ASM Sc. J., 14, 2021

https://doi.org/10.32802/asmscj.2020.576

Unsupervised Text Feature Extraction for Academic Chatbot using Constrained FP-Growth

Suraya Alias[1*], Mohd Shamrie Sainin[1], Tan Soo Fun[1], Norhayati Daut[1] and Teoh Lee Sheng[1]

[1]*Faculty of Computing and Informatics, Universiti Malaysia Sabah*

In the edge where conversation merely involves online chatting and texting one another, an automated conversational agent is needed to support certain repetitive tasks such as providing FAQs, customer service and product recommendations. One of the key challenges is to identify and discover user's intention in a social conversation where the focus of our work in the academic domain. Our unsupervised text feature extraction method for Intent Pattern Discovery is developed by applying text features constraints to the FP-Growth technique. The academic corpus was developed using a chat messages dataset where the conversation between students and academicians regarding undergraduate and postgraduate queries were extracted as text features for our model. We experimented with our new Constrained Frequent Intent Pattern (cFIP) model in contrast with the N-gram model in terms of feature-vector size reduction, descriptive intent discovery, and analysis of cFIP Rules. Our findings show significant and descriptive intent patterns was discovered with confidence rules value of 0.9 for cFIP of 3-sequence. We report an average feature-vector size reduction of 76% compared to the Bigram model using both undergraduate and postgraduate conversation datasets. The usability testing results depicted overall user satisfaction average mean score is 4.30 out of 5 in using the Academic chatbot which supported our intent discovery cFIP approach.

Keywords: chatbot; unsupervised; constrained FP-Growth; feature extraction; intent pattern

I. INTRODUCTION

A chatbot is an automated conversational agent that can communicate and respond to the user's request using both voice and text. This emerging solution currently being applied to assist customers in product recommendation, providing FAQs, healthcare support and smart home assistant (Hardalov *et. al.,* 2018; Laranjo *et. al.*, 2018; Nuruzzaman & Hussain, 2018; Pereira & Díaz, 2018). Nevertheless, as the education era becomes digitalized, online communication platforms, i.e. Telegram, Facebook, WeChat, and WhatsApp are more flexible for disseminating information against the conventional face to face method. This triggers the adoption of chatbot agents in the academic-related domain as student's assistant in managing university life (Dibitonto *et al.,* 2018) and acts as an academic advisor (Ho *et. al.,* 2018; Nwankwo, 2018). The chatbot agent has been used as library assistance (Meincke, 2018) and is useful to support repetitive university-related FAQ from students (Ranoliya *et al.,* 2017).

In general, the architecture of chatbot development can be classified into a rule and corpus-based approach with Artificial Intelligence technique. ELIZA was a classic example of a chatbot developed based on pattern-rules using regular-expressions (Weizenbaum, 1983). In comparison, corpus-based chatbot learns to converse with users based on example conversation between humans or broad corpus (film review and dialogue, twitter replies) (Serban *et al.,* 2018).

Recently, commercialized chatbot solutions have been in the market that implements AI and Machine Learning technique with NLP to naturally converse with the user by providing personalized solution based on user's intention such as Siri by Apple, Google Assistant and ALEXA by IBM.

*Corresponding author's e-mail: suealias@ums.edu.my

However, most of the standard chatbot engines are language-dependent, data-driven, and less focused in the academic domain. Hence, in this paper, we present a corpus-based chatbot model for academic domain based on the improvised FP-Growth (PrefixSpan) algorithm (Adamo, 2012). The basis of "divide and conquer" in the FP-Growth method is improvised in our model by implementing text features constraints which are the unique term constraint, adjacency term order constraint and sequence length constraint. Our goal is to discover users' intention using unstructured social messaging corpus dataset as text feature representation for our model.

This paper is a continuation study from Alias *et al. (*2019a & 2019b) for the initial undergraduate dataset and preliminary findings in Academic chatbot domain. Also, this paper thoroughly evaluates and compare the performance of the new Constrained Frequent Intent Pattern (cFIP) model using new postgraduate data in the Academic Corpus dataset against existing language models.

II. RELATED WORKS

Almansor & Hussain (2019) and Yan (2018), they generally categorized the conversational agents into two categories that are the task-oriented chatbots that perform certain tasks such as recommends and reserves best eating and travelling spot nearby. Another category is a non-task oriented chatbot, where the agents can converse socially with human for entertainment, assistance and companion. Related work of rule-based chatbot and corpus-based using Artificial Intelligence method are discussed in the next section. Following that, we discuss on text representation model for intent identification using social dataset.

A. Rule-Based Chatbot

Most of the rule-based chatbot is influenced by the early ELIZA and PARRY systems. The ELIZA chatbot engine uses hand-crafted rules such as regular-expression pattern rules and templates to match user's query to respond accordingly. The basis of the rules is to transform the response based on the associated keyword. Then, another chatbot named PARRY (Colby, 1981) has been improvised with richer language capabilities where it passed the Turing test in 1972 in the area of psychology to study schizophrenia patients. However, since the rule-based approach does not rely on training data for learning, thus opens the gap to overcome the issues when the human-chatbot conversation becomes out of the rule's context. The next-generation chatbot applied AI and NLP such as the A.L.I.C.E chatbot (Wallace, 2009), where the system applied complex pattern-matching technique and was built on top of AIML (Artificial Intelligence Mark-up Language). AIML language is based on XML structure used to design the chatbot's conversation flow which is made of element objects such as topic and categories. Each category has elements such as "pattern" and "template" to match the user's input query and generate an appropriate response based on matching template.

Some related chatbot systems that implement the AIML technique in the academic domain are Ghose & Barua (2013) and Yamaguchi *et al.* (2018). Previously, Ghose & Barua (2013) implements AIML with a graph-based Information Repository for their University FAQbot to assist the undergraduate students as academic advisor. Their topic-specific FAQbot produces better results against BaseBot in terms of topic switching ratio and dialogue correction rate performed by the user. On the other hand, Yamaguchi *et al.* (2018) automatically generate AIML rules using social media Twitter dataset. The chatbot queries and response sets are based on the informal conversation on Twitter and their findings highlight that their chatbot can engage with the user's conversation to a certain extend. Both researchers agreed that the topic switching issue revolves from the nature of how human communicates that somehow becomes a challenge for the chatbot to keep up in the current conversation.

This AIML approach is suitable for QA chatbot which is related to our proposed work, however manual generation of all possible patterns has scalable issues and may limit the capabilities of our proposed work. Thus, we proceed to review the corpus-based and AI-driven method.

B. Corpus-Based AI Chatbot

A corpus-based chatbot learns by example from human-human conversations or broad corpus to be able to respond naturally and provide solutions to users (Serban *et al.,* 2018). The current motivation in this area is to build an intelligent AI chatbot with Machine Learning technique

powered with NLP engine. Researchers (Jeong & Seo, 2019) expanded their chatbot knowledgebase using social media Twitter dataset to improve the chatbot response accuracy to be similar to human dialogue response. They tested their RuT approach against commercial bot application and produced encouraging results.

One of Deep Learning approach is a Sequence to Sequence model (Seq2Seq) where it has been used extensively in the area of Speech and Language Processing. The Seq2Seq model is based on the technique of Recurrent Neural Network, for instance, the Long Short-Term Memory (LSTM) approach. By using the Seq2Seq-LSTM model, Arsovski *et al.* (2019) proposed a new methodology for extracting knowledge from a big noisy dataset as the basis for question answering. The main highlight of the proposed work is designing automated knowledge reuse and sharing conversational model for QA. Meanwhile, related to the academic domain, Chandra & Suyanto (2019) developed their QA chatbot using the Seq2Seq model on top of Whatsapp conversations dataset from the university admission QA dataset. The total of conversation lines is 2,903, where the questions are feed as the input and the answers as the target sentences.

Another recent work by Cuayáhuitl *et al.* (2019) incorporates the ensemble-based technique with value-based Deep Reinforcement Learning using a finite state action set representation. Their concern is on facilitating the chatbot response to unseen-data, however, training with ensemble agents has shown better results as compared to single-agent as reflected in their findings.

C. Intent Identification using FP-Growth

The key to providing related feedback and response for a conversational agent is to understand the user's intention during the conversation.

Sequential Pattern Mining (SPM) approach has been widely used for transactional data for discovering user's pattern and behaviour. In the area of Text Mining and Information Retrieval, the discovered Frequent Pattern can be used to represent the document's content since the natural sequence order of words in the text are preserved. The underlying concept of the FP-Growth algorithm is to divide the search space based on the first discovered frequent pattern. Next, the local search list is conquered recursively to find the subsequent frequent pattern (Adamo, 2012).

The N-gram language model is commonly applied in most studies for word representation and speech modelling. Nonetheless, this model has the issues of data sparsity and vector size problems due to large size of word sequence probability generation (Kim *et al.,* 2012).

Thus, in this work, we outline an unsupervised Text Feature Extraction technique for Academic Chatbot using Constrained FP-Growth to identify and represent user's intent pattern called Constrained Frequent Intent Pattern (cFIP) from mining unstructured social messaging dataset.

cFIP representation is an enhanced version of the text representation model named Frequent Adjacent Sequential Pattern (FASP) developed by Alias *et al.* (2018a) and Alias *et al.* (2018b) with text feature constraints. The FASP approach has been experimented in the area of Text Summarization, Sentence Compression and finding Document Similarity. The prior results of Malay and English news dataset experiments also have shown very competitive results.

III. ACADEMIC CHATBOT MODEL

In this section, to develop our academic chatbot model, firstly, some basic chatbot definitions are described.

Utterance: An utterance is a sentence that was uttered by a user using text or speech in a conversation. For example, a message line of "*Dr, may I know when is the due date for the final report?*" or "*We want to make an appointment with Dr and asking a bit info about CS project*".

Intent: User's intention during a conversation is called an intent such as asking or requesting certain action. From the example, it can be seen that the student intends to set an appointment with the coordinator to get course information and asking for the submission schedule. Sample intent name that can be set to represent this type of query is "getCourseInfo" or "setAppmnt" which consist of a verb and a noun phrase.

Entity: An entity is a word or phrases that support an intent description to be specific so that certain actions can be performed. Based on the same utterance example the entity found is "computer science project" or "CS", a short

form of Computer Science, one of a course enrolled by the students.

A. Academic Corpus Development

For our conversation corpus, we downloaded WhatsApp messages logs with the consent from the undergraduate and postgraduate academic coordinators. The initial undergraduate conversation dataset with preliminary findings can be referred to Alias *et al.* (2019a).

Table 1 describes the corpus dataset that consists of 24 conversation files from both undergraduate and postgraduate, that includes 12 individual chats (personal) and 3 group chats. The social messages dataset contains a mix of English and Malay words, with short forms, emoticons and informal words. The goal of this research is to extract and discover users' intent patterns from the developed corpus as text features for our academic chatbot intents, entities and responses.

Table 1. Academic Corpus Dataset (Undergraduate vs Postgraduate)

Item	**Topic**		**Total**
	Postgraduate	**Undergraduate**	
Total WhatsApp files	12	12	24
Chat-type	2 groups, 10 personal	1 group, 11 personal	3 groups, 21 personal
Number of conversation lines	1702	537	2239
Number of terms	4022	1834	5856

A sample of WhatsApp log excerpt from multiple postgraduate chat conversation is illustrated in Table 2. Literally, from the conversation, we can see similar intentions querying about the Computer Science Project. The senders are asking about the course information on the topic of submission, presentation timeline and also seeking for advice. From the conversation, we can see the academic redundant queries issues, which can be curbed using automated response to assist the academicians. Furthermore, rather than manual creation of the list of intents and entity using the existing chatbot framework, our unsupervised chatbot model is trained on top of conversation corpus where new intents can be discovered that will benefit the chatbot implementation procedure.

Table 2. WhatsApp log excerpt on postgraduate students conversation with the possible intent

#	**Sample Conversation Extract**	**Possible Intent**
1	For the computer science project thesis submission, can i delay it?	Request Course Info, Request Advice
2	We want to make appointment with Dr n asking a bit info bout Cs project	Request Course Info, Set appointment
3	I'm unable to finish my project. I'm aware of the consequences.	Request Advice
4	May i know when can we get the schedule for the progress presentation	Request Course Info
5	Dr, may I know when is the due date for the final report?	Request Course Info

B. Constrained Frequent Intent Pattern Discovery

The steps for cFIP discovery include:

1. Preprocessing
2. *prefixIntent* Initialization
3. cFIP Generation
4. cFIP Rules Discovery

1. Preprocessing

For each chat messaging files, the preprocessing steps include data cleansing, tokenization, removal of invalid characters and punctuation and sentence splitting. We also remove basic stopwords for Malay and English.

Intent Pattern Discovery

Problem Definition:

Given Frequent Intent Pattern cFIP is a sequence of frequent terms uttered in a conversation, the support frequency or sup of pattern cFIP denoted as sup(cFIP) is defined as the number of sequence occurrences within the online messaging dataset.

2. prefixIntent Initialization

A *prefixIntent* is a 1-sequences list that consists of unique frequent terms denoted as $\alpha = \{t_1, t_2, t_k..t_n\}$ that meets the predefined minimum support *min_sup* threshold. In this step, for each conversation sentence *cs* in message line *n*, we calculate the utterance support *uSupp(P)* for each term *t*. The weight *w* for every *t* is the total occurrence of *uSupp(P)* in this feature-vector implementation. The *uSupp(P)* weight for each *t* is then sorted and only Top-k of 1-sequences is initialized as the *prefixIntent*. In this setting, top k=30% is used for the cFIP generation.

3. Frequent Intent Pattern Generation

In this cFIP generation step, the search space for each *cs* is divided using the *prefixIntent* list. Next, the *cs* list is conquered recursively using each term in the *prefixIntent α* list. The cFIP generation is performed on each term in the *prefixIntent* α initialized earlier. The current *α* is joined with next adjacent $t_{(k+1)}$ denoted as β based on the sequence order in the original text. We apply text features pattern constraints to the cFIP generation as follows:

1. Unique term constraint

The unique term constraints refer to the terms that belong to *prefixIntent* list, where only the listed frequent unique terms are used to generate the cFIP sequence combination. The purpose of this constraint is to curb the problem of generating insignificant combinations of sequence patterns and maintain a limited projection of the vector data set. The basic idea is to maintain each new sequence generation so that the next $t_{(k+1)}$ or β in the conversation sentence *cs* is constrained as one of the elements in the *prefixIntent* (or $\beta \in \alpha$) list.

A new $cFIP_m$ is generated only if the current β value is frequent and unique in the inspected document. The equation for unique item constraint can be written as $C_{item}(FIP)$ given in Equation (1).

$$C_{item}(cFIP) \equiv (\forall_t: 1 \leq t \leq len(cs); cs[t] \in \alpha) \quad (1)$$

Let α is *prefixIntent*, cFIP is a set of patterns to be mined and *cs* is a conversation sentence with the order of terms *t*. For each term *t*, it must be equal or greater than 1 and less than the length of the *cs*. The item *t* in *cs* must also belong to the element α. For example, to generate a cFIP of 2-sequence containing "computer science", the term "computer" must first meet the constraint of unique frequent item C as below:

$$if\ sup\ (X) \geq\ \sigma\ \wedge\ C_{item}(X) = true$$

2. Adjacency term order constraint

Adjacency term order constraint refers to the constraint on the next order of words *(k+1)* in the sequence of a conversation sentence. This constraint is considered crucial when dealing with natural language because it affects the meaning of words being uttered by the user. This constraint is applied to preserve the semantics of the term sequence so that the generated cFIP can be expressed as a feature set of rules that describes the content of a document (conversation in this case). The adjacency term order constraint is written as $C_{adjs}(cFIP)$ in Equation (2). Suppose t_{k+1} in cs or β that follows the sequential order of word in *cs*; β must be one of the elements in the *prefixIntent* set as described in unique item constraints and has been reviewed for each cFIP sequence generation by taking into account the sentence size constraints.

$$C_{adjs}(cFIP) \equiv (\forall_{cs}: \beta = t_{k+1}; 1 \leq \beta \leq len(cs); \beta \in \alpha) \quad (2)$$

For instance, the cFIP 2-sequence of "computer science" where the prefix α is "computer" and β is "science" is generated from the utterance order "computer science project" of the original conversation excerpt in Table 2.

The term at the adjacent sequence position that is "science" is said to correspond to the cFIP generation rule when it meets the constraint the $C_{adjs}(cFIP) = true$.

C. Sequence Size Constraint

This constraint refers to the length limit or size of a pattern. The sequence size constraint for cFIP determined in this work is based on Bigram (2-sequence of cFIP) or Trigram (3-sequence of cFIP) sizes. This constraint is applied to limit the size of cFIP generation and to speed up the mining process by focusing to discover only significant textual patterns. The sequence length constraint is represented as $C_{len}(cFIP)$.

For instance, to generate up to cFIP of 3-sequence or Trigrams, it can be written as:

$$C_{len}(cFIP) \equiv (len(FIP) \geq 3) \quad (3)$$

Generating cFIP for each inspected document up to the term vector size of m is based on Equation 4 where it should fulfil all the stated constraints in Equation (1)-(3):

$$cFIP_{\mathrm{m}} = \alpha \cup \beta;\ \text{where}\ sup(cFIP) \geq \sigma \ \&\&\ C(cFIP) = true \quad (4)$$

4) cFIP Rules Discovery

The algorithm to generate the cFIP with text feature constraints is illustrated in Figure 1 where the support and confidence value for the discovery of the cFIP rule is calculated by referring to the basic Sequential Rules Mining in Equation 5. The value of *min_sup* and *min_conf* threshold is predefined by the user. The output from this process is a term feature vector represented by cFIP with support and confidence value.

$$X \rightarrow Y\ is\ a\ Sequential\ Rule\ if;\ conf\ (X \rightarrow Y\ ) \geq \min\ \_conf; where\ support\ (X \rightarrow Y) = support\ (X \cup Y)\ and\ conf(X \rightarrow Y) = \frac{support\ (X \cup Y)}{support(X)} / \quad (5)$$

```
Algorithm 1: cFIP(cs, σ, k, l)
Input: list of conversation sentences cs,
       min_sup σ,
       top frequent_term k,
       sequence_size l
Output: Term_Vector cFIP with support and confidence

for(cs:S)
α <- frequent(top(term(k), 1_sequences)))
        sort(α)
        for(α:cs)
                generate_frequent_sequence(len, α)
                β <- next_adjacent(k+1)
                cFIPm = β∈α? α∪β:next
return cFIPm
calculate_supConf(cFIPm)
add_vector(MxN,weight(cFIPm)>σ && (unique(cFIPm))
```

Figure 1. cFIP with text feature constraints

IV. RESULTS AND DISCUSSION

In this paper, we experimented using two academic topics that refer to the undergraduate and postgraduate social conversation using the developed chatbot corpus. The total of conversation lines is 2239 with a total of 24 WhatsApp files of 3 groups and 21 individuals. Our cFIP model was implemented using Java language, meanwhile, the N-gram model was implemented using Python NLTK (Natural Language Toolkit). For consistency, the term frequency threshold for each model was set to $\geqslant 2$, to avoid generating insignificant word combination with the intuition that any important conversation should be mentioned more than once.

In this section, we present and compare our findings based on the term-feature vector size dimension, descriptive intent pattern discovery, analysis of the discovered cFIP rules and usability testing results.

A. Term-feature Vector Size

Figure 2 illustrates the term vector size comparison between the two models (N-gram vs cFIP) using the postgraduate and undergraduate social conversation dataset. Similar results have been found for both datasets using the N-gram; whereby exponential term vector size generated when moving from Unigram to Bigram. However, since we had limited the threshold on the frequency weighting in this experiment, we can see that the Trigram generation has been reduced, indicating lesser Trigram has been found, to avoid the data sparsity problem.

In contrast with the cFIP findings, for both datasets, we can see reduced term vector size generation from 1-sequences until 3-sequences frequent patterns. We relate these findings based on the textual constraints implemented on the cFIP generation which are the unique term constraint, adjacency term order constraint and sequence size constraint to filter unmeaningful patterns upfronts. An average of 76% term feature-vector size reduction was found in this experiment when compared to the Bigram model using both undergraduate and postgraduate conversation dataset.

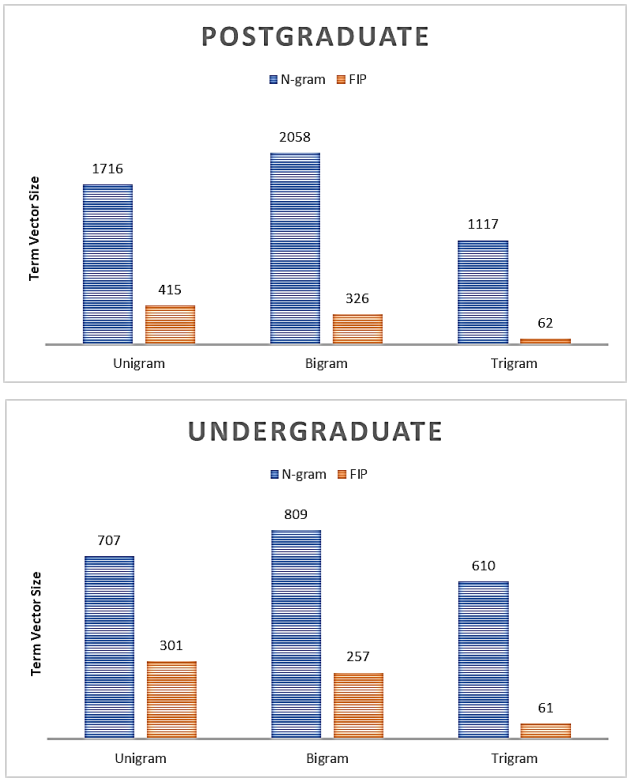

Figure 2. Term feature vector size comparison between the two models (N-gram vs cFIP) using the postgraduate and undergraduate social conversation dataset.

In this experiment, we use the min_sup threshold $\geqslant 2$ with the intuitive to discover the cFIP has occurred or being uttered more than two (2) times during the social conversation (regardless of the sender). For our model, 619 cFIP discovered for the undergraduate and 803 for the postgraduate dataset generation from 1-sequences until 3-sequences. The sample of cFIP discovered during this experiment is shown in Table 3 using the format of "<cFIP>:uSupp(P)". For instance, referring to the postgraduate dataset of cFIP of 1-Sequence, some of the highest terms that were mentioned were "submit" that was mentioned 66 times, "register" 36 times and "schedule" with 15 times that is related to user's intention in a conversation. In the next 2-sequence, the phrase "presentation schedule" was asked 6 times. Then, the query regarding "computer science project" was conversed 18 times by the user was found in cFIP of 3-Sequence. This indicates very frequent related entities were found using cFIP regarding computer science project presentation schedule or submission as mentioned by the students in Table 2 WhatsApp log excerpt on the postgraduate conversation.

Table 3. Sample cFIP discovery for the postgraduate dataset

Sequence Length	Postgraduate
1-seq	<(submit:66),(computer:37),(science:39),(project:43), (presentation:20),(schedule:15),(register:36) >
2-seq	<(computer)(science):32>, <(science)(project):21>, <(presentation)(schedule):6>
3-seq	<(computer)(science)(project):18>, <(science)(project)(presentation):4>

B. Descriptive Intent Discovery

In this section, this research compares and discusses on the Descriptive Intent Discovery between the cFIP and N-Gram model. Table 4 shows our comparison findings using cFIP of 3-sequences and Trigram for the top-10 representation result for the postgraduate dataset. In this experiment settings, the weight used for cFIP is *uSupp(P)* and for the N-gram, the setting refers to frequency weighting of the grams. The *uSupp(P)* value refers to the occurrence of the pattern per-conversation line which is counted as once regardless of how many times it was mentioned. This approach normalized the term frequency and useful in avoiding overweighing certain frequent terms in the text.

We focus on postgraduate cFIP and N-gram discoveries in Table 4. From our observation, it is found that the 1st intent consists of the sequence "computer science project" which is one of the courses in the postgraduate program. It can be seen that the top frequent intent is also related to frequent entities discovered in the 7th cFIP that is regarding the "science project presentation". Immediately from the findings, we can visualize that there exists an inter-relationship between the cFIP discovered that can be analysed at the very early stage of the intent discovery. This discovery can assist in the process of creating the intent and entities of our academic chatbot development which will be explained later in the next section.

Meanwhile, for the N-gram model, the top N-gram text features extracted seems to be rather generic and out of focus such as "fki ums labuan". The important text features are hardly related to each other and this can be considered as one of the limitations of the N-gram approach, where it creates a repetitive possible sequence combination of each gram such as the phrase "ums ums best" that opens for further investigation in the next research.

Table 4. Top 10 cFIP and Trigram representation for postgraduate social conversation dataset

	Postgraduate	
#	**cFIP**	**N-Gram**
1	**computer science project**	kerani fki ums
2	**universiti malaysia sabah**	**fki ums labuan**
3	sekian terima kasih	**computer science project**
4	**information retrieval class**	**universiti malaysia sabah**
5	**ktru meeting room**	joined using invite
6	class this week	using invite link
7	**science project presentation**	ums ums best
8	pelajar pascasiswazah yg	ums best luck
9	know this week	florence felo ums
10	joined using this	message chat call

From this comparison, we can conclude that even though the term vector size dimension for the cFIP representation has been reduced, it is still capable of identifying and extracting the top descriptive intent from the user's conversation. However, in contrast with the N-gram representation, with large term vectors generations, noisy data may be included which makes it difficult to identify meaningful and interesting user's intent from social conversation.

C. Analysis of cFIP Rules Discovery

In this section, we present the analysis of our cFIP rules discovery. The strength of the rules refers to the support and confidence value metrics. In Frequent Pattern Mining, the support value of a rule $X \rightarrow Y$ reflects the probabilistic estimation, to see if the rule discovered is applicable in the transaction.

Meanwhile, the confidence value reflects the predictability of the rule. As such a low confidence value discovered, we cannot reliably infer rule *Y* from *X*. The findings of cFIP rules and confidence value in this work is used to describe the user's intention during the conversation.

Referring to Table 5, Let *X* represents the term "computer" with the *uSupp* of 37. It indicates the term was mentioned 37 times. Next, *Y* is a cFIP of 2-sequences represented by "computer → science" with the support of 32. It indicates that the pattern has been mentioned 32 times. A cFIP rule can be generated from here, for instance, if the *prefixIntent* is "computer"; then the next term "science" can occur together in adjacent within the conversation with 0.9 confidence value. Next, if the term "project" follows the antecedent of cFIP of 2-sequence "computer science", we can assume the probability of 0.6 confidence value that the cFIP of 3-sequence that consists of "computer science → project" occurs in the conversation. We can also state that if both terms "science" and "project" occur together in adjacent during the conversation, the term "presentation" can occur with 0.2 probability, respectively.

Table 5. Discovery of constrained cFIP rules using the postgraduate dataset

cFIP (X → Y)	Support value	Conf value
computer	37	-
computer → science	32	0.9
computer science → project	18	0.6
science	39	-
science → project	21	0.5
science project → presentation	4	0.2
information	9	-
information→ retrieval	6	0.7
information retrieval → class	6	1

Another example of cFIP rule is "information retrieval class" with the *uSupp* of 6. Let *X* is "information retrieval", given *Y* is "information retrieval class". The discovered cFIP rule has a confidence value of 1. The high confidence highlights the importance of the discovered rule where in this research it refers to one of the courses in the postgraduate studies.

From the comparison of experiments results, our cFIP model can represent user's intent pattern in social dataset automatically and independently. It performs better from the N-gram model in terms of reduced term vector size for both undergraduate and postgraduate social conversation dataset. It can also discover the top interesting intents to reflect essential topic being discussed by the students. For example, intent such as "getCourseInfo()" can be set with entities, i.e. "Computer Science project" and "Project presentation" that is related to the user's query.

D. Usability Analysis

In this work, the intents and entity discovered from mining social conversation dataset using the constrained FP-Growth method were used to develop the knowledge base for our Academic chatbot prototype using IBM Watson Assistant framework as illustrated in Figure 3.

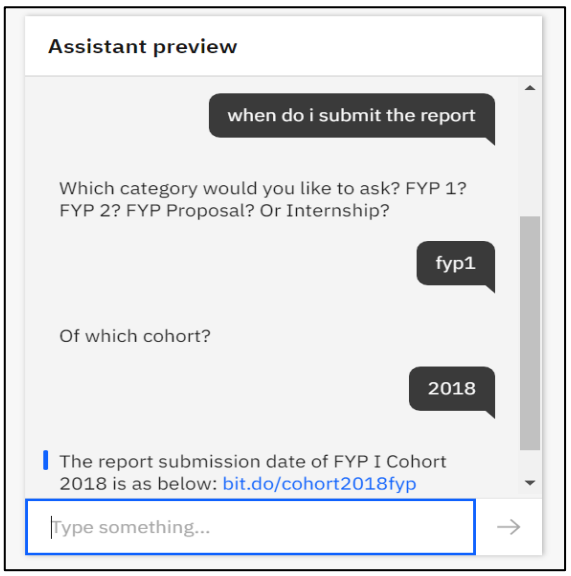

Figure 3. Sample Academic Chatbot conversation

For the usability testing and analysis, we conducted the ASQ: After Scenario Questionnaire to perceive the model's usability and ease-of-use (Lewis, 1995). The respondents

(students) must complete their interactions with the academic chatbot using a similar scenario when communicating with the academic advisor. A group of 53 random respondents participated in this experiment with a confidence level of 95% and 10% of margin error settings. The usability score is using the Likert scale of 1-5, indicating the value 5 of strongly agree and 1 strongly disagree.

The AQS statements with the average mean and standard deviation results are depicted in Table 6. The results showed that the overall user satisfaction average mean score is 4.30 out of 5 with 0.5 standard deviation. This reflects that the students agreed and were satisfied with the usability of the academic chatbot model in assisting academic-related queries and disseminating information automatically. A deeper feedback analysis also revealed that the students stated that the chatbot implementation is good for day-to-day usage and a very helpful assistant.

Table 6. ASQ Usability results for Academic chatbot

#	ASQ Statements	Avg Mean	Avg Standard Deviation
1	Overall, I am satisfied with the ease of completing the task using Academic chatbot in this scenario	4.09	0.53
2	Overall, I am satisfied with the Academic chatbot response (accuracy, time, and interface) to complete the task in this scenario	4.02	0.45
3	Overall, I am satisfied with the support information (chatbot replies and interactions) when completing the task	4.79	0.53
	Overall Score	**4.30**	**0.50**

V. CONCLUSION

In this paper, we presented the work of our Academic chatbot model to automate the intents and entities discovery in an academic conversation between student and academician. Our unsupervised text feature extraction methods named cFIP with text feature constraints is compared with the N-gram model to represent user's intent of undergraduate and postgraduate students' social messages conversation. From the experiment results, we found descriptive intent pattern can be discovered using cFIP representation which can be used as text features. The cFIP representation is data-driven with reduced term vector size makes the approach lightweight for mining social dataset. We relate these findings based on the textual constraints implemented on the cFIP generation which are the unique term constraint, adjacency term order constraint and sequence size constraint. Moving forward, we will improve the knowledge base of our chatbot prototype model by mining dataset from different resources related to Academic domain to provide technical support and academic recommendation.

VI. REFERENCES

Adamo, JM 2012, Data mining for association rules and sequential patterns: sequential and parallel algorithms, Springer Science & Business Media.

Alias, S, Mohammad, SK, Gan, KH & Ping, TT 2018a, 'MYTextSum: a Malay text summarizer model using a constrained pattern-growth sentence compression technique', eds Alfred, H, Iida, AA, Ag, Ibrahim & Y, Lim, in Computational Science and Technology, Singapore, Springer Singapore, pp. 141-150.

Alias, S, Mohammad, SK, Hoon, GK & Ping, TT 2018b, 'A text representation model using Sequential Pattern-Growth method', Pattern Analysis and Applications, vol. 21, no. 1, pp. 233-247.

Alias, S, Sainin, MS, Fun, TS & Daut, N 2019a, 'Identification of conversational intent pattern using Pattern-Growth technique for academic chatbot', in International Conference on Multi-disciplinary Trends in Artificial Intelligence, Springer, Cham, pp. 263-270.

Alias, S, Sainin, MS, Fun, TS & Daut, N 2019b, 'Intent pattern discovery for academic chatbot-a comparison between n-gram model and frequent Pattern-Growth method', in 2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS).

Almansor, EH & Hussain, FK 2019, 'Survey on intelligent Chatbots: state-of-the-art and future research directions', in Conference on Complex, Intelligent, and Software Intensive Systems, Springer, Cham, pp. 534-543.

Arsovski, S, Osipyan, H, Oladele, MI & Cheok, AD 2019, 'Automatic knowledge extraction of any Chatbot from conversation', Expert Systems with Applications, vol. 137, pp. 343-348.

Chandra, YW & Suyanto, S 2019, 'Indonesian Chatbot of university admission using a question answering system based on sequence-to-sequence model', Procedia Computer Science, vol. 157, pp. 367-374.

Cuayáhuitl, H, Lee, D, Ryu, S, Cho, Y, Choi, S, Indurthi, S & Kim, J 2019, 'Ensemble-based deep reinforcement learning for chatbots', Neurocomputing, vol. 366, pp. 118-130.

Dibitonto, M, Leszczynska, K, Tazzi, F & Medaglia, CM 2018, 'Chatbot in a campus environment: design of LiSA, a virtual assistant to help students in their university life', in International Conference on Human-Computer Interaction, Springer, Cham, pp. 103-116.

Ghose, S & Barua, JJ 2013, 'Toward the implementation of a topic specific dialogue based natural language chatbot as an undergraduate advisor', in International Conference on Informatics, Electronics & Vision (ICIEV), IEEE, pp. 1-5.

Hardalov, M, Koychev, I & Nakov, P 2018, 'Towards automated customer support', in International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer, Cham, pp. 48-59.

Ho, CC, Lee, HL, Lo, WK & Lui, KFA 2018, 'Developing a Chatbot for college student programme advisement', in 2018 International Symposium on Educational Technology (ISET), 31 July-2 August 2018, IEEE, pp. 52-56.

Jeong, SS & Seo, YS 2019, 'Improving response capability of chatbot using twitter', Journal of Ambient Intelligence and Humanized Computing, pp. 1-14. doi: 10.1007/s12652-019-01347-6.

Kim, HD, Park, DH, Lu, Y & Zhai, C 2012, 'Enriching text representation with frequent pattern mining for probabilistic topic modeling', in Proceedings of the American Society for Information Science and Technology, vol. 49, no. 1, pp. 1-10. doi: 10.1002/meet.14504901209.

Laranjo, L, Dunn, AG, Tong, HL, Kocaballi, AB, Chen, J, Bashir, R & Lau, AY 2018, 'Conversational agents in healthcare: a systematic review', Journal of the American Medical Informatics Association, vol. 25, no. 9, pp. 1248-1258.

Lewis, JR 1995, 'IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use', International Journal of Human-Computer Interaction, vol. 7, no. 1, pp. 57-78.

Meincke, D 2018, Experiences building, training, and deploying a Chatbot in an academic library.

Nuruzzaman, M & Hussain, OK 2018, 'A survey on chatbot implementation in customer service industry through deep neural networks', in 2018 IEEE 15th International Conference on e-Business Engineering (ICEBE), 12-14 October 2018, China, pp. 54-61.

Nwankwo, W 2018, 'Interactive advising with bots: improving academic excellence in educational establishments', American Journal of Operations Management and Information Systems, vol. 3, no. 1, pp. 6.

Pereira, J & Díaz, Ó 2018, 'Chatbot dimensions that matter: lessons from the trenches', in International Conference on Web Engineering, Springer, Cham, pp. 129-135.

Ranoliya, BR, Raghuwanshi, N & Singh, S 2017, 'Chatbot for university related FAQs', in International Conference of Advances in Computing, Communications and Informatics (ICACCI), IEEE, pp. 1525-1530.

Serban, IV, Lowe, R, Henderson, P, Charlin, L & Pineau, J 2018, 'A survey of available corpora for building data-driven dialogue systems: the journal version', Dialogue & Discourse, vol. 9, no. 1, pp. 1-49.

Wallace, RS 2009, 'The anatomy of ALICE', in Parsing the Turing Test, Springer, Dordrecht, pp. 181-210.

Weizenbaum, J 1983, 'Eliza - computer program for the study of natural language communication between man and machine', Communication of the ACM, vol. 26, no. 1, 23-28. doi:10.1145/365153.365168

Yamaguchi, H, Mozgovoy, M & Danielewicz-Betz, A 2018, 'A Chatbot based on AIML rules extracted from twitter dialogues', in FedCSIS Communication Papers, pp. 37-42.

Yan, R 2018, 'Chitty-chitty-chat bot: deep learning for conversational AI', in IJCAI, vol. 18, pp. 5520-5526.