“Dr. Detective”: combining gamification techniques and crowdsourcing to create a gold standard in medical text

Anca Dumitrache[1,3], Lora Aroyo[1], Chris Welty[2], Robert-Jan Sips[3], and Anthony Levas[2]

[1] VU University Amsterdam
anca.dumitrache@student.vu.nl, lora.aroyo@vu.nl
[2] IBM Watson Research Center, New York
cawelty@gmail.com, levas@us.ibm.com
[3] CAS Benelux, IBM Netherlands
robert-jan.sips@nl.ibm.com

Abstract. This paper proposes a design for a gamified crowdsourcing workflow to extract annotation from medical text. Developed in the context of a general crowdsourcing platform, *Dr. Detective* is a game with a purpose that engages medical experts into solving annotation tasks on medical case reports, tailored to capture disagreement between annotators. It incorporates incentives such as learning features, to motivate a continuous involvement of the expert crowd. The game was designed to identify expressions valuable for training NLP tools, and interpret their relation in the context of medical diagnosing. In this way, we can resolve the main problem in gathering ground truth from experts – that the low inter-annotator agreement is typically caused by different interpretations of the text. We report on the results of a pilot study assessing the usefulness of this game. The results show that the quality of the annotations by the expert crowd are comparable to those of an NLP parser. Furthermore, we observed that allowing game users to access each others' answers increases agreement between annotators.

Keywords: crowdsourcing, gold standard, games with a purpose, information extraction, natural language processing

1 Introduction

Modern cognitive systems require human annotated data for training and evaluation, especially when adapting to a new domain. An example of such system is Watson QA [1] developed by IBM, that won the Jeopardy TV quiz show against human competitors. To tune its performance, Watson was trained on a series of databases, taxonomies, and ontologies of publicly available data [2]. Currently, IBM Research aims at adapting the Watson technology for question-answering in the medical domain, which requires large amounts of new training and evaluation data in the form of human annotations of medical text. Two issues arise in this context: (1) the traditional way of ground-truth annotations is slow, expensive and generates only small amounts of data; (2) in order to achieve high inter-annotator agreement, the annotation guidelines are too restrictive. Such practice has proven to create over-generalization and brittleness [3], through losing the sense of diversity in the language, which leads to the fact that natural language processing tools have problems in processing the ambiguity of expressions in text, especially critical in medical text.

The diversity of interpretation of medical text can be seen at many levels; as a simple example, consider the sentence, "Patients exhibiting *acute tailbone pain* should be examined for extra bone nodules." Human experts disagree routinely on whether "acute tailbone pain", "tailbone pain", or "pain" is the primary term in this sentence. Proponents of "tailbone pain" argue that there is a medical term for it (*Coccydynia*) making it primary, others argue that it is pain which is located in the tailbone. Traditional methods of gathering ground truth data for training and evaluation fail to capture such interpretation diversity, leading us to the innovative Crowd Truth approach [4] providing context for this work.

Our analysis led us to believe that the diversity of interpretation occurs at two levels, depending on whether the context is being considered. Term identification, as exemplified in the example above, may be done independent of the clinical context, for example when processing a textbook for background knowledge. However, in the presence of a particular patient, the role of the location and duration modifiers (e.g. tailbone, acute, resp) may or may not be important. We also observe that context-independent tasks tend to require less expertise, allowing us to use a lay crowd more effectively.

These two types of annotation tasks can be performed by two different types of crowds in order to optimize the time, effort and the quality of the final result. Given the experience [4, 5] with defining micro-tasks for the general crowd via crowdsourcing platforms such as Amazon Mechanical Turk[4], or CrowdFlower[5], in this paper we focus on method to engage a crowd of medical experts to be able to resolve *Semantic Ambiguity* in medical text. Annotating complex medical text could be a time consuming and mentally taxing endeavor, therefore the monetary incentive might not be sufficient for attracting a crowd of experts. However, providing a tailored experience for medical professionals through features such as e-learning, and competition with peers, could serve as additional motivation for assembling the right crowd for our task. This can be accomplished by incorporating gamification features into our application.

In this paper, we propose a gamified crowdsourcing application for engaging experts in a knowledge acquisition process that involves domain-specific knowledge extraction in medical texts. The goal of such text annotations is to generate a gold standard for training and evaluation of IBM Watson NLP components in the medical domain. First, we position our work in the context of already existing games with a purpose, crowdsourcing and other niche-sourcing initiatives. Then we outline our approach by focusing on the gaming elements used

[4] www.mturk.com

[5] www.crowdflower.com

as incentives for medical experts, in the context of the overall game application architecture. We show how this gaming platform could fit together with a micro-task platform in a joint workflow combining efforts of both expert and non-expert crowds. Next, we describe the experimental setup to explore the feasibility and the usability of such an application. Finally, we discuss the results of the pilot run of our application, and we identify the points of improvement to bring in future versions.

2 Related Work

In recent years, crowdsourcing has gained a significant amount of exposure as a way for creating solutions for computationally complex problems. By carefully targeting workers with gaming elements and incentives, various crowdsourcing applications were able to garner a significant user base engaged in their tasks. The ESP Game [6] (later renamed Google Image Labeler) pioneered the field by implementing a gamified crowdsourcing approach to generate metadata for images. The reCAPTCHA [7] application combined the CAPTCHA security measure for testing human knowledge with crowdsourcing, in order to perform text extraction from images. The gamified crowdsourcing approach has been employed successfully even in scientific research, with applications such as Galaxy Zoo [8] using crowd knowledge to perform image analysis and extract observations from pictures of galaxies. All of these systems employ mechnisms for a continuous collection of a large amount of human annotated data.

A crowdsourcing framework by [9] introduces 10 design points for Semantic Web populating games. In the context of our research, of a particular interest are: identifying tasks in semantic-content creation, designing game scenarios, designing an attractive interface, identifying reusable bodies of knowledge, and avoiding typical pitfalls. As not all crowdsourcing tasks are suitable for redesign as part of a gamified platform, identifying which of these tasks could engage successfully medical expert crowd is of a key importance to our research. It is also crucial to involve mechanisms to optimize the ratio of time spent and quality and volume of the output [9]. External knowledge sources for annotations (e.g. vocabularies, NLP parsers) can be used to target the work of the players to problems that are too complex to be handled only by computers [9]. Finally, in order to ensure the quality of the answers, unintentional mistakes of the users need to be avoided through clear instructions in the interface [9].

Gamification as applied to text annotation crowdsourcing is an emerging field in different domains. For instance, the Phrase Detective project [10] uses gamified crowdsourcing for building anaphoric annotation ground truth. The input documents are general purpose, and the crowd is not specialized. Two interesting features we considered for Dr. Detective as well, (1) the need for a user training task to improve the usage of the application, and (2) understanding of the user profile (e.g. players can examine a considerable variation in their interaction styles, abilities or background knowledge.

The Sentiment Quiz [11], played through various social networking platforms, employs crowdsourcing to evaluate accuracy of sentiment detecting algorithms over sentences, and to create a lexicon of sentiments in various languages. The requirements for user incentives in *Dr. Detective* were based on the analysis provided by Sentiment Quiz, e.g. for scoring, high score board, and level-based goals, as well as for enhancing the crowd output through statistical methods applied in the disagreement analytics.

However, neither the Sentiment Quiz, nor the Phrase Detective applications actively seek out to capture the ambiguity in language. Phrase Detective even tries to enforce agreement, by awarding additional points for annotators that agree with the ground truth. Neither do most applications in the domain study the effect of using specialized crowds to perform the information extraction tasks. Our goal is to build an end-to-end gamified crowdsourcing platform that can capture disagreement between annotators, while catering specifically to experts in the medical field.

3 "Crowd-Watson" Architecture: The Game Perspective

In this section, we describe the architecture for *Dr. Detective* [6] – an application for engaging experts in knowledge extraction tasks for creating ground truth annotations in medical texts. We start by framing *Dr. Detective* as part of the general *Crowd-Watson*[7] framework for crowdsourcing medical text annotation [12]. Then, we tackle the challenge of tailoring the application to a specialized crowd of medical professionals, through a study of possible motivating factors. Finally, we describe how gamification elements were integrated with the crowdsourcing workflow.

The *Crowd-Watson* framework supports the composition of crowd-truth gathering workflows, where a sequence of micro-annotation-tasks can be executed jointly either by the general crowd on platforms like CrowdFlower, or by specialized crowd of domain experts on gaming platform as Dr. Detective. Crowd-Watson framework focuses on micro-tasks for knowledge extraction in medical text. The main steps involved in the *Crowd-Watson* workflow are: **pre-processing** of the input, **data collection**, **disagreement analytics** for the results, and finally **post-processing**. These steps are realized as an automatic end-to-end workflow, that can support a continuous collection of high quality gold standard data with feedback loop to all steps of the process. The input consists of medical documents, from various sources such as Wikipedia articles or patient case reports. The output generated through this framework is annotation for medical text, in the form of concepts and the relations between them, together with a collection of visual analytics to explore these results. The architecture of this application, and the way its components interact with each other, can be seen in Figure 1. In this paper, we focus on those aspects of the architecture that relate to the *Dr. Detective* gaming platform for data collection. A full description of the *Crowd-Watson* architecture is available at [12].

[6] `http://crowd-watson.nl/dr-detective-game`

[7] `http://crowd-watson.nl`

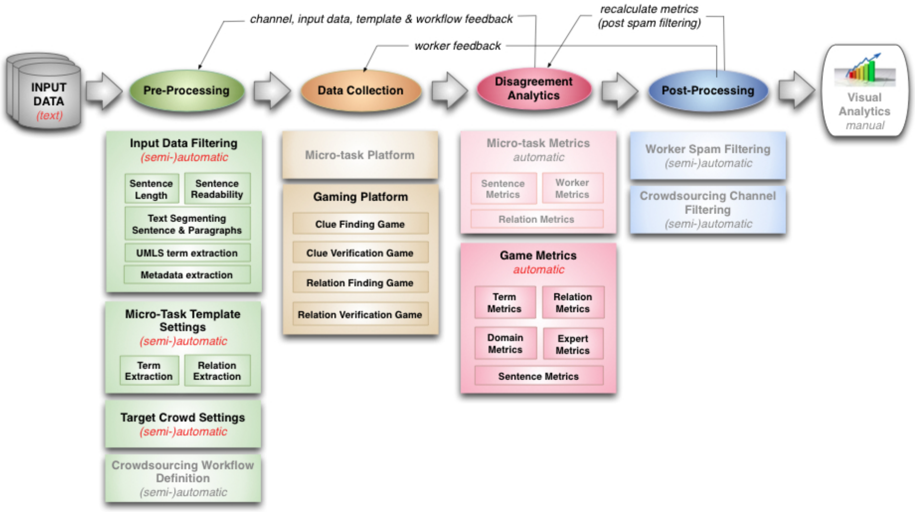


Fig. 1. Crowd-Watson Framework Design *(the highlighted components are the ones related to the Game Platfom)*

3.1 Pre-Processing for the Game Platform

Typically, the input is available in an unstructured format (e.g. simple text). As part of the **input data filtering** step, additional metadata, such as the specialization field in which it was published or, for case reports, the diagnosis of the patient, can be extracted from these documents. In addition, some annotation can also be generated automatically, by mapping the text to the UMLS vocabulary of biomedical terminology, classification, and coding standards [13]. The UMLS parser can be used to identify both concepts and relations, however, as a fully automated approach, it suffers from the typical issues of NLP techniques [14], such as lack of contextual awareness, and limited ambiguity processing capabilities. Nevertheless, UMLS annotations can be employed as a good baseline for measuring the efficiency of the crowdsourced answers.

The workers are asked to perform a series of annotation tasks on the input documents. The purpose of these tasks is creating annotation in the form of concepts and the relations between them. We define these tasks according to four **micro-task templates**:

1. *Term extraction* – the task of identifying all the relevant terms in a text, where a term refers to a set of words that forms a coherent medical concept;
2. *Term categorization* – the task of classifying a medical term into an appropriate category, such as the concepts in the UMLS thesaurus;
3. *Relation extraction* – the task of identifying whether or not a relation exists between two medical terms;

4. *Relation categorization* – the task of classifying a medical relation into an appropriate category (or set of categories), such as the relations in the UMLS thesaurus.

The workers on *Crowd-Watson* consist of both an expert crowd, and a general crowd. Each of these crowds interacts with the input documents on a specialized platform – for the general crowd, regular crowdsourcing micro-tasks have been constructed on CrowdFlower, whereas the expert crowd employs the *Dr. Detective* application for solving tasks tailored to their profile. The tasks can be solved by both the general, and the expert crowd. The **target crowd setting** step entails picking the difficulty level of the task according to the level of expertise of the crowd. For instance, when discussing term extraction, the general crowd could reliably find demographic terms, as they do not require significant medical knowledge, whereas the expert crowd can focus on annotating more difficult terminology.

3.2 Game Disagreement Analytics

After the input data is formated and filtered appropriately through the *pre-processing* components, it is sent to the *data collection* component to to gather either expert annotation (through the gaming platform) or lay crowd annotations (through the micro-task platform). Next, the annotation results are analyzed with a set of content and behavior-based metrics, to understand how the disagreement is represented in both cases [15, 16], and to assess the quality of the individual workers, and the quality of the individual and overall crowd truth results.

To track the individual performance of a user in the crowd, the expert metrics were developed. For each sentence in the input, the performance of the worker can be measured as a set of vectors, according to the task they solved on that input. Such a vector is composed of 0 and 1 values, such that for each answer a user annotated in that sentence, there is a 1 in the corresponding position, whereas answers that were not picked by the user are set to 0. These answer vectors can also be measured at the level of the domain.

At the level of the sentence, a set of task-dependent sentence metrics were also defined. For either term extraction or relation extraction, any sentence can be expressed as a sentence vector – the sum of all the individual user vectors on that sentence, for that task. Furthermore, an added layer of granularity can be introduced by considering the categories for the terms and relations. This representation can then be used to define appropriate metrics for sentence clarity, what the popular answers were, how disagreement is represented, and similarity of annotation categories and domains.

The prime role of the disagreement analytics in the gaming platform are to provide explicit measures for the quality and completeness of the final result; to identify gaps of missing types of annotations; or to discover possible contradictions and inconsistencies. This is opposed to the micro-task disagreement analytics, which follow the same approach but apply to filters for spam identification.

4 Data Collection: Gaming Platform

In order to collect data from a crowd for medical experts, it is imperative to find the necessary motivators for engaging them into contributing. To this end, we have performed a series of qualitative interviews with medical students and professionals. The purpose was to identify what requirements and features would the medical crowd be interested in seeing in a crowdsourced application, and how this application could be built to help in their work. These interviews established incentives for crowd labor [17], such as competition, learning, and entertainment in the context of working in the medical field, as well as documents that the medical crowd would be interested in reading.

After discussing with 11 people in the medical field (2 professionals, 3 lecturers, 5 students), we were able to identify several key requirements to incorporate into the gaming platform:

- at the level of the input, the interviewees expressed their interest in **reading medical case reports**;
- **learning** about their field, through targeted micro-tasks and extended feedback on their answers, was the most significant motivator;
- the interviewees expected the tasks to challenge their **problem-solving** skills;
- **competition** with peers emerged as a secondary motivator;
- the tasks need to be fun to solve, making **entertainment** as another secondary motivator;
- medical professionals have difficult schedules, and would prefer to have **flexibility** in the time required to engage with the application;

In order to attract users to the application, a goal that is seen as useful by the players needs to be firmly established. As *learning* proved to be the most relevant incentive from the interviews, we focused the goal of the application on this, while also trying to incorporate the *problem-solving* requirement. We developed the concept of a **clue-finding game**, where the text annotation tasks were put in the context of searching for clues in the history of a patient. For instance, when performing the task of term extraction on a patient case report, the user can annotate any of these three **clue types**:

1. the term is a clue *leading* to the final diagnosis of the case;
2. the term is a false clue that is *irrelevant* to the final diagnosis of the case;
3. the term is a *normal condition* that does not influence the final diagnosis of the case.

The clue types can be used as an incentive, involving users with the task they are solving by redesigning it as a medical puzzle, but it can also be used to generate additional annotation. The annotations retrieved from the general crowdsourcing approach are dependent on the context of the sentence where they were identified, so by asking the expert crowd to find meta-relations at the level of the document, we can generate knowledge that is valid generally for the domain. This kind of task cannot be solved simply with the use of contextual information, and requires background knowledge of the field, therefore making it suitable for an application targeted at experts.

The qualitative interviews helped us identify the extrinsic motivators for engaging the medical crowd. After the goal of the application was established, the final step was translating the user incentives into concrete features for building the *Dr. Detective* gaming platform.

4.1 Difficulty

In order to support the user learning experience and introduce flexibility in task solving, we define the concept of difficulty. This refers to the combination of skill and time required for reading the document, and then performing the annotation task. While it is difficult to hypothesize on the comparative difficulty of performing annotations, the difficulty of the document can expressed as syntactic and semantic difficulty. The syntactic difficulty expresses the effort need for reading the document in three components: the *number of sentences* in the document (NoS), the *number of words* (NoW), and the *average sentence length* (ASL). The semantic difficulty expresses the effort needed for understanding the text in two components: the *number of UMLS concepts* present in the document ($NoUMLS$), and the *readability* of the document ($SMOG$). The SMOG [18] formula for computing readability was employed, as it is often recommended for use in evaluating healthcare documents [19]. Therefore, for every document D, its difficulty is defined as the norm of the normalized five-component vector:

$$difficulty(D) = \|(NoS, NoW, ASL, NoUMLS, SMOG)\|.$$

4.2 Scoring

In order to develop the *competition* incentive, a scoring system was devised, to reward players for their work. Through viewing a high score board, they are also encouraged to compete against each other.

We want to reward users when they perform in a way that is beneficial to us. We want to collect the correct answers to the task, therefore, selecting a high-consensus solution should yield more points. This strategy could, however, make users rely entirely on the answers of others. Therefore, in order to encourage a wider answer set and capture semantic ambiguity, we need to give points for newly discovered answers. Users should also be penalized for giving wrong answers. We also want to encourage users to return to the application, and keep playing. Finally, in order for users to solve tasks in increasing difficulty, scoring needs to be proportional to the difficulty for solving the task [20]. Based on this, for each user U solving a task T on document D, we developed the following scoring components:

- *popular*(U, D, T): the points users receive if they make annotations that were previously selected by at least one other user; we also want to reward partial answers, in order to capture ambiguity;
- *consecutive*(U): the points users gain the more consecutive tasks they solve;

- $disovered(U, D, T)$: the points users receive if they are the first to discover an answer, if it is then selected by at least one other user;
- $wrong(U, D, T)$: the points users lose if their answers are not selected by any other user.

Based on this analysis, we developed the following scoring formula:

$$\begin{aligned} score(U, D, T) = difficulty(D)\cdot \\ \cdot (popular(U, D, T) + consecutive(U) \\ + discovered(U, D, T) - wrong(U, D, T)). \end{aligned}$$

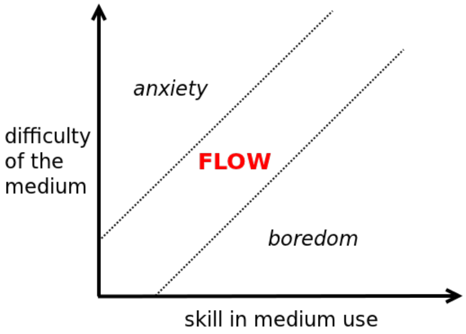


Fig. 2. Game flow as an expression of skill and difficulty

4.3 Immersion

In order to develop the *entertainment* incentive, the crowdsourcing application needs to provide immersion inside the task-solving experience. Immersion is based on the concept of game flow [21], which states that at every point in the game, the difficulty needs to be proportionate with the skill required to solve the task. Skill at playing is acquired by the user as they solve more tasks. If the difficulty is disproportionately large compared to the skill, it will cause anxiety for the user, whereas if the difficulty is too small, the user will be bored. Immersion is achieved when skill and difficulty are proportionally balanced, as illustrated in Figure 2.

Immersion is considered when choosing the next document that the user will be asked to solve as part of the game. When a user solves a task on D_i, the document they will be asked to solve next needs to have a higher difficulty in order to avoid boredom, but the increase needs to be low enough to avoid anxiety. Therefore, we define the set of possible documents that occur after D_i as:

$$\begin{aligned} next(D_i) = \{D_j\ |difficulty(D_j) = min(difficulty(D_i) - difficulty(D_t), \\ \forall t \neq i \text{ where } difficulty(D_t) \geq difficulty(D_i))\} \end{aligned}$$

4.4 Levels

Finally, in order to satisfy the constraint for *flexibility*, game levels were implemented to quantify the skill required for solving the tasks. As skill is proportional with difficulty, we defined the game levels by quantifying the difficulty metric previously described into three intervals:

1. easy: $\{D \mid difficulty(D) \in [0, 2]\}$,
2. normal: $\{D \mid difficulty(D) \in [3, 4]\}$,
3. hard: $\{D \mid difficulty(D) \in [5, 6]\}$.

These levels should enable users to plan which task they want to solve in accordance to the time they have at their disposal, while also providing a goal-based incentive of progressing in their skill [20].

5 Experimental Setup

In order to test the feasability of the *Dr. Detective* setup, we implemented a version of the workflow described in Section 3, and set up a pilot run involving a crowd of medical professionals. As part of our pilot run, we performed an initial evaluation of both the quality of the answers, and the user enjoyment as part of this gamified crowdsourcing platform. The goal of this experiment can be described as three questions, which will be discussed as part of our results:

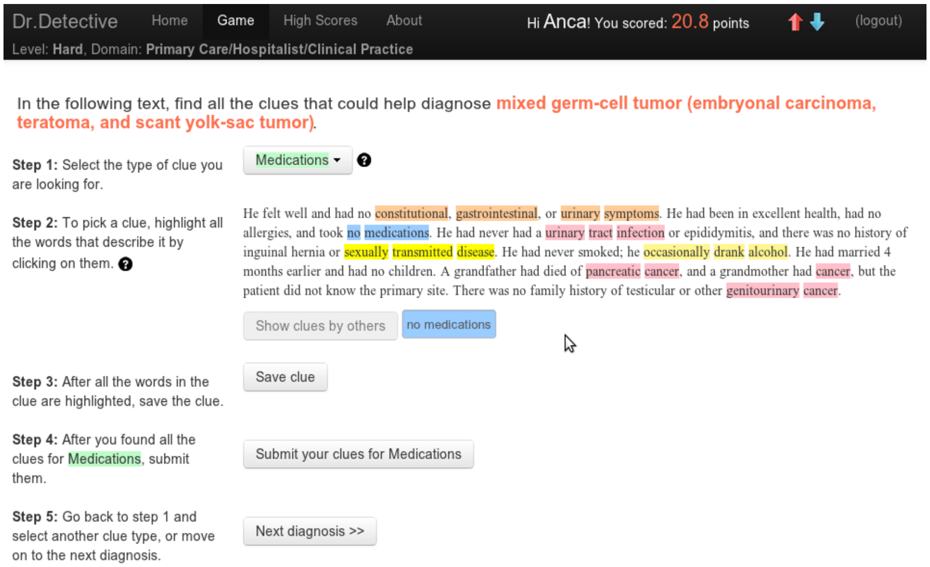
1. How do the answers annotated by the crowd compare to those found by the UMLS parser?
2. Does having access to the answers of other users stimulate diversity of opinion?
3. Did users experience immersion in the gaming experience?

In order to answer these questions, we set up two versions of the game, one in which users had the ability to see the answers of others, and one in which they did not. In addition, some of the gaming elements that would ensure the users keep in the state of game flow (high scores board, next document selection mechanism, levels) were only limited to the full version of the game. We constructed an experiment where the users would play both versions of the game, then answer a questionnaire on their experiences. The details of this experimental setup are described in this section.

5.1 Input

Based on a suggestion in the qualitative interviews, the input was selected from clinical cases published in the New England Journal of Medicine[8]. 10 documents were picked out of four of the most popular specialties (Hematology/Oncology, Nephrology, Primary Care/Hospitalist/Clinical Practice, Viral Infections). The diagnosis was extracted from each document, based on a string matching procedure performed on the text marked in “diagnosis” section headings (e.g. clinical diagnosis, pathological diagnosis etc.). The documents were split into paragraphs, to increase the ease of reading, and the difficulty metrics (described in

[8] `www.nejm.org`



Level: Hard, Domain: Primary Care/Hospitalist/Clinical Practice

In the following text, find all the clues that could help diagnose mixed germ-cell tumor (embryonal carcinoma, teratoma, and scant yolk-sac tumor).

Step 1: Select the type of clue you are looking for.

Step 2: To pick a clue, highlight all the words that describe it by clicking on them.

He felt well and had no constitutional, gastrointestinal, or urinary symptoms. He had been in excellent health, had no allergies, and took no medications. He had never had a urinary tract infection or epididymitis, and there was no history of inguinal hernia or sexually transmitted disease. He had never smoked; he occasionally drank alcohol. He had married 4 months earlier and had no children. A grandfather had died of pancreatic cancer, and a grandmother had cancer, but the patient did not know the primary site. There was no family history of testicular or other genitourinary cancer.

Step 3: After all the words in the clue are highlighted, save the clue.

Step 4: After you found all the clues for Medications, submit them.

Step 5: Go back to step 1 and select another clue type, or move on to the next diagnosis.

Fig. 3. Screenshot from the "Dr. Detective" game

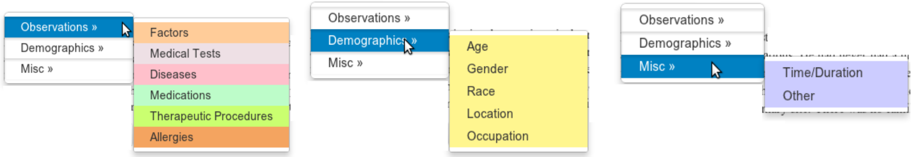


Fig. 4. Term types in the game.

Section 4.1) were then applied to each paragraph. Finally, we selected a set of 20 paragraphs, with the values in the difficulty vector uniformly distributed to represent a broad range of text types, to use for the game, as we wanted to ensure that all of the text would be annotated in the limited time frame of the experiment run.

5.2 Task

The micro-task templates (described in Section 3.1) selected for this pilot were (1) term extraction, and (2) term categorization. Based on how relevant they are at describing patient case reports, 3 meta-types, each with a set of term types taken from UMLS, were selected and implemented in the interface for the categorization task. These term types are based on factor categories given to domain experts during the expert annotation phase for Watson. The type selection menu can be seen in Figure 4. In total, 13 term types were available for the users to annotate. As most interviewers expressed their interest in a problem-solving application, we decided to set the clue type user seek as part of the application (described in Section 4) to (1) the term is a clue *leading* to the final diagnosis of the case. Finally, in order to encourage the diversity of opinion, and therefore capture ambiguity, we allowed users to look at the answers of others for the task they are solving. This feature was made available through a button, which the users could choose to press in order to toggle the other answers. The scoring formula (described in Section 4.2) ensures that users are motivated to find new answers even in this circumstances, through the use of discovery bonus points. The users could access the details of how their score was computed through a hover notification in the menu. An example of how this task was presented to the users as part of the *Dr. Detective* interface can be seen in Figure 3.

5.3 Users

The pilot run of the *Dr. Detective* game had 11 participants in total, with 10 players engaging with the full game version, and 7 engaging with the simple version. In total, 155 annotation sets were collected, with each paragraph solved as part of 2 to 7 different game rounds. In addition, 6 players completed the feedback questionnaire.

6 Results and Discussion

In keeping with the research questions defined in the previous section, we first analyzed how the answers from the crowd compare to the results of the UMLS parser. We selected the top three paragraphs that were played the most, and compared the answers to the term list generated by the UMLS MetaMap parser [9] for the same paragraphs. Fig. 5 shows the crowd was able to identify the majority of the words annotated with UMLS. Additionally, Fig. 6 shows that around one third of the terms in UMLS had a full match with terms annotated by the crowd. Factoring in the partial term matches, the crowd was able to identify most of the

[9] http://metamap.nlm.nih.gov/

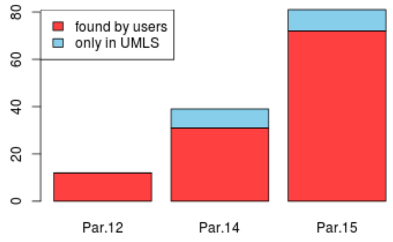


Fig. 5. Words in UMLS for the 3 most popular paragraphs in the game

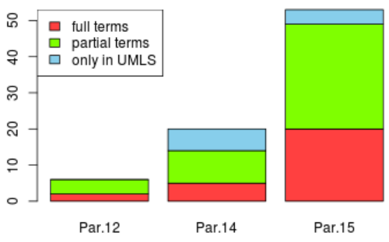


Fig. 6. Terms in UMLS for the 3 most popular paragraphs in the game

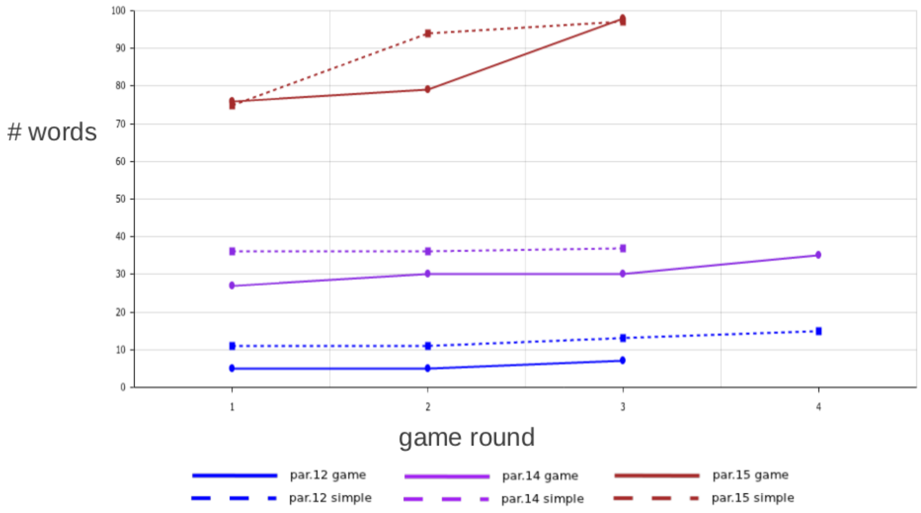


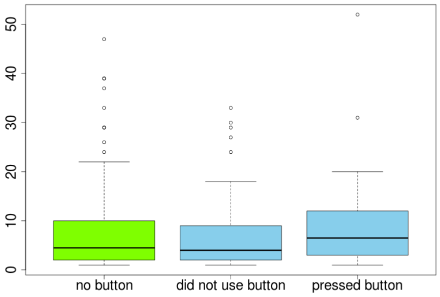
Fig. 7. Number of words for the 3 most popular paragraphs, after each round of each game version

UMLS terms. This shows the efficiency of the crowd answers is quite high, enough for the crowd to be considered as a viable alternative to automated named-entity recognition, provided that enough users give their input for a paragraph.

Next, we look at how diversity of opinion was expressed by the game users. Specifically, we are interested in finding out whether being able to see the results of other people will stimulate disagreement, or rather make users select each other's answers. In other to achieve this, we look at how the answers per paragraph varied according to the version of the game that the user played.

Fig. 7 shows how the number of new words per paragraph increases after each round of the game, for the top three paragraphs. Each version of the game seems to follow the same progression in the rate of new words identified, with the first users finding most of the words, and then only slight increases as the paragraph is played by other people. However, the simple version of the game seems to constantly feature a higher total word count, as opposed to the full game version. The same trend was observed both for the number of new types, and the number of distinct terms. This seems to indicate that the full game version was less encouraging for collecting a wide array of terms.

In order to rule out an issue related to some other feature in the full game version, we looked at how the behavior of pressing the button to view other answers affected the output. Out of 67 game rounds played in the full version, this button was only pressed in 18 of the rounds, so it appears this was not a popular feature to begin with. Fig. 8 shows that, actually, users tended to annotate more words in total when they pressed. However, as evidenced in Fig. 9, the ratio of new words to total words in this case was much lower than when the button was not pressed. Additionally, it appears there is not much difference



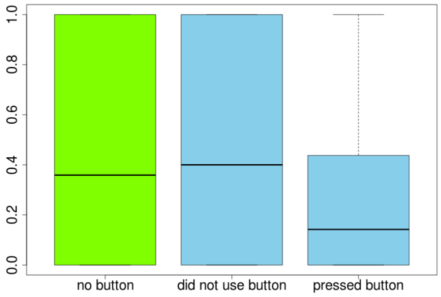


Fig. 8. Ratio of total words per round, grouped by the use of the button to view the answers of others

Fig. 9. Ratio of new to total words, grouped by the use of the button to view the answers of others

between the simple version of the game, and the full version, but where the users chose not to look at the answers of others. Therefore we can infer that having access to all the answers makes the crowd act more conservative, selecting less new words, but rather choosing to validate the answers of others.

When looking at the answers in the questionnaire related to the usefulness of seeing other people's annotations, we found that most people (67%) were ambivalent to having the option of checking their answers. Some users reported using this feature as a tool for better understanding the task, while others claimed it validated the answers they had already chosen. Overall, it seems that having access to all the other answers makes users less likely to find and annotate new words, which could mean a loss in the ambiguity of the annotation. It also provides an unfair advantage to the first users to annotate a paragraph, as their score would likely keep increasing as other people keep selecting their answers.

Finally, we analyzed whether immersion in the game occurred for the users involved, and how each individual game feature was rated. The flow of the game was reported to be good, with 83% of the users saying they were neither too bored, or overwhelmed. Most users found the levels to be a useful addition, with 50% being satisfied with the level progression, and 33% being ambivalent to it. However, some users pointed out that they expected more challenge from the advanced level. As the difficulty is currently computed only based on textual metrics, the game could potentially get boring for users. For this reason, domain difficulty should be incorporated in future versions of the game. The scoring part of the game was less well received, with 83% of the users declaring they found the way their score is computed only somewhat clear. Therefore, in future game versions, a more detailed scoring breakdown should be implemented, with users being able to access the history of the cases they solved. Finally, most users reported to have enjoyed the game, and expressed an interest in returning to play, provided they can solve more difficult cases and get more feedback. The full game version was almost universally preferred by the users.

7 Conclusion and Future Work

This paper proposes a design for *Dr. Detective* – a gamified crowdsourcing platform to extract annotation from medical text. *Dr. Detective* was developed in the context of *Crowd-Watson*, a general crowdsourcing framework for extracting text annotation by engaging both a general crowd, and a domain expert crowd. The gaming platform was designed taking into account the requirements of the expert crowd, and illustrating their implementation in a clue finding game. Specific gamification elements were incorporated, such as difficulty, scoring, immersion, and levels. A first version of *Dr. Detective* was implemented and tested. The pilot run showed that the quality of the results of the crowd are comparable to those of an NLP parser. Allowing users to see the answers of others resulted in increased agreement, and thus decreased the desired diversity in answers. The overall user feedback for the application was positive. However, it was clear that users desire more complex challenges in order to keep them engaged.

An important next step is to define and test disagreement metrics that are specific to the gaming environment. As we have seen in previous research, a promising starting point are the disagreement metrics developed for the data collected through the micro-task platform. We also plan to further test how each of the gaming features performs individually, in order to fine-tune the application to understand better their influence on the quality and volume of the end result, as well as to adapt best to the needs of the users. Finally, we will explore how to further integrate the gaming and the micro-task crowdsourcing workflows, by using the output from one workflow to enhance the input for the other (e.g. ask one crowd to perform the term extraction, and the other crowd the relation extraction), or by asking one crowd to validate the output of the other crowd.

Acknowledgements

The authors would like to thank Kathrin Dentler and Dr. Petra Wolffs for their help with finding participants for both the interviews and the application pilot run, as well as the students and medical professionals who were involved in these activities, and who provided their feedback.

References

1. Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A.A., Lally, A., Murdock, J.W., Nyberg, E., Prager, J., Schlaefer, N., Welty, C.: Building Watson: An Overview of the DeepQA Project. AI Magazine **31** (2010) 59–79
2. Kalyanpur, A., Boguraev, B., Patwardhan, S., Murdock, J.W., Lally, A., Welty, C., Prager, J.M., Coppola, B., Fokoue-Nkoutche, A., Zhang, L., et al.: Structured data and inference in DeepQA. IBM Journal of Research and Development **56**(3.4) (2012) 10–1
3. Aroyo, L., Welty, C.: Harnessing disagreement for event semantics. Detection, Representation, and Exploitation of Events in the Semantic Web 31
4. Aroyo, L., Welty, C.: Crowd Truth: Harnessing disagreement in crowdsourcing a relation extraction gold standard. WebSci2013. ACM (2013)

5. Inel, O., Aroyo, L., Welty, C., Sips, R.J.: Exploiting Crowdsourcing Disagreement with Various Domain-Independent Quality Measures. Technical report, VU University Amsterdam (July 2013). `http://crowd-watson.nl/tech-reports/20130702.pdf`
6. Von Ahn, L., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the SIGCHI conference on Human factors in computing systems, ACM (2004) 319–326
7. Von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M.: recaptcha: Human-based character recognition via web security measures. Science **321**(5895) (2008) 1465–1468
8. Lintott, C.J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M.J., Nichol, R.C., Szalay, A., Andreescu, D., et al.: Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. Monthly Notices of the Royal Astronomical Society **389**(3) (2008) 1179–1189
9. Siorpaes, K., Hepp, M.: Games with a Purpose for the Semantic Web. Intelligent Systems, IEEE **23**(3) (2008) 50–60
10. Chamberlain, J., Poesio, M., Kruschwitz, U.: Phrase Detectives - A Web-based Collaborative Annotation Game. In: Proceedings of I-Semantics. (2008)
11. Scharl, A., Sabou, M., Gindl, S., Rafelsberger, W., Weichselbraun, A.: Leveraging the Wisdom of the Crowds for the Acquisition of Multilingual Language Resources. In: Proc. 8th LREC - International Conference on Language Resources and Evaluation. (2012)
12. Lin, H.: Crowd Watson: Crowdsourced Text Annotations. Technical report, VU University Amsterdam (July 2013). `http://crowd-watson.nl/tech-reports/20130704.pdf`
13. Bodenreider, O.: The unified medical language system (UMLS): integrating biomedical terminology. Nucleic acids research **32**(suppl 1) (2004) D267–D270
14. Goldberg, H.S., Hsu, C., Law, V., Safran, C.: Validation of clinical problems using a UMLS-based semantic parser. In: Proceedings of the AMIA Symposium, American Medical Informatics Association (1998) 805
15. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. In: AAAI 2013 Fall Symposium on Semantics for Big Data (in print). (2013)
16. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Crowd Truth Metrics. Technical report, VU University Amsterdam (July 2013). `http://crowd-watson.nl/tech-reports/20130703.pdf`
17. Tokarchuk, O., Cuel, R., Zamarian, M.: Analyzing crowd labor and designing incentives for humans in the loop. IEEE Internet Computing **16**(5) (2012) 0045–51
18. McLaughlin, G.H.: SMOG grading: A new readability formula. Journal of reading **12**(8) (1969) 639–646
19. Doak, C.C., Doak, L.G., Root, J.H.: Teaching patients with low literacy skills. AJN The American Journal of Nursing **96**(12) (1996) 16M
20. Von Ahn, L., Dabbish, L.: Designing games with a purpose. Communications of the ACM **51**(8) (2008) 58–67
21. Sherry, J.L.: Flow and media enjoyment. Communication Theory **14**(4) (2004) 328–347