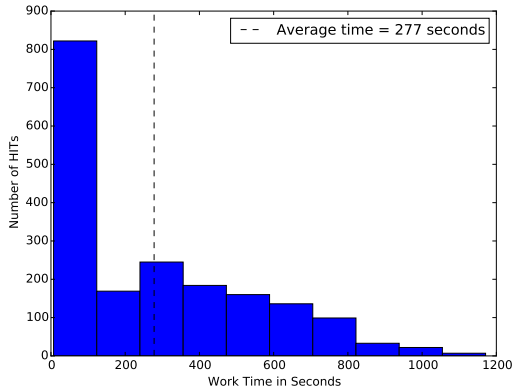
We found that the CoreNLP predicted sentiment



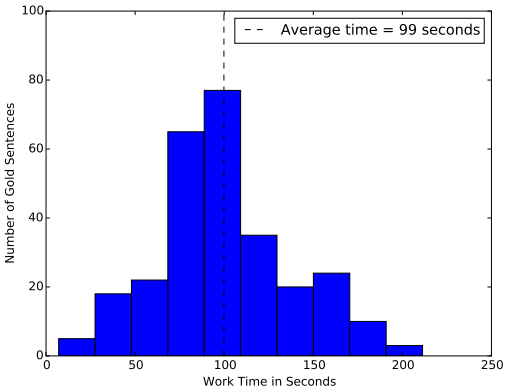
(a) Sentences classified by worker



(b) Classification accuracy by worker



(c) HIT work times



(d) Control sentence work times

Figure 2: Histograms of classification quantity, accuracy, and work time by worker

Top-5 quickest control sentences	
Sentence	Average Time (sec)
Normal appearance of cochlear nerves.	7.08
The vestibular nerves are seen and appear normal.	14.17
The middle ear cavities, mastoid air cells, and other paranasal sinuses are clear.	14.17
Probable adhesion of the anterior aspect of the head of the left malleus to the anterior wall of the attic and anterior portion of the scutum.	26.17
CLINICAL INDICATION: 6-year-old female with sensorineural hearing loss on the left, evaluate for malformation or cochlear nerve aplasia.	33.5
Top-5 slowest control sentences	
Sentence	Average Time (sec)
Narrowed left IAC with absent left cochlear nerve.	211.3
There is slight dysmorphic appearance of the cochlea with a slightly prominent vestibule and proximal semicircular canal.	205.5
The semicircular canals and cochlea are unremarkable.	182.3
The vestibular nerve is seen but there is absence of a cochlear component.	176.0
There is dilated cystic appearance of the vestibule, cochlea and the proximal portion of the superior and posterior semicircular canals.	173.0

Table 3: Top-5 'quickest' and 'slowest' control sentences based on average work time.

of our gold standard dataset sentences does correlate with the class labels (Spearman's ρ=0.49) after making the word substitutions in table 2. If we give a *normal* label to sentences predicted to have *Neutral*, *Positive*, or *Very Positive* sentiment by CoreNLP and a label of *abnormal* to the others, the sentiment analysis model correctly identifies the sentence class 70.4% of the time.

But there is a significant difference in the accuracy of the sentiment prediction model in classifying our sentences for quickly-classified gold sentences and slowly-classified gold sentences. If we estimate the average classification time for each gold sentence to be 1/3 of the total HIT work time averaged over each HIT in which that gold sentence appeared, the distribution of gold sentence classification times is shown in table 2d. We arbitrarily choose 60 seconds as a cutoff between 'quick' sentences and 'slow' sentences, and find that the sentiment prediction model 76.4% accurate in classifying 'quick' sentences and only 69.5% accurate in classifying 'slow' sentences. The correlation between predicted sentiment and sentence class for 'quick' sentences is ρ=0.65, and for 'slow' sentences ρ=0.47. These statistics seem to indicate that 'quick' sentences more closely resemble *positive* and *negative* sentiment sentences than 'slow' ones. Qualitatively, the five slowest control sentences appear significantly harder to classify than the five quickest control sentences (table 3).

References

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 254–263

Table 4: Tooltip example sentences provided to workers, by anatomical structure.

Structure	Class	Example Sentence
Scutum	Normal	The scutum is patent.
		There is no evidence of bony erosion of the ossicles or the scutum.
	Abnormal	The scutum is thickened.
		The scutum is at least partially absent.
Mastoid	Normal	The mastoid air cells are well-pneumatized and clear.
		The mastoid air cells are patent.
	Abnormal	Abnormal signal intensity in the left mastoid air cells, as described.
		As previously seen the right mastoid air cells are under pneumatized.
EAC	Normal	The EAC is patent.
		Minimal soft tissue stranding is present in the right middle ear, otherwise the EAC and the middle ears appear within normal limits.
	Abnormal	The medial bony floor of the EAC which normally shows focal thinning appears slightly thinner than on the left side.
		Because of the EAC atresia the tympanic membrane is absent and consequently the handle of the malleus is foreshortened and thickened.
Auditory Canal	Normal	On the right side, the external auditory canal is patent.
		The internal auditory canal is unremarkable.
	Abnormal	The right internal auditory canal is severely hypoplastic, but is patent.
		The external auditory canal is atretic as well as the tympanic membrane.
Vestibules	Normal	The vestibule is patent.
		The semicircular canal, vestibule, and internal auditory canal on the left are unremarkable.
	Abnormal	The right vestibule is enlarged compared to the contralateral vestibule.
		The right vestibule is enlarged.
Ossicles	Normal	The ossicles are patent.
		The middle ear structures, including the ossicles are normal.
	Abnormal	The ossicles are abnormal on the right with abnormal globular shape and positioning.
		The ossicles are markedly dysplastic.
Cochlea	Normal	The vestibulo-cochlear nerves are patent.
		The inner ear structures, including the semicircular canal, vestibule and cochlea are unremarkable.
	Abnormal	The cochlea is mildly dysplastic with incomplete partition of middle and apical turn.
		IMPRESSION: Mild dysplasia of the cochlea bilaterally.
Vestibular Aqueduct	Normal	The vestibular aqueduct is patent.
		The cochlea, vestibular aqueducts and vestibules are normal in appearance and configuration bilaterally.
	Abnormal	Bilaterally enlarged vestibular aqueducts, see comments.
		The vestibular aqueducts of the temporal bones, bilaterally, appear enlarged on the axial T2 images.
Malleus	Normal	A normal malleus and incus was identified bilaterally.

Table 4 – continued from previous page

Structure	Class	Example Sentence
		The head of the malleus is patent.
	Abnormal	Normal tympanic membranes is not seen on the right and is absent on the left, and the left malleus is abnormal in morphology and orientation.
		The ossicles appear abnormal in orientation with apparent fusion of the malleus and incus to each other as well as with the wall of middle ear cavity.
External Canal	Normal	The external canal and tympanic membrane are unremarkable.
		The external canal is normal in size.
	Abnormal	Imaging of the temporal bones on the right reveals narrowing of the external canal to the level of the tympanic membrane, where there is apparent bony narrowing.
		In the left ear, the external canal is clear with a mildly thickened and retracted tympanic membrane.
Incus	Normal	The incus and malleus are patent.
		No evidence of erosion in the incus and malleus.
	Abnormal	The middle ear ossicles on the right, particularly the orientation of the head of the malleus to the body of the incus, are abnormal.
		The handle and the long process of the incus is markedly hypoplastic if not absent.
Stapes	Normal	The stapes is patent.
		FINDINGS: Inferior to the facial nerve in the region of the oval window, there is a tiny soft tissue mass abutting the footplate of the stapes without evidence of dehiscence.
	Abnormal	IMPRESSION: Dysplastic appearing left incus and malleus, abnormal articulation between the incus and the malleus and absent stapes on the left.
		The stapes is thickened.
Semicircular Canal	Normal	The inner ear structures, including the semicircular canal, vestibule and cochlea are unremarkable.
		The semicircular canals are patent.
	Abnormal	The lateral semicircular canal is not formed and the proximal portion is bulbous.
		The semicircular canals are present but appear slightly widened.
Tympanic Membrane	Normal	The tympanic membrane is unremarkable.
		The external canal is clear and the tympanic membrane is patent.
	Abnormal	The tympanic membrane is thickened.
		The external auditory canal is atretic as well as the tympanic membrane.