IST-Africa

IST-Africa 2011 Conference Proceedings
Paul Cunningham and Miriam Cunningham (Eds)
IIMC International Information Management Corporation, 2011
ISBN: 978-1-905824-24-3

Stop Words for "Dr Math"

Laurie BUTGEREIT[12], Reinhardt A BOTHA[2]
[1]*Meraka Institute, CSIR, Pretoria, South Africa*
Tel: +27-12-841-3200, Fax: + 27-12-841-4720
[2]*Nelson Mandela Metropolitan University, Port Elizabeth, South Africa*

Abstract: "Dr Math" is a facility where primary and secondary school pupils can use MXit on their cell phones to get help with their mathematics homework. Pupils use an abbreviated "MXit lingo" leaving out most vowels and substituting various numerals and symbols for common sounds. Topic spotting in these conversations is required for two reasons. From a pedagogical point of view, it would be beneficial to determine if high numbers of pupils were needing help with similar topics. This could indicate a lack or gap in the educational materials. Another important use would be to attempt to pick up any inappropriate conversations between pupil and tutor which may not adhere to the code of conduct signed by the tutors or to the ethical clearance of the original "Dr Math" project. This paper describes work in spotting topics in conversations between pupils and tutors. This work is to first determine which words can be safely ignored by the future topic spotters. These words which can be ignored are known as "stop words".

Keywords: Dr Math, MXit, topic spotting

1 Introduction

Much has been written decrying the poor state of mathematics education in South Africa [1-3]. South African universities are often required to provide bridging courses to prepare the new first year university students for actual first year university education.

"Dr Math" is a project which attempts to help solve this crisis in mathematics education by linking primary and secondary school pupils to university students for help with their homework. The pupils use MXit on their cell phones. The university students have a web interface. A software platform, C³TO (Chatter Call Centre/Tutoring Online), links the two groups together attempting to ensure the anonymity of all participants. The project does have an ethical clearance. All tutors sign a code of conduct which limits the topics they are allowed to discuss with the pupils. All conversations are recorded for research, quality, and security purposes and are spot checked by administrators. The existing "Dr Math" project and the C³TO software platform have been extensively reported previously [4].

Two requirements have arisen where it has become important to spot topics in the conversations between the pupils and the tutors.

The first requirement is for security reasons. Although all conversations are recorded, there are now thousands of pupils using "Dr Math" and it is no longer feasible to manually check the recorded conversations for inappropriate conversation.

The second requirement is for pedagogical reasons. The author has acted as a "Dr Math" tutor for many, many hours. From a human point of view, it is very apparent to a tutor when course materials fail in explaining a specific topic or when a teacher has lost a class of pupils. It is a common occurrence that a tutor may get five to ten similar conversations on similar topics during the same hour.

This paper discusses the problems encountered with topic spotting in the "Dr Math" conversations and describes work in defining the appropriate "stop words" for the

conversations. For the scope of this research, only the second requirement (spotting mathematical topics) will be discussed. In addition, besides limiting the topic spotting to subjects in mathematics, the scope of this research was further reduced to subjects in mathematics in the school syllabus.

2 Examples of MXit Conversations

In order for the reader to appreciate the extent of the problem, some samples of conversations will be presented. In these samples, "Dr Math" identifies messages from the tutor and "Pupil" identifies message from the pupils.

Dr Math: hi, how may I help
Pupil: hello can u help wit word sumz
Dr Math: yes, wats the prob
Pupil: okay it says: a certain numba is increasd by 7, it will be equal 2 13 decreasd by dat numba, wat is the numba? so my equation is x + 7 = 13 - x wher did i go wrong?
Dr Math: hmm, let me c
Dr Math: its correct, now take the -x to the other side
Pupil: ohkay so it become x + x = 13 - 7? Ryt?
Dr Math: yip
Pupil: 2x = 6
Dr Math: therefore x = …
Pupil: Oh... Thanx x=3 lol yeah thanx
Dr Math: :)

In addition, sometimes the tutors also have difficulties in understanding the pupils. This was especially true when the tutors were not native English speakers.

Pupil: EloW
Dr Math: helo! How can I help u 2day?
Pupil: hw cman i find beta if cos 2 beta = -0,5
Dr Math: what?
Pupil: Find x if cos 2 x = -0,5
Dr Math: is it (2x) or cos squared?
Pupil: its actuly find theta if cos 2 theta = -0,5
Dr Math: what is the cos^-1 of -0.5
Pupil: I dnt key in da minus 4 0,5 ryt?
Dr Math: yes
Pupil: its 60
Dr Math: so because it is negative what do u do?
Pupil: Key in cos^1 0,5 then get da answer then find da ref angle
Dr Math: 180-60
Pupil: 120
Dr Math: ya

And some pupils can be extremely energetic in asking for help:

Pupil: Its lyk 12 and n on da top plus one x 9 nd 2 on da tp wit n nxt t 2 nd minus 1 nd dat divided by 36 n on da tp x 8 1 on da tp 1 minus n

3 Ethics and Safety of Minor Children

The pupils interacting with "Dr Math" are usually minor children. "Dr Math" is an open service and no parental permission is required to ask "Dr Math" for help.

As mentioned previously, the "Dr Math" project has an ethics approval from the Tshwane University of Technology. As part of that ethics approval, all tutors are required to sign a code of conduct which controls their interaction with the minor pupils. In addition, the tutors sign an informed consent which allows Meraka Institute to record their conversations with the minor pupils and report thereon. All conversations between the tutors and the pupils are recorded.

The C³TO provides a number of mechanisms to keep all participants anonymous. C³TO attempts to hide all cell phone numbers. All pupils are required to create a nick name for them. Initial string handling is implemented to ensure that nick names do not include cell phone numbers.

In addition, the minor pupils are given daily reminders that all their conversations are recorded for quality, security and research purposes. The message that they receive when they first register with "Dr Math" is:

> *This service is hosted at Meraka Institute (www.meraka.org.za). All conversations are recorded for quality, security, and research purposes. Never give out personal information over Mxit or chat.*

Subsequently, every day when they first connect to a tutor, they receive the message:

> *Never give out personal details to Dr Math. All conversations are recorded for security, quality and research purposes. Just a sec while we swap you to Dr Math....*

It is these recorded log files which form the basis of this research.

4 Data

This project is based on conversations recorded during "Dr Math" during the academic year for 2010. A total of 12817 conversations containing a total of 82289 lines were extracted from the recorded log files for the "Dr Math" project during the academic year of 2010.

Many of these conversations, however, could not be construed to be conversations in the English sense of the word because they contained just a few messages back and forth between tutor and pupils. Many of these were, in fact, one-line conversations which were simply messages that were sent from a pupil when no tutor was available. From a mathematical point of view, these messages usually did not contain any mathematical concepts; however, they are still valid MXit messages and were used to help determine "stop words".

5 Spelling

Substantial work has already been done in the area of topic spotting. Wiener, Pedersen and Weigend used neural networks in analysing documents from Reuters newswire stories (corpus 22173) [5] as did Wiener [6]. Liu and Chua also used the Reuters newswire stories (corpus 21578) as data for their semantic perceptron net [7]. In these three examples, the topic spotting was being done on grammatically correct English documents where misspellings were rare.

In the collection of conversations with "Dr Math", however, misspellings were the norm and not the exception. For example, the term attempting to be "arithmetic" was found spelled 15 different ways:

arith arithematic arithemetic arithemtic arithmatic	arithmdtic arithmet arithmetc arithmetical arithmetiec	aritmetic arth arthemetic arthemtic arthmetic

And an important word in high school mathematics classes trying to be "quadratic" was spelled incorrectly 18 times

quaddratic quadractc quadractic quadrad quadradic quadratc	quadratiba quadrativ quadratric quadriletric quadrotic quadrtc	quadtracti quadtratic quadtrc quardratic quatratic qudratic

Although it could be argued that the attempt "quadriletric" was attempting to be "quadrilateral" and not "quadratic".

When dealing with a chat medium (as opposed to traditionally written news articles), Schmidt and Stone identified a number of problems to be overcome including lack of proper capitalisation, odd punctuation, shorthand notation, emoticons as well as misspellings [8].

6 Stop Words

One of the first steps to spotting topics is to define the "stop words" (often also called "stopwords"). The term is often used with search engines and indexers and in such cases is defined as a word that has the same likelihood of occurring in those documents and are not relevant to the query [9]. Common stop words in English include words such as "the", "a", and "an". It is not feasible to just use a list of English stop words when attempting to spot topics in the "Dr Math" conversations. There are two important exceptions:

1. Some English stop words are critical to spotting specific topics in "Dr Math" conversations. For example, the common English word "a" can be easily ignored in normal English text. In "Dr Math" conversations, however, the word "a" could indicate the concept of area in a phrase such as "a = pi r sqrd" or it could indicate a turning point in a parabola in a phrase such as "tp eq -b ova 2 a".
2. There are, in fact, common MXit terms (which are not used in English) which could be considered to be stop words. These are greeting words such as "awe", and "sup", parting words such as "g2g", and common MXit spellings of common English words such as "da" (meaning "the"), "dat" (meaning "that"), and "gud" (meaning "good").

The next section will discuss how a list of "stop words" could be created using the existing "Dr Math" corpus.

7 "Dr Math" Stop Words

In order to determine the appropriate "stop words" when attempting to spot mathematical topics of conversation with "Dr Math", a statistical analysis was done on the 12817 conversations.

The conversations were broken into words and the words were then counted. In this context, a word was deemed to be any combination of letters which was delimited by a space, a symbol or a digit. The inclusion of the digit as a word delimiter was necessary in order to pick up words in expressions such as "x2+5x+6" where the pupil means "x squared plus five times x plus six". All one letter words were not considered as stop words because they are often integral pieces of mathematical formulae. These 12817 conversations contained a total of 20517 unique words. At this point the concept of a word included both mathematical terms and non-mathematical terms both spelled correctly and incorrectly.

As expected, words such as "the", "you", "is", "to" and "and" were some of the most common words and have no mathematical meaning. But even in this case, the word "to" is often an appropriate misspelling of the word "two" and should not be considered a "stop word". The following table shows the number of times some common words occurred in the 12817 conversations:

Word	Number of occurrences
the	21003
you	14961
is	11251
to	10125

Important words unique to chat conversations included:

Word	Number of occurrences
hey	1604
hw	1182
dnt	1124
ur	1022
bt	927
knw	875
don	815
cn	804
dat	801
lol	781
helo	682
wit	638

This analysis also showed the frequency of common misspellings for words. For example:

Word	Number of occurences
hypotenues	16
hypot	12
hypotenus	6
hypoteneus	6
hypoteneuse	5
hypotonues	4
hypotensue	3
hypothenusa	2
hypoteuse	2

Including one occurrence each of

hypotnus
hypotneus
hypothenuse
hypotenys
hypotenuses
hypotenuese
hypotensues
hypotencuse
hypotanur

After analysing the number of occurrences and the unique ways to spell words, it was decided that in order to be considered a stop word, the word had to appear at least 3 times in the corpus of conversations. This number, 3, is subject to change in the future depending on the accuracy of the results.

8 Topic Specific Vocabulary

After the list of "stop words" was created, a list of specific mathematical words were created. This list contained words such as "circle", "line", "point", etc. This list contained words which were common to all aspects of the school mathematics syllabus including terms in the topics of

2D geometry	functions	number theory
3D geometry	general terms	probability
algebra	graphs	quadratics
calculus	greek	sequences and series
equations	line	sets
exponents	linear programming	statistics
financial mathematics	logic	transformations
fractions	logarithms	trigonometry

These words needed to be extracted from the potential "stop words" which were automatically generated from the historical data. But it was important to also remove common misspellings of these words.

After manually reviewing the potential list of "stop words" common trends were visible on how pupils shortened words in MXit conversations. These trends included:

1. Omitting vowels: so the term "circle" would become "crcl"
2. Swapping consonants: so the term "crcl" would become "crlc"
3. Having an extra consonant from a previous or subsequence word: so the term "crlc" could become something like "dcrlc" if the previous word was "and" or it could become "crlcn" if the subsequent word was "and"
4. A trailing "er" at the end of a word often became "a": so "over" and "never" were often written "ova" and "neva"
5. Double consonants were often changed to one consonant: so the term "compass" would eventually become just "cmps"
6. Some common strings of characters were replaced by other strings would sounded similar: so "like" was often typed "lyk" and "right" was often typed "ryt" and and "the" and "there" became "da" and "der"

These rules did not include common spellings where digits were inserted into the word. For example:

1. the string sounding like "ate" was often expressed with an 8 giving words such as "gr8" for "great"
2. Often zeroes replaced the letter oh and ones replaced the letter ell.
3. The digits 4 and 2 often represented "for" and "four" or "to" and "too"

For the scope of this research, however, the author only attempted to cater for the first six trends listed above. The word and number combinations will be dealt with in further research.

A utility was written which would take a properly spelled English mathematics term and attempt to find all the common misspellings in the list of potential "stop words". The algorithm would take a term such as "triangle" and successfully remove the words

atriangle	triangl	triangular
traingels	triangle	tringle
traingle	triangles	trng
triange	trianglez	trngl
triangels	triangls	trngle

from the list of potential "stop words". It also mistakenly removed the words "turing" and "turng" from the list of "stop words" but in view of the fact that these were, in fact, "stop words" wrongly removing something from the list was not a critical problem. It would just cause extra processing at a later stage.

At this stage of our research, words which were one short in consonants in the middle of the word or had one extra consonant in the middle of the word remained in the "stop word" list. That means that words such as

triagle
triagles
triagular

An efficient mechanism needs to be found to find such common misspellings.

9 Testing

The "Dr Math" conversations from the 2011 were used as a test bed for the "stop words". It is important to note at this stage in the research, no attempt is being made yet to spot any specific mathematics topic. Only the irrelevant words are being removed from the conversations. In addition, for testing of the "stop words" all symbols and numerals were removed from the text.

For this initial testing. the algorithm described previously created a list of "stop words" of 4709 words. This was based on the assumption mentioned in section #7 that a word had to appear at least 3 times in order to be classified as a potential "stop word". If the number 3 is changed, then the number of "stop words" which exclude mathematical terms also changes.

These words where then automatically removed from all conversations in the untrained data. For example, the original conversation:

hi
hi there, any math questions for me today?
rowan plans to buy a car for r 125000,00.he pays a deposit of 15% & take out a bank loan for the balance the bank changes 12,5% p.a compound monthly
calculate tie value of the loan borrowed 4rm the bank
yes and what do you need to do, find the total cost?
total cost is r125000,00 i thnk
that is the selling price, the price the consumer pays is much more if he finances it. ok so what is the 15% deposit. how much is that?
is that 15% of 125000? did you do it on a calcluator?
ja ds
ok so how much is left to pay after the deposit?

Was distilled down to:

there math rowan plans buy car deposit bank loan balance bank changes compound calculate tie loan rm bank and find total cost total cost that consumer finances deposit that that calcluator deposit

And the conversation:

imprpr frctions pls
do u mean like 12/5 ?
yebo
well, an improper fraction is one that has its numerator greater than its denominator.
so how do u mk it prpr
oh so 2 2\5 is da prpr frctn?
12/5 is 12 divided by 5
good
fraction is da same as divid?
yes 2 and 2 fifths or 2 2/5
so prpr is lyk mxd frction
yes

became:

imprpr frctions mean improper fraction one that numerator greater denominator mk prpr prpr frctn divided fraction same divid and fifths or prpr mxd frction

This conversation about trigonometry:

can u help me 2day? 83
yes! how can i help you today?
trig ratios for sin cos and tan
do u want to knw these ratios?
yes pls
sinx/cosx=tanx, cosx/sinx=cotx, ok?
th u
any other question?
i mean da ratios lyk opp\adj and stuff lyk dat

ok, sinx=opp/hyp, cosx=adj/hyp, and tanx=opp/adj, ok?
yes dats wat i nee

became

day trig ratios sin cos and tan ratios sinx cosx tanx cosx sinx cotx other question mean ratios opp adj and sinx opp hyp cosx adj hyp and tanx opp adj

From these few examples, it is clear that by removing the unnecessary "stop words" from the conversation, the mathematical topic can be clearly recognized humanly. This research is the first step in topic spotting. The next step in the research will be to automatically recognize the mathematical topic from the resulting words.

10 Conclusion

Communication between two people is filled with common words such as "is", "are", "the", and "and". These words are called "stop words" and are routinely removed by search engines when indexing documents.

MXit is a chat facility available over cell phones. MXit conversations are characterised by special MXit-vocabulary such as "lyk" (for "like"), "sup" (for "what's up"), and "ova" (for "over"). In addition, because of the fast pace of MXit conversations, there is a high degree of misspellings.

This paper described the first step in spotting mathematical topics in these MXit conversations. This step was to remove non-mathematical terms from the conversations distilling out just the mathematical terms which could lead to the actual topic being discussed. This was done by doing a statistical analysis of historical conversations between "Dr Math" and pupils during the 2010 academic year. These words were then compared against common terms in mathematics. This comparison took into account various spelling changes that MXit users commonly employ such as changing a trailing "er" in a word to an "a" (for example, changing the word "over" to "ova") or changing the suffix "tion" to "shun" (for xample, changing "fraction" to "fracshun").

References

[1] B. Fleisch. (2008, *Primary Education in Crisis: Why South African School Children Underachieve in Reading and Mathematics* .

[2] N. Yeld, C. Bohlmann, A. Cliff, R. Prince and G. Van Der Ross, "National benchmark tests project as a national service to higher education(draft copy)," Higher Education South Africa, 2009.

[3] J. Engelbrecht and A. Harding. (2008, The impact of the transition to outcomes-based teaching on university preparedness in mathematics in south africa. *Mathematics Education Research Journal 20(2)*, pp. 57-70.

[4] L. Butgereit and R. A. Botha, "C^{3}TO: An architecture for implementing a chat based call centre and tutoring online," in *IST-Africa 2010 Conference Proceedings*, 2010.

[5] E. Wiener, J. O. Pedersen and A. S. Weigend. A neural network approach to topic spotting. Presented at Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval.

[6] E. D. Wiener. (1995, *A Neural Network Approach to Topic Spotting in Text* .

[7] J. Liu and T. S. Chua. Building semantic perceptron net for topic spotting. Presented at Proceedings of the 39th Annual Meeting on Association for Computational Linguistics.

[8] A. P. Schmidt and T. K. M. Stone. (1993, Detection of topic change in IRC chat logs.

[9] W. J. Wilbur and K. Sirotkin. (1992, The automatic identification of stop words. *J. Inf. Sci. 18(1)*, pp. 45.