An Evaluation of Learning Analytics To Identify Exploratory Dialogue in Online Discussions

Rebecca Ferguson[1], Zhongyu Wei[2], Yulan He[3], Simon Buckingham Shum[3]

[1] Institute of Educational Technology
[3] Knowledge Media Institute
The Open University
Milton Keynes, MK7 6AA, UK
+44-1908-654956
rebecca.ferguson@open.ac.uk

[2] Information Systems
Dept. Systems Engineering & Engineering Management
The Chinese University of Hong Kong
Hong Kong
(852) 3943 8461
zywei@se.cuhk.edu.hk

ABSTRACT

Social learning analytics are concerned with the process of knowledge construction as learners build knowledge together in their social and cultural environments. One of the most important tools employed during this process is language. In this paper we take exploratory dialogue, a joint form of co-reasoning, to be an external indicator that learning is taking place. Using techniques developed within the field of computational linguistics, we build on previous work using cue phrases to identify exploratory dialogue within online discussion. Automatic detection of this type of dialogue is framed as a binary classification task that labels each contribution to an online discussion as exploratory or non-exploratory. We describe the development of a self-training framework that employs discourse features and topical features for classification by integrating both cue-phrase matching and *k*-nearest neighbour classification. Experiments with a corpus constructed from the archive of a two-day online conference show that our proposed framework outperforms other approaches. A classifier developed using the self-training framework is able to make useful distinctions between the learning dialogue taking place at different times within an online conference as well as between the contributions of individual participants.

Categories and Subject Descriptors

K.3.1 [**Computers and Education**]: Computer Uses in Education – *collaborative learning, distance learning.*

General Terms

Measurement, Design.

Keywords

computational linguistics; cue-phrase matching; discourse analytics; educational dialogue; exploratory dialogue; learning analytics, educational assessment; *k*-nearest neighbour; MaxEnt; self-training framework; social learning analytics; social learning; SocialLearn; synchronous dialogue

LAK '13, April 08 - 12 2013, Leuven, Belgium

1. INTRODUCTION

Learning analytics are concerned with the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs [1]. As learning is a complex process, encompassing both knowledge construction and identity formation [2], researchers are unable to measure learning itself. Instead, they take measures such as group performance, behavioural engagement and student grades as proxies for learning [3-5]. However, these proxies reveal little about 'the dynamic processes involved in the joint creation of meaning, knowledge and understanding' [6].

Learner dialogue has the potential to be more revealing, because language functions as a psychological tool directed towards the mastery of mental processes [7]. In forms of talk in which knowledge is made publicly accountable and reasoning is visible, it is possible to observe learning taking place as speakers negotiate and express shifts in understanding.

Mercer and his colleagues identified three social modes of thinking employed in the classroom (see (§2.1): disputational, cumulative and exploratory talk [8-10]. Of these, exploratory talk is most characteristic of an educated discourse. In classroom contexts, exploratory talk can be employed by teachers and taught to students, thus producing measurable improvements in their learning achievements [11]. Although these forms of dialogue were first identified in face-to-face classrooms, learners also employ them within online text-based discussion [6, 12, 13].

A previously reported pilot study analysed synchronous text chat that took place during an online conference and identified cue words and phrases that are indicative of exploratory dialogue [14-16]. This suggested that learning analytics could be developed to distinguish different types of contribution within text chat, and to support learner engagement in fruitful learning discussions. However, the manual approach employed in the pilot study was time consuming and was not capable of identifying all relevant cue phrases.

The research reported here therefore uses methods from computational linguistics, the interdisciplinary field that deals with statistical and rule-based modeling of natural language from a computational perspective. These methods are used to assess the pilot study [16] and to identify the associated challenges, to

develop and test a self-training framework for analysis of online textual discussion, and to develop prototypes of learning analytics with the potential to support both learners and teachers.

The paper is organised as follows. We review related work in the fields of educational research and computational linguistics in order to clarify the research challenge (§2). We then set out our experimental method and, in so doing, introduce self-training frameworks, the *k*-nearest neighbours approach, *n*-grams and cue phrases (§3). We then analyse our results (§4) before describing prototypes that show how our method of exploratory dialogue detection could be employed to support learners and teachers (§5). We go on to outline ethical considerations of visual analytics (§6) before concluding with a consideration of the significance and originality of our study and the identification of possibilities for future research and development (§7).

2. DIALOGUE ANALYSIS

2.1 Exploratory dialogue

An extensive programme of work by Mercer and his colleagues has identified three social modes of thinking employed in the classroom: disputational, cumulative and exploratory talk [6, 8, 9, 11-13, 17-20]. Disputational talk is unproductive, characterised by individuals restating their own point of view while rejecting or ignoring the views of others. Cumulative talk is potentially more constructive; speakers build on each other's contributions, adding their own information and constructing a body of shared knowledge and understanding, but they do not challenge or criticise each other's views.

Exploratory talk is more characteristic of an educated discourse because it involves constant negotiation. Explanations and reasoning are made explicit where necessary and all participants make critical evaluations in order to reach joint conclusions. Mercer and Littleton provide a clear description of its use in a school environment:

Exploratory talk represents a joint, coordinated form of co-reasoning in language, with speakers sharing knowledge, challenging ideas, evaluating evidence and considering options in a reasoned and equitable way. The children present their ideas as clearly and as explicitly as necessary for them to become shared and jointly analysed and evaluated. Possible explanations are compared and joint decisions reached. By incorporating both constructive conflict and the open sharing of ideas, exploratory talk constitutes the more visible pursuit of rational consensus through conversation [11, p62].

Studies have shown that, in classroom contexts, teachers can employ exploratory talk and teach it to students, thus producing measurable improvements in learning achievement [19, 21-23].

Exploratory dialogue has also been found to support learning in online textual discussions [12, 13]. The classification of these three types of dialogue is well suited to large-group and informal discussion in which participants may not be working through a single argument, but are likely to be engaged in a more 'messy' process that intertwines multiple strands of argument and varying personal goals. For this reason, we prefer 'exploratory dialogue' to 'collaborative reasoning', which is specifically a taught approach [24]; to teacher-led 'effective discourse' [25]; to 'accountable talk', which is concerned primarily with content-focused talk [26]; and to 'argumentative knowledge construction', which assumes a goal of convergence on a joint solution [27].

Previous studies of exploratory dialogue have employed sociocultural discourse analysis to locate and study examples [28]. This method combines detailed analysis of dialogue in specific events with comparative analysis across a sample of cases.

Our pilot study applied this method to the synchronous discussion related to a two-day online teaching and learning conference (see §3.3 for dataset details) and identified 94 cue words or phrases that could be indicative of exploratory dialogue, including:

Critiques	eg However, I'm not sure, maybe
Discussion of resources	eg Have you read, more links
Evaluations	eg Good example, good point
Explanations	eg Means that, our goals
Explicit reasoning	eg Next step, relates to, that's why
Justifications	eg I mean, we learned, we observed
Others' perspectives	eg Agree, here is another [15]

These cue phrases could be used to distinguish meaningfully between conference sessions and to support evaluation of those sessions. However, this was a labour-intensive approach that made little use of the data-crunching power of computers and it was therefore not possible to identify all possible discourse cues signaling the presence of exploratory dialogue.

2.2 Automatic detection of exploratory dialogue

The automatic detection of exploratory dialogue is closely related to dialogue act detection, an area of computational linguistics. A dialogue act is the meaning of an utterance at the level of illocutionary force [29]. In other words, it is the function of a sentence, or part of a sentence, within the dialogue. 'Hi', 'Hello' and 'Good morning' are different words and phrases, but all function as the dialogue act of greeting. Other dialogue acts include: question, thank, introduce, suggest, feedback, confirm and motivate [30]. Based on detailed analysis of extensive annotated datasets, some dialogue act tag-sets have emerged as pseudo-standards in this area [31]. These large annotated datasets and tag sets are used to train classifiers that can distinguish between different dialogue acts.

The detection of exploratory dialogue in online discussion can be seen as a binary classification problem, requiring a classifier able to label each turn in the dialogue as exploratory or non-exploratory. Our pilot study highlighted three difficulties associated with the development of a classifier for exploratory dialogue.

1. The annotated dataset is limited. Although there are many online discussions, there are very few annotated corpora developed for the detection of exploratory dialogue. This rules out most supervised training methods.
2. Text classification problems are typically topic driven. In the case of exploratory dialogue, the focus is on dialogue features that are not topic dependent.
3. Despite this focus on dialogue features, the topic of the dialogue is relevant when identifying discussion that is off-topic and should not be classified as part of an ongoing exploratory dialogue. Therefore, both discourse and topical features should be considered when identifying exploratory dialogue.

The research challenge for this study was, taking these difficulties into account, to develop a classifier capable of discriminating between exploratory and non-exploratory contributions to online text dialogue.

3. METHOD

To address the three challenges outlined above, we propose a Self-training from Labelled Features (SELF) framework to carry out automatic detection of exploratory dialogue from online content. Our proposed SELF framework makes use of a small set of annotated data and a large amount of un-annotated data. In addition, it employs both cue-phrase matching and *k*nn-based [32] instance selection to incorporate discourse and topical features into classification model training. The SELF framework makes use of self-learned features instead of pseudo-labelled instances to train classifiers by constraining the model's predictions about unlabeled instances. It avoids the incestuous bias problem of self-training approaches that use pseudo-labelled in stances in the training loop. This problem arises when instances are consistently mislabeled, which makes the model worse instead of better in the next iteration.

3.1 Self-training frameworks

Self training is a form of semi-supervised learning [33] that makes use of both labelled and unlabeled data for training. In self training, a supervised model is first trained using labelled data. The trained model then assigns a pseudo-label to each instance of unlabeled data. These pseudo-labels are associated with a confidence value indicating how certain the model is about this identification. Instances with high confidence values are retained, classed as pseudo-labelled instances because they have been identified by the model rather than by humans, and automatically merged into the existing set of annotated data. The process is repeated, using the expanded dataset, until a given endpoint. This may be reached when there is no improvement in performance, when few or no labels are changed when the process is run, or when the process has run a specified number of iterations.

However, the instance-based self-training method suffers from an incestuous bias problem because adding mislabeled instances to the training pool can degrade model performance, progressively reducing the performance of the classifier as the training process is repeated. We therefore employed a feature-based self-training approach, training a classification model using labelled features instead of labelled instances. While it is possible to employ any classifier – for example the Naïve Bayes [34] or Support Vector Machines [35] – we only focused on training the Maximum Entropy (MaxEnt) model [36] from labelled features, using the Generalised Expectation (GE) criteria [37], and left the investigation of other classifiers as future work. The self-training loop derived labelled features from pseudo-labelled instances and added these self-labelled features to the original labelled feature set in order to re-train the model. The feature-based self-training approach calculates word-class association probabilities by averaging over many pseudo-labelled examples. This has a smoothing effect, making this more tolerant to class prediction errors than an instance-based approach and thus avoiding the incestuous bias problem.

3.2 Running a self-training classifier with a *k*-nearest neighbours approach

A *k*-nearest neighbours approach works on the basis that turns in the discourse are likely to have the same classification as those closest to them in terms of their topical features. When using this approach, a section of dialogue is first assigned a pseudo-label. These labels are described as pseudo because they are temporary – the classifier may or may not decide they are correct. The classifier then assesses the probability that the pseudo-label is correct by checking the labels of a number (k) of the nearest neighbours of the pseudo-labelled instance.

A benefit of this approach is that it makes use of local topical information within the dialogue to improve classification accuracy. When the classifier method is applied to a specific piece of dialogue, *k*nn provides a way of increasing the salience of domain-specific vocabulary. This can reduce the errors introduced by pseudo-annotated instances generated by the classifier.

The combined process begins, as with the self-training framework described in §3.1, by training the GE MaxEnt classifier on manually annotated data, then running unlabeled data through the classifier. The classifier assigns each unlabeled turn in the dialogue p ($p_1, p_2, \ldots p_n$) a pseudo-label (l) and a confidence value (c) for that label.

For each instance of p, the classifier examines its set, pn_i (pn_{i1}, pn_{i2}, … pn_{ik}), of k-nearest neighbours in terms of topical features. These nearest neighbours have the pseudo-label set ln_i (ln_{i1}, ln_{i2},… ln_{ik}) and confidence values cn_i (cn_{i1}, cn_{i2},… cn_{ik}). The support value of p_i is indicated by S_i which is computed using this equation.

$$S_i = \frac{\sum_{j=1}^{k} \delta(l_i = ln_{ij}) cn_{ij}}{k}$$

Here, $\delta(x)$ is an indicator function that takes a value of 1 if x is true, 0 otherwise.

Only the pseudo-labels with a support value above a value (R) determined by the researchers are considered to be correct. Instances with a lower support value for their pseudo-labels are discarded. This provides a new list of pseudo-annotated instances, (p'1, p'2, …p'$_n$) – the labeling of these is not only based on their features, but has also been checked against their local context. This list is merged into the annotated dataset, and the revised dataset is used to retrain the classifier.

3.3 Dataset

For the pilot study, data were collected from *Elluminate*, a web conferencing tool that supports chat alongside video, slides and presentations. The focus was on the synchronous text-based discussion related to a two-day online teaching and learning conference organised by The Open University in 2010 (OUC2010). The *Elluminate* text chat in four conference sessions, each between 150 and 180 minutes in length (24,530 words in total) was investigated. During these four sessions, 233 participants logged in to the *Elluminate* sessions at one or more times and 164 of these contributed to the synchronous discussion. The majority of participants were higher education researchers and practitioners from around the world, although most were UK-based [for more information on the dataset, see 15].

Each contribution to the OUC2010 text chat was considered to be a turn in the dialogue. There were 2,636 of these, containing 6,789 distinct word tokens in all. Turns in the dialogue were typically short, containing a mean average of 10.14 word tokens.

1.1 Category	1.2 Description	1.3 Examples
1.4 Challenge	1.5 A challenge identifies that something may be wrong and in need of correction	1.6 calling into question 1.7 calling to account 1.8 contradicting 1.9 disputing 1.10 finding fault with 1.11 proposing revision 1.12 putting forward an opposing view 1.13 raising an objection
1.14 Evaluation	1.15 An evaluation has a descriptive quality	1.16 appraising 1.17 assessing 1.18 expressing in terms of something already known 1.19 judging
1.20 Extension	1.21 An extension builds on, or provides resources that support, discussion	1.22 applying idea to a new area 1.23 increasing the range of an idea 1.24 linking to, developing or providing related resources 1.25 requesting additional resources to support understanding 1.26 taking the same line of argument further
1.27 Reasoning	1.28 Reasoning is the process of thinking an idea through.	1.29 asking questions *about content* 1.30 changing position in the light of arguments presented 1.31 explaining *1.32* inferring 1.33 justifying your position 1.34 reaching a conclusion 1.35 working ideas out in a logical manner

Table 1: Coding scheme for sub-categories of exploratory dialogue. Dialogue turns coded in any of these categories were also coded as exploratory. All other turns were coded non-exploratory

For this study, we constructed an additional, un-annotated dataset from three massive open online courses (MOOCS), LAK11, LAK12 and CHANGE. Again, synchronous text-based discussion from *Elluminate* sessions was collated. In this case, 49 sessions were considered. During these sessions, 1152 participants contributed 10,568 turns to the dialogue. Turns in the dialogue were once again short, containing a mean average of 9.24 word tokens.

Each of the *Elluminate* sessions included in the dataset was an open, public event. Participants signed in using recognisable names, recognisable online identities, role titles (for example, moderator) and a variety of other pseudonyms. Participants were aware that the sessions would be archived and would be made openly available for replay online. In many cases, participants were also aware that the archived sessions would be used as research data. We therefore consider the records of these sessions to be in the public domain.

3.4 Data preparation

We hired two postgraduate students to annotate a subset of OUC2010, using the coding scheme set out in Table 1, above. Each coder received one morning's training. Their task was to classify each turn in the dialogue as exploratory or non-exploratory. All exploratory turns were also assigned one or more sub-category labels (challenge, evaluation, extension, reasoning). Dialogue transcripts were presented in chronological order so that annotators could make decisions based on contextual information.

As in many learning environments, participants did not only learn about subject matter, but also about the tools available to them (such as *Elluminate*) and about their fellow learners. Only dialogue related to subject matter was coded as exploratory (for example 'I don't think your microphone is working' was not classed as evaluation).

Cohen's Kappa coefficient was used to assess pairwise agreement between coders making category judgments, correcting for expected chance agreement [38]. Inter-annotator agreement was 0.5977 for the binary classification exploratory / non-exploratory. This was taken to be moderate agreement, meaning that this coding was reliable enough for the data to be used to train a classifier.

Inter-annotator agreement was only 0.3887 for the multi-class classification into sub-categories. Agreement on sub-categories was therefore considered to be unreliable and these were not used to train the classifier.

In order to increase reliability, only turns in the dialogue that were classified in the same way by each coder were included in the annotated dataset used to train the classifier. The coding work therefore provided 2087 coded turns in the dialogue, 1417 coded as exploratory and 670 coded as non-exploratory.

As OUC2010 was a two-day conference, with an afternoon and a morning session each day, the annotated dataset was divided into four subsets of roughly equal size, based on date and time of day: OU22AM, OU22PM, OU23AM and OU23PM.

3.5 Discourse features incorporating cue phrases

In order for the self-training process to run, dialogue turns had to be presented as features. In this study, the features were *n*-grams – *n*-character slices of a longer string [39], including one-word segments (unigrams), two-word segments (bigrams) and three-word segments (trigrams). As an example, the phrase 'that is why' would form three unigrams ('that' 'is' 'why'), two bigrams ('that is' and 'is why') and one trigram ('that is why'). These three types of *n*-gram were initially employed because the cue phrases identified in the pilot study included unigrams, bigrams and trigrams. All three types were retained because preliminary

experiments showed that combining unigrams with bigrams and trigrams gave better performance than using any one of them alone, or any two of them together.

The 94 cue phrases identified in the pilot study were found to have high precision (see Table 2, below) when used in a classifier – when they identified as a dialogue turn as exploratory it was almost always a turn that a human classifier also identified as exploratory. However, they missed many exploratory turns and therefore had low recall performance. This meant they were not suitable for use as the only training elements for a classifier. Nevertheless, due to their high level of accuracy, they were used to identify exploratory dialogue within the un-annotated data in order to improve the accuracy of un-annotated dataset handling.

3.6 Experimental set-up

The study compared seven approaches to the development of an exploratory dialogue classifier. The aim was to explore the overall effectiveness of the proposed self-training framework (described as approach 7, below), and the effectiveness of the two integrated components, cue-phrase matching and *k*NN-based instance selection. The seven approaches were:

1. Cue-phrase labeling (CP) Developed manually in the pilot study, this approach searches turns in the dialogue for cue phrases

2. Supervised MaxEnt (MaxEnt) A supervised MaxEnt classifier, trained using annotated data

3. Generalised Expectation (GE) A MaxEnt classifier, trained using labeled features based on GE criteria. Labelled features are accepted if the probability that they are associated with one category exceeds 0.65

4. Self-training features (SF) The feature-based self-learning framework, without cue-phrase matching or *k*NN instance selection, described in §3.1.

5. Self-training features, including *k*NN (SF+KNN) *k*NN-based instance selection is used to select the pseudo-labelled instances from which the self-labelled features are derived

6. Self-training instances, including *k*NN and cue phrases (SI+CP+KNN) Documents labelled by an initial classifier are taken as training examples for a subsequent classifier

7. Self-training features, including *k*NN and cue phrases (SF+CP+KNN) Our proposed method. The feature-based self-learning framework with *k*NN described in §3.2, incorporating the cue-phrase matching method

In each run of the experiment, one of the four sections of the annotated dataset OUC2010 was used as a test set, and all or part of the remaining annotated dataset was used to train the classifier. The un-annotated dataset was used for self-training. In order to evaluate performance, all possible training/testing combinations were tested, and the results of these runs were averaged. Where cue phrases were added, this was done using the same exploratory/non-exploratory ratio that was present in the initial training set.

Four evaluation criteria were used to evaluate the experiment outcomes.

Accuracy: How many of the classifier's decisions were correct? Calculated by dividing the number of correctly identified turns in the dialogue by the total number of turns in the dialogue.

Precision: How many of the turns classified as exploratory were actually exploratory? Calculated by dividing the number of exploratory turns by the number of turns classified as exploratory.

Recall: How many exploratory turns are classified as exploratory? Calculated by dividing the total number of exploratory turns by the number of exploratory turns correctly identified by the classifier.

F1: A weighted average of the precision and recall. F1 values range from 0 (completely inaccurate) to 1 (completely accurate).

4. RESULTS

4.1 Overall performance

Table 2 shows the exploratory dialogue results obtained from the OUC2010 dataset by using the seven methods identified above. In each case, half a session was selected from the four conference sessions for training purposes. The total number of dialogue turns used for training therefore ranged between 220 and 330. The cue-phrase method (CP) provided the greatest precision, over 95%. At the same time, it had the lowest recall value, only 42%. The manually identified cue phrases were accurate indicators of exploratory dialogue, but they missed over half the instances of exploratory dialogue.

Approach	Accuracy	Precision	Recall	F1
CP	0.5389	**0.9523**	0.4241	0.5865
MaxEnt	0.7886	0.8262	0.8609	0.8301
GE	0.7658	0.7753	**0.8717**	0.8017
SF	0.7659	0.7572	0.8710	0.8062
SF+KNN	0.7701	0.7865	0.8539	0.8148
SI +CP+KNN	0.7888	0.8273	0.8602	0.8302
SF+CP+KNN	**0.7924**	0.8083	0.8688	**0.8331**

Table 2: Exploratory dialogue classification results, with the best result for each measure highlighted in bold

Top 10 exploratory features

feature	exploratory probability	non-exploratory probability
httpurl	exploratory:0.9999999998924732	non-exploratory:1.0752688169730606E-10
learning	exploratory:0.9999999998863637	non-exploratory:1.136363636105372E-10
i-think	exploratory:0.9999999997727272	non-exploratory:2.272727271694215E-10
also	exploratory:0.9999999997727272	non-exploratory:2.272727271694215E-10
agree	exploratory:0.9999999997435897	non-exploratory:2.5641025627876394E-10
online	exploratory:0.9999999996551724	non-exploratory:3.448275859690844E-10
social	exploratory:0.9999999996428571	non-exploratory:3.5714285688775506E-10
new	exploratory:0.9999999996428571	non-exploratory:3.5714285688775506E-10
students	exploratory:0.9999999996153845	non-exploratory:3.8461538431952657E-10
content	exploratory:0.9999999996153845	non-exploratory:3.8461538431952657E-10

Top 10 non-exploratory features

Feature	exploratory probability	non-exploratory probability
bye	exploratory:3.448275859690844E-10	non-exploratory:0.9999999996551724
sound	exploratory:4.999999995E-10	non-exploratory:0.9999999995
hi	exploratory:4.999999995E-10	non-exploratory:0.9999999995
see-you	exploratory:9.090909074380164E-10	non-exploratory:0.9999999990909091
room-num	exploratory:1.111111108641975E-9	non-exploratory:0.9999999988888887
good-morning	exploratory:1.428571424489796E-9	non-exploratory:0.9999999985714286
you-later	exploratory:1.6666666611111112E-9	non-exploratory:0.9999999983333333
see-you-later	exploratory:1.6666666611111112E-9	non-exploratory:0.9999999983333333
volume	exploratory:1.999999992E-9	non-exploratory:0.999999998
me-to-room	exploratory:1.999999992E-9	non-exploratory:0.999999998

Table 3: Examples of exploratory and non-exploratory features, together with the probability that these unigrams, bigrams and trigrams fall into the non/exploratory category

The original self-labelled features method (SF) yielded very similar results to the GE method, and both performed worse than the supervised classifier (MaxEnt). SF+KNN outperformed SF

alone, showing the effectiveness of employing a *k*-nearest neighbours approach.

Our proposed self-training features method, including *k*-nearest neighbours and cue phrases (SF+CP+KNN) outperformed all others in terms of accuracy and F1 value. The instance-based alternative (SI+CP+KNN) received the second-highest values for accuracy and F1, but these results were only marginally better than those generated by the MaxEnt approach.

Table 3 provides examples of the results obtained using our proposed self-training features method, including *k*-nearest neighbours and cue phrases.

4.2 Varying the size of the training set

We also explored the influence of the amount of training data on the accuracy of these approaches, and the effectiveness of incorporating cue phrases and the *k*-nearest neighbours approach. To do this, we varied the size of the annotated dataset from an eighth of a session to one session and compared the performance of different approaches. Figure 1 shows the results for the four approaches for which the amount of training data could be manipulated.

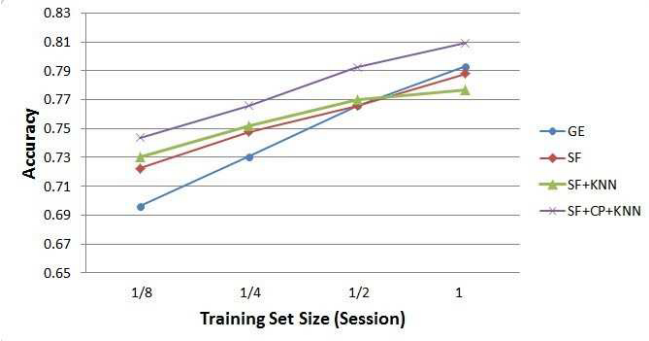

Figure 1: Variation in accuracy depending on the size of the training set

The performance of all approaches increased steadily as the size of the training set was increased. Again, our proposed approach, SF+CP+KNN, consistently out-performed the others. As the size of the training set increased, the accuracy of the GE approach rose rapidly, exceeding both SF and SF+KNN when the size of the training set reached one session. This was expected; supervised classifiers are typically more accurate than self-trained classifiers if a large enough annotated database is available for training.

Incorporating both cue phrases and *k*-nearest neighbours improved the self-training features method significantly. In Figure 1, the crosses indicating SF+CP+KNN are consistently higher than either the diamonds representing the self-training features method alone or the triangles representing the self-training method incorporating *k*-nearest neighbours without cue phrases. However, the *k*-nearest neighbours component did not prove to be stable in its influence on effectiveness. When the size of the training dataset reached one session, the *k*-nearest neighbours (+KNN) component degraded performance when compared to the self-training framework (SF).

4.3 Varying *k* in *k*-nearest neighbours instance selection

In order to explore the impact of *k*, the number of neighbours in *k*NN-based instance selection, on the performance of our proposed framework, we varied its value in SF+CP+KNN. During this part of the experiment, we used half a session of the annotated dataset to train the classifier. As shown in Table 4, the best performance is achieved when *k* is set to 3; this was the value for *k* used in the experiment that gave the results reported in Table 2.

k	Accuracy	Precision	Recall	F1
1	0.7868	0.8007	0.8666	0.8282
3	0.7924	0.8083	0.8688	0.8331
5	0.7881	0.8005	0.8685	0.8292
7	0.7586	0.7505	0.8640	0.8001

Table 4: Performance of proposed method using different values for k

5. APPLICATION OF EXPLORATORY DIALOGUE DETECTION

Automatic detection of exploratory dialogue has many potential uses in online learning environments. In the case of online interactions that occur over long periods of time, such as conferences and MOOCs, it could be used to enable learners to focus on the most productive sections of an extensive resource. Teachers could use the distribution of exploratory dialogue within a learning session as a means of evaluating the learning that had taken place. On an individual level, both the volume of exploratory dialogue contributed by a learner and the ratio of exploratory to non-exploratory dialogue could provide a basis for self-reflection or for guided improvement.

In §5.2 and §5.3, we describe two prototype implementations developed to facilitate more effective learning on The Open University's *SocialLearn* platform. All analysis described in these sections is based on session OU23AM of the OUC2010 dataset.

5.1 SocialLearn

SocialLearn (sociallearn.open.ac.uk) is a social media space tuned for learning. It has been designed to support online social learning by helping users to clarify their intention, to ground their learning and to engage in learning conversations [16]. The system's architecture includes a recommendation engine; a pipeline designed to process data and output it in a form suitable for analysis by *SocialLearn* recommendation services.

Our exploratory dialogue detection module (EDDM) has been integrated within a development build of *SocialLearn*. There it can be used to process online dialogues such as online discussions in order to generate recommendations and contribute to user profiles. When a dialogue is run through EDDM, each turn in the dialogue is assigned a rating in the range -100 to +100 to indicate the likelihood that it is exploratory (with a positive-value rating) or non-exploratory (with a negative value). These ratings can be used to generate timeline visualisations and user visualisations.

5.2 Timeline visualisation

The EDDM timeline visualisation presents an overview of the distribution of exploratory dialogue over time. Learners could use such a visualisation to focus on particular sessions of a discussion, while teachers could use it to explore the reasons for this distribution.

A straightforward way of generating such a visualisation would be to plot the ratings of each turn in the dialogue, sorted into chronological order.

However, synchronous online dialogue is noisy, with many different conversations intertwined and overlapping, so this

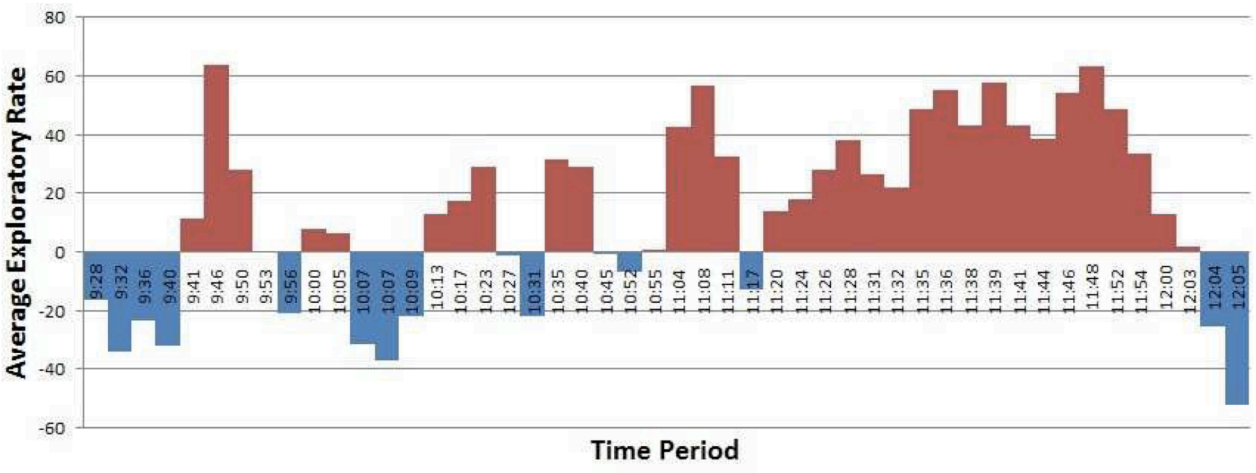

Figure 2: Timeline visualisation of the conference session OU23AM with the granularity m *set to 10*

approach generates a very volatile time series. In order to derive a more consistent signal, we smoothed the series by selecting a representative time point every *m* transcripts and then calculating a rating for that time period by averaging over the window of *m* transcripts. Figure 2 (above) shows a 150-minute conference session in 10-message groupings (note that these are shown with time stamps, but the time intervals are not equal). There are 48 representative time points of which 15 (represented in blue) have negative ratings. It is clear that the dialogue exchanges at the beginning and end of the session are mainly non-exploratory. This is not surprising, as at the beginning of the conference people introduced themselves and exchanged greetings, while at the end of the conference they thanked the speakers, said goodbye and left the session. What is more interesting is the intensive engagement in exploratory dialogue in the final 40 minutes of the session, a finding that was supported when the conference was examined manually.

5.3 User visualisation

A user visualisation (Figure 3) provides a view of the contribution of each participant to the online discussions. This could provide teachers with a tool for monitoring and supporting engagement, and a personalised version of this view could be used to support student self-reflection. In this visualisation, each user is plotted in a two-dimensional space defined by the total volume of user contributions and the number of turns rated as contributions to exploratory dialogue.

In Figure 3, each point represents a user. The gradient of the solid line could be set at any level to differentiate between participants in a learning group. Here it represents the level at which five posts in every six are exploratory. In the conference section visualised in this figure, a high proportion of the dialogue was exploratory, with around one in three of those participants who made more than five contributions to the dialogue represented on or near the solid line. As highlighted, the most productive participant contributed over 50 times, and 47 of these contributions were classified as exploratory. At the other end of the spectrum was a participant who contributed 15 times without engaging in any exploratory dialogue.

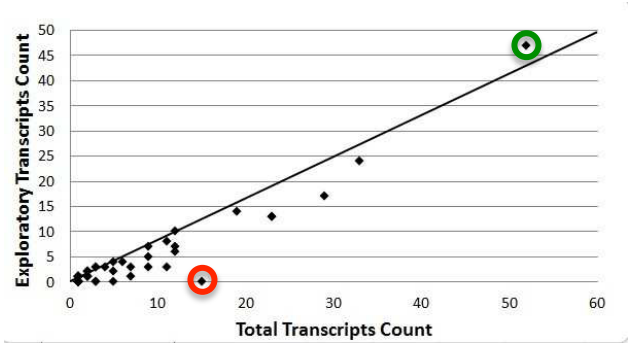

Figure 3: User visualisation of session OU23AM in OUC2010

6. DISCUSSION

6.1 Ethical considerations of visual analytics

As discussed in [16], our social learning analytics perspective draws attention to instilling reflective, self-regulating qualities in learners, and to the importance of providing analytics tools feedback to them, not just the institution tracking them. Given the two visual analytics proposed here, we must therefore consider whether it is appropriate to give all learners access to them, bearing in mind the principles of good assessment for learning.

6.1.1 Visual Literacy

One issue (equally applicable to educators and administrators) is literacy in the visual language. We propose that the bar graph timeline visualisation (Figure 2) is relatively simple to interpret but there are, of course, issues as to where thresholds are set for classification and on what scale the bars are rendered. The user visualisation (Figure 3) requires an understanding of why the threshold line is set at that angle, and whether this should remain fixed, or if it can be varied to support exploratory enquiry for different kinds of learner or course. If we were to take the Purdue *Signals* approach [40], we would not consider exposing this level of detail to untrained users, but would provide only a very simple feedback mechanism such as red/amber/green lights, with no

rationale. This has demonstrated its power and ease of adoption, but remains opaque to the end user.

6.1.2 Assessment for Learning

If we imagine visual analytics such as these properly embedded in a learning environment, the graphs would be directly linked to the underlying source texts and learner profiles. The research base underpinning the work on designing formative *assessment for learning* [41] emphasises the importance of providing motivating feedback, and clear guidance on the next steps to take to improve, in order to help develop self-directed learners. If learners are performing very poorly, according to these metrics, we would argue that they should be informed of this, and that they should be guided to examples of better contributions.

What remains to be tested is how learners respond when presented with such analytics, and whether they do indeed explore them in ways that advance their learning. Another issue is whether learners who are not performing well feel exposed by social learning analytics, which render visible in an uncompromising way the contribution patterns of individuals or (to take another example) who is 'on the edge' of a social network. It might be argued that social learning analytics of this sort could have a destructive impact on the individuality and creativity of learners who feel pressured into conforming to what the analytics have deemed to be 'good' learning behaviours. Ultimately, this comes down to the external validity of these proxies.

6.1.3 Participatory design and deployment

The participatory design (PD) movement has a long tradition within human-centred computing and ethics [cpsr.org/issues/pd]. We hypothesise that PD that involves learners and educators from the start may be particularly important for social learning analytics (cf. [42]), given the sensitivity of social processes to observation and quantification (even formative assessment, not high-stakes summative assessment).

Arguably, but not yet proven, the process of engaging a community of learners in reflecting on what such tools can or should do, and how they might be used, is a pedagogically effective strategy, since it introduces very explicitly to learners what 'good' could or should look like.

7. CONCLUSION

This paper has been the outcome of productive multivocality in the middle space within which learning and analytics intersect. Our interest in the possibilities offered by social learning analytics prompted us to combine machine learning methods and natural language processing techniques with theoretical frameworks that have been developed in the classroom using qualitative methods. In addition, this work draws on insights from psychology – particularly Vygotsky's conceptualization of language as a conceptual tool with which we simultaneously interpret and construct our experience of the world [7] and from English literature – particularly Ong's understanding of how writing enlarges the potentiality of language and restructures thought [43].

This paper is significant in that it has proposed and tested a self-training framework for the detection of exploratory dialogue within online discussions. This self-training framework employs cue phrases to make use of discourse features for classification. It also uses a *k*-nearest neighbours instance selection approach to draw on topical features and thus reduce the number of wrongly labelled instances introduced by use of a self-training method. Experimental results show that this approach out-performs six alternative methods. In the future, this approach could be used to provide visualisations and prompts that would support learners to reflect on and develop their learning dialogue. This approach could also be applied to recordings of long conference sessions or presentations to highlight areas of interest to learners.

This paper is original in that it pioneers the development of automatic exploratory dialogue detection. In order to develop our self-training framework, we have built OUC2010, the first annotated corpus for exploratory dialogue detection. We also developed prototype visualisations on the *SocialLearn* platform that show how this method could be used to produce learning analytics visualisations to support both learners and teachers. In the future, we intend to develop and test these prototypes.

In the research reported here, we focused only on the use of *n*-grams. Future studies could explore other features, such as the position of dialogue transcripts within a session. As illustrated in Figure 2, dialogue turns at the beginning and end of a dialogue session are likely to be non-exploratory. In addition, if one turn in the dialogue is exploratory, this increases the likelihood that the next turn will also be exploratory, as participants respond to a challenge or develop a train of thought. Contextual information could therefore be used to increase the effectiveness of the self-training framework. Another interesting area for future study would be to explore automatic ways of expanding the cue phrase list and combining it with machine learning methods for the detection of exploratory dialogue.

8. ACKNOWLEDGMENTS

We gratefully acknowledge The Open University for making this work possible by resourcing the *SocialLearn* project.

9. REFERENCES

[1] SoLAR, *Open Learning Analytics: An Integrated & Modularized Platform*. White Paper, Society for Learning Analytics Research. 2011.

[2] Sfard, A., On two metaphors for learning and the dangers of choosing just one. *Educational Researcher*, 27, 2, (1998), 4-13.

[3] Goodman, P. S. and A, D. L., Methodological issues in measuring group learning. *Small Group Research*, 42, 4, (2011), 379–404.

[4] Koch, J., Does distance learning work? A large sample, control group study of student success in distance learning. *E-Journal of Instructional Science and Technology*, 8, 1, (2005).

[5] Axelson, R. D. and Flick, A., Defining student engagement. Change: *The Magazine of Higher Learning*, 43, 1, (2010), 38-43.

[6] Littleton, K. and Whitelock, D., The negotiation and co-construction of meaning and understanding within a postgraduate online learning community. *Learning, Media and Technology*, 30, 2, (2005), 147-164.

[7] Vygotsky, L. S., The instrumental method in psychology. In: R. W. Rieber and J. Wollock (Eds.), *The Collected Works of L. S. Vygotsky*. Plenum Press. (Original work written 1924-1934), New York, 1997.

[8] Wegerif, R. and Mercer, N., Computers and reasoning through talk in the classroom. *Language and Education*, 10, 1, (1996), 47-64.

[9] Mercer, N. and Wegerif, R., Is 'exploratory talk' productive talk? In: P. Light and K. Littleton (Eds.), *Learning with Computers: Analysing Productive Interaction*. Routledge, London and New York, 1999.

[10] Mercer, N., Littleton, K. and Wegerif, R., Methods for studying the processes of interaction and collaborative activity in computer-based educational activities. *Technology, Pedagogy & Education*, 13, 2, (2004), 195-212.

[11] Mercer, N. and Littleton, K., *Dialogue and the Development of Children's Thinking*. Routledge, London, 2007.

[12] Ferguson, R., *The Construction of Shared Knowledge through Asynchronous Dialogue*. PhD, The Open University, Milton Keynes. http://oro.open.ac.uk/19908/ 2009.

[13] Ferguson, R., Whitelock, D. and Littleton, K., Improvable objects and attached dialogue: new literacy practices employed by learners to build knowledge together in asynchronous settings. *Digital Culture and Education*, 2, 1, (2010), 116-136.

[14] Ferguson, R. and Buckingham Shum, S., Social Learning Analytics: Five Approaches. In: *LAK12: 2nd International Conference on Learning Analytics and Knowledge* (30 April - 2 May) (Vancouver, Canada, 2012). ACM.

[15] Ferguson, R. and Buckingham Shum, S., Learning analytics to identify exploratory dialogue within synchronous text chat. In: *Proc. 1st Int. Conf. on Learning Analytics and Knowledge* (27 Feb – 1 Mar) (Banff, Canada, 2011). ACM.

[16] Buckingham Shum, S. and Ferguson, R., Social learning analytics. *Educ. Technology & Society*, 15, 3, (2012), 3-26.

[17] Mercer, N., Words & Minds: *How We Use Language To Think Together*. Routledge, London, 2000.

[18] Mercer, N., Developing dialogues. In: G. Wells and G. Claxton (Eds.), *Learning for Life in the 21st Century*. Blackwell Publishers, Oxford, 2002.

[19] Mercer, N., Wegerif, R. and Dawes, L., Children's talk and the development of reasoning in the classroom. *British Educational Research Journal*, 25, 1, (1999), 95-111.

[20] Wegerif, R., Using computers to help coach exploratory talk across the curriculum. *Computers & Education*, 26, 1-3, (1996), 51-60.

[21] Rojas-Drummond, S. and Mercer, N., Scaffolding the development of effective collaboration and learning. *Int. Jnl. Education Research*, 39, (2003), 99-111.

[22] Rojas-Drummond, S., Perez, V., Velez, M., Gomez, L. and Mendoza, A., Talking for reasoning among Mexican primary school children. *Learning and Instruction*, 13, (2003), 653-670.

[23] Wegerif, R., A dialogic understanding of the relationship between CSCL and teaching thinking skills. *Computer-Supported Collaborative Learning*, 1, (2006), 143-157.

[24] Anderson, R. C., Chinn, C., Waggoner, M. and Nguyen, K., Intellectually stimulating story discussions. In: J. Osborn and F. Lehr (Eds.), *Literacy for All*. Guildford Press, New York, 1998.

[25] Mezirow, J., Transformative learning: theory to practice. *New Directions for Adult and Continuing Education*, 74, (1997), 5-12.

[26] Michaels, S., O'Connor, C. and Resnick, L. B., Deliberative discourse idealized and realized: accountable talk in the classroom and civic life. *Studies in the Philosophy of Education*, 27, (2008), 283-297.

[27] Weinberger, A. and Fischer, F., A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers & Education*, 46, (2006), 71-95.

[28] Mercer, N., Sociocultural discourse analysis: analysing classroom talk as a social mode of thinking. *Journal of Applied Linguistics*, 1, 2, (2004), 137-168.

[29] Austin, J. L., *How To Do Things with Words*. Clarendon Press, Oxford, 1962.

[30] Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. and Meteer, M., Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26, 3, (2000), 339-373.

[31] Kral, P. and Cerisara, C., Dialogue act recognition approaches. *Computing and Informatics*, 29, (2010), 227–250.

[32] Cover, T. and Hart, P., Nearest neighbor pattern classification. *EEE Transactions on Information Theory*, 13, 1, (1967), 21-27.

[33] Zhu, X., *Semi-supervised learning literature survey*. University of Wisconsin. Available online from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.103.1693, Madison, 2005.

[34] Domingos, P. and Pazzani, M., On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29, 2-3, (1997), 103-130.

[35] Cortes, C. and Vapnik, V., Support-vector networks. *Machine Learning*, 20, 3, (1995), 273-297.

[36] Nigam, K., Lafferty, J. and McCallum, A., Using maximum entropy for text classification. *Proc. IJCAI-99 Workshop on Machine Learning for Information Filtering*, (Stockholm, Sweden, 1999), 61-67.

[37] Druck, G., Mann, G. and McCallum, A., Learning from labeled features using generalized expectation criteria. In: *Proc. 31st ACM SIGIR Conf.* (Singapore, 2008). ACM. 595-602.

[38] Carletta, J., Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22, 2, (1996), 249-254.

[39] Cavnar, W.B. and Trenkle, J. M. N-gram-based text categorization, *Symposium on Document Analysis and Information Retrieval* (Las Vegas, 1994), 161-175.

[40] Pistilli, M. D., Arnold, K. and Bethune, M., Using Academic Analytics to Promote Student Success. *EDUCAUSE Review Online*, July/Aug, (2012).

[41] ARG, *Assessment for Learning: 10 Principles*. Assessment Reform Group (assessment-reform-group.org). 2002.

[42] Govaerts, S., Duval, E., Verbert, K. and Pardo, A., *The Student Activity Meter for awareness and self-reflection*. ACM. 2012.

[43] Ong, W. J., *Orality and Literacy: The Technologizing of the Word*. Methuen, London, 1982.