Exploiting Disagreement Through Open-Ended Tasks for Capturing Interpretation Spaces

Benjamin Timmermans(✉)

VU University, Amsterdam, The Netherlands
b.timmermans@vu.nl

Abstract. An important aspect of the semantic web is that systems have an understanding of the content and context of text, images, sounds and videos. Although research in these fields has progressed over the last years, there is still a semantic gap between data available of multimedia and metadata annotated by humans describing the content. This research investigates how the complete interpretation space of humans about the content and context of this data can be captured. The methodology consists of using open-ended crowdsourcing tasks that optimize the capturing of multiple interpretations combined with disagreement based metrics for evaluation of the results. These descriptions can be used meaningfully to improve information retrieval and recommendation of multimedia, to train and evaluate machine learning components and the training and assessment of experts.

Keywords: Semantic interpretation · Multimedia · Crowdsourcing · Disagreement

1 Introduction

The semantic web has contributed to improving the usability of websites through semantic metadata, but this has not proven effective for media such as images, sounds and videos. The problem is that although every day more and more media content is shared online, the amount of metadata available is still limited [1]. Often some metadata is made available by the author, there is still a semantic gap between the available metadata and how it is perceived by humans. As a result, the metadata does not give a good representation of the actual content and context. The underlying systems need to have an understanding of this context, the human perspectives and opinions in order to provide meaningful search and discovery of information.

Recent work on video search engines has shown the existence of a semantic gap in video content [2]. In order to bridge this gap, it is essential for these systems to gain an understanding of the actual content. Search in social platforms like YouTube is limited by the annotations made by the uploader of a video. This means that the content you are looking for can be there, but that there is no representation of it in the metadata. The current improvements in information extraction methods will contribute to solving this problem, but the problem

H. Sack et al. (Eds.): ESWC 2016, LNCS 9678, pp. 873–882, 2016.
DOI: 10.1007/978-3-319-34129-3_56

remains that sounds, images and videos are prone to have multiple interpretations because they can be represented in multiple ways [3]. This means more interpretations of what can be heard are needed, in order to form a full spectrum of meaningful representations. In addition, gold standards for sounds are limited in size, and are often homogeneous because they only contain single interpretations [4]. Rich information can be obtained by aggregating multiple of these interpretations, which can be seen as collective intelligence [5].

In crowdsourcing, annotations are obtained from a large crowd of people using small microtasks. Often these tasks consist of the crowd workers selecting the right answer to a question. Capturing the complete answer space of possible interpretations has proven to be difficult with annotation tasks. Guidelines in tasks often focus on only one semantic interpretation in order to increase the inter-annotator agreement [6]. Also, no common reference system for defining the answers may be available, because for many tasks the answers simply cannot be predefined [7]. It can also be that there simply is no standardization in the categorization available [4]. Experts can help to define the reference system, but their views are biased and often it is difficult to define who the "experts" are in tasks such as general question-answering, interpretations of multimedia or sentiment analysis of texts. It may even be possible that the answer space is infinite, because the boundaries of the space are unknown or cannot be known.

The CrowdTruth[1] initiative [8–11] has been investigating how annotations can be collected on a large scale using crowdsourcing based on disagreement, rather than artificially forcing agreement through the inter-annotator agreement. The results have shown that crowdsourcing is a quick and cheaper way to solve the problem of scale, lack of experts and lack of interpretations [12]. Gathering many annotations from a lay crowd rather than few annotations from experts gives a closer representation of reality. Crowdsourcing has proven to be a successful tool for bridging the semantic gap [13]. Yet the question remains how crowdsourcing tasks can be designed if the interpretation space of possible answers is unknown, such that interpretations are not limited by the design of a task. Such open-ended tasks in which the user is less restricted in the answer that can be given can help gather annotations that better represent what is heard in a sound or what can be seen in a video. For instance for improving search or recommendation through content-based indexing.

In Sect. 2 the state of the art is described. Next, the problem statement and contributions of this research are formulated. In Sect. 4, the approach of this research is presented, followed by the preliminary results in Sect. 5. Last follow the evaluation plan for the approach and the conclusions of this paper.

2 State of the Art

There are many questions to which there is no single answer, or the answer is unknown. In such case an open ended task can be used to allow a multitude of possible answers to be annotated, without there being one correct answer. This

[1] http://crowdtruth.org.

does not restrict an annotator in giving a predefined answer, but has shown to give the worker freedom to answer according to their own interpretation [14]. This makes these tasks efficient for gathering a wide range of human interpretations, and that the resulting ground truth has a low influence of the reference system or answer options.

The problem of interpretation and answer spaces of unknown or large size is addressed in a study on object localization in images [15]. It is claimed that probabilistic inference on the ability to solve a task is necessary, because there is a significant difference in the ability of a person to perform a cognitive task. Related, another study found that if the data spans multiple domains, it is more likely that the answer space changes across different tasks [16]. If there is agreement between different workers in an open-ended task, it can be seen as a stronger signal because the answer space is much larger [17]. The chance of workers selecting the same answer in a small answer space is much higher, and having no overlap increases the difficulty of evaluating the answers.

Designing a clear and more detailed crowdsourcing task is needed in order to avoid misunderstanding and result in more substantial answers [18]. For open-ended tasks this is more important, because closed tasks have better defined rules and guidelines than open ended tasks. As a result, the interpretation of every aspect of an open-ended task plays a more important role. This feature of open-ended tasks is often exploited for teaching purposes, because the task can be approached in more ways where the difficulty plays less of a role. This has been found to result in a more diverse and larger group of answers that can be given than with closed questions, while providing valuable feedback to the teacher about the level of understanding of the students [19]. In a study on describing images through crowdsourcing [20] it was found that for efficient task designs, the resulting annotations should be as simple as possible while being as meaningful as possible. The simple annotations can also improve the detection of low quality annotations, because they have a more unified structure. This combination of an open-ended design with clear guidelines and simple meaningful annotations can be used to gather the interpretations.

The multitude of interpretations resulting from open-ended crowdsourcing tasks has also shown to be more difficult to evaluate [21], because these tasks can be highly subjective. Furthermore, if multiple answers can be correct majority voting as quality measure is not appropriate. Instead a peer-review approach should be used where workers verify each others work [22]. In [23] an open-ended crowdsourcing task was used to describe in words the differences between two images. The resulting lexicon was found to be comparable in quality to one created by experts using the different interpretations. For spam detection, open-ended questions have also been found to have the advantage that the response time is an indicator for lies and deception [24]. A study has been done on the use of open tasks [7] using free-response formulation with infinite outcome spaces. However, it was shown that although there are many crowdsourcing tasks that can match the quality of expert annotations, tasks with an infinite interpretation space where the boundaries of possible answers were unknown are not yet used in

practice. Although these results have been found to be more difficult to evaluate, the design can result in high quality annotations.

There has been an increasing interest in research on the nature of sounds, and how people perceive sounds. The content that is perceived in sounds can be described in three categories: the source of the sound, attributes of the sound and the environment of the sound [25]. The perception differs from visual interpretation, because the awareness of sound sources are not as direct as when the source can be visually identified. Also, visible entities such as images differ from sound because they fully exist over time, while a sound always begins and ends like an event does. During this, its audible features change over time and do not have to fully exist at any given time [26]. This makes the sounds prone to have multiple interpretations, which makes the inter-annotator agreement a difficult metric to for instance measure the emotion in a sound [3]. Although our recent work on CrowdTruth has shown that annotator disagreement should be used as a signal, it is still considered a consensus problem [27,28]. This is specifically present in the ambiguity of sounds, which is why it is used in this study.

In music information retrieval, the features that influence how humans perceive music can be categorized into four factors [29]: (1) The content of what is heard in the music, discriminated using low to high level features like harmony or rhythm. (2) The context of the music such as lyrics, video clips, artist info and semantic labels. (3) The user listening to music, such as demographics and experience. (4) The context the user is in, such as mood and temporal context. The need for common representations in music using linked open data is discussed in [30], where they developed a semantic audio analysis interface for extracting content features of music. Another automatic tag classification of music was found to be improved in [31] by using the Music Information Retrieval Evaluation eXchange[2] datasets. An example of these datasets is majorminer.org, which is created through its own music labeling game. The goal of the game is to label music with words that people agree on. Several games with a purpose exist for annotating music, such as Listen Game and TagATune. The latter resulted in the Magnatagatune dataset, which according to [31] contains 21642 songs annotated with 188 unique tags. These findings from music information retrieval can be applied in this study for sound and video annotations.

3 Problem Statement and Contributions

The identified problem is that multimedia annotations are sparse, homogeneous, do not represent everything that can be heard or seen. Furthermore, crowdsourcing tasks are designed to stimulate agreement, while an open-ended approach is necessary to capture the full spectrum of subjectivity in human interpretations. Here it does not matter whether these are subjective or objective annotations, as they are both relevant interpretations. Based on this, the main research questions is: *Are open-ended crowdsourcing tasks a feasible method for capturing the*

[2] http://www.music-ir.org/mirex/.

interpretation space of multimedia? In order to answer this, we need to investigate how to evaluate the quality of the results, how to efficiently design such tasks and how this can solve the growing need for improved classification and semantic understanding of the content of multimedia.

1. The first step is to assess the quality of annotations captured using open-ended tasks. This is done under the hypothesis that open-ended tasks contribute to a larger interpretation space than closed tasks. The question to be answered is: *How can the quality of results of open-ended tasks be measured?* In order to answer this, the following sub questions are asked:
   - How can crowdsourcing quality measures be improved for open-ended crowdsourcing tasks which have no clear answer? The quality of open-ended tasks has shown to be more difficult to measure because these tasks can be highly subjective [21].
   - How can the confidence that people performing crowdsourcing tasks have in their answers be measured? With a multitude of answers that people provide, it is expected that they are more confident in some answers than others.
2. The second step is to assess the design of open-ended tasks for gathering large interpretation spaces. This is done under the hypothesis that design features such as pre-selected answers or visual clues have a positive effect on the efficiency of open-ended crowdsourcing tasks while maintaining the openness of the task. The research question asked is: *Can open-ended crowdsourcing tasks efficiently generate reliable ground-truth data?* This is investigated using the following sub questions:
   - How can constraints be used in open-ended tasks to improve the detection of low quality results? By constraining the tasks, the results may be able to be validated better.
   - How can existing automated feature extraction methods be used to optimize the use of the crowdsourcing tasks?
   - How can the threshold be measured for gaining a clear distribution of answers? If the interpretation space is unknown, a measure has to indicate when there is a clear distribution.
3. The third step is to assess the usability of the captured interpretation spaces. The hypothesis is that a larger interpretation space generated through open-ended tasks results in improved descriptions of the content, which lead to better search and discovery of the multimedia. The research question is: *How can a ground truth with a large interpretation space of what can be heard or seen in multimedia improve their search and discovery?*
   - How can the annotated human interpretation space of multimedia be combined with the context and content factors of both the entity and user which influence the human perception?
   - Can the novel ground truth data improve the indexing of multimedia on existing platforms? By analyzing the ground truth improvements, we can measure if and how the quality of search results can be improved.

The main contribution of this paper is to investigate how open-ended crowdsourcing tasks can be used for gathering training data on things for which the interpretation space is unknown. Because it is unclear what the classifications are, designing crowdsourcing tasks becomes more difficult. The results should lead to more efficient crowdsourcing tasks resulting in higher quality representations of what can be heard in a sound or seen in a video. This will lead towards systems that have a more human friendly interaction through improved search and discovery of multi-representational entities.

4 Research Methodology and Approach

First, the quality of annotations captured through open-ended tasks has to be evaluated. In order to test this experiments will be performed with different open and closed approaches that represent different gradations of constraints. The least constrained design is free-text input The results will be evaluated using the CrowdTruth disagreement-based metrics [10] by transforming the annotated answers into a vector space representation. This allows for the evaluation of the quality of the annotations, the task, and the people performing annotations [32]. The metrics will have to be improved for these open-ended tasks in order to support interpretation spaces for which the size is unknown. As a result the vector space representation may be infinite, but an estimation of its size may be possible to deal with the unknown dimensions. The confidence people have in their answers can be tested through several approaches such as asking people to order their answers by level of confidence or using peer-reviews as done in [19]. This can then be compared to measures such as time spent on the task, the amount of times listened to a sound or video and the order the annotations were made in.

The second step of the approach is to find the optimal design of open-ended tasks for gathering large interpretation spaces. The task should have a clear design to avoid misunderstandings [18], but several features such as free text input, auto-completion or gamification can be used to reach the most efficient task in terms of total cost and time spent to complete it. Another method is to present an answer choice of pre-defined probable answers, but allow users to add more options. This can be extended by showing the options other users have added. By adding constraints to the open-ended tasks, the detection of low quality workers can be improved. This can again be done by testing and comparing features that are normally present in closed tasks. For instance a two-step task like presented in [6], where the user is allowed to provide self-contradicting answers. This has proven a useful measure, but requires at least two steps.

Another method of optimizing the tasks is by using automated feature extraction. Several studies have shown that high level descriptions such as abstract can be extracted [30,31]. By giving predefined descriptions through distant supervision, the annotator is forced to focus on making low-level annotations. Experiments have to show whether this approach works and does not bias the users

into providing only high level descriptions that are similar to the predefined descriptions. Finally, an experiment has to be performed to compare measurements that indicate when there is a clear distribution of answers. In other words: how do you know when you have enough annotations?

For instance, short clear sounds may need less descriptions than long ambiguous sounds. This means that instead of requiring a fixed number of annotators, it can be dynamic depending on the complexity or ambiguity of the thing to annotate. This can not only help save cost by reducing the amount of needed annotations, but also increases the captured annotations for the most ambiguous examples. Content descriptive features such as length, pitch and melody can be extracted can be tested for determining the optimal threshold.

The third step in the approach is to test the usefulness of the captured interpretations. First, the annotations of sounds and videos are placed within the context and content the users and items. For instance, some annotations may describe sound content while other describe the context, which can be categorized following the research described in [29]. Combining the different factors with the context and content of the user is expected to improve the performance of search results. These are to be further investigated by testing whether the indexing of sounds with the ground truth built through crowdsourcing in this research improves the results. This requires an information retrieval evaluation that can be performed using measures such as precision and recall and by assessing users.

5 Preliminary or Intermediate Results

In [6] we describe our initial work on crowdsourcing semantic interpretations for open-domain questions answering. The goal of this study was to gather training data on open-domain questions for IBM Watson, as part of the Crowd-Watson collaboration between IBM Research and the VU University Amsterdam. The problem with open-domain questions is that it is difficult to define who experts are, and both the question and the answer can be highly ambiguous. For Watson to better understand why a text passage justifies the answer to a question, we used multiple crowdsourcing tasks to map question and answer pairs and disambiguate terms. The results showed that CrowdTruth is an efficient approach for gathering ground truth data on open-domain questions that can have multiple interpretations and answers. By designing the crowdsourcing tasks with limited constraints, we found that self-contradicting answers were an effective measure for identifying low quality annotations.

In [14] we describe and publish the VU sound corpus[3], which is a continuation on the work of [33]. The gathered corpus consists of fine-grained annotations of 2000 short sound effects in the Freesound database[4]. The annotations were obtained through a simple free-text input form, where per sound 10 crowd workers were asked to describe with keywords what they heard in a given sound. Due to the open-ended design of the task there was high disagreement between

[3] http://dx.doi.org/10.5281/zenodo.35508.
[4] http://freesound.org.

the annotators, which increased the difficulty of detecting low quality results. The perspectives of crowd annotators proved to be important, because there was a large distinction in descriptions made by the author of a sound and the crowd. These descriptions are essential for having effective search and discovery of sounds.

6 Evaluation Plan

The approach presented in the previous chapter is assessed through the three hypotheses and associated research questions. The first evaluation is the quality of the gathered crowdsourcing annotations. The use of open-ended tasks should result in a larger interpretation space than if closed tasks were used. The second evaluation is of the design of the open-ended tasks. Alternating between different features should conclude which features result in a higher quality of results. Furthermore, the quality of the task can be evaluated through the cost and time of the crowdsourcing task to complete. The third evaluation is of the performance of the captured interpretations, which can be measured through precision and recall and by assessing users. This will show the effectiveness of the larger interpretation space of sounds.

7 Conclusions

Information systems should always have a semantic understanding of their content. This is exemplified in this study through sounds and videos, because their annotations are sparse and homogeneous. Furthermore, there is a semantic gap between the available descriptions and what can be heard or seen in those sounds and videos. The existing approaches can be improved because they stimulate agreement between annotators or do not deal with the fact that the interpretation space is unknown.

This study aims to investigate how open-ended crowdsourcing tasks and disagreement-based metrics can be used to capture the complete human interpretation space of multimedia. The approach is to first investigate how the quality of results for open-ended crowdsourcing tasks can be measured. Next, the design of these open-ended tasks is assessed for gathering large interpretation spaces, and their usability for improving the search and discovery of multimedia. The preliminary experiments have shown that the approach cam be feasible, but evaluation through applying the ground truth is necessary.

Acknowledgements. I would like to thank Dr. Lora Aroyo for her supervision, Dr. Matteo Palmonari for his guidance, Emiel van Miltenburg for our collaborative work and Robert-Jan Sips for his support.

References

1. Nixon, L., Troncy, R.: Survey of semantic media annotation tools for the web: towards new media applications with linked media. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC Satellite Events 2014. LNCS, vol. 8798, pp. 100–114. Springer, Heidelberg (2014)
2. Jiang, L.: Web-scale multimedia search for internet video content. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, WSDM 2016, p. 701. ACM, New York (2016)
3. Aljanaki, A., Wiering, F., Veltkamp, R.C.: Emotion based segmentation of musical audio. In: Proceedings of the 15th Conference of the International Society for Music Information Retrieval (ISMIR 2014) (2015)
4. Campos, G., Quintas, J.: On the validation of computerised lung auscultation. In: Proceedings of the International Conference on Health Informatics (BIOSTEC 2015), pp. 654–658 (2015)
5. Singh, P., Lasecki, W.S., Barelli, P., Bigham, J.P.: Hivemind: A framework for optimizing open-ended responses from the crowd. Technical report, URCS Technical Report (2012)
6. Timmermans, B., Aroyo, L., Welty, C.: Crowdsourcing ground truth for question answering using crowdtruth. In: WebSci (2015)
7. Lin, C.H., Mausam, M., Weld, D.S.: Crowdsourcing control: moving beyond multiple choice. In: Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence (2012)
8. Inel, O., et al.: CrowdTruth: machine-human computation framework for harnessing disagreement in gathering annotated data. In: Mika, P., et al. (eds.) ISWC 2014, Part II. LNCS, vol. 8797, pp. 486–504. Springer, Heidelberg (2014)
9. Aroyo, L., Welty, C.: Measuring crowd truth for medical relation extraction. In: AAAI 2013 Fall Symposium on Semantics for Big Data (2013)
10. Soberón, G., Aroyo, L., Welty, C., Inel, O., Lin, H., Overmeen, M.: Measuring crowdtruth: disagreement metrics combined with worker behavior filters. In: Proceedings of 1st International Workshop on Crowdsourcing the Semantic Web (CrowdSem), ISWC, pp. 45–58 (2013)
11. Inel, O., Aroyo, L., Welty, C., Sips, R.-J.: Domain-independent quality measures for crowd truth disagreement. J. Detect. Representation Exploit. Events Semant. Web, 2–13 (2013)
12. Aroyo, L., Welty, C.: Truth is a lie: 7 myths about human annotation. AI Mag. **36**(1), 15–24 (2015)
13. Macanas, J., Ouyang, L., Bruening, M.L., Muñoz, M., Remigy, J.C., Lahitte, J.F.: Development of polymeric hollow fiber membranes containing catalytic metal nanoparticles. Catal. Today **156**(3), 181–186 (2010). doi:10.1016/j.cattod.2010.02.036
14. van Miltenburg, E., Timmermans, B., Aroyo, L.: The VU sound corpus: adding more fine-grained annotations to the freesound database. In: LREC 2016 (2016)
15. Salek, M., Bachrach, Y., Key, P.: Hotspottinga probabilistic graphical model for image object localization through crowdsourcing. In: Twenty-Seventh AAAI Conference on Artificial Intelligence (2013)
16. Kurve, A., Miller, D.J., Kesidis, G.: Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention. IEEE Trans. Knowl. Data Eng. **27**(3), 794–809 (2015)

17. Lasecki, W.S., Homan, C., Bigham, J.P.: Architecting real-time crowd-powered systems. Human Comput. **1**(1), 69 (2014)
18. Liu, D., Bias, R.G., Lease, M., Kuipers, R.: Crowdsourcing for usability testing. Proc. Am. Soc. Inf. Sci. Technol. **49**(1), 1–10 (2012)
19. Sullivan, P., Clarke, D., Clarke, B.: Using content-specific open-ended tasks. In: Sullivan, P., Clarke, D., Clarke, B. (eds.) Teaching with Tasks for Effective Mathematics Learning, vol. 104, pp. 57–70. Springer, New York (2013)
20. Ooi, W.T., Marques, O., Charvillat, V., Carlier, A.: Pushing the envelope: solving hard multimedia problems with crowdsourcing. MMTC e-letter **8**(1), 37–40 (2013)
21. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2013)
22. Schulze, T., Nordheimer, D., Schader, M.: Worker perception of quality assurance mechanisms in crowdsourcing and human computation markets. In: Proceedings of 19th Americas Conference on Information Systems, AMCIS 2013, pp. 1–11 (2013)
23. Maji, S.: Discovering a lexicon of parts and attributes. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012 Ws/Demos, Part III. LNCS, vol. 7585, pp. 21–30. Springer, Heidelberg (2012)
24. Walczyk, J.J., Roper, K.S., Seemann, E., Humphrey, A.M.: Cognitive mechanisms underlying lying to questions: response time as a cue to deception. Appl. Cogn. Psychol. **17**(7), 755–774 (2003)
25. Nudds, M., O'Callaghan, C.: Sounds and Perception: New Philosophical Essays. Oxford University Press, Oxford (2009)
26. O'Callaghan, C.: Objects for multisensory perception. Philos. Stud. **173**(5), 1269–1289 (2016). doi:10.1007/s11098-015-0545-7
27. Ekeroma, A., Kenealy, T., Shulruf, B., Hill, A.: Educational and wider interventions that increase research activity and capacity of clinicians in low to middle income countries: a systematic review and narrative synthesis. IBM J. Res. Dev. **3**, 120 (2015)
28. Boland, M.R., Miotto, R., Gao, J., Weng, C.: Feasibility of feature-based indexing, clustering, and search of clinical trials. Methods Inform. Med. **52**(5), 382–394 (2013). doi:10.3414/ME12-01-0092
29. Schedl, M., Widmer, G., Knees, P., Pohle, T.: A music information system automatically generated via web content mining techniques. Inform. Process. Manage. **47**(3), 426–439 (2011). dx.doi.org/10.1016/j.ipm.2010.09.002
30. Allik, A., Fazekas, G., Dixon, S., Sandler, M.: Facilitating music information research with shared open vocabularies. In: Cimiano, P., Fernández, M., Lopez, V., Schlobach, S., Völker, J. (eds.) ESWC 2013. LNCS, vol. 7955, pp. 178–183. Springer, Heidelberg (2013)
31. Seyerlehner, Klaus, Schedl, Markus, Sonnleitner, Reinhard, Hauger, David, Ionescu, Bogdan: From Improved Auto-Taggers to Improved Music Similarity Measures. In: Nürnberger, Andreas, Stober, Sebastian, Larsen, Birger, Detyniecki, Marcin (eds.) AMR 2012. LNCS, vol. 8382, pp. 193–202. Springer, Heidelberg (2014). doi:10.1007/978-3-319-12093-5_11
32. Aroyo, L., Welty, C.: The three sides of CrowdTruth. J. Human Comput. **1**, 31–34 (2014)
33. Lopopolo, A., van Miltenburg, E.: Sound-based distributional models. In: IWCS 2015, p. 70 (2015)