Agreement Behavior of Isolated Annotators for Maintenance Work-Order Data Mining

Emily M. Hastings[1], Thurston Sexton[2], Michael P. Brundage[3], and Melinda Hodkiewicz[4]

[1] *University of Illinois, Urbana, IL 61801, USA*
ehstngs2@illinois.edu

[2,3] *National Institute of Standards and Technology, Gaithersburg, MD 20814, USA*
thurston.sexton@nist.gov
michael.brundage@nist.gov

[4] *University of Western Australia, Crawley WA 6009, AUS*
melinda.hodkiewicz@uwa.edu.au

ABSTRACT

Maintenance work orders (MWOs) are an integral part of the maintenance workflow. These documents allow technicians to capture vital aspects of a maintenance job, including observed symptoms, potential causes, and solutions implemented. MWOs have often been disregarded during analysis because of the unstructured nature of the text they contain. However, research efforts have recently emerged that clean MWOs for analysis. One such approach is a tagging method which relies on experts classifying and annotating the words used in the MWOs. This method greatly reduces the volume of words used in the MWOs and links words, including misspellings, that have the same or similar meanings. However, one issue with this approach and with the practical usage of data-annotation tools on the shop-floor more generally is the usage of only one expert annotator at a time. How do we know that the classifications of a single annotator are correct, or if it is, for example, feasible to divide the tagging task among multiple experts? This paper examines the agreement behavior of multiple isolated experts classifying and annotating MWO data, and provides implications for implementing this tagging technique in authentic contexts. The results described here will help improve MWO classification leading to more accurate analysis of MWOs for decision-making support.

1. INTRODUCTION

Maintenance Work Orders (MWOs) are one of the main records of activities that occur during a maintenance event. MWOs often include data such as problems observed, corrective actions taken, potential causes, necessary parts, and time of machine breakdown. The information in these MWOs is useful in maintenance decision making; for example, it can be used in failure mode identification, problem spot identification, and calculating more accurate mean time to repair (MTTR) and mean time between failure (MTBF) metrics, which can improve maintenance strategy and efficiency (Sexton, Hodkiewicz, Brundage, & Smoker, 2018). However, the MWO data, in its raw form, is often too unstructured, informal, and filled with jargon for immediate analysis. To combat this issue, researchers have used Natural Language Processing (NLP) techniques to extract important information out of the natural language descriptions within the MWOs. Examples of recent work in this area include (Smoker, French, Liu, & Hodkiewicz, 2017; Sexton, Brundage, Morris, & Hoffman, 2017; Brundage, Morris, Sexton, Moccozet, & Hoffman, 2018; Sexton et al., 2018; Brundage, Kulvatunyou, Ademujimi, & Rakshith, 2017).

One promising area in this space involves "tagging" the MWOs, by assigning "tags" to concepts of importance in the maintenance domain (Sexton et al., 2018). For example, a MWO might contain the text "Replaced the hydraulic hose and fixed the leaking valve." The important concepts to a maintenance practitioner might be the problem that was observed, the items that were addressed, and the solutions that were provided. In this example, the problem would consist of "leak", the items are "hydraulic hose" and "valve", and the solutions are "replaced" and "fixed". Researchers at the National Institute of Standards and Technology (NIST) have created an open-source, free toolkit called Nestor[1] to aid in

[1] https://github.com/usnistgov/nestor

this tagging process by estimating the a priori importance of concepts in the corpus, and helping annotators link together potential cases of domain-specific abbreviations, misspellings, and synonyms. Once tagged, this data can then be used to apply the previously-inaccessible knowledge stored in the MWOs to improve the manufacturing process by, for example, diagnosing and addressing problems faster (Sexton et al., 2017).

The current research on this tagging approach to MWO analysis focuses primarily on introducing the method and on issues of the quality of the MWOs themselves, and has not yet examined more practical issues related to the deployment of the approach in authentic maintenance contexts. For example, this research assumes that a single person will annotate a given dataset through tagging. Indeed, it is likely the case that a single person would be assigned this task in an authentic maintenance environment that uses this technique. However, this assumption raises questions that need to be addressed. For example, is a single isolated tagger reliable or experienced enough to properly tag the dataset and produce data usable for future analysis? In addition, it may not always be the case during the use of a tagging tool that a single individual tags the entire dataset. This tagger may have only limited time to devote to the task, and so it may be necessary for additional taggers to contribute their own work later. In this alternate situation, can multiple taggers achieve sufficient agreement with each other to produce usable data and warrant splitting the task? Since the information gained from analyzing the tagged MWO data will be used in important maintenance decisions, it is crucial that these tags be accurate and reliable. In the absence of a gold standard for tagging, it is necessary to utilize an alternative validation method to answer these questions.

This paper seeks to address these issues by investigating the agreement behavior of multiple isolated experts tagging MWO data. We conducted an experiment in which six annotators independently tagged a single dataset using the Nestor toolkit. By having multiple people tag the data, agreement on classifications of the different words in MWOs (e.g., "replace" is a solution) and the aliases assigned to them (e.g., "replace," "replaced," and "repalce" refer to the same concept) could be measured to assess the level of consensus reached and the viability of using a tagging approach to analysing MWO data in practice. We found that the six annotators achieved high levels of agreement, lending support to the use of a tagging approach to clean MWO data for analysis. We also identified several opportunities for improvement of the approach; for example, performing the tagging task in an environment that supports real-time feedback or collaboration could further improve the level of consensus achieved.

1.1. Background on Crowdsourcing

These issues of agreement among multiple taggers are common in the domain of crowdsourcing, where complex jobs are broken down into smaller, granular tasks and completed individually by many people, whose work is aggregated to create a final solution (Surowiecki, 2005). A common application of crowdsourcing is generating labeled training data to be used for machine learning algorithms, for example, labelling the contents of images (Nowak & Rüger, 2010) or the emotion of speech assets (Tarasov, Delany, & Cullen, 2010). Crowdsourcing has been shown to be an effective method for generating high-quality labels cheaply and efficiently (Hsueh, Melville, & Sindhwani, 2009; Snow, O'Connor, Jurafsky, & Ng, 2008; Ambati, Vogel, & Carbonell, 2010).

The concept of inter-annotator agreement or reliability is crucial to this line of work, as it can give researchers a clearer idea of whether multiple annotations are needed for each piece of data, whether non-expert annotators are capable of providing reasonable labels, and similar knowledge (Nowak & Rüger, 2010; Brew, Greene, & Cunningham, 2010). Some common metrics of agreement used in this context are accuracy (e.g., (Nowak & Rüger, 2010; McCreadie, Macdonald, & Ounis, 2010)), the κ-statistic (Fleiss, 1971; Randolph, 2005), and correlation coefficients such as Kendall's τ (Kendall, 1938).

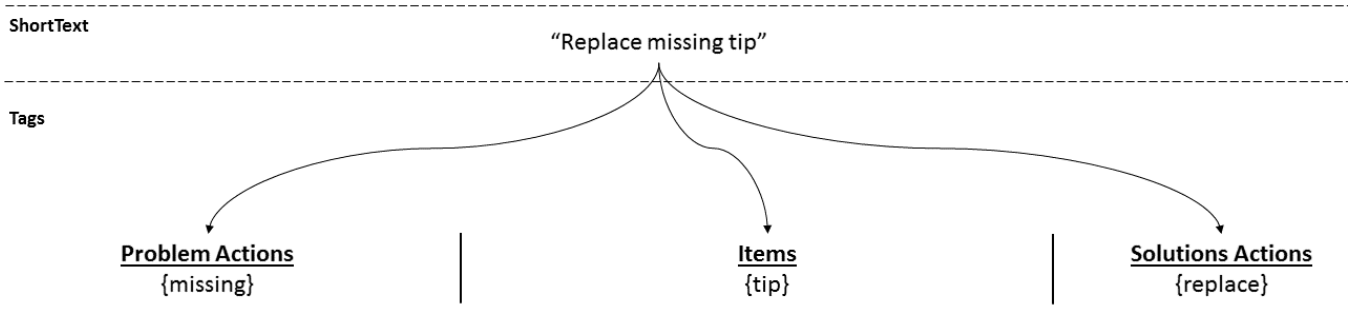

Figure 1. Taken from (Sexton et al., 2018), to illustrate the procedure used both in that work and in the case study presented here, for tagging a MWO with Nestor. Tokens are extracted from the original MWOs, and annotators are tasked with mapping each to an alias and a classification, which together form a "tag".

We ground our work in this body of prior literature and present a study reporting the levels of agreement reached by multiple experts tagging MWO data. We also provide insights into how to best measure agreement in this context.

1.2. The Tagging Tool

As described in (Sexton et al., 2018), the Nestor toolkit, used here to characterize machine-assisted tagging of MWOs, requires input of existing data to be processed, along with an assumed tag schema that represents the possible types of tag classifications within the corpus of MWOs. Here we utilize the toolkit default schema—namely, *Item*, *Problem*, and *Solution* tags, along with *Unknown* tags where some ambiguity exists without more context.

The mapping task performed by an annotator within Nestor thus consists of taking a list of extracted "tokens" — the word-level strings of text that were estimated to be statistically important per the Nestor back-end — and giving them an alias and a classification (see Figure 1). Importance in this case is given by the tokens' term-frequency/inverse-document frequency (TF-IDF) score, a common NLP metric (Leskovec, Rajaraman, & Ullman, 2014). Tokens are presented in decreasing order of importance to facilitate the efficiency of the annotator's task, and potentially related tokens that likely share an alias are recommended by a fuzzy string match. The final output is then a sort of "dictionary" that can parse out useful tags from a large variety of more informal, jargon-filled or misspelled text documents.

2. EXPERIMENTAL DESIGN

To investigate the agreement behavior of multiple users tagging MWO data, we conducted a study with 6 expert annotators from NIST and University of Western Australia. Using the "Research Mode" of the Nestor toolkit, participants tagged MWO data from a publicly available dataset about excavation machinery, which is included with the tool[2]. Research mode was used because it allows the tool to periodically and automatically save the classifications and aliases assigned so far, in order to provide insight into trends in agreement over time.

Participant volunteers were to tag the dataset using the tool's "Single-word Analysis" section for approximately 30 minutes. This task length was chosen in order to allow participants to tag a sufficient amount of data for analysis while reducing the risk of performance loss due to fatigue. Prior work has shown that factors such as vigilance decrement can lead to reduced performance during long tasks after approximately 30 minutes, especially when the tasks require discrimination based on a standard held in memory, as the tagging does (Mackworth et al., 1950; Parasuraman & Davies, 1976; Parasuraman, 1979). The task was also limited to this length in order to minimize intra-study training effects, which can occur as participants become better at experimental tasks over time. Due to the design of the tool, the highest rate of annotation occurs at the beginning of the task (with the most important tokens) (Sexton et al., 2018), and so a large amount of tagging can be completed while training effects are at their lowest.

[2]The dataset can be found at https://prognosticsdl.ecm.uwa.edu.au/pdl/ labeled *ExcavatorBucketFailures*. If using this data, please provide the proper citation.

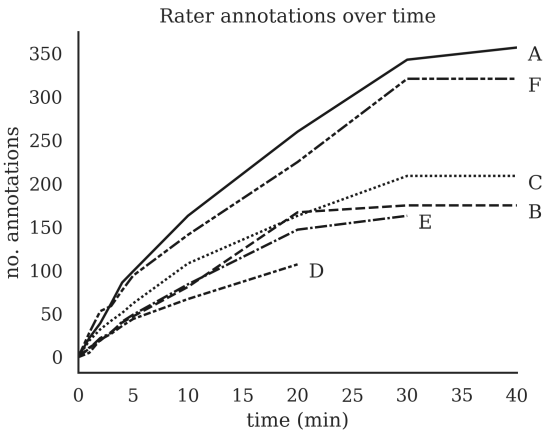

Figure 2. Participants' annotations over time.

3. RESULTS

On average, participants each completed 265 annotations during the allotted time, with a low of 175 and a high of 357. Figure 2 shows participants' progress over time. Interestingly, although participants had been instructed to perform the tagging task for 30 minutes, there was some variability in the actual amount of time spent tagging. We also observed differences in tagging rate between participants, as might be expected.

3.1. Agreement Measure

To measure the level of agreement achieved by taggers we use Fleiss' Multi-rater Kappa statistic (κ_{Fleiss}) (Fleiss, 1971). κ_{Fleiss} is given by

$$\kappa_f = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}\,, \tag{1}$$

where $\bar{P}$ represents the proportion of overall observed agreement, and $\bar{P}_e$ represents the proportion of agreement between raters expected by chance.

$\bar{P}$ is given by

$$\bar{P} = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij^2} - Nn\right), \tag{2}$$

and $\bar{P}_e$ is given by

$$\bar{P}_e = \sum_{j=1}^{k} \left(\frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \right)^2 , \quad (3)$$

where N is the number of cases, n is the number of ratings per case, and k is the number of rating categories. It is important to note that when calculating κ_{Fleiss}, not all raters need to provide a rating for every case.

Fleiss suggests that values of κ_{Fleiss} less than 0.4 indicate low agreement, values between 0.4 and 0.7 indicate good agreement, and values over 0.7 indicate excellent agreement, although these conventions vary. Other research suggests that only values of κ_{Fleiss} above roughly 0.7 indicate good agreement (Tarasov et al., 2010).

3.2. Alias Agreement

We first consider the raters' agreement on the alias assigned to each token. Participants achieved a κ_{Fleiss} of 0.85, which indicates a high level of agreement.

In addition to this analysis of the entire set of aliases produced, we also calculated the agreement on only the most important tags. These aliases were those that were associated with tokens the tool ranked in top 1 % in terms of statistical importance (recall that the tool ranks tokens' importance via TF-IDF scores). For this set, κ_{Fleiss} was 0.81, which also indicates good agreement.

We further analyzed changes in the level of agreement over time, as shown in the top section of Figure 3. The figure shows these trends for both the full set of tokens and the reduced set of important tokens. See the left side of Figure 4 for a visualization of the agreement levels for some specific high-importance aliases and tokens. The values in that matrix are the number of raters who assigned the given alias/classification to the token (n_{ij} in Eqs. 2,3). Recall that not all raters necessarily give a tag for every token.

In general, we see that agreement starts very high, when participants have tagged only a few tokens, and decreases as time goes on (although it remains high overall). This is likely due to the fact that the tool presents tokens in decreasing order of importance. Participants may be more likely to agree on how to classify the earlier, more important words, than they are on later words that occur more infrequently in the MWOs.

Another potential explanation for the dropoff in agreement is synonym cases. As can be seen in Figure 4, disagreement is most common in cases where different aliases have been assigned to words with the same meaning (e.g., "hose" and "line"). These pairs of words occur frequently in the data, especially in the important token set, which could help explain the differences in patterns between the full set of tokens and the reduced, popular set.

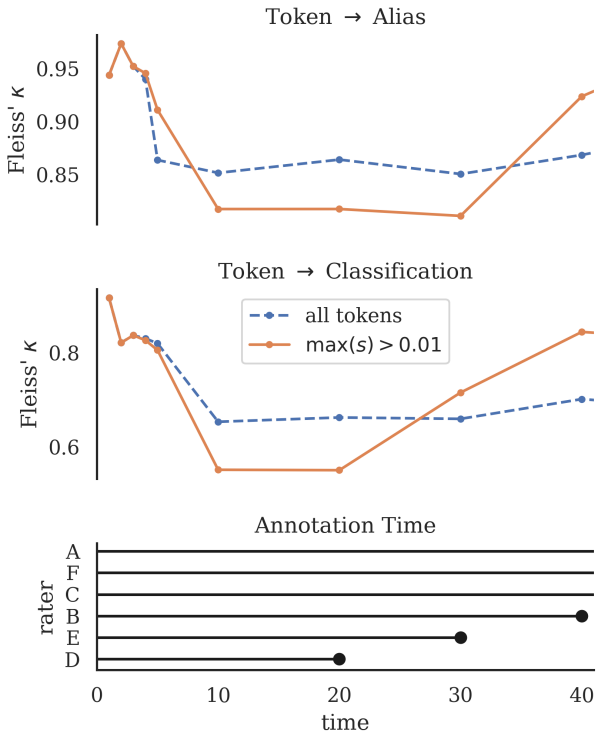

Figure 3. Agreement trends over time. Top: Alias agreement. Middle: Classification agreement. Bottom: Time progression of the task, showing participants still actively tagging.

Interestingly, we see in Figure 3 that agreement improves (especially for the important tags) after the participants who tagged for a shorter amount of time had finished. This trend points to how individual differences in annotators can have an impact on the tags assigned to tokens and the level of agreement reached.

3.3. Classification Agreement

In addition to measuring agreement on the aliases assigned to tokens, we also calculated κ_{Fleiss} for the classification of tokens into the concept categories used by Nestor. Inter-rater agreement at 30 min for token classification was 0.66, which indicates a good level of agreement, although lower than the other values of κ_{Fleiss} we measured.

As with the alias analysis, we additionally considered the participants' agreement on the classification of the most popular tokens. Here, κ_{Fleiss} was 0.72, which again indicates good agreement. See the right side of Figure 4 for a visualization of the classification agreement for a selection of the popular tokens.

We also repeated the analysis of changes in agreement over time. These results are shown in the middle section of Figure 3. Again, the figure shows these trends for both the full

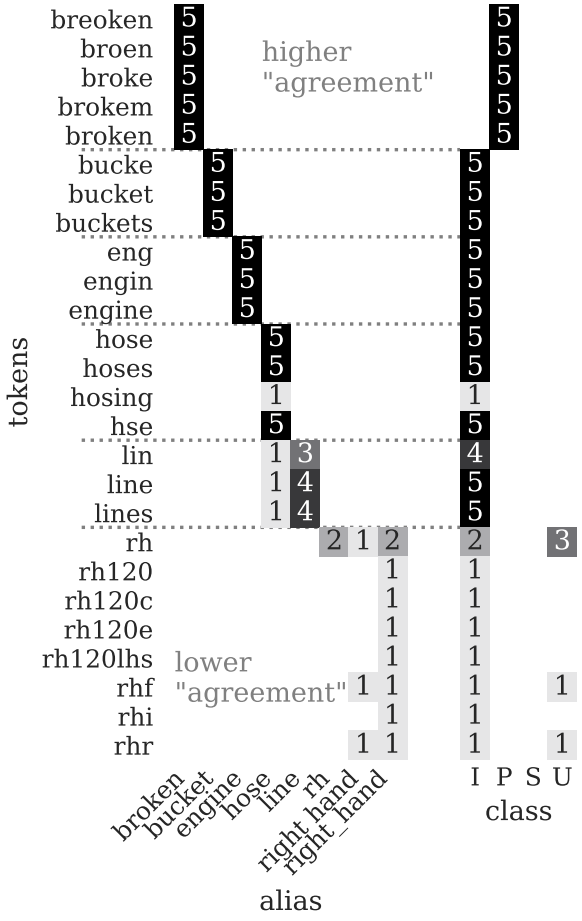

Figure 4. Matrix of rater agreement for a selection of high-importance tags and their corresponding tokens.

set of tokens and the reduced set of important tokens. The general trends are the same as described previously for the alias analysis.

One notable observation about the classification results is that the category which seemed to cause the most disagreement was "Unknown." We saw that many of the tokens had at least one participant label them as "Unknown," however these labels did not achieve the same kind of consensus that those given for the "Item" category did.

4. DISCUSSION

In this study, we examined the agreement behavior of multiple independent experts tagging MWO data for analysis. We found that the annotators were able to achieve high levels of agreement ($\kappa_{Fleiss} = 0.85$ for alias assignment and $\kappa_{Fleiss} = 0.66$ for concept classification on the set of all tags).

This finding has two main implications. First, since the aliases and labels assigned by a single expert are similar to those assigned by the entire group of experts, users of tools like Nestor can be reasonably confident that a single user can produce a set of valid tags that can be used in future analysis. Second, in the case that a single user is not tagging the entire dataset, users can again be reasonably confident that having multiple taggers will not compromise the quality of the set of tags produced.

However, our results also indicate a few opportunities to further improve agreement and the quality of the tags produced. We observed the following potential modes of disagreement:

1. Synonym token sets lead to split decisions among annotators, depending on domain ambiguity *See "hose" vs. "line" in Fig. 4. Some prefer retain specificity, while others tend to generalize.*
2. Token shorthand, ambiguity, and abbreviation is likely to be classified as "Unknown" by a subset of users. *This increases for tokens with many low-frequency variants, that not all users reach or understand (see "rh" in Fig. 4).*
3. Compound or multi-token aliases can hard to classify consistently. *For instance, some users retain "right hand" as two distinct concepts put together, while others chose "right_hand" as a single conceptual modifier.*

It is important to note the role that domain familiarity might be playing in the occurrence of these observed disagreements. As seen in Fig. 2, the speed of annotation, differs significantly among annotators. Additionally, the agreement level increased dramatically with the end of D and E annotators' participation (Fig. 3). This must be kept in mind, as such annotations performed in the field will likely come from analysts needing to prepare data, but not necessarily having

domain-specific expertise needed for high-agreement, reliable tagging.

In all of these modes, performing the tagging task in a collaborative environment that can offer real-time feedback, rather than tagging individually, could help resolve disagreements. For example, a user might be able to see how previous annotators tagged a token before finalizing their own decision, which may lead to greater convergence. This is especially true if the hypothesized familiarity differences are at play.

This concept of simultaneous groups of users arriving at a shared vocabulary at the most relevant level of abstraction for information retrieval is central to the idea of a *folksonomy* (Peters, 2009). These "folk"-taxonomies are built on the idea that annotation should reflect the vocabulary of a user-base (here, the technicians and operators storing information for future retrieval in MWOs). Folksonomies allow for convergence on a shared vocabulary in collaborative settings, and can facilitate communication between users, which is an ideal scenario for time-constrained practitioners on the shop-floor.

A limitation of this study is that we considered only one method of measuring agreement. By using κ_{Fleiss} we measured what could be called the "nominal agreement" between tokens and tags; i.e., did the exact label assigned to this token by Expert A match the exact label assigned by Expert B? An alternative way of conceptualizing agreement, particularly for aliases, would be to focus less on the exact alias assigned to a token, but rather on the group of tokens assigned the same alias, regardless of what that alias is.

One way to implement this more "topological" agreement measure would be to frame the annotation process as a graph and quantify disagreement using metrics like Graph Edit Distance (GED). Formulating crowd-sourced tagging efforts as graphs has a long and rich history in the folksonomy literature, where ranking the quality of folksonometric annotations through topology has been done with success previously (Hotho, Jäschke, Schmitz, & Stumme, 2006).

For instance, we could view an annotation session as the construction of two bi-partite graphs of the form $G = (V, E)$, where the vertex sets V can be split into two disjoint sets (e.g. the token nodes and the alias nodes, or the alias nodes and the classification nodes). The edges E in a bipartite graph only exist between sets; e.g. an "edge" between each of several tokens and their representative tag means that the tag is synonymous with all of its connected tokens.

In this way, such a bipartite graph would represent the annotations made by a user in a session of the experiment. The graph edit distances between each user's annotations would be high if the users disagreed on the the set of edges or the size of each node-set. This allows more flexibility in user annotation naming, valuing pattern similarity instead. Perhaps more interestingly, these GEDs could be quickly disaggregated and updated in realtime, with user similarity scores estimated over many time-steps for any of a number of local subgraphs, just as e.g. Eksombatchai et al., 2018 do for recommending pins/boards in the Pinterest "bipartite graph."

5. CONCLUSIONS AND FUTURE WORK

In this work we presented a preliminary study examining the agreement behavior of multiple isolated experts tagging MWO data. The results of the study have implications for implementing the tagging technique for MWO data analysis in authentic maintenance contexts: the annotators had high levels of agreement, suggesting that tagging by a single expert or by multiple experts are both feasible approaches. In addition, we identified potential opportunities for improvement of the tagging technique and tools that implement it. For example, performing the tagging task in a collaborative environment supporting real-time feedback could further improve the level of agreement achieved.

Domain knowledge could be an important factor in the level of agreement reached by multiple taggers. For example, users with less manufacturing experience may be more likely to tag concepts as "Unknown" than more experienced users, who might possess the background to apply a more appropriate classification, leading to a lower level of agreement between multiple taggers. Future work can further tease apart the effects of prior domain or tagging experience on agreement and investigate workflows that explicitly utilize differing levels of experience, such as those in which more experienced users train novices.

It is also possible that characteristics of the MWO dataset, such as complexity or domain, could impact the agreement among taggers. Future work could further investigate these factors' effect on agreement and whether our results generalize to other datasets.

Finally, the information that is provided from the MWO tagging needs to be incorporated into the maintenance decision workflows. This analysis is ongoing and will be further studied in future work.

ACKNOWLEDGMENT

Thanks to Dr. William Bernstein (NIST), Dr. Madhusudanan N (NIST), Mr. Michael Hoffman (NIST), Mr. Drew Georgiades (UWA), Ms. Emily Low (UWA), and Mr. Toby Griffiths (UWA) for their efforts annotating MWOs.

DISCLAIMER

The use of any products described in this paper does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that products are necessarily the best available for the purpose.

REFERENCES

Ambati, V., Vogel, S., & Carbonell, J. G. (2010). Active learning and crowd-sourcing for machine translation. In *LREC* (Vol. 1, p. 2).

Brew, A., Greene, D., & Cunningham, P. (2010). Using crowdsourcing and active learning to track sentiment in online media. In *ECAI* (pp. 145–150).

Brundage, M. P., Kulvatunyou, B., Ademujimi, T., & Rakshith, B. (2017). Smart manufacturing through a framework for a knowledge-based diagnosis system. In *ASME 2017 12th international manufacturing science and engineering conference* (pp. V003T04A012–V003T04A012).

Brundage, M. P., Morris, K., Sexton, T., Moccozet, S., & Hoffman, M. (2018). Developing maintenance key performance indicators from maintenance work order data. In *ASME 2018 13th international manufacturing science and engineering conference* (pp. V003T02A027–V003T02A027).

Eksombatchai, C., Jindal, P., Liu, J. Z., Liu, Y., Sharma, R., Sugnet, C., ... Leskovec, J. (2018). Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference* (pp. 1775–1784).

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Folkrank: A ranking algorithm for folksonomies. In *Workshop information retrieval 2006 of the special interest group information retrieval.*

Hsueh, P.-Y., Melville, P., & Sindhwani, V. (2009). Data quality from crowdsourcing: a study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing* (pp. 27–35).

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. Retrieved from `http://www.jstor.org/stable/2332226`

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets*. Cambridge university press.

Mackworth, N. H., et al. (1950). Researches on the measurement of human performance. *Researches on the Measurement of Human Performance.*(268).

McCreadie, R. M., Macdonald, C., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. In *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (cse 2010)* (pp. 31–38).

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on multimedia information retrieval* (pp. 557–566).

Parasuraman, R. (1979). Memory load and event rate control sensitivity decrements in sustained attention. *Science*, *205*(4409), 924–927.

Parasuraman, R., & Davies, D. R. (1976). Decision theory analysis of response latencies in vigilance. *Journal of Experimental Psychology: Human Perception and Performance*, *2*(4), 578.

Peters, I. (2009). Folksonomies. indexing and retrieval in web 2.0. In (pp. 162–164). Walter de Gruyter.

Randolph, J. J. (2005, October). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. In *Joensuu learning and instruction symposium.* Joensuu, Finland.

Sexton, T., Brundage, M. P., Morris, K., & Hoffman, M. (2017). Hybrid datafication of maintenance logs from AI-assisted human tags. In (p. 1-8). IEEE Big Data 2017.

Sexton, T., Hodkiewicz, M., Brundage, M. P., & Smoker, T. (2018). Benchmarking for keyword extraction methodologies in maintenance work orders. In *PHM society conference* (Vol. 10).

Smoker, T. M., French, T., Liu, W., & Hodkiewicz, M. R. (2017). Applying cognitive computing to maintainer-collected data. In *System reliability and safety (ICSRS), 2017 2nd international conference on* (pp. 543–551).

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 254–263).

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Tarasov, A., Delany, S. J., & Cullen, C. (2010, October). Using crowdsourcing for labelling emotional speech assets. In *W3C workshop on emotion ML.* Paris, France. doi: 10.21427/D7RS4G