IST-Africa 2012 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2012
ISBN: 978-1-905824-34-2

Automated Topic Spotting Provides Added Efficiency in a Chat Based Tutoring Environment

Laurie BUTGEREIT[12], Reinhardt A BOTHA[2]
[1]*Meraka Institute, CSIR, Pretoria, South Africa*
Tel: +27128413200, Fax: +27128414720, Email: lbutgereit@meraka.org.za
[2]*Nelson Mandela Metropolitan University, Port Elizabeth, South Africa*

Abstract: Dr Math is a mobile, online tutoring project which allows primary and secondary school pupils to contact tutors in mathematics using text based chat systems on their cell phones. The tutors use traditional Internet based workstations. Dr Math is extremely popular with school pupils and Dr Math tutors are often chatting to twenty pupils concurrently. It is important that the tutors are as efficient as possible. This paper describes the implementation of an automated topic spotter in a mobile chat-based tutoring environment. The tutors, although well versed in mathematics, were not all home language English speakers. They often struggled with the cryptic abbreviations which the pupils used in their queries. In addition, the number of pupils asking questions far exceeded the number of tutors available. It became important to ensure that the tutors were as efficient as possible. The automated topic spotter would process the cryptic questions, determine the general mathematical topic, and automatically provide supporting material to the tutors. The learning gained on this project could be easily reimplemented in other domains besides education including health and public services.

Keywords: MXit, Dr Math, chat, topic spotting

1. Introduction

Dr Math is a mathematics tutoring project hosted at Meraka Institute. It has been running since January, 2007. Primary and secondary school pupils can use text based chat protocols on their cell phones to contact volunteer tutors to receive help with their mathematics homework. The volunteer tutors use traditional Internet based work stations. Because of the popularity of the Dr Math project, tutors are often chatting concurrently with twenty or more school pupils about mathematics.

This paper describes an effort to make the project more efficient by adding automated topic spotting to assist the tutors. The automated topic spotter monitors the conversation between the tutor and pupil, determines the mathematical topic being discussed, and then provides hyper links to supporting documentation to assist the tutor. Although this paper specifically discusses a mathematics tutoring environment, the lessons learned and the technology developed could be easily used in a project dispensing information about health matters or dispensing information about government services such as grants and pensions

2. Dr Math

Dr Math is a project which has been running at Meraka Institute since 2007. It links primary and secondary school pupils to volunteer tutors in order to obtain help with their mathematics homework. The primary and secondary school pupils use MXit (a text chat protocol) on their cell phones. The tutors use traditional Internet based workstations. The majority of the volunteer tutors are university students; however, some of the tutors are more mature and are employed in industry.

MXit is a text based communication system which is extremely popular with young people. People who use MXit need to install a special MXit client program on their cell phones and can then communicate with other MXit users (and some other peered services) at an extremely low cost.

The Dr Math software which was developed at Meraka Institute in conjunction with Nelson Mandela Metropolitan University, C^{3}TO (Chatter Call Centre/Tutoring Online), linked the school pupils with the tutors ensuring that the safety and security of the minor children was always protected and that the tutors were used as efficiently as possible[1].

The Dr Math project does have an ethics clearance. This ethics clearance was required due to the fact that the pupils were minor children and the project did not insist on receiving written parental permission from the parents of the children who asked for help with their mathematics homework. The tutors all signed codes of conduct where they agreed not to discuss any illegal activities. In addition, the tutors signed a document agreeing that all their conversations could be recorded. The pupils received daily messages informing them that all their conversations were recorded and that they should not give out any personal information over the system.

At the time of writing this paper, there are over thirty thousand pupils registered with Dr Math. Although these thirty thousand pupils do not all use Dr Math all the time, whenever a tutor logs in, the tutor is soon chatting with twenty to fifty pupils concurrently about various mathematical topics. A typical tutoring session is a very fast paced with tutors helping pupils with a wide range of topics and covering a wide range of grade levels.

The Dr Math tutors are all volunteers. The majority of the tutors come from tertiary institutions in South Africa. The University of Pretoria (UP) supplies the majority of the volunteer tutors. The African Institute for Mathematical Studies (AIMS) also supply a high number of volunteer tutors. An interesting situation developed with the students from AIMS. Many of the AIMS tutors were French speaking students where English was a second language. They are from Central Africa and were studying advanced mathematics in South Africa. This means that their mathematics knowledge and skills were excellent. But they often did not understand the questions posed to them – especially if the question was posed in the cryptic lingo used in the MXit instant messaging. Some of the tutors are professionals who are employed in industry. The majority of these tutors are based in South Africa but there is also a growing number of tutors based in Europe and North America.

Dr Math supports two major instant messaging protocols for use by the mobile pupils. It uses XMPP (eXtensible Messaging and Presence Protocol) which is also known as Jabber and is the protocol which is supported on normal Google mail accounts. Dr Math also uses MXit. MXit is a specific protocol which was developed by a South African company, MXit Lifestyle. This protocol is extremely compact and is ideal for use on a cell phone. At the time of writing this paper, MXit boasts millions of users. The majority of the pupils who contact Dr Math do so via MXit.

A number of Dr Math tutors are older, more mature volunteers. These older volunteers often have never used MXit or any chat protocols on their cell phones. Because of that, they often do not understand the cryptic language used by teenagers on their cell phones. These volunteers are similar to the French speaking volunteers from Central Africa. Their mathematics knowledge is excellent but they have problems communicating with the pupils.

It was important to make these tutors more efficient by enhancing the C³TO platform. This paper specifically describes the topic spotting facility which was added to C³TO and how it enhances the efficiency of the volunteer tutors.

3. MXit Lingo

Languages change over the course of time. As people migrate, the languages they know and understand also migrates. Citizens who temporarily visit another country for employment purposes come back with new vocabulary. People who live next to each other but speak other languages share words between themselves. South Africa has eleven official languages and words and phrases such as *yebo, bakkie* and *ja* are shared across language boundaries. This is the normal way that spoken languages develop.

Written language is often influenced by the tools that are used to write that language. Cuneiform's wedge shaped formations were created when reed styluses were pressed into moist clay tables [2]. The Greek alphabet is populated with letters consisting of either straight lines or slight curves and was originally written with knives in wooden blocks [3]. The graceful and rounded Chinese characters were written using brushes on rag cloth [4].

Written language has moved from a hand written medium to a printed medium. And now, with the advent of cell phones, the written language is undergoing another change with people writing SMSs (text messages) or chatting on MXit as they walk down the street.

In the case of the Dr Math project, questions to Dr Math often look like:

> hi there simltanes eq pls?
>
> how do i get da form of a st line from 2 pts?
>
> how do u bisect an angle wif just a composs and rler?
>
> determine d general solutn of d equatn sin(x+10degreez)=0,75
>
> hw du u prove that a triangle z a right angle triangle?
>
> hv prbmz wth finatial matsh
>
> hi is f(.x) lyk y=x ?

The use of MXit lingo was not mandated by the Dr Math project. It was just the way young people chatted when using MXit on their cell phones. Messages sent via MXit over cell phones are characterised by

1. Specialised vocabulary such as *awe* and *sup* being greetings.
2. Shortage of vowels
3. Numerals and symbols being used in place of some sounds (such as *l8r* representing the word *later* and *th@* representing the word *that*)
4. Some numerals being used in place of similar shaped letters (such as *h3llo* representing the word *hello* and *he1p* representing the word *help*)
5. Some recognisable new spelling conventions
6. Short messages

Three problems present themselves to the tutors:

1. If the tutor is not a home language English speaker, then these questions are almost incomprehensible
2. An older, more mature tutor may not be familiar with MXit lingo. In such cases, the questions are also almost incomprehensible
3. In some cases, the tutor may not be familiar with the exact topic. This is especially true with engineering or science university students tutor. These types of tutors may not be familiar with questions about financial mathematics such as calculating simple or compound interest.

It is important that the tutors in these three situations be given timeous supporting documentation by the platform in order to make them more efficient.

4. Literature Review

NLP (Natural Language Processing) refers to computer systems that analyse, attempt to understand or produce one or more human languages. The language could either be textual or verbal; however, normally when spoken language is involved, the terms “speech recognition” and/or “speech synthesis” are used [5].

NLP research explores how computers can be used to understand and/or manipulate natural language in order to do useful things. The foundations of NLP lie in a number of disciplines including (but not limited to) computer sciences, information sciences, linguistics, and artificial intelligence [6].

A search engine is the common term for an information retrieval system. Search engines started to appear on the Internet in 1994 [7]. Search engines typically process the documents or web pages before they are needed and create an index of specific terms. Users can then type in a query. The query is executed against the index and a link to the appropriate document or web page is produced. The initial document processing includes such steps as deleting stop words, term stemming, and term weight assignment [8].

Lucene is an open source indexing and search engine application. Documents (or web pages) are initially processed and an index of search terms is created for later access. Lucene provides a facility where stemmers in different languages can be written [9].

Extensive work has already been done in indexing and searching the Internet. In 1998, Brin and Page, two PhD candidates in Computer Science at Stanford University, described their prototype of a large-scale search engine which was designed to crawl and index the Internet efficiently [10]. Brin and Page's major contribution was not in the indexing of the documents but in the relative ranking of the pages against each other. The name of their prototype was Google.

Text mining refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [11]. According to Tan, text mining consists of two distinct phases. Phase one, text refining, transforms free-form text documents into a chosen intermediate form. Phase two, knowledge distillation, deduces patterns or knowledge from the intermediate form.

An interesting and useful example of text mining and summarization is reported by Hu and Liu in which a merchant solicited customer reviews of its products via a web page. These reviews were then automatically mined for information and the information was summarized [12].

Substantial work has already been done in the area of topic spotting in well formatted English language documents. Wiener, Pedersen and Weigend have used neural networks in analysing documents from Reuters newswire stories Corpus 22173 [13]. Liu and Chua also used Reuters newswire stories Corpus 21578 as data for their semantic perceptron net [14]. The Reuters Corpus 21578 was also used by Jo using an alternative approach [15]. In these examples, the documents were grammatically correct English documents where misspellings were rare. In addition these corpora have thousands (and even tens of thousands) of documents.

The Maui project is another example of topic spotting in documents. Maui automatically identifies main topics in text documents and was a PhD project of O Madelyan sponsored by Google [16].

With the global advent of the Internet and email, there has been an increase in systems that allow the technically naïve Internet user to construct a personalised system for filtering email messages [17]. Many of these systems worked on simple keyword spotting rules. Others used more traditional text mining techniques [18].

However, by nature, email messages are short documents which often contain misspellings, special characters, and abbreviations. This brings an added challenge for text classifiers forcing them to cope with “noisy” input data [19].

The term *chat* describes Internet based communication between two people who are both online at the same time and attentive to the conversation. These are often very fast paced conversations where the textual messages are sent line by line from one person to the other – and in some cases even character by character.

Schmidt and Stone found that chat conversations have the added complexity of written words which attempt to mimic speech [20]. For example:

> *wwwhhhaaattt aaarrreee you saying???!!!*

Dong, Hui, and He [21] identified a number of common language usage differences between chat messages and traditional documents including:

1. Acronyms formed by extracting the first letters of a sequence of words such as *lol* (laughing out loud), *brb* (be right back) and *g2g* (got to go)
2. Short forms which are distinct from acronyms such as *l8r* (later) and *nvm* (nevermind)
3. Misspelling of terms occurs at a higher rate in chat conversations than in traditional published text documents due to the real-time and informal nature of the chat conversations.

Dong, Hui, and He also found that found the message length to be quite short with 91.5% of the chat messages being less than 50 bytes.

Mobile text conversations are conversations which are typed on mobile phones or cell phones. Until recently, most mobile phones or cell phones consisted of a numeric keypad with a handful of additional keys to cater for special symbols. Some mobile phones have predictive text dictionaries and some don't. Even when phones do have dictionaries, some users prefer not to use the predictive text feature. Knoblock, Lopresti, Roy and Subramaniam have collated some of the work that has been done in topic spotting in extremely noisy text such as SMS lingo [22].

MXit language or MXit lingo has become a type of *lingua franca* among South Africa's youth [23]. Millions of messages are sent daily using this abbreviated MXit lingo.

Much has been written in both the academic press and the popular press arguing whether MXit language will enhance or harm literary skills [23-25]. The jury is still out on this debate.

5. Topic Spotting Methodology

In order to successfully spot topics in these mathematical conversations held in MXit lingo, a number of steps need to be taken. Some of these steps need only be done initially and others need to be done repeatedly. The initial steps include:

1. Creating a list of mathematics vocabulary for each topic to be recognized.
2. Creating a stemmer to remove trailing suffixes and some prefixes from MXit lingo words.
3. Creating a list of stop word (or words that can be safely removed from any conversation).

During the development of the topic spotting facility, school textbooks were scoured looking for common vocabulary on specific mathematics topics. For example, the topic financial mathematics contained the vocabulary *simple, compound, interest, annuity, deposit, rate,* etc. The topic pythagoras contained the vocabulary *legs, hypotenuse, squared, sum,* etc.

A stemmer was written to remove unnecessary suffixes from the ends of the words. For example, in English, the word *dog* and the word *dogs* refer to the same concept. The suffix at the end (in this case a trailing *s*) merely gives additional numerical information about the dog. For example, in MXit, a trailing *z* can also be used to indicate a plural thus the stemmer also had to equate the word *dogz* to the word *dog.* The new spelling rules and conventions along with a technical description of how this stemmer has been created has already been reported by the authors [26]; however, a couple examples of stemming rules will be itemised here.

> *1.*A trailing *-er* can be replaced by a trailing *-a;* therefore, the stemmer to handle MXit based conversations needed to be able to recognize a word such as *numba* as being the same as the word *number.* The word *over* could be written as *ova*. In view of the fact that these suffixes could be compounded, a word such as *numbaz* (which combines the MXit *-a* suffix and the *-z* suffix) could be encountered to represent the word *numbers.*
>
> *2.*Many suffixes are written phonetically. The trailing *-tion* or *-sion* suffix could be written as *-shun* and the trailing *-ing* suffix could be written as *-tin or -tn*
>
> *3.*Numerals could be used in place of letters when the shape of the number is similar to the shape of the letter. For example, the trailing *-ed* suffix might be written as *-3d* and the trailing *-ing* suffix might be written as *-1ng*

Stop word are words which can be safely removed from a sentence or conversation without altering its basic underlying meaning. For example, in the English sentence “The boy played soccer in the street”, the words *the* and *in* can be safely removed from the sentence without detracting from its general meaning. MXit lingo also has stop words which can be safely removed from a conversation. These stop words include words like *sup, pls, da, wif,* etc. The method used to create this list of stop words has already been reported by the authors [27]; however, a brief overview of how it was done will be given here.

The historical data from previous years of Dr Math conversations was parsed and sorted. The mathematical words which were defined for specific mathematical topics were automatically removed from the historical data. This removal was not simply a string removal. The removal had to take cognisance of the various MXit spelling conventions and had to also be processed by the stemmer. In other words, a single term in the mathematical vocabulary such as *factor* would delete numerous words out of the historical data including:

facta
factors
facters
facterz
factaing
factoring
factrin
factazation
factorizin
fakta
faktring
factorisation

Once these three steps were taken, a conversation which originally looked like:

> Pupil: i nd hlp pls
> Dr Math: can I help today
> Pupil: yes pls
> Dr Math: ask me
> Pupil: i want to knw bwt tha pie equation and hw to work it out.

can be reduced to just the words *pie* and *equation* by just stemming and removing stop words.

The next steps taken are used to determine the specific mathematical topic under discussion. This was done with the use of N-grams. N-grams are collections of N sequential characters in a word, sentence, or document [28]. The N can be various values. If N is 2, the N-gram is often called a bi-gram. If N is 3, the N-gram is often called a tri-gram. N-grams are used to identify patterns. They are often used when attempting to suggest a correct spelling when an incorrect spelling is used.

Consider the word *triangle*, the N-grams of length 3 are ***t, *tr, tri, ria, ian, ang, ngl, gle*, *le**, and *e*** where * indicates the space at the beginning or end of the word. If the word were now misspelled as *triagnle*, the N-grams of length 3 would be ***t, *tr, tri, ria, iag, agn, gnl, nle, le**, and *e**.* Notice that six of the N-grams are identical between the two words. By using N-grams, it is possible to calculate a *similarity* value between two strings (which could be individual words, or sentences, or entire documents). This *similarity* ratio is the proportion of identical N-grams to total N-grams.

Previous research by Cavnar and Trenkle shows that conversations on specific topics typically have words which occur more often than other words [28]. This implies that text with two similar topics of discussion will have similar N-gram profiles. N-grams have been used to classify text in many languages besides English including Turkish [29] and Arabic [2].

Testing has shown that using N-grams of length 3 or 4 on MXit based conversations where stop words have been removed and stemming has already occurred coupled with using both the first choice topic match and the second choice topic match provided an approximately 90% hit rate on mathematical topics. It also provided an extremely high hit rate on topics while the conversation was in progress.

6. Tutoring Web Interface

As mentioned previously, the tutors use a traditional Internet based workstation and have a web interface to the C³TO tutoring system. This interface was augmented so that once the mathematical topic was determined, a hyperlink to supporting information was given to the tutor. Thus if the pupil asked a question such as:

wats da diff between simpl intrs and compund intrst?

Then the tutoring interface could provide a hyperlink to the Wikipedia page which provides a good description of the difference between the two types of interest and the formulas for both calculations. Or if the pupil asked:

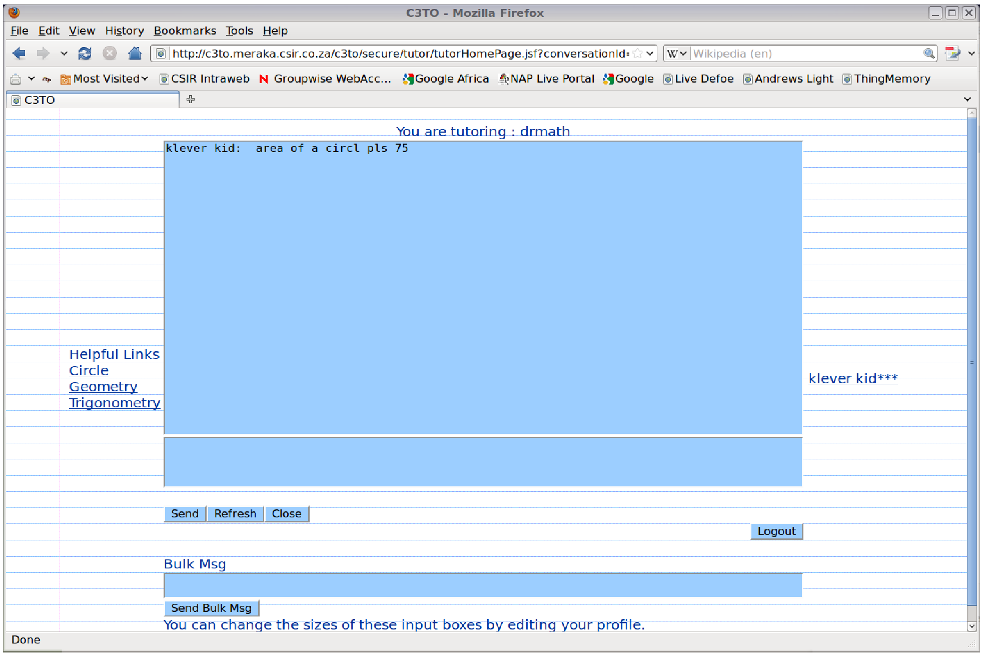
wat iz an asymtote?

Then the tutoring interface could provide a link to an appropriate page describing asymptotes and hyperbolas.

By providing these hyperlinks, the tutoring interface was making it easier for the tutors (who are, after all, volunteers) to support the pupils. The tutors would not have to waste time by asking the pupil to clarify the meaning of words such as *diff* and *intrs.* In addition, if the tutor happened to be an engineering student and was unfamiliar with the exact formula for various interest calculations, then the tutor would be able to obtain that information more efficiently by merely clicking on the provided link instead of having to search for the information itself.

7. Results

In situations where the pupil opened the conversation with an appropriate question, the topic spotter managed to spot nearly all the topics for which it was configured. Illustration 1 shows the topic spotter correctly spotting a question about circles. As can be seen from this illustration, the topic spotter makes three suggestions. These suggestions are presented as links. The links can easily be configured to point to appropriate webpages using an administrative interface.



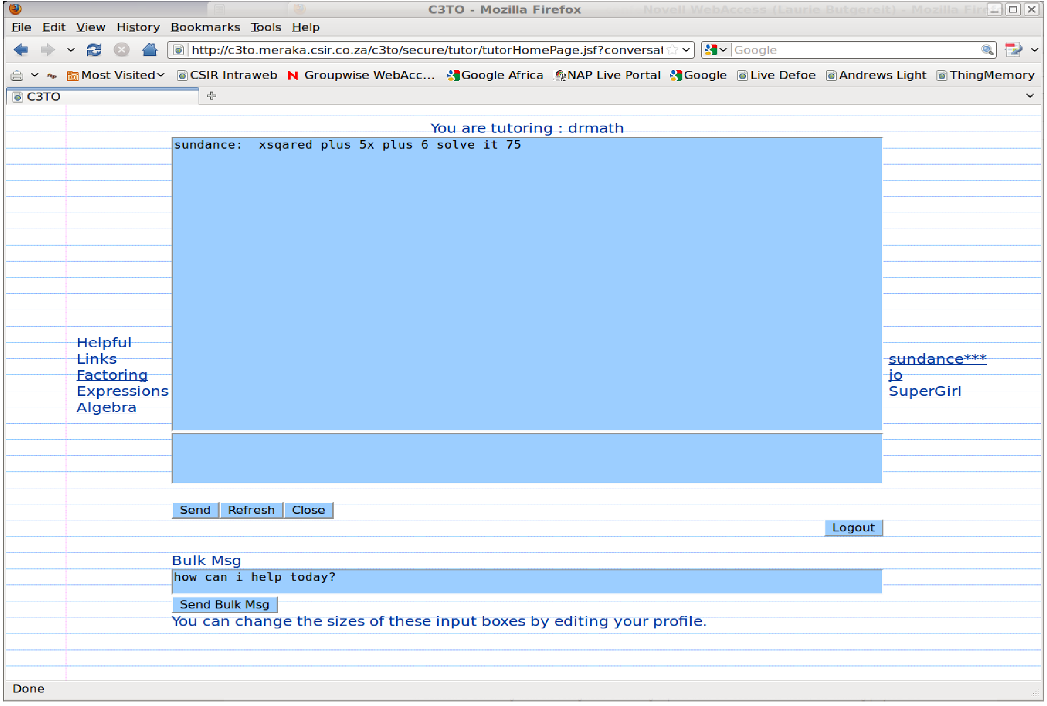
You are tutoring : drmath

```
klever kid:  area of a circl pls 75
```

You can change the sizes of these input boxes by editing your profile.

Illustration 1: Example spotting topics about circles

Illustration 2 shows a slightly more interesting example where the pupil opened the conversation with a verbal description of an equation by typing in “xsquared plus 5x plus 6 solve for x”. The topic spotter correctly determined that this was a conversation about factoring expressions and algebra and presented the appropriate links.



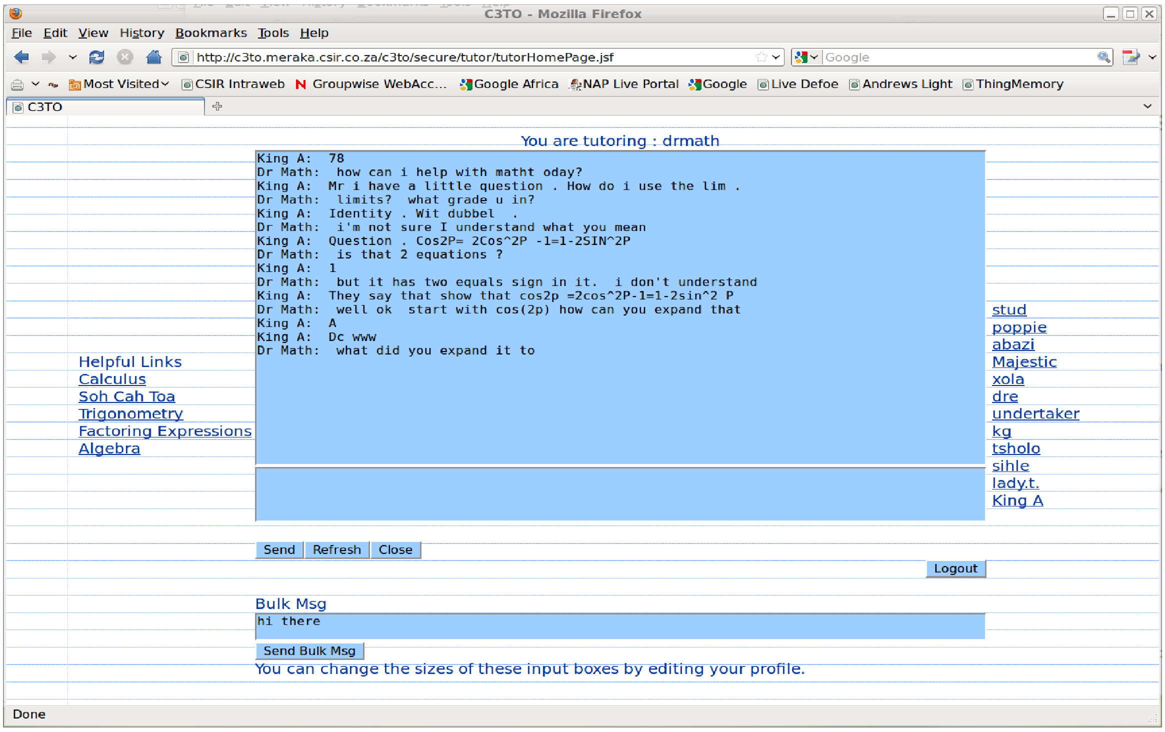
You are tutoring : drmath

```
sundance:  xsqared plus 5x plus 6 solve it 75
```

```
how can i help today?
```

You can change the sizes of these input boxes by editing your profile.

Illustration 2: Example spotting the topic of factoring expressions



You are tutoring : drmath

```
King A:   78
Dr Math:  how can i help with matht oday?
King A:   Mr i have a little question . How do i use the lim .
Dr Math:  limits?  what grade u in?
King A:   Identity . Wit dubbel  .
Dr Math:  i'm not sure I understand what you mean
King A:   Question . Cos2P= 2Cos^2P -1=1-2SIN^2P
Dr Math:  is that 2 equations ?
King A:   1
Dr Math:  but it has two equals sign in it.  i don't understand
King A:   They say that show that cos2p =2cos^2P-1=1-2sin^2 P
Dr Math:  well ok  start with cos(2p) how can you expand that
King A:   A
King A:   Dc www
Dr Math:  what did you expand it to
```

Bulk Msg

hi there

You can change the sizes of these input boxes by editing your profile.

Illustration 3: Example spanning multiple topics

As the conversation progresses, a pupil and tutor may change the topic of conversation. Illustration 3 is an example of a conversation which spans multiple topics. That particular conversation started with the topic of calculus (indicated by the keyword *limits*). It then migrated to the topic of trigonometry (indicated by the keywords *sin* and *cos*). Theconversation then moved onto topics of algebra and factoring expressions (indicated by ords such as *expand* and *equals signs*).

There are, however, two known problems with the topic spotter. The first problem is when there is, in fact, no mathematical topic to the conversation or the mathematical topic has not been configured. In view of the fact that topic spotter is being used to assist tutors, the fact that it currently classifies some topics as mathematical topics which are not, is not a major problem. The volunteer tutor will be intelligent enough to ignore these results. However, if the topic spotter was used in other applications (such as automatically attempting to route questions to an appropriate tutor) then those false results would be a problem. More work needs to be done in this respect and we expect to report positively on that at the presentation of this paper.

8. Other Applications

In this particular project, the tutors provided help with mathematics and the topic spotter was specifically configured to look for mathematical topics. The methodology, however, is easily replicable for other domains.

Any educational type topic can easily be configured because there are existing textbooks with vocabulary. In other words, if a similar tutoring system in the subject of physics were to be configured, then vocabulary for various physics topics would need to be obtained. This vocabulary would include words such as *acceleration*, *velocity*, *speed*, *vector*, and *scalar*. Or if the tutoring system was to be configured for a biology class, then vocabulary such as *phylum*, *species*, *genus*, etc, would need to be used to configure the topic spotter.

The tutoring system and topic spotter could also be used to help to dispense information about government services, child welfare grants, pensions, etc. In such case the part of the "tutor" would be taken by an information officer at the specific government department. The topics would need to be defined and proper vocabulary obtained.

The stemmer that was developed, however, was specific for English based conversations being held over MXit. If this application was to be ported to another human language, another stemmer would need to be written to cater for the cryptic chat based lingo used in that particular human language.

9. Future Research

At the time of writing this paper, the vocabulary and topics are manually created by a domain expert. Future research could perhaps automate that process by scanning online encyclopaedias for vocabulary lists on specific topics. Future research must allow the topic spotter to automatically add more words to the stop word list or to the vocabulary.

10. Conclusion

Many projects and platforms which provide critical information to the public are not viable, profitable projects. Projects operating in this area of market-neglect include projects in education and health. With the economic problems currently hitting the first world, developing countries are finding that it has become necessary for under-funded projects to be more efficient.

This paper described a topic spotting facility which was added to a mobile chat based tutoring platform. This topic spotting facility analysed the incoming questions from the pupils and provided timeous links to supporting documentation to the tutor. This supporting documentation helped in three major situations. One situation was where the tutor was not a native English speaker and did not understand the cryptic words used. Another situation was where the tutor was a mature, older person who was not familiar with the cyptic lingo used by the teenagers over MXit on cell phones. A third situation was where the tutor did not have the formula committed to memory as in the case of engineering students being asked financial questions. By providing timely links to supporting documentation, the tutors were able to answer the questions more efficiently.

Acknowledgements

The author would like to thank the Rupert Family Trust for the Rupert Gesinstigting Award supporting her PhD studies.

References

[1] L. Butgereit, "C^{3}TO: a Scalable Architecture for Mobile Chat Based Tutoring," 2011.
[2] C. B. F. Walker. *Cuneiform* 1987.
[3] J. L. Myres. The order of the letters in the greek alphabet. *Man 42*pp. 110-114. 1942.
[4] B. Laufer. A THEORY OF THE ORIGIN OF CHINESE WRITING. *American Anthropologist 9(3),* pp. 487-492. 1907.
[5] J. F. Allen. "Natural language processing," in *Encyclopedia of Computer Science* (4th ed.)Anonymous 2003, .
[6] G. G. Chowdhury. Natural language processing. *Annual Review of Information Science and Technology 37(1),* pp. 51-89. 2003.
[7] C. Schwartz. Web search engines. *Journal of the American Society for Information Science 49(11),* pp. 973-982. 1998.
[8] E. Liddy. How a search engine works. *Searcher 9(5),* pp. 38-45. 2001.
[9] E. Hatcher and O. Gospodnetic. *Lucene in Action* 2004.
[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Networks ISDN Syst. 30(1-7),* pp. 107-117. 1998.
[11] A. H. Tan. "Text mining: The state of the art and the challenges," in *Proceedings of the PAKDD 1999 Workshop on Knowledge Discovery from Advanced Databases*Anonymous 1999, .
[12] M. Hu and B. Liu. Mining and summarizing customer reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discover and Data Mining, August 22-25, 2004, Seattle, Washington, USA* pp. 168-177. 2004.

[13] E. Wiener, J. O. Pedersen and A. S. Weigend. A neural network approach to topic spotting. *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval, April 24-26, 1995, Las Vegas, Nevada, USA 332*1995.
[14] J. Liu and T. S. Chua. Building semantic perceptron net for topic spotting. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistic, July 9-11, 2001, Stroudsburg, Pennsylvania, USA* pp. 378-385. 2001.
[15] T. Jo. Profile based algorithm to topic spotting in Reuter21578. *Proceedings of the Intelligent Computing 5th International Conference on Emerging Intelligent Computing Technology and Applications, Ulsan, South Korea, September, 2009* pp. 252-257. 2009.
[16] O. Medelyan, "Maui-Indexer," vol. 2010, 2009.
[17] W. W. Cohen. Learning rules that classify E-mail. *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access* pp. 18-25. 1996.
[18] S. Zelikovitz. Mining for features to improve classification. *Proceedings of MLMTA 2003, International Conference on Machine Learning, Models, Technologies and Application, Jun 23-26, Las Vegas, Nevada, USA* 2003.
[19] H. Berger and D. Merkl. A comparison of text-categorization methods applied to N-gram frequency statistics. *AI 2004: Advances in Artificial Intelligence 3339*pp. 287-326. 2005.
[20] A. P. Schmidt and T. K. M. Stone. Detection of topic change in IRC chat logs. 1993.
[21] H. Dong, S. C. Hui and Y. He. Structural analysis of chat messages for topic detection. *Online Information Review 30(5),* pp. 496-516. 2006.
[22] C. Knoblock, D. Lopresti, S. Roy and L. V. Subramaniam. Special issue on noisy text analytics. *International Journal on Document Analysis and Recognition 10(3),* pp. 127-128. 2007.
[23] S. Vosloo. The effects of texting on literacy: Modern scourge or opportunity? *Shuttleworth Foundation* pp. 2-6. 2009.
[24] K. C. Wei. The impact of using net lingo in computer mediated communication on off-line writing tasks. 2007.
[25] D. M. Considine. LINKING THE LITERACIES: Teaching & learning in a media landscape. *Wisconsin State Reading Association Journal 44(5),* pp. 49-53. 2004.
[26] L. Butgereit and R. A. Botha, "A Lucene Stemmer for MXit Lingo," *Proceedings of ZA WWW 2011, Sept 14 - 16, Johannesburg, .*
[27] L. Butgereit and R. A. Botha, "Stop Words for "Dr Math" *,In IST-Africa 2011 Conference Proceedings, Paul Cunningham and Miriam Cunningham (Eds), IIMC International Information Management Corporation, 2011, ISBN: 978-1-905824-24-3*.
[28] W. B. Cavnar and J. M. Trenkle. N-gram-based text categorization. *Ann Arbor MI 48113*pp. 4001. 1994.
[29] A. Güran, S. Akyokuş, N. G. Bayazıt and M. Z. Gürbüz. Turkish text categorization using N-gram words. *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications, June 29 - July 1, 2009, Trabzon, Turkey* 2009.