# Evaluating Sequence Discovery Systems in an Abstraction-Aware Manner

Eoin Rogers, Robert J. Ross, John D. Kelleher

# Evaluating sequence discovery systems in an abstraction-aware manner

Eoin Rogers, Robert J. Ross, and John D. Kelleher

Applied Intelligence Research Centre, Dublin Institute of Technology, Dublin, Ireland
eoin.rogers@student.dit.ie
robert.ross@dit.ie
john.d.kelleher@dit.ie

**Abstract.** Activity discovery is a challenging machine learning problem where we seek to uncover new or altered behavioural patterns in sensor data. In this paper we motivate and introduce a novel approach to evaluating activity discovery systems. Pre-annotated ground truths, often used to evaluate the performance of such systems on existing datasets, may exist at different levels of abstraction to the output of the output produced by the system. We propose a method for detecting and dealing with this situation, allowing for useful ground truth comparisons. This work has applications for activity discovery, and also for related fields. For example, it could be used to evaluate systems intended for anomaly detection, intrusion detection, automated music transcription and potentially other applications.

## 1 Introduction

*Activity discovery* (AD) refers to the unsupervised discovery of plausible human activities in unannotated datasets composed of sensor readings of human subjects. AD is itself a sub-field of *activity recognition*, the recognition of activities from sensor readings in a supervised manner. These technologies have potential applications in the automatic labelling of activity recognition datasets and building profiles of normal and abnormal behaviour.

Evaluating activity discovery systems in a fair manner is a major challenge for the field. A major reason for this is that pre-annotated ground truths, often used to evaluate the performance of such systems on existing datasets, may exist at different levels of abstraction to the output of the output produced by the system. We propose a method for detecting and dealing with this situation, allowing for useful ground truth comparisons.

Activity discovery is equivalent to a number of challenging problems that are known in the wider computing literature. One good example would be anomaly detection algorithms used for such applications as intrusion detection in the field of computer security. Security practitioners have already made use of machine learning algorithms for this task [1], and our work could help evaluate such systems fairly.

The layout of this paper is as follows. Section 2 investigates prior work in this area. Section 3 discusses in detail the general problem with ground truth-based methods being applied to this problem. We introduce the concept of *activity abstraction* in more detail in section 4, before utilising it to produce an activity discovery evaluation metric presented in section 5. We detail experiments carried out to evaluate the metric in section 6, and present the results of these before concluding in section 7.

## 2 Prior work

A number of existing approaches to evaluating activity discovery systems have already been proposed in the literature. Cook & Krishnan [2] provide a good overview of existing approaches, and we refer the interested reader to this reference, rather than repeat its contents in detail here.

Our notation for this section will be relatively standard: we assume the input dataset $D = \langle d_1, d_2, \ldots, d_L \rangle$ is an ordered sequence of *sensor events* drawn from an alphabet $\Sigma$. An activity discovery system is modelled as a mathematical function $g$, which takes the dataset (or a subset of the dataset) as input and returns a set of activities $Y = g(D)$. There are a number of forms that $Y$ could take, and in order to keep our discussion as general as possible, we refrain from privileging one over the other. In the simplest case, each element of $Y$ may only be a non-contiguous subset of the events in $D$. Alternatively, each element of $Y$ may in fact be an ordered non-contiguous sequence of events, or even a grammar- or state machine-like object that could allow for the learning of complex activities with optional and mandatory elements, complex rules relating to the allowed ordering of elements within activities, or even probabilistic activity rules.

### 2.1 Stability-based metrics

When we evaluate any machine learning system, we are usually interested in determining the degree to which the learned model is *generalised* (that is, the degree to which it can be applied to similar but unseen data). Many authors propose the use of similar criteria for the evaluation of activity discovery systems. These take the form of measures of *stability*, where the dataset $D$ is split into training and test subsets, $B$ and $C$ respectively, such that $D = B \cap C$, and the system is evaluated by demonstrating that some property of the system is stable across both subsets. For example, the Cook & Krishnan book [2] mentioned previously outlines two stability-based metrics: *predictive* and *compressive* stability. Predictive stability measures that the activities seen in the training set are also observed in the test set with about the same frequency. By contrast, compressive stability measures the degree to which the compression ratio achieved on the training and test set is roughly equivalent. The idea of evaluating activity discovery (and related) systems via the use of compression ratios is an idea that shows great promise, and has already seen use in the wider ML community in the form of *perplexity* [5]. A related concept, although one we will skip over here due to space constraints, is *minimum description length* [8].

### 2.2 Ward et al.'s error analysis technique

Moving away from Cook & Krishnan's proposals, we feel that the contents of [10] could be relevant to the task of evaluating activity discovery systems. This paper does not propose an evaluation metric, but rather an *error analysis method*, in other words a means to detect the types of errors a system under analysis seems to make consistently. The core mechanism proposed in the paper is presented in Figure 1, which is a figure taken from the paper itself. The three sub-figures correspond to three stages in the method itself. Here, the ground truth is depicted as pale dotted lines, and the prediction output as darker bold lines.

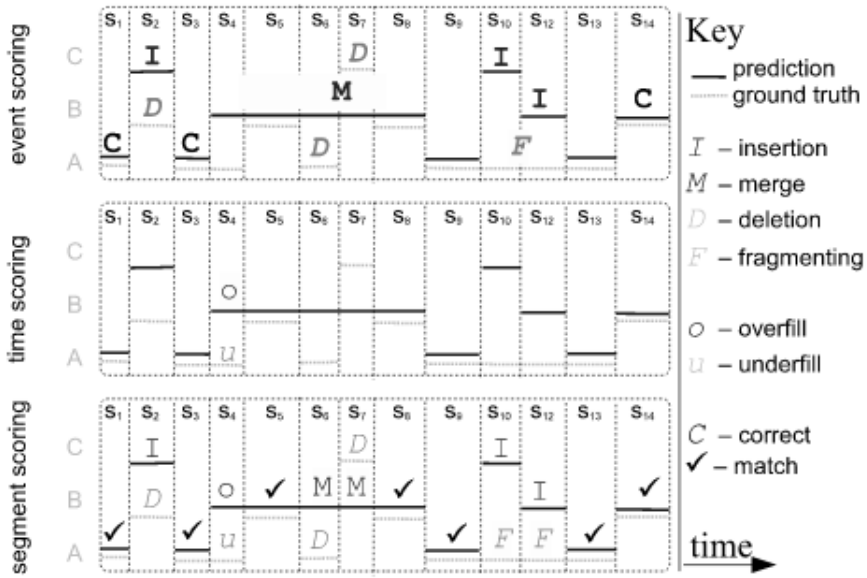


**Fig. 1.** An illustration, taken from [10], of the three stages involved in the proposed error analysis method. The first stage evaluates segments according to whether they match with an activity in the ground truth, irrespective of how accurately they may do this. The second evaluates segments according to the precision of the start and end of activities, and the fourth combines these into a finished per-segment evaluation.

The first stage (which the authors of the paper call *event scoring*) looks at each event/ground truth pair, and counts the amount of event insertions, deletions, merges and fragmentations that are observed. The second stage, *time scoring*, computes the temporal correspondence between the system output and the ground truth by counting overfills and underfills. Finally, the output from the two previous steps are combined to create the final output, which is called

the *segment scoring*, and can be seen at the bottom of Figure 1. Although simple, this error analysis technique is a major inspiration behind the metric that will be proposed later in this paper.

## 3 Ground truth-based metrics

While reading section 2, one thing that may strike some readers as unusual is the fact that most of the metrics proposed (Ward's is an exception) seem to refrain from using ground truths as a gold standard with which to compare the output of the activity discovery system under evaluation. Although a perfectly valid way of evaluating machine learning models in the general case, there are two major reasons why one might be suspicious of ground truth comparisons for activity discovery. The first of these is that, by definition, and activity discovery system must be *unsupervised*, that is it trains without making use of any sort of output data in the dataset. The entire point of the *discovery* of activities is to provide a way for the detection of plausible activities in *unannotated datasets* without any ground truth. In a real-world use case, it is quite possible that the model will therefore be trained on a dataset for which no ground truth to compare against exists, and so we have to find a means of evaluation that can be used even in these kinds of situations.

The second issue with ground truth-based evaluation is the *subjectivity inherent in the output of any activity discovery process*. Although the behaviours reflected in the sensor stream may be objective and leave no room for subjective interpretation, the same cannot be said for the activities detected in the stream. For example, the point at which an activity can be said to start and end is arbitrary. Consider the hypothetical case of a sensor stream in a house where an activity corresponding to *making dinner* takes place every evening. One could say that this activity begins when the resident(s) of the house enter the kitchen to cook, or when they turn on the oven, or when they first put food into the oven. Different activity discovery systems (and indeed human annotators) may well use different boundaries for their activities in this manner, and one cannot privilege one annotation over the other. By extension, it is also possible for entire activities to be (in a sense) subjective. For instance, what if one argues that the resident entering the kitchen to cook does not constitute part of the *making dinner* activity, but rather an activity in its own right, perhaps called *preparing to make dinner*? This issue provides major challenges for the evaluation of these systems. Note that we are speaking of a very particular kind of subjectivity. Intuitively, any system which fails to find a consistent activity every evening around dinner time in our hypothetical house seems to be objectively wrong in some way, since it cannot pick up a real pattern that exists in the data. But aspects of the pattern (its size, constitution, cardinality and so forth) are subjective in a way that makes comparison to a ground truth seem like an inherently unfair approach to evaluation.

Any proposed activity discovery evaluation metric must take these issues into consideration. Failure to do so could result in an unfair evaluation that biases in

favour of certain systems and against others without justification. Nonetheless, if a ground truth is available, it would be sensible to make use of it, even if only in addition to, rather than instead of, the unsupervised evaluation metrics mentioned in section 2.

## 4 Instances, types and abstractions

Many of the subjective differences noted in section 3 can be *attributed to differences in the level of abstraction that the various systems we are looking at are outputting*. Say $g$ and $h$ are activity discovery models, $Y_g$ is the set of activities output by $g$ (where each $y \in Y_g$ is a subset of $D$), and likewise $Y_h$ is the set of activities output by $h$ (where each $z \in Y_h$ is a subset of $D$). Note that since we don't annotate these activities as instances or types we presume they could be either. We formally represent this scenario as:

$$g(D) = Y_g \tag{1}$$

$$h(D) = Y_h \tag{2}$$

Suppose that for a particular $y \in Y_g$ and a particular $z \in Y_h$, we find that $y \subset z$, in other words $y$ is strictly a subset of $z$ (i.e. $\forall i(i \in y \Rightarrow i \in z)$, but $\exists i(i \in z \wedge i \notin y)$). We say that $z$ is thus a *more abstract* version of the activity $y$: everything in $y$ is also in $z$, but the reverse is not true. To make this more concrete, we can imagine $y$ being an activity like *making dinner*, and $z$ being a more abstract version of the same activity like *having dinner*, which contains *making dinner* in its entirety, in addition to other sensor events covering the consumption of the dinner, and perhaps cleaning up after. Notationally, we represent this scenario as $y \prec z$, which can be read as "$y$ is less abstract than $z$", or "$y$ precedes $z$".

To complicate matters further, we have to resist the temptations we may have at this point to claim that activity discovery model $g$ is less abstract than $h$, simply because it output a less abstract activity in one instance. It is entirely possible that multiple levels of abstraction are interleaved in the output of our models, i.e. it may be possible to find activities for which $g$ finds a more abstract version than $h$. Unless all activities found by $g$ are less abstract than or equal to all activities found by $h$, we should refrain from talking about the abstraction of entire models. We will later use this concept in Section [**?**] as a component in our metric.

## 5 A proposal for a new metric

By combining the insights from section 4 with the error analysis from [10] (see section 2.2), we believe that we can propose an evaluation metric for activity discovery systems using ground truths that (at least to some extent) bypasses the

second issue discussed in section 3. The idea is to use abstraction to get around the issue of subjectivity. Suppose our dataset $D$ has an associated ground truth $G$, and our discovered activities $Y$ contain an activity called *making dinner*, but the ground truth only recognises an activity called *having dinner*. The intuition is that if *making dinner* $\prec$ *having dinner*, we can mark each instance of *making dinner* as correct if it overlaps with an instance of *having dinner* in the ground truth. This is similar to Ward et al.'s proposal of marking merges, fragmentations, overfills and underfills, but rather than treating these as a sort of error, we instead allow them to be seen as correct once the types of the activities match.

We will formalise this intuition by first proposing a simple means for evaluating an activity discovery system, which will have the flaws described in Section 3. We will then modify the definition to match our proposal. Recall that $D$ is a dataset, and $G$ is the associated ground truth. We assume $|G| = |D|$, and that each element $g \in G$ is a sequence of $k$ Boolean values, where $k$ is the number of activities in the ground truth. Thus, $G_{ij}$ is true iff activity $j$ is true or active for the $i$th event in the dataset. We will also commit to a specific structure for the output $Y$, since not doing so would make our formalism needlessly abstract. We feel that the formalism can be easily adapted for other output structures and formats, although we will not attempt to prove this here. We model $Y$ as a matrix, such that each value $Y_{ij}$ represents the probability that activity $j$ is true or active for the $i$th event. This maps closely to the probabilistic output of the topic modelling based system we will be using for our experiments, yet to be discussed in Section 6 below. Given a particular ground activity $g$, an index into the output activities $y$, and a real-valued threshold value $t$, we can define the true positives of our AD system to be:

$$TP_{gyt}(G, Y) = \sum_{i=1}^{L} \mathbb{1}(g \in G_i \wedge Y_{iy} \geq t_y) \tag{3}$$

Where $\mathbb{1}$ is an indicator function that evaluates to 1 if the Boolean formula passed to it is true, and 0 if it is false. The threshold $t$ is a meta-parameter, and we compute a different value of $t_y$ for each proposed activity in $Y$. It will hopefully be clear to the reader how this could be extended to compute false positives and true and false negatives also. From here, we can obviously calculate F-measures for the system.

A diagrammatic example of our proposal is shown in Fig. 5. Here, each of the horizontal lines labelled A to E represent a single channel of information. Channel A is a ground truth, as found in an annotated dataset. Channel B represents the *raw output* of an activity discovery system for a particular event type. The output overlaps to a degree with the ground truth. We are proposing extending the length of channel B to match channel A, as shown in channel C (which is the *extended output*). We can formalise this by modifying Equation 3 as follows:

$$TP_{gyt}(G,Y) = \sum_{i=1}^{L} \mathbb{1}((g \in G_i \vee g \in G_{i-1} \vee g \in G_{i+1}) \\ \wedge (Y_{iy} \geq t \vee Y_{(i+1)y} \geq t_y \vee Y_{(i-1)y} \geq t_y)) \tag{4}$$

Now, rather than strictly requiring that the probability of activity $y$ during event $i$ at least meet our threshold, we look to the events immediately before and after the current ($i$th) event, and we will also accept event $i$ as a valid true positive if one of its neighbours are also a true positive. The ground truth is similarly extended in the same manner. We actually repeat this computation *as many times as needed for the true positive value to stop increasing.* Thus, we are willing to extend the length of both the ground truth (channel A in our diagram) and the output until their respective lengths match.

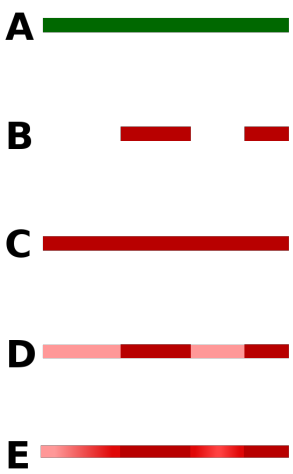


**Fig. 2.** If channel A is an output, and channel B is the output from a system under evaluation, we propose extending B to match A (*extended output*, channel C), optionally making the extensions values less than 1. In channel D (*staircase output*) we use a small value for the extensions, and we use a reducing gradient for channel E (*gradient output*).

Some people may object to the presented proposal on the basis that it is making the evaluation *too easy* for the activity recognition system. For this reason, channels D and E can be used as alternatives to C for comparison to the ground truth (channel A). In these cases, the darkness of the colour corresponds to its magnitude, with the number 1 being as dark as channels B and C, and lower numbers (closer to zero) being represented with a lighter number. Most

evaluation metrics (raw similarity, F-measures and so on) work by counting the number of matches between two binary channels. For example, F-measures build a confusion matrix, and match a False from both the ground truth and output channels as a true negative, a False from the ground truth and True from the output as a false positive and so on. We are proposing to use non-binary, fuzzy values instead of these binary comparisons, so that we would increment the counts for the confusion matrix by a number between zero and one. For channel D (the *staircase output*), we use a value of 1 for the true overlaps, and a smaller value (0.4, or $\frac{2}{5}$ in our experiments, see below) for the extensions. Formally, this becomes:

$$\begin{aligned} TP_{gyt}(G,Y) = \sum_{i=1}^{L} \mathbb{1}(g \in G_i \wedge Y_{iy} \geq t) + \\ \frac{2}{5}(G_{i-1} \vee g \in G_{i+1} \vee Y_{(i+1)y} \geq t \vee Y_{(i-1)y} \geq t) \end{aligned} \tag{5}$$

Here, we now have 2 indicator functions: the $\mathbb{1}$ function from previously, and a new $\frac{2}{5}$ function, which returns a value of $\frac{2}{5}$ if its input is true, and false otherwise.

Finally, in channel E (the *gradient output*), the extensions don't have a fixed value, but rather have a value of $1 - (0.001 \times n)$, where $n$ is the number of events away from the true overlaps, but cannot have a value below zero. Again, we define an indicator function which is suitable for this purpose, but rather than calling it $\mathbb{1} - (.00\mathbb{1} \times \mathbb{n})$, we instead give it the more succinct (but less descriptive) name $\mathbb{f}$.

$$\begin{aligned} TP_{gyt}(G,Y) = \sum_{i=1}^{L} \mathbb{f}((g \in G_i \vee g \in G_{i-1} \vee g \in G_{i+1}) \\ \wedge (Y_{iy} \geq t \vee Y_{(i+1)y} \geq t_y \vee Y_{(i-1)y} \geq t_y)) \end{aligned} \tag{6}$$

Note that in Figure 5, we are only showing the extensions applied to the output channels. The ground truth channels should also be extended according to the above process.

At this point, the usual performance metrics used to evaluate ground truth-based systems can be employed. This could include raw percentage accuracy measures, or preferable a more sophisticated metric like F-measures.

## 6 Experiments and results

In order to try to determine if our proposed metric is useful, we used an activity discovery system that was presented previously by the same authors [6]. We refer the interested reader to the cited paper for a detailed explanation of how this system works, but in summary we split the dataset $D$ up into $L - w + 1$

subsets using a sliding window of length $w$ and run each window through a topic modelling algorithm as if it was a single document. This allows us to compute a probability distribution over topics for all events in the dataset. We threshold these values to assign each event to zero or more activities, using the $t_y$ threshold previously mentioned in Section 5. In effect, this threshold is the prior over activities. For each ground truth activity $g$ and output activity $y$, we compute the candidate threshold value $t_{gy}$ that comes closest to making $c_{gy} = \|P(g \in G_i) - P(Y_{iy} \geq t_{gy})\|$ (the difference between the ground truth and output activity probabilities) equal to zero. The final threshold $t_y$ is then simply the threshold that has the minimal $c_{gy}$ value over all $g$s, i.e. $t_y = argmin_{t_{gy}} c_{gy}$. This thresholding gives us a dataset of 10 channels, consisting of 5 ground truths and 5 discovered topics (outputs). We then compute the F1 score for each $(ground\ truth, topic)$ pair for each of the 4 types of evaluation shown in Fig. 5. Each ground truth is then associated *with the single topic that scores highest with it according to the extended F1 score.*

We present here the result of the experiment described above on two different datasets. The first of these datasets was generated by the author using a state machine probabilistically moving from state to state and emitting events, with some events being more common than others for each state. The results of this experiment are shown in in Table 6 below. The first two columns show the $(ground\ truth, topic)$ pairs, and the remaining columns show the raw F1 score (i.e. the score calculated *without* using our method), the extended F1, the staircase F1 and the gradient F1 respectively. The results show an interesting pattern: for each row, the raw F1 score is substantially lower than the equivalent scores computed with our proposed method. Bearing in mind that the only difference between these metrics are that the latter three take the concept of abstraction into account in the manner described above, we take this as evidence that our metrics are a fairer way to evaluate such systems. The raw F1 score is unfairly penalising the system for what could actually be valid disagreements over abstraction levels and the start and end times of activities, while our method does not do so.

**Table 1.** Performance metrics gathered by our experiment on an artificial dataset

| Topic | Label | F1 | Extended F1 | Staircase F1 | Gradient F1 |
|---|---|---|---|---|---|
| Activity A | Topic 2 | 0.6385 | 0.9521 | 0.9865 | 0.9896 |
| Activity B | Topic 3 | 0.2269 | 0.9211 | 0.9834 | 0.9853 |
| Activity C | Topic 1 | 0.3159 | 0.876 | 0.9619 | 0.977 |
| Activity D | Topic 4 | 0.1146 | 0.8426 | 0.8923 | 0.8994 |
| Activity E | Topic 0 | 0.01835 | 0.1428 | 0.8053 | 0.8192 |

In order to evaluate the metric on a more challenging dataset, we repeated the experiment on the SCARE corpus [9]. SCARE is an annotated corpus of human actions in a 3D game-like environment. This dataset has already been converted

to the necessary binary-event-based format that our system expects [7], so we used this version of the dataset. The results of this experiment are presented in Table 6. Again, one can see a substantial improvement in performance when our metric is employed. Note that the SCARE corpus is extremely challenging: it is unusual for activity recognition systems to obtain a score greater than about 0.6, let alone activity discovery systems, which must produce their output without access to the ground truth. This metric could not only give a fairer means to evaluate activity discovery systems, but potentially a fairer means to evaluate corpora used also, by highlighting excessively narrowly defined activities in a corpus's ground truth.

**Table 2.** Performance metrics gathered by our experiment on an artificial dataset

| Topic | Label | F1 | Extended F1 | Staircase F1 | Gradient F1 |
|---|---|---|---|---|---|
| goal˙move˙box | Topic 1 | 0.1349 | 0.6441 | 0.6831 | 0.8309 |
| goal˙move˙rebreather | Topic 5 | 0.3057 | 0.9484 | 0.9575 | 0.9589 |
| goal˙move˙quad | Topic 2 | 0.1111 | 0.623 | 0.5625 | 0.5472 |
| goal˙move˙silencer | Topic 4 | 0.05674 | 0.5812 | 0.8176 | 0.8444 |
| goal˙move˙picture | Topic 3 | 0.07292 | 0.5128 | 0.6491 | 0.6379 |
| null˙goal | Topic 0 | 0.0 | 0.0 | 0.0 | 0.0 |

Before moving on to the conclusion, we also present Figure 3. This consists of a selection of visualisations of the output of our system (shown as red bars in the upper half of the images) compared to the associated ground truth (shown as blue bars in the lower half of the images) running on the SCARE corpus. The complete images are of course extremely wide, and cannot therefore be reproduced in full here. However, the extracts show real-world examples of the issues that we were highlighting in this paper. Figure 3(a) shows a typical example of a length mismatch between the output and ground truth events. Here, the output is more conservative than the ground truth, and assumes that the activity both starts later and ends sooner. As noted in Section 3, a human directly annotating this dataset could have a legitimate disagreement with the annotator of the ground truth, which would lead to the system being given an unreasonably poor score. Figure 3(b) shows an example of the output activity starting later, but ending later also. Again, most metrics would penalise a system which did this, which is unfair, since the sensor event that allowed the system to recognise that the activity has started may occur after the ground truth annotation declares the onset of the activity. One interesting aspect of this is that one could argue that the quality of the output should depend on how far to the left the ground truth continues for. If the ground truth activity is very long, but was only picked up for a short while, that could indicate that the output activity was a spurious co-incidence, rather than the system finding the activity at all. This presents a case to use the staircase or gradient based versions of the metric that we have discussed. Another interesting scenario is shown in Figure 3(c). Here the ground

truth shows a strange pattern: an activity comes to an end, there is a short pause, and then it resumes. This presents some questions about the annotation of the dataset: would it be fair to just bridge the gap, and say that there was simply one long instance of the activity rather than two short instances? If the output from the system was to bridge the gap, would it be fair to give it a poorer or better score for doing that? Ones view on these issues could lead one to therefore argue *against* using the staircase or gradient variants of the metric. Again, it pays to be aware that there is in fact a degree of subjectivity on this evaluation problem, and assuming something is set in stone can lead to unreasonable conclusions.

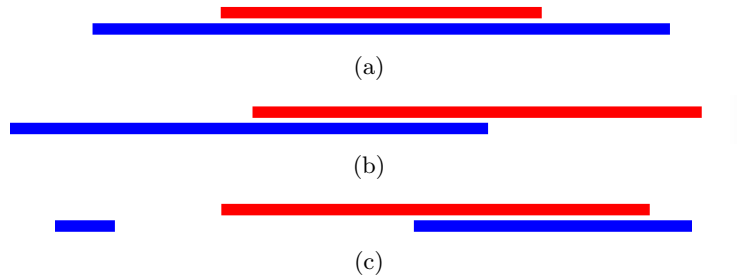


**Fig. 3.** A sample of a side-by-side comparison of an output and the ground truth. This illustrates some of the issues that we have discussed in this paper.

## 7 Conclusion

This paper has presented an argument as to why current means of evaluating the performance of existing activity discovery systems by comparison of outputs to ground truth may be construed as unfair and misleading, due to valid disagreements in abstraction level. We have proposed an extension to existing methods that we feel remedies this issue. Our experiments show that an existing activity discovery system gets a better result from our metrics. Since our metric only aims to resolve the abstraction issue, we feel that there is an argument to be made that our metrics are a fairer way to evaluate activity discovery systems, and thus help progress the state-of-the-art in the field. We are not, however, claiming that our metric is a one-size-fits-all panacea to the activity discovery evaluation problem: rather, it would be better served by using it in conjunction with other metrics (for instance, the stability-based metrics detailed in Section 2.1. We advocate the use of a suite a disparate metrics to illuminate the strengths and weaknesses of activity discovery systems.

## References


1. Buczak, A.L. and Guven, E.: A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), pp.1153-1176 (2016)
2. Cook, D.J. and Krishnan, N.C.: Activity learning: discovering, recognizing, and predicting human behavior from sensor data. John Wiley & Sons. ISBN: 978-1-119-01024-1, pp.121-124 (2015)
3. Heinz, J. and Rogers, J.: Estimating strictly piecewise distributions. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 886-896). Association for Computational Linguistics (2010)
4. Heinz, J. and Rogers, J.: Learning subregular classes of languages with factored deterministic automata. In Proceedings of the 13th Meeting on the Mathematics of Language (MoL 13) (pp. 64-71) (2013)
5. Jelinek, F., Mercer, R.L., Bahl, L.R. and Baker, J.K.: Perplexitya measure of the difficulty of speech recognition tasks. The Journal of the Acoustical Society of America, 62(S1), pp.S63-S63 (1977)
6. Rogers, E., Kelleher, J.D. and Ross, R.J.: Using topic modelling algorithms for hierarchical activity discovery. In Ambient Intelligence-Software and Applications7th International Symposium on Ambient Intelligence (ISAmI 2016) (pp. 41-48). Springer. (2016)
7. Ross, R. and Kelleher, J.: A comparative study of the effect of sensor noise on activity recognition models. In International Joint Conference on Ambient Intelligence (pp. 151-162). Springer. (2013)
8. Rissanen, J.: Modeling by shortest data description. Automatica, 14(5), pp.465-471 (1978)
9. Stoia, L., Shockley, D., Byron, D., Fosler-Lussier, E.: SCARE: A Situated Corpus with Annotated Referring Expressions. Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC08) (2008)
10. Ward, J.A., Lukowicz, P. and Tröster, G.: Evaluating performance in continuous context recognition using event-driven error characterisation. In LoCA, Vol. 3987, pp. 239-255 (2006)