International Journal of Electrical and Computer Engineering
Vol.6, No.4, Month 2016
ISSN: 2088-8708

Inferring Student's Chat Topic in Colloquial Arabic Text using Semantic Representation

Faisal T. Khamayseh
Department of Information Technology and Computer Engineering
Palestine Polytechnic University

Article Info

Article history:

Received Jun 12, 201x
Revised Aug 20, 201x
Accepted Aug 26, 201x

Keyword:

Arabic chat
Arabic corpora
Semantic net
Palestine Arabic corpus
Colloquial analysis

ABSTRACT

Since the colloquial Arabic is now widespread it is required to describe the collection and classification of a multi-dialectal corpus of Arabic. Nowadays, colloquial multi-dialectal comes in almost country based forms such as Egyptian, Iraqi, Levantine, Tunisian, etc. This paper discusses a new method for analyzing the conversation of the educational chat room using Corpus for Palestinian Arabic and Stanford Tagger tool. This method represents the key words using semantic net-like representation to obtain the main subjects of the conversation. The main subject of the chat is obtained using the proposed method which achieves a high accuracy. Using Arabic Corpus, Stanford Tagger and percentage of keywords will assure more accuracy. The study also examines the effect of pivot-words distribution based on occurrences and *betweeness* values of the pivots throughout the text. This study examines some of the characteristics of the texts written in colloquial Arabic dialect and analysis of the free expressive Arabic statements. The results show that the core subject of the chat can be determined by combining both the occurrences and the distribution of the word through the conversation.

Copyright © 201x Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Faisal T. Khamayseh,
Department of Computer Science
College of Information Technology and Computer Engineering,
Palestine Polytechnic University,
Department of Computer Science, Hebron. P.O. Box 198. Palestine.
Email: faisal@ppu.edu.ps

1. INTRODUCTION

Social networking and social media platforms increase rapidly in types and in the huge number of users. They increase in their usage and their huge number of documents such as Skype, WhatsApp, Twitter, Facebook, Viber, IRC, Blogs, Myspace, just to mention a few. Each of these networks provides chat platform for the large number of users. Some platforms exist to serve some specific scopes such as studies and research, while others are shared with the followers on various social media platforms such as open conversation rooms. Specific groups benefit from open platforms to form their closed social network, and others may benefit more from specific configured platforms such as LMS e-classes on Moodle, Illuminate, etc.

Nowadays conversation on social media skipped the standard grammatical rules in almost all languages. As in most of the current languages, Arabic language has two forms; the standard and the colloquial. The standard form is subject to the firm rules that syntactically cover all forms of written and spoken statements. Colloquial Arabic is widely used as spoken language and lately is being widely used as written language especially in mobile messaging and web social media. Some recent attempts focus on analyzing the rule-free text and building some rules (rooting). A considerable work done by [1]-[4] in developing Arabic Ontology to define the formal specification of the concepts of Arabic words related to

Palestinian spoken and written conversations. People may use their social colloquial text while chatting on some open social environment such as student e-classes and academic student to student portals. Search engines do not fully support Arabic language in acceptable form while available search engines are typically limited to keyword searches based on structures and rules of the language and do not take in account the semantics of the content [5]. The challenge comes in the correct analysis of both forms the standard Arabic and the colloquial Arabic texts. Despite the demands for the correct use of the standard language phrases especially the written text, the dialects of colloquial impose themselves in the current internet world.

There are many educational chat rooms that contain different students and various conversations in different subjects. Some of current universities use educational websites to support the discussions between students and teachers. This conversation has advantages for the evaluation of the student acknowledgments concerning specific subject. In this paper, we suggest a new method to extract the main subjects of the conversations that students engaged in. This method, based on analyzing the chat of the students, depends on converting the words of the chat to the equivalent gloss in Corpus for Palestinian Arabic, then the system computes the higher percentage of words that exist in the given conversation.

2. BACKGROUND

The new emerging spelling errors and other inaccurate terms are perceived as an acceptable act in online communication. Text-based chat in now available as one of the most valuable communication tools in web. As a matter of fact, the normal chat conversations do not conform to language rules. Different languages have different levels of colloquial words in spoken or written conversations.

2.1 Chat Text Analysis

Recently, there has been proposed different modified tools and algorithms focusing on ontology analysis such as using lexico-syntactic patterns [6]. Recent researches attempt to achieve better analysis of chat text and network representation of chat-log data [7]. Contemporary analysis acknowledges the characteristics of chat messages and proposes an indicative term-based categorization approach for chat topic detection [8]. The benefit of these analytical studies and investigations is to enable tracing the relay chat of people and avoid the vast amount of continuous monitoring effort. Although this analysis may not be 100% accurate in outcomes, it provides evidences and assists in accomplishing the desired goals.

Arabic instant chats also lack of analysis especially the colloquial spoken/written Arabic conversation. Like most languages, it is easier to analyze standard Arabic since the standard statements are governed by the fixed structures and rules. Although there are many studies deal with standard Arabic, still there is much lack in phonology, morphology, and syntax analysis. Arabic language is characterized in its complexity due to rich "root-and-pattern" morphology and ambiguity. This is due to the absence of short vowels for most Arabic texts. Moreover, productive clitics and affixes of Arabic words and some root letters can be hard to guess if one or two root letters are long vowels or belong to letters' affixes or clitics [9].

A notable effort by [10] presents a statistical study of clitics in Arabic language to show the distribution of clitics and examine the performance of the used tokenizer. They applied clitics tokenization on a large Arabic corpus and showed that a reduction of 24.54% in a number of unique tokens could be achieved. As any other language, deep or shallow syntactic analysis of free conversations requires large corresponding corpora [20].

2.2 Corpus for Palestinian Arabic

Some minor differences among the Palestinian spoken dialects exist while sharing the same linguistic assets. The main differences exist in phonology and lexicon preferences that vary among major historical areas of Palestine. Many words in Corpus are annotated in the context as they have different annotations in different contexts [11]. The study defined the annotation of a word as a tuple: <Raw (Unicode), Raw (Buckwalter), CODA (Unicode), CODA (Buckwalter), Lemma, Buckwalter POS, Gloss and Analysis>. The gloss of each word in the chat is required to obtain the equivalent word in English Corpus in order to speed up the annotation process using MADAMIRA tool for morphological analysis and disambiguation of MSA and EGY. MADAMIRA tool is chosen because of assumption that EGY/MSA and PAL share many orthographic and morphological features. As a recent work, [11] constructed a corpus consisting of words from Palestinian different recourses and presented different pilot studies to select the best tool to speed up their annotation process.

2.3 Arabic Corpora

In the form of the standard language words and morphological and syntactic structuring rules, Arabic language has very strict and powerful rules. In the form of spoken language, Arabic encompasses many dialects scattered all over Arabic world areas. Different referencing Corpora are now available to

enable computing linguistics and words' analysis including *Adjir Corpora*, *Tashkeela*, *Arabic Word Corpora*, *Alwatan*, *OSAC*, to mention a few [12].

Arabic Corpus currently is one of the most language areas of research. Many recent studies handle investigate annotated linguistic resource which shows the Arabic grammar, syntax and morphology for Arabic words. Some linguistic studies and analysis serve the mother standard language and just few of them serve local dialects [12]-[16]. Various dialects share many morphological and syntactic structures especially in closed countries that share some common traditions such as Levantine area, Gulf area, Maghreb (Morocco) countries. The similarity in local dialects makes the studies fairly limited to some areas.

2.4 Social Conversations

Factors such as participants, topic, function of interaction, and the value of the interaction affect the level of dialectic conversation [17]. Social media and social channels also affect the language and now become major resources of popping up uncontrolled words. Social media has prompted a powerful subtle revolution in conversations. Educational platforms are not so far from this accelerating linguistic revolution. However, there is urgent need for analytical studies to assess reality, the effect and values of social conversations especially in the context of learning.

2.5 Educational Chat Rooms

Educational chat room is the use of technological tools in learning and for sharing information via text with groups of students that offer a real-time transmission of text. Chat rooms enable many students to converse with each other in the same conversation from websites. Students in an educational chat room are generally connected via a shared interest of education. Chat room, which is intended for students' conversations, usually possesses rules and instructions that they require students to follow. Commonly used chat rooms are not moderated so students may chat freely, which may lead to long useless conversation [18].

2.6 Stanford Tagger

Stanford Tagger is a piece of software that reads text in some language, in our case English, and assigns parts of speech to each word, such as noun, verb, adjective, etc. All user requirements are processed using the Stanford tagger by writing the statements in the text area provided for that purpose. In addition, Stanford Tagger uses a set of tags to describe different components of a statement. Some improvements in the features, parameters, and learning methods give small incremental gains in POS tagging performance, in addition to splitting certain categories, part-of-speech and phrasal categories, and parsing with the resulting split-category Treebank grammar [19].

Stanford Tagger tokenizes the statements and uses a large number of tags. Since the input text is in Colloquial Arabic that is somehow different and far from standard Arabic, the text has to go through special corpus to produce the standard Arabic or English equivalence. The required tags that this study uses in the analysis phase is therefore produced using Stanford Tagger. Table 1 shows these tags that are produced using Stanford Tagger to be used in the proposed approach.

Table 1. The Stanford Tags

Tag	Description	Tag	Description
CC	conjunction, coordinating	**VB**	verb, base form
IN	preposition or conjunction, subordinating	**VBD**	verb, past tense
JJ	adjective or numeral, ordinal	**VBN**	verb, past participle
NN	noun, common, singular or mass	**VBP**	verb, present tense, not 3rd person singular
NNP	noun, proper, singular	**VBZ**	verb, present tense, 3rd person singular

3. ANALYSIS OF GLOSS CHAT AND RANK OF PIVOT WORDS

In applied methodology, the paper proposed an approach to get or decide the main subjects of the conversation between students based on the extracted Arabic chat text keywords. First step requires inserting the free conversations of all students in a selected controlled learning blog. In order to narrow the text analysis, a single topic is selected. Students' statements are then converted to the corresponding gloss using one of available Arabic Corpora mainly the Palestine Arabic Corpus. In this step, a limited temporary corpus has been constructed based on [1], [11] since the real corpus is not available up to the date of submitting this paper for publish. Then, we analyze the gloss of the words to get the tags using Stanford Tagger. We accept the words that have a noun and adjective tags (NN, DTNN, JJ). After that, the algorithm computes the

percentage of the count of each noun and the adjective words by counting the occurrences of them in the chat text. As a decision, the high percentage ranked words are chosen according to the most possible conversation topics. The following algorithm summarizes the main tagging and ranking steps which illustrated in figure 1.

Algorithm Pseudo code: Tagging and ranking chat gloss keywords

- Convert chat text to gloss list using Corpus for Palestinian Arabic (CPA).
- Obtain tags corresponding to gloss words using Stanford tagger.
- Determine the set of candidate words including nouns and adjectives.
- Count the occurrences of each key word.
- Place keywords in such order.
- Obtain the decision from high occurrence values.

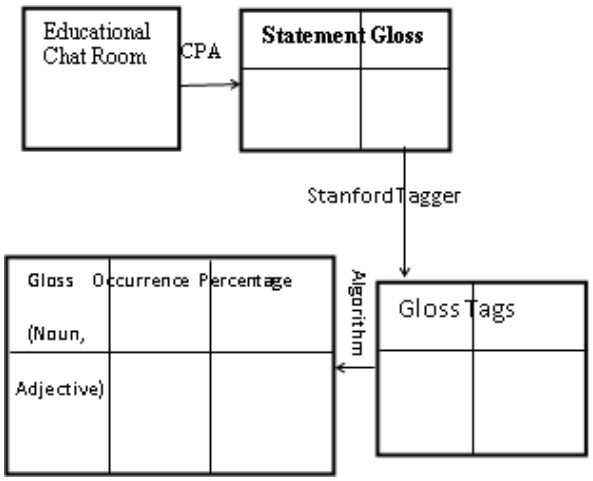


Figure 1. Methodology block diagram

The following example illustrates the proposed method using a conversation for three students.

Students chat I:

الامتحان صعب جدا	Ibrahim
تأخرت على الامتحان عشان المواصلات	Obada
فش وقت كفاية احل فيه الامتحان	Ali

In the first step, we convert the statements of students' conversation to the corresponding gloss words as in table 2:

Table 2. Corpus Gloss

Chat Statement	Gloss
الامتحان صعب جدا	Exam, Hard, Very
تأخرت على الامتحان عشان المواصلات	Late, On, Exam, Because, Transport
فش وقت كفاية احل فيه الامتحان	No, Time, Enough, Solve, In, Exam

The second step obtains the tags of the gloss words using Stanford Tagger as listed in table 3:

Table 3. Stanford Tags

Chat Gloss	Exam	Hard	Very	Late	On	Exam	Because	Transport	No	Time	Enough	Solve	In	Exam
Tag	NN	JJ	PRT	VBN	IN	NN	PRT	NN	PRT	NN	PRT	VBN	IN	NN

In the last step, the system applies a count-algorithm that takes the list of word-strings and counts the occurrences for each word of the form noun or adjective in all conversation, as in table 4:

Table 4. Percentage of tagged words of chat I

Gloss	Tag	Occurrences	Percentage
Exam	NN	3	50%
Hard	JJ	1	16.7%
Transport	NN	1	16.7%
Time	NN	1	16.7%

Table 4 shows that the candidate keys which are the ones with high counts. While counting the occurrences of words provides a good hint towards a good decision about the context of the conversation, it may also mislead the decision. Figure 2 illustrates the main pivot words and their rates based on their occurrences in the conversation.

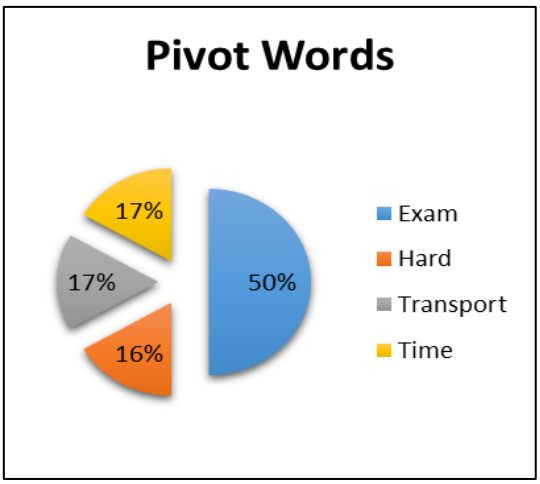


Figure 2. Occurrences of pivot words of chat I

The counts of key spoken words in longer conversation lead to more correct analysis with lower error or misleading results. The example in student chat II displays the same conversation statements on the same topic between other regular students in a different chat group. Some slight update has been made on the conversation of students using eclass.ppu.edu, such as having longer conversation in the chat:

Student chat II:

مدرس الخوارزميات عين الامتحان قبل كم يوم	Ahmed
بس ما اختصر من مادة الامتحان ولا اشي	Sara
يا ريت لو نعرف نأجلو كمان اسبوع	Omar
هاد مدرس امتحاناتو خبط والخوارزميات صعبة جدا	Sara
انا خايف ارسب في الامتحان زي ما رسبت في الامتحان الي قبلو	Ahmed
بكرة حشوف المدرس اذا بيعينو بعد الامتحانات العامة	Omar
هذا اذا برد علينا رح يزيد مادة ورح يجيب لنا خوارزمية ما بتنفهمش	Ahmed
فهمت	Omar

Tokenizing the chat text using Stanford Tagger, MADAMIRA or any other tool results in many errors or tagging mistakes due to the high occurrences of unknown words. None standard Arabic spoken or written language gets many of these unknown words day by day. We update the gloss of unknown Arabic

colloquial words manually based on the Palestinian Arabic Corpus [1], [11]. Table 5 shows the total number of occurrences of each pivot word in the text of chat II. When the chat text is a bit small, tagging and word analysis is high error prone. The analysis shows that the key *Exam* occupies the first place in the conversation due to its high occurrence count.

Table 5. Nouns and objectives of students chat II

Pivot word/Tags	Occurrences	Rank
Instructor: [NN,NN,DTNN]	3	17.6%
Algorithm: [DTNN,DTNN,JJ]	3	17.6%
Exam: [DTNN,DTNN,DTNN,DTNN,DTNN,NNP]	6	35.3%
Day: [NN]	1	6%
Material: [NN,NN]	2	11%
Week: [NN]	1	6%
General: [DTJJ]	1	6%

Figure 3 shows the different occurrences of the pivot words in the chat. Selected words are only the nouns and adjectives since such types have the main influences in the conversation.

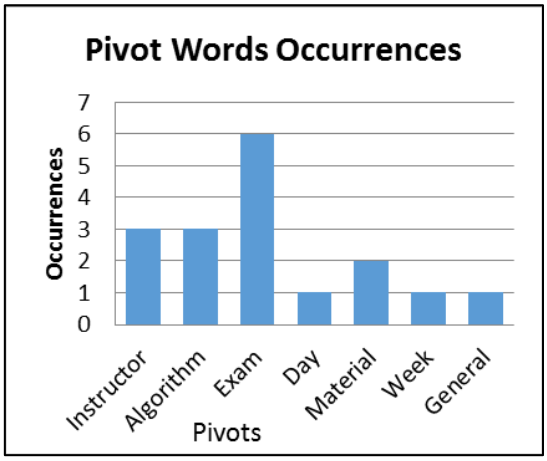


Figure 3. Occurrence of pivot words of chat II

Results show that the decision of inferring the subjects of the chat is taken according to the higher percentage of all words (Nouns, Adjectives). In the previous examples the students' chats about the exam occupy the first possible rank because of the resulted highest percentage. In reality, this does not exclude other lower ranked topics; rather the highest rate is the most possible. Since the highlighted gloss words are nouns and adjectives, there is no need to analyze and compute occurrences of other words. This is possible by the deliberate ignorance of tags other than nouns and adjectives. However, pronouns and abbreviations may lead to more accuracy in results. Many factors have to be considered to ensure more accurate results such as scope, time of conversation, previous chats, participants, etc. For example, it is more likely college students talk about course material and exams during the last two weeks of the semester when using university e_class. These factors should be taken in consideration when analyzing and counting the occurrences.

4. CONSECUTIVE REPETITIVE WORDS AND DISTRIBUTION FACTORS

To avoid misleading consecutive word repetition, analyst may consider only one of the repetitive words. For example, the chat statement "*…unfortunately I have been studying math, math, math and math during last days. I had no time to study on compiler exam*", may not necessarily mean that the topic is about math, it could be about compiler exam, in which the conversation is more likely about. Furthermore, to strengthen the algorithm, researchers may study and investigate the relationship graph model among the tokens. The relay of the key words in the conversational text as well as the relay of pronouns help in determining the conversational context with reference to keywords. Usually the distribution of the key words throughout the chat text has a stronger indication than the consecutive repetition that occurs in some limited parts. This case is illustrated in figure 4 representing the students' conversation taking about a laboratory exam of biology course.

Students chat III:

Sara	متى امتحان مختبر الأحياء النصفي ؟
Omar	للأسف بكرة امتحان برمجة نظري وبعدو امتحان مختبر برمجة وبعدو تسليم مشروع البرمجة وبعدو الخميس مختبر الاحياء.
Sara	يا سلام. كل الاسبوع برمجة في برمجة في برمجة .
Ahmed	مختبر الاحياء النصفي سهل وما يحتاج تقلق كتير.
Sara	هو صباحي؟
Ahmed	لا. بعد محاضرة الاساليب بنقدم مختبر الاحياء النصفي.

It is obvious that students talk about biology in the first place even the word “programming” counts more. This can be represented using semantic net-like where occurrences of each gloss are to be considered. Going through the semantic net looking for higher counts give an indication of the chat topic. The distribution of keys over the chat text as illustrated in the semantic net shows a stronger indication despite having less occurrence count of some repetitive weights. To achieve this, it is helpful to scan the text first looking for small set of high occurrences of pivot words, then compute the *betweeness* values of identical words forming set of landmark representations. The best representation combines properties including high occurrences with large *betweeness* values between landmark words. These properties are extremely important and therefore should be considered in defining the chat topic. Figure 4 shows the semantic net of the above example. The pronoun is also considered with the diamond symbol representing the equivalence words. All other words such as connectors and verbs are deliberately dropped due to the weak participation in the analysis phase.

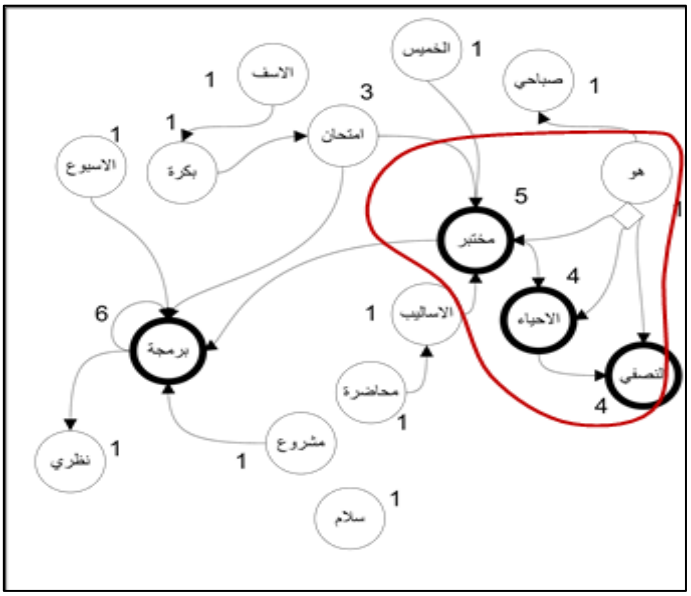


Figure 4. Semantic net-like of pivot words of chat III

General algorithm: Inferring the core of the chat using semantic-net analysis
Input: Chat text.
Process: Computing the weights of key words using semantic net and occurrences counts.
Output: Chat core.

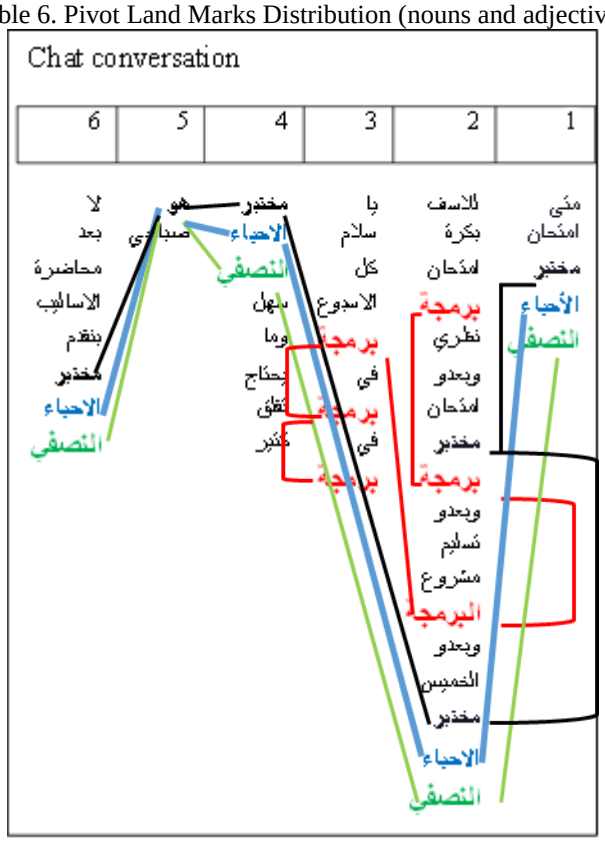
- Convert chat text to gloss list using Corpus for Palestinian Arabic.
- Obtain tags corresponding to gloss words using Stanford tagger.
- Determine the set of candidate words including nouns and adjectives.
- Construct the semantic net representation.
- Count the occurrences of each pivot word.
- Determine the set of key words with the high occurrences.
- Determine the *betweeness* value of pivot words.
- Place keywords in such order to obtain the decision.

According to the semantic net graph shown figure 4 and table 6; the candidate landmark words with their distribution counts are:

Lab: مختبر : count = 4+1(Pronoun) =5 , Programming: برمجة : count=6+1(Pronoun) =7,

Biology: الاحياء : count=4+1(Pronoun) =5, Midterm: النصفي : count=4+1(Pronoun) =5

Table 6. Pivot Land Marks Distribution (nouns and adjectives)



It is obvious that the word Biology expands over the whole chat in which it combines the a high count (may not be maximum) and the large *betweeness* value. The word "Programming" counts more but has very low *betweeness* value. The tradeoff between count of words and *betweeness* values gives a strong indication of determining the text core.

Table 6 shows the landmark distribution of pivot words with different *betweeness* values in adjacency list of the chat words. Figure 4 shows the consecutiveness of pivot words while table 6 shows the power of distribution of these pivot words. This is important step where the word that occurs over the chat has higher possibility to be selected than the word that occurs in some limited parts of the text. Another feature is that the words that occur together have very high probability if they occur in different parts of the chat text. In the above example, table 6 shows that the best landmark of related words are Biology, lab and midterm. The question that the text starts with is about the Biology lab midterm exam. This also should be taken in consideration. Several conversations have been analyzed where different university student groups were asked to be involved in the chat blogs using available e-class templates. It is important to mention that the sizes of the conversations used in this study are relatively small of maximum 18 lines. Table 7 summarizes the analysis of these outcomes of 4 groups ranging from 3 to 7 students chatting several times using the e-learning templates.

Table 7. The percentage of correctness of inferring chat topic

Group/Lines	G1(2students)	G2(3 students)	G3 (4 students)	G4(5 students)	G5 (6 students)	G6 (7 students)
7 Lines	50	55	70	74	76	84
9 Lines	60	65	72	77	78	86
12 Lines	70	70	75	80	80	87
18 Lines	73	74	80	82	86	91

Figure 5 illustrates the correctness percentage as chat lines grow in the conversation. It is clear that the correctness is clear as the conversation lines grow. However very long conversations may talk about different major topics where this study did not analyze yet.

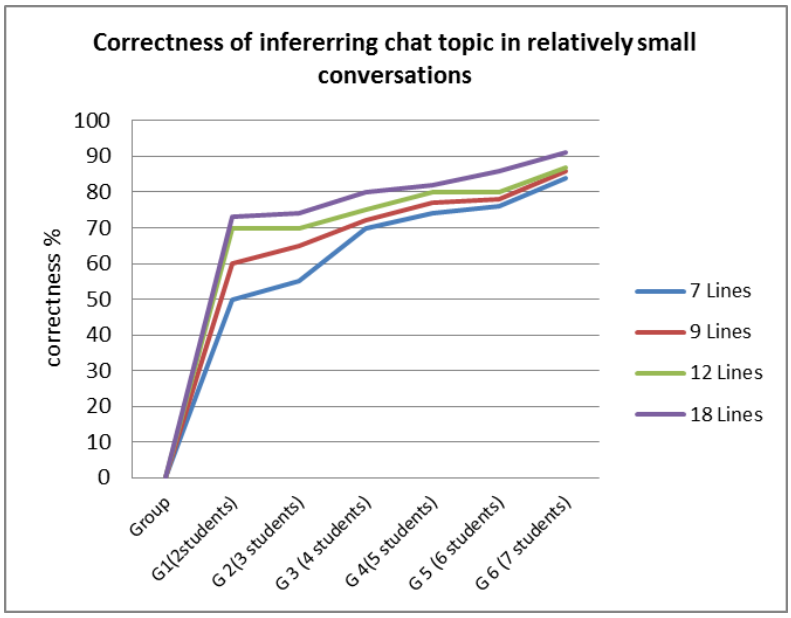


Figure 5. Correctness of inferring chat topic

5. CONCLUSION

This paper proposes an approach to infer the subject of educational student chat. The study suggests an important way to determine the core of the colloquial Arabic student chat text. This approach starts with extracting the equivalent gloss of each word in English language using the Corpus of Palestinian Arabic. The second step obtains the tags of each word using Stanford Tagger while focusing on the nouns and adjectives which occur in the conversation. After that, the approach counts the occurrences of each word in all conversations to decide the main subject of the written conversation based on the higher percentage of the occurrences of noun and adjective words.

Semantic analysis and the concern of other words help in achieving higher accuracy. The distribution of the pivot words to cover the conversation has very high impact on deciding the core of the text. Many other factors such as time of the conversation, starting statements and the used platform may have direct impact on the accuracy of determining the nature of the content. Although human interventions should be abandoned, it also help in obtaining better outcomes such as in correcting colloquial corresponding gloss and tags, pronouns, and unexplained terms.

Further studies and experiments should rely on real comprehensive colloquial corpus. Analyses of the distribution of all types of words and the analysis of the semantic relations of more long chats should be taken in account to obtain more accurate results.

REFERENCES

[1] M. Jarrar, *et al*., " Building a Corpus for Palestinian Arabic: A Preliminary Study," *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing. Association for Computational Linguistics (ACL)*, pp. 18-27, Oct 25, 2014, Doha, Qatar. ISBN: 978-1-937284-96-1.

[2] M. A. Helou, *et al.*, "Towards Building Lexical Ontology Via Cross-Language Matching," *Proceedings of the 7th Conference on Global WordNet. Global WordNet Association*, pp. 346-354, Jan 2014, Tartu, Estonia. ISBN: 7492329949978.

[3] A. L. Soares, *et al.*, "OnToContent 2014 PC Co-Chairs Message," *Proceedings of The International Workshop on Ontology Content and Evaluation (OnToContent 2014)*. *In OTM 2014 Workshops*, pp. 575, Oct 2014, LNCS:8842, Springer. ISBN: 978-3-662-45549-4.

[4] M. Altantawy, *et al*., "Morphological Analysis and Generation of Arabic Nouns: A Morphemic Functional Approach," *Proceedings of the International Conference on Language Resources and Evaluation*, LREC 2010, 17-23 May 2010, Valletta, Malta.
[5] L. Al-Safadi, *et al.*, "Developing Ontology for Arabic Blogs Retrieval," *International Journal of Computer Applications*, vol. 19, No. 4, pp. 0975 – 8887, April 2011.
[6] E. Karyawati, *et al.,* "Ontology-based Why-Question Analysis Using Lexico-Syntactic Patterns A.A.I.N," *IJECE International Journal of Electrical and Computer Engineering*, vol. 5, No. 2, pp. 318~332, April 2015.
[7] S. Tavassoli and K. A. Zweig, "Analyzing the activity of a person in a chat by combining network analysis and fuzzy logic," *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 2015, ASONAM '15.
[8] S. C. H. Haichao Dong and Y. He, "Structural analysis of chat messages for topic detection," *Online Information Review*, vol 30, No. 5, pp. 496--516, 2006.
[9] M. Sawalha, "Open-Source Resources and Standards for Arabic Word Structure Analysis: Fine Grained Morphological Analysis of Arabic Text Corpora," PhD. Thesis. School of Computing. University of Leeds, 2011.
[10] F. Alotaiby, *et al.*, "Clitics in Arabic language: a statistical study," *Proceedings of Pacific Asia Conference on Language, Information and Computation* 24, (PACLIC 24). Sendai, Japan, pp. 595-602, 2010.
[11] M. Jarrar, et al., "Building a Corpus for Palestinian Arabic: a Preliminary Study," *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Association for Computational Linguistics*, pp. 18–27, *ANLP 2014*, Oct 25, 2014, Doha, Qatar.
[12] W. Zaghouani, "Critical Survey of the Freely Available Arabic Corpora," *Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop*. Rejkavik, Iceland, pp. 26-31, May 2014.
[13] N. Habash, "Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies," *Journal of Machine Translation*, Vol. 24, Issue 3-4, pp. 285-289, Dec 2010.
[14] A. Al-Thubaity, *et al.*, "New Language Resources for Arabic: Corpus Containing More Than Two Million Words and a Corpus Processing Tool," *Proceedings of the Asian Language Processing (IALP) Conference*, pp.67-70, 2013.
[15] K. Almeman and M. Lee, "Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words," *Proceedings of The First International Conference on Communications, Signal Processing, and their Applications (ICCSPA'13)*, Sharjah, UAE, 12-14 Feb. 2013.
[16] W. Salloum, and N. Habash, "ADAM: Analyzer for Dialectal Arabic Morphology," *Journal of King Saud University - Computer and Information Sciences.* Vol. 26, No. 4, pp. 372–378, Dec 2014.
[17] S. Ervin-Tripp, "An Analysis of the Interaction of Language, Topic, and Listener. American Anthropologist," pp. 66: 86–102, 1964, doi:10.1525/aa.1964.66.suppl_3.02a00050.
[18] M. Hanini, *et al.*, "Text Modeling in Adaptive Educational Chat Room," *International Journal of Computer Applications*, vol. 103, No. 5, pp. 33-37, 2014.
[19] N. Arman and S. Jabbarin, "Generating Use Case Models from Arabic User Requirements in a Semiautomated Approach Using a Natural Language Processing Tool," *Journal of Intelligent Systems*, vol. 24, No. 2, pp. 277-286, 2015.
[20] M. A. S. Hazaa, *et al*., "Automatic Extraction of Malay Compound Nouns using a Hybrid of Statistical and Machine Learning Methods," *IJECE International Journal of Electrical and Computer Engineering*, vol. 6, No. 3, June 2016.

BIOGRAPHY

Dr. Faisal Khamayseh is a Computer Science assistant professor. He received his BS in Computer Information – Advanced Computer Careers, from Southern Illinois University, USA 1992, and MS in Computer Science from same university in 1995. He received his PhD in Computers and Information Systems from the College of Computers and Information, Helwan University, Egypt, in 2009.Currently working at Palestine Polytechnic University as instructor and head of Dept. of Information Technology and as instructor of MS in Informatics. Dr. Khamayseh is a researcher in software engineering research group (SERG) at college of Information Technology and Computer Engineering. He is interested in Computer Algorithms, Software Engineering and E-learning.
My LiveDNA is 970.11840.