A model to identify mathematics topics in MXit lingo to provide tutors quick access to supporting documentation

Authors:
Laurie Butgereit[1,2]
Reinhardt A. Botha[1]

Affiliations:
[1]Institute for ICT Advancement, Nelson Mandela Metropolitan University, South Africa

[2]Meraka Institute, Council for Scientific and Industrial Research, Pretoria, South Africa

Correspondence to:
Laurie Butgereit

Email:
lbutgereit@meraka.org.za

Postal address:
PO Box 290, Lanseria 1748, South Africa

Dates:
Received: 30 Sept. 2011
Accepted: 02 Nov. 2011
Published: 25 Nov. 2011

How to cite this article:
Butgereit, L., & Botha, R.A. (2011). A model to identify mathematics topics in MXit lingo to provide tutors quick access to supporting documentation. *Pythagoras*, *32*(2), Art. #59, 11 pages. http://dx.doi.org/10.4102/pythagoras.v32i2.59

Dr Math™ is a mobile, online tutoring system where learners can use MXit™ on their mobile phones to receive help with their mathematics homework from volunteer tutors. These conversations between learners and Dr Math are held in MXit lingo. MXit lingo is a heavily abbreviated, English-like language that is evolving between users of mobile phones that communicate using MXit. The Dr Math project has been running since January 2007 and uses volunteer tutors who are mostly university students who readily understand and use MXit lingo. However, due to the large number of simultaneous conversations that the tutors are often involved in and the diversity of topics discussed, it would often be beneficial to provide assistance regarding the mathematics topic to the tutors. This article explains how the µ model identifies the mathematics topic in the conversation. The model identifies appropriate mathematics topics in just over 75% of conversations in a corpus of conversations identified to be about mathematics topics in the school curriculum.

Introduction

Dr Math™ is an on-going project hosted at the Meraka Institute[1] which enables primary and secondary school learners to converse with tutors about their mathematics homework (Butgereit, 2011). The learners use MXit on their mobile phones and the tutors use traditional Internet-based computer workstations. MXit is a communication system which uses Internet technologies over mobile phones to provide text-based communication (Chigona, Chigona, Ngqokelela, & Mpofu, 2009). Whether because of the small mobile phone screen, the small keypad, or the fast pace of MXit-based conversations, an abbreviated form of English has developed when communicating using MXit, which is often called MXit lingo.

Although the English term 'lingo' is considered by some to be a slang term for a dialect of a language or the vocabulary of a specific industry or body of knowledge, the terms 'SMS lingo', 'Net lingo' and 'IM lingo' are often encountered in academic literature. In this project we have therefore standardised on the term 'MXit lingo' to describe the specialised vocabulary, spelling, syntax and grammar used when communicating using MXit as a medium.

The learners who converse with Dr Math usually ask their mathematics questions in this MXit lingo. Some of the questions asked by learners are straightforward. For example:

> i wnt 2 knw hw 2 fnd th nth term tht they always ask abt

Other questions can be quite complicated:

> nadeem and jeny keep fit by skiping. nadeem cn skp 90 tyms pe min, when he starts training. each week he increase dis by 5 per min. jeny starts wt 60per min n increase dis by 10 per week. after hw many weeks wl deir number of skipz b the same

The Dr Math business model uses volunteer tutors to answer these questions. Most tutors come from universities in South Africa and are typically fairly familiar with MXit lingo. However, situations do arise where tutors do not actually understand what the learners are asking; for example, the following (from now on, messages by learners will be set in bold to distinguish them from the messages from the tutors):

> **o is the centre. e.o =squareroot of=2 and oe is perpendicular todf.fynd de**
> sorry pls try write the question clearly i dont understand what you are asking
> **it z a triangl n a circle**
> which points are on the circle
> **line e.o=root3**

This article describes ongoing work on Project µ. The term µ, with its pronunciation of 'mu', represents the phrase 'MXit Understander'. Next we consider the objective of Project µ and the related research problem and questions in more depth.

1.The Meraka Institute is an operating unit of the Council for Scientific and Industrial Research (CSIR), South Africa, focused on information and communication technology.

Research problem and objectives

A number of situations have occurred where tutors do not understand the questions posed in MXit lingo. Such situations are becoming more common, due to a number of reasons:

- Increasingly, a large number of the tutors are not South African. These tutors are primarily graduate students from Central Africa who are studying mathematics in South Africa. The mathematics knowledge of these tutors is excellent, but their home language is French. They often cannot understand what is being asked in MXit lingo.
- A growing number of tutors are more 'mature' South Africans from industry who have never used MXit or instant messaging. Although their mathematics knowledge is also excellent and they are usually home language English speakers, they are not familiar with the cryptic MXit lingo.
- A handful of tutors are citizens and/or residents in Europe and North America. Again, these tutors have excellent mathematics knowledge and English knowledge, but are often not used to some of the specific mathematics vocabulary used in South Africa (such as 'surd' in the place of 'radical') and are not familiar with local words that pepper conversations (such as 'howzit', 'ja', and 'yebo').

Dr Math has become extremely popular, and tutors are typically chatting with 20–30 learners concurrently. From Dr Math's modest start, when only 20–30 learners were expected to participate, Dr Math now has over 30 000 registered learners. Often a tutor needs to quickly look up some formula or definition. A good example is when tutors from an engineering background are asked questions about financial mathematics (e.g. L.B. always needs to look up the formula for compound interest). If topics can be automatically identified, supporting documentation can be timeously presented to the tutor by the platform, thereby reducing tutor response times.

The research question, therefore, asked: How can mathematics topics be spotted in Dr Math conversations? The research objective was: To create a topic spotter which can timeously identify mathematics topics in conversations between Dr Math tutors and learners.

Research design

Project μ adopted a design science paradigm. Baskerville (2008) argues that design science can be considered a research paradigm. Design science changes the state of the world through purpose-driven development of artefacts, and thus researchers are comfortable with alternative realities. Knowledge is gained through the construction and validation of artefacts (Vaishnavi & Kuechler, 2007, pp. 16–19).

Design science is characterised by having an artefact as primary output. March and Smith (1995) propose four possible artefacts as outputs of design science research: constructs (which provide the vocabulary in which problems and solutions are defined and communicated); models (which use constructs to represent real-world situations); methods (which define solution processing, algorithms, and 'best practices'); and instantiations (which are implementations of constructs, models, and methods).

This article presents two artefacts: the μ model, which describes an executable process model used to identify mathematics topics, and the μ topic spotter, which is an instantiation of the μ model in the Dr Math tutoring platform.

The interplay between model and implementation is important in the generation of new knowledge. Vaishnavi and Kuechler (2007, p. 12) emphasise that 'the circumspection process is especially important in understanding design science research because it generates an understanding that could only be gained from the specific act of construction'. The μ model has been refined, based on lessons learned whilst constructing the μ topic spotter. Hevner, March, Park and Ram (2004) established seven requirements for good design science research, which will be used at the end of this article to evaluate the research results.

Ethical considerations

Project μ uses conversations obtained from minor children. These conversations were part of the Dr Math project. The Dr Math project has received ethics clearance from the Tshwane University of Technology. The Tshwane University of Technology is not involved with the Dr Math project in any way and could take an objective view of the project prior to issuing ethics clearance. All conversations between learners and tutors are recorded for security, quality, and research purposes.

The minor children receive daily messages from the Dr Math project which say: 'Never give out personal details to Dr Math. All conversations are recorded for security, quality and research purposes.'

Tutors sign codes of conduct where they agree not to discuss any illegal activities with the learners. Tutors also supply copies of identification such as copies of ID books, passports, or driver's licences. The tutors also sign informed consent documents agreeing that their tutoring conversations can be used for research purposes. Whenever tutors log in to the Dr Math system, they receive messages which say: 'By logging into this website, you agree that all tutoring conversations are recorded for security, quality and research purposes.'

From these messages and documents, it is clear that all participants have given their consent for their conversations to be used for research purposes. In addition, participation in the Dr Math project is completely voluntary, from the point of view of both the learners and the tutors. Any participant could resign from the project at any time. All identities were hidden from all participants.

The μ model

The μ model consists of four major phases or sections.

Phase 1 of the μ model consists of analysing and using initial historical data to create an initial configuration of μ.

Figure 1 shows that just as human infants need to encounter human language in order to learn language (Barinaga, 1997), the μ model needs historical conversations and certain mathematics information in order to start processing. The μ model uses historic conversations between tutors and learners. These historic data consist of textual recordings of conversations between tutors and learners between January 2010 and July 2011. Phase 1 only happens when there is a need to reconfigure the system; this may be when new trends in MXit lingo are observed and new words are identified, as feedback from Phase 4 indicates in Figure 1.

The μ model must be instantiated in a computer program. Every time the μ program is started, Phase 2 of the μ model is executed. In Phase 2 the configuration created in Phase 1 is read by μ, providing it with a basis for determining the topics of conversations between tutors and learners. Only minor processing of this configuration information happens at this point in time.

As learners and tutors begin to converse, Phase 3 of the μ model is executed. Phase 3 processes the conversations according to the configuration which was read in Phase 2, and attempts to determine the topics of the conversations. During Phase 3 new items may be encountered which the μ model does not understand. These could be new words, new spellings, or new contractions. In the context of the Dr Math project, this could also include new topics in mathematics which might be added to the school curriculum.

Phase 4 of the μ model takes the new words, spellings, contractions or topics and adds them to the configuration of μ. It is important to the μ model that these changes can be integrated back into the model itself, as languages change and evolve over time.

We now consider each of these four phases in more detail.

Phase 1: Initial creation

The initial creation phase of the μ model consisted of three steps:

1. creating stemmers
2. compiling mathematics vocabulary
3. identifying stop words.

Step 1: Creating stemmers

A stemmer is a utility which removes suffixes from the ends of words, leaving just the root stem of the word. Stemmers are often used in Internet search engines and other information retrieval systems (Hatcher & Gospodnetic, 2004). Stemmers take words such as *factor*, *factoring*, *factorisation*, and *factored* and remove the suffixes *-or*, *-ing*, *-isation*, and *-ed* to obtain just the root or stem of the word – *fact*. Besides being able to stem both American English and British English, this stemmer must also be able to handle new MXit lingo suffixes. For example, the word *facta* is the MXit equivalent of the English word *factor*, illustrating the MXit lingo suffix *-a* which can

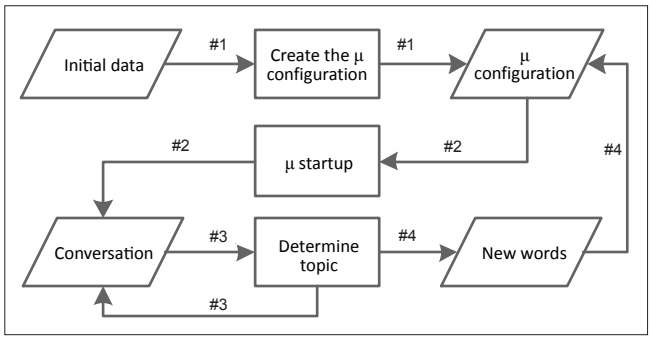

FIGURE 1: Overview of μ model processing.

replace the normal *-er* or *-or* English suffix. The stemmer must be able to remove multiple suffixes from words in order to handle a word such as *factazashun*, where the *-shun* suffix is the MXit lingo equivalent of the English *-tion* or *-sion* suffix. We have previously reported full details about MXit spelling conventions and MXit stemming (Butgereit & Botha, 2011a).

Stemmers can also operate at the beginning of a word. For example, the words *equality* and *inequality* only differ by the prefix of *in-*. Another example is the word pair *internal* and *external*; the only difference between them is the prefixes attached to the beginning of the stem. The terms 'pre-stemming' and 'post-stemming' will be used to differentiate between these two types of stemming.

Post-stemming is absolutely critical to the μ model, and will be explained in more detail when describing the actual μ implementation. Pre-stemming is also catered for in the μ model, but is not as important. Throughout the rest of this document the term stemming will thus refer to post-stemming.

Step 2: Compiling mathematics vocabulary

A mathematics vocabulary must be compiled, which must contain the words and terms which are common to mathematical conversations, such as 'parallel', 'factor', 'sum', 'sin' and 'expression'. Compilation of this vocabulary could be done manually or by automatically extracting mathematical terms from pre-tagged conversations. These vocabulary lists are created in proper English format.

The mathematics vocabulary needs to be classified into various topics and subtopics. This means that words such as 'sin', 'cos' and 'tan' need to be classified as terms in the topic of 'trigonometry', and words such as 'parallel' and 'perpendicular' need to classified as terms in the topic of 'geometry'. Mathematics terms can belong to more than once topic. For example, the term 'hypotenuse' could exist in both the 'trigonometry' topic and the 'geometry' topic.

In addition, the relationship between various topics and subtopics needs to be defined. For example, the topic of 'parabola' could be a subtopic in the major topic of 'algebra'. Subtopics could belong to more than one major topic; for example, the subtopic 'parabola' could also belong to the major topic 'graphs'. As with the compilation of the

mathematics vocabulary, the determination of topics and subtopics could be done manually or in an automated manner.

Step 3: Identifying stop words

Stop words are those which can be removed from a sentence without altering the major idea of the sentence. The expression 'stop words' is that which natural language processing practitioners use to describe these extraneous words. Stop words are identified as words which have the same likelihood of occurring in documents not relevant to a topic as in those which are relevant to the topic (Wilbur & Sirotkin, 1992). For example, in the sentence

> the sin of an angle is equal to the ratio of the opposite side to the hypotenuse

the words *of*, *an*, *is*, *to*, and *the* can be safely removed from the sentence, leaving just *sin*, *angle*, *equal*, *ratio*, *opposite*, *side* and *hypotenuse*.

Stop words for this new MXit lingo must be determined. These include words such as 'sup', 'awe', and 'howzit', which are common greetings in MXit lingo. As with the mathematics vocabulary and the mathematics topics, the compilation of stop words could be done manually or in an automated manner.

Phase 2: On start-up of µ

Since µ is an executable model, some initial processing must be done on the configuration files created in Phase 1. Figure 2 shows the aspects of processing.

During start-up, the stop words (which have been created in Phase 1) are processed by both the pre-stemmer and post-stemmer. In addition, the mathematics terms are also processed by both stemmers. After stemming, µ then reads the unstemmed stop words, stemmed stop words, unstemmed mathematics terms, stemmed mathematics terms, and a configuration of the relationship between mathematics topics, subtopics, and the mathematics terms. This information is only read at start-up time.

Phase 3: µ processing of each message

As the name suggests, Phase 3 of the µ model is where the majority of the processing takes place. Whenever a message from a learner enters the system, the steps in this phase are executed. Figure 3 shows processing for both Phase 3 and Phase 4. All processing is assumed to belong to Phase 3 unless clearly marked as Phase 4.

As each message from a learner is received by µ, stop words are immediately removed from the conversation. This removes complete words such as 'sup' and 'howzit'. The remaining words are then processed by both the pre-stemmer and post-stemmer. This stemming step changes words such as 'factorisation' to 'fact'. The stemmers cater for the English and American spelling of terms, so 'factorization' will also be stemmed to 'fact'.

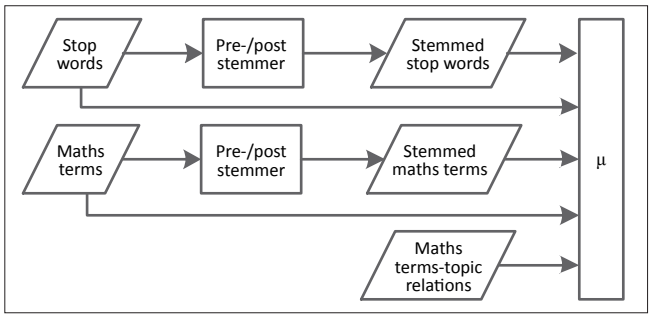

FIGURE 2: Start-up processing.

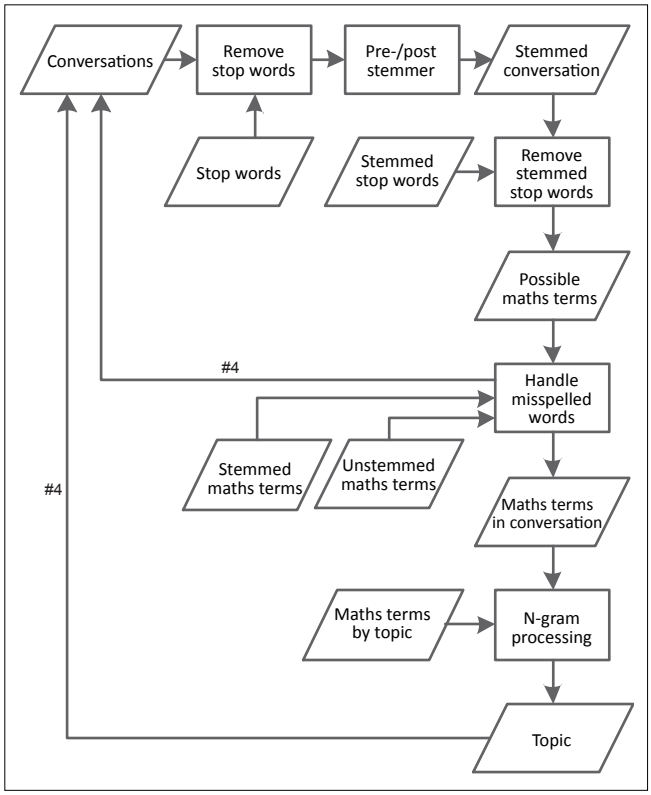

FIGURE 3: Phase 3 and phase 4 processing.

The stemmed conversation is then compared against the stemmed stop words. This is to cater for situations where the original stop word had a suffix. For example, it could be that the stop word was 'looked' and the conversation now being processed held the word 'looking'. By making this comparison with the stemmed stop words, this word can also be removed as extraneous to the conversation about mathematics.

At this point, there should only be stemmed words about mathematics in the remaining text. The next step is to look for unique misspellings of mathematics words which have not yet been encountered. This would cater for situations where, for example, the original mathematics term was 'transform' but the learner typed in 'trasnform' or, taking into account the fact that the stemmer has already executed, the learner may have typed in something like 'trasnformashunz'. This attempt to find misspelled words in MXit lingo uses algorithms similar to finding normal misspelled words in word processing systems.

Once all the mathematics terms have been extracted or distilled from the conversation, N-gram processing is used against the mathematics terms and their relationships to determine the topic of the conversation. N-gram processing will be discussed in more detail when the model instantiation is presented.

Phase 4: µ feedback loop

The µ model provides for a feedback loop where newly encountered spellings of words could be added to the µ configuration. In such a situation a word such as 'trasnform' could be added to the configuration files if it starts to appear often.

Model instantiation

To facilitate experimentation and to serve as proof of concept, a specific instantiation of the µ model was created. To avoid confusion, the term 'µ model' will be used when describing the model and the term 'µ topic spotter' will be used to describe the specific instantiation of that model.

Creating a post-stemmer

A post-stemmer utility was written which catered for American English, British English, and MXit lingo. A sample routine which caters for plurals is listed below:

```
public String singular(String word) {
String stem = word;
int length = word.length();
if (length > 4 && word.endsWith("ies") ) {
    stem = [something]
}
else if ( length > 4 && word.endsWIth("iez") ) {
    stem = [something]
}
else if ( length > 3 && word.endsWith("es") ) {
    stem = [something]
}
else if ( length > 3 && word.endsWith("ez") ) {
    stem = [something]
}
else if ( length > 3 && word.endsWith("s") ) {
    stem = [something]
}
else if ( length > 3 && word.endsWith("z") ) {
    stem = [something]
}
return stem;
}
```

This sample code removes the normal English plural suffixes of *-s*, *-es*, and *-ies*. It also, however, removes the common MXit suffixes for plurals which are *-z*, *-ez*, and *-iez*. We have previously reported an in-depth discussion of the stemming facilities of the µ model (Butgereit & Botha, 2011a).

Selecting mathematics terms

For the specific instantiation of the µ model for integration into the Dr Math tutoring platform, mathematics topics were subdivided into topics and subtopics. The topics were algebra, geometry, trigonometry, calculus, statistics, financial mathematics, number theory, logarithms, graphs, measurement, and sequences and series. For the scope of this project, three of the terms had specific definitions. The topic 'statistics' included probability and data handling. The term 'number theory' indicated the way numbers worked, including the differences between integers, natural numbers, whole numbers, real numbers, imaginary numbers, rational and irrational numbers, et cetera. The term 'number theory' also included concepts such as prime numbers, factoring, lowest common denominator, highest common factor, et cetera, but did not cover topics such as Euler's Theorem, Fermat's Theorem, Waring's problem, or Riemann's Hypothesis. It referred only to concepts of how numbers work within the scope of the school syllabus. The term 'graphs' referred to drawing curves on a set of axes. It did not refer to the higher mathematical concept of 'graph theory'.

Subtopics were also defined. These subtopics were parabolas, circles, exponents, functions, hyperbolas, lines, quadratics, solving for *x*, factoring expressions, simultaneous equations, inequalities, prime numbers, fractions, scientific notation, Pythagoras, transformations, parallel lines with transversal, sin/cos/tan, double angles, compound interest, simple interest, effective and nominal interest, and percentages.

The subtopics were classified into one or more topics. For example, the subtopic of 'parabola' was classified under the topic 'algebra' and the topic 'graphs'. The subtopic 'circle' was classified under three topics: 'geometry', 'algebra', and 'graphs'.

Creating a pre-stemmer

Pre-stemming was not as important in the specific instantiation of the µ model necessary for Dr Math. However, there was one prefix which needed to be specifically handled because of its high occurrence rate in conversations about mathematics.

The exercise of collecting mathematics terms netted 568 common terms used in conversations about mathematics. Of those 568 terms, 31 (or approximately 5%) began with the prefix *in-*. These words beginning with the prefix *in-* spanned a number of mathematics topics and included the words income, increment, inequality, infinity, inflection, insolvency, instalment, integer, integral, intercept, interest, interior, interquartile, interval, and investment. No other prefix (such as *con-*, *pri-*, *per-* or *sub-*) occurred with such a high percentage.

Special processing was done with words beginning with the prefix *in-*. The *in-* was stripped from the beginning of the word. It is important to note that the µ model has facilities to cater for any number of prefixes. In the specific case of the µ topic spotter instantiated for the Dr Math project, only one prefix was specifically processed.

Removing stop words

Historic data from Dr Math were used to 'prime' the μ topic spotter. A total of 17 413 conversations between tutors and learners, recorded between January 2010 and July 2011, were used as the basis of the μ topic spotter. These conversations consisted of a total of 25 715 unique words. An elementary statistical analysis of these historical data was done. Of those 25 715 unique words, nearly half of them (12 969) occurred only once. These were often words which had no relation to the mathematics conversation at all but represented extra sounds a person might be making, such as laughter, kisses, anger, confusion or exasperation. For example:

hahahahahahaha
mwaaaaa
aaaaaaarrrrrrrrrrrgggggggggggghhhhhhhhh
bwhahahaha
heloooooooooooooo
hmmmmmmmmmm
xoxoxoxoxoxoxo

The stop words would be removed from the message as each message is received. Considering how often it needs to be done, the removal of stop words must be a fast operation. It was therefore necessary to reduce the number of stop words. We decided that only words which had occurred more than once during the period January 2010 to July 2011 would be eligible to be stop words. This reduced the number of potential stop words to 12 746.

The next step was to automatically remove any of the mathematics terms from the potential stop words, taking into account as many of the MXit spelling conventions as possible. For example, *calculatd*, *calculate*, *calculated*, *calculater*, calculates, *calculatin*, *calculating*, *calculation*, *calculations*, *calculationz*, *calculatns*, *calculator*, *calculators*, and *calculatr* were removed. This reduced the number of stop words by approximately 13.5% to 11 015. At this point manual intervention was necessary to remove the last vestiges of mathematics terms from the stop word list, and it was manually reduced to 10 478 words.

We have previously reported our research on stop words (Butgereit & Botha, 2011b).

Processing the conversations

As the conversation between learner and tutor grew, the stop words were removed. This means that a conversation which looked like:

hw do i do transformation geometry
ooh well what is the questions
pleas explain transformation to me
what grade u in?
so what are the types of questions they ask you.
they asked me to determine t(-1;3) under the translation (x; y) into (x; y+2)
using a cartisian plane
so the trasnformation is (x;y) to (x;y+2) so you put (-1;3) as the x and y. what do you get
you get (3; 7)
no you get (-1, 5) x remains the same but y gets 2 added to it
oh, yes now i understand, so if i hv to translate t(3 ; 5) under the translation (x+4; y+2), i will get (7;
7) ?
yes
ohk thanks, let me do some practically peace out
ok

would be simply reduced to

transform geomet transform grad under transl cart plan trasnform add transl under transl

by removing the stop words.

Correcting misspellings

Once the stop words are removed, the remaining words are compared against the expected mathematics terms. In this example, the stem word 'trasnform' (which originated as the word 'trasnformation') is found not to match any stemmed mathematics term. The μ topic spotter works with so few words that every word is important.

In the μ model N-gram processing is used to determine which mathematics term is the best possible candidate match for the word 'trasnform'. After extensive experimentation it was determined that N-grams of length four would be used to attempt to find the best match for a stemmed word which has slipped through the stop word removal but does not match a stemmed mathematics term.

N-grams are collections of N sequential letters in a word, sentence, document or file (Cavnar & Trenkle, 1994). The value of N can vary depending on the specific application. This means that there may be N-grams of length 2 (often called bi-grams) or of length 3 (often called tri-grams). Figure 4 provides all the possible N-grams of length 4 for the word 'transform', where * indicates leading or trailing blanks.

N-grams have been used in text classification or categorisation in many languages besides English, including Arabic (Khreisat, 2006) and Turkish (Güran, Akyokuş, Bayazıt & Gürbüz, 2009). In addition, N-grams have been used to attempt to identify actual authors of specific segments of documents by comparing N-grams in a document where the author is known to N-grams in documents by specific authors (Kešelj, Peng, Cercone & Thomas, 2003).

In this particular case the string 'trasnform' was also converted into N-grams of length 4 and a similarity ratio was calculated for each mathematics term. Figure 5 shows

Transform
t **tr *tra tran rans ansf nsfo sfor form orm* rm** m

FIGURE 4: N-grams of length 4 for the word *transform*.

Trasnform
t **tr *tra tras rasn asnf snfo nfor form orm* rm** m

FIGURE 5: N-grams of length 4 for the word *trasnform*.

the N-grams of length 4 for the misspelled word 'trasnform'. N-gram processing calculates the similarity between two strings. The similarity is defined as the ratio of the number of common N-grams divided by the number of the union of N-grams. The calculation of the union of the N-grams can be done two different ways. As can be seen in Figure 4, there are 12 N-grams of length 4 in the word 'transform' and, as can be seen in Figure 5, there are 12 N-grams of length 4 in the word 'trasnform'. Some implementations of N-grams would calculate the union as being 24. Other implementations of N-grams would calculate the union to be the union of unique N-grams. In such a case, the union would be 17. In the case of the μ topic spotter, the union was calculated from total N-grams (not unique N-grams) in terms of the misspelling corrector.

Table 1 shows the similarity values between two possible matches for the string 'trasnform'.

Remembering that not all words that remain after stop word processing are, in fact, mathematics terms, it was not a simple matter of just taking the best choice for the word. A cut-off value for the N-gram similarity needed to be determined. After experimentation the value of 0.290 was used, which indicated that at least 29% of the N-grams were the same. In other words, the word with the highest similarity value which was over 0.290 was used as the properly spelled word, taking into account MXit stemming. Common misspellings of mathematical terms were extracted from the historical conversations and tested by the misspelling corrector.

In the case of the conversation about geometric transformations, in the set of remaining words after stop word processing the word 'transform' was listed three times.

Determining the topics

Once just the important words were distilled from the conversation, N-gram processing was carried out again. This second time, however, all of the important words (including the words which had spelling corrected) between the learner and tutor were converted to N-grams of length 4 and compared against the collections of various mathematics terms which had been classified into various topics and subtopics.

Table 2 shows the similarity values in this particular example of discussions about transformations.

In this particular case, the subtopic 'Transformations' with the highest similarity ratio was, in fact, the correct subtopic. However, N-gram processing is not an exact science. Often, when the term with highest similarity after N-gram processing is wrong, the second- or third-ranking topic is correct. Therefore, consider a case where the highest similarity is not necessarily the best match for the conversation:

i ned help wif parabola graphs
ok what is the formula
xsqrd plus 3x plus 2
can you factor that?
no

TABLE 1: Various N-gram similarities for *trasnform*.

Word	Common	Union	Similarity
trapezium	4	24	0.170
transform	7	24	0.291

TABLE 2: N-grams for topic determination.

Topic and/or subtopic	Similarity
Transformations	0.466
Parallel lines with transversals	0.313
Geometry	0.309
Algebra	0.257
Trigonometry	0.242
Circle	0.198
Inequalities	0.191
Parabola	0.185
Soh-cah-toa	0.130
Double Angle Formula	0.051

u need 2 integers which mutliply up to 2 but add up to 3 wot are dey
1 n 2?
yes well done the factors are (x+1) and (x+2) do you know what the roots are
yes -1 n -2

As each line of the conversation was received from the learner, N-gram processing was executed. So, for example, when the first line (*I ned help wif parabola graphs*) was received by μ from the learner, the topic similarities were: Graphs (0.141), Quadratic (0.131), and Parabola (0.120). When the second line (*xsqrd plus 3x plus 2*) was received by μ from the learner, the topic similarities were: Quadratic (0.144), Graphs (0.141), and Parabola (0.123).

It is important to point out that the N-gram processing is being done only when messages are received from the learner; however, the processing is done on the entire growing conversation including the tutor portion. That means that when the second line of the conversation from the learner is processed, the growing conversation includes the word *formula* from the tutor's response (*ok what is the formula*). This is to ensure that the insights of the tutor are also used in providing possible help.

When the third line from the learner (simply the word *no*) is received, the growing conversation also includes the word *factor* (or *fact* after stemming) from the tutor's response (*can you factor that?*). At this stage, the topic similarities were: Graphs (0.178), Factoring (0.157), and Quadratic (0.156).

The next line received from the learner (*1 n 2*?) is the guess of the factors as being one and two. The growing conversation now includes the terms *integers* (stemmed to *teg*) and *add*. It is interesting to note that the string *mutliply* failed to be recognised as the word *multiply*. There are two reasons for this. Since the spelling corrector works with stemmed words, it compares *multipl* and *mutlipl*, rather than *multiply* and *mutliply*. Of the 20 common N-grams, five are shared, giving a similarity ratio of 25%. This similarity is under the cut-off of 29%. However, even with ignoring *mutliply*, the topic similarity values were: Graphs (0.207), Algebra (0.202), and Measurement (0.164).

When the last line was received from the learner, adding the term *root* to the growing list of mathematics terms, the similarity ratios were: Factoring (0.234), Algebra (0.222), and Graphs (0.221).

However, it is clear that this conversation is really not about graphs, despite the fact that the learner specifically asked about graphs. The conversation is really about factoring. For this reason the three highest-ranking topics are displayed to the tutor whenever a new line is received from the learner.

Presenting the topic to the tutor

When the μ topic spotter was integrated into the Dr Math tutoring platform, the Dr Math system administrator or domain expert created a list of websites or web pages which held good supporting information on each of the specific mathematics topics supported by the μ topic spotter.

Figure 6 shows a sample tutoring screen where the learner asked:

> **area of a circl**

(Note that the learner is asked by the Dr Math system to include the number 75, to be certain that the message comes from an actual human and thereby protect the tutors from MXit-based spam. These numbers are stripped from the messages for our analysis.)

In the left column of the screen in Figure 4, the three best guesses as to the topic of the mathematics conversation are displayed as links to supporting documentation. In this particular case the top three choices were circle, geometry, and trigonometry. By clicking on the link the tutor was directed to a webpage which may provide the tutor with assistance in helping the learner.

Evaluation

Evaluation of the μ model and the μ topic spotter involved testing the topic spotter on conversations which happened after July 2011. The evaluation did not include formal user evaluation from the tutors, although some feedback was received that suggested that the μ topic spotter does identify relevant topics.

For the evaluation we selected conversations from random days in August and September 2011, yielding 1399 conversations between learners and tutors. These conversations were manually reviewed and categorised into mathematical topics and subtopics by L.B. This was done prior to using the μ topic spotter on the conversations.

We acknowledge that some bias may be present in this evaluation as an independent person unrelated to the project was not used to categorise the conversations in the corpus used. However, the extensive tutoring experience of L.B.

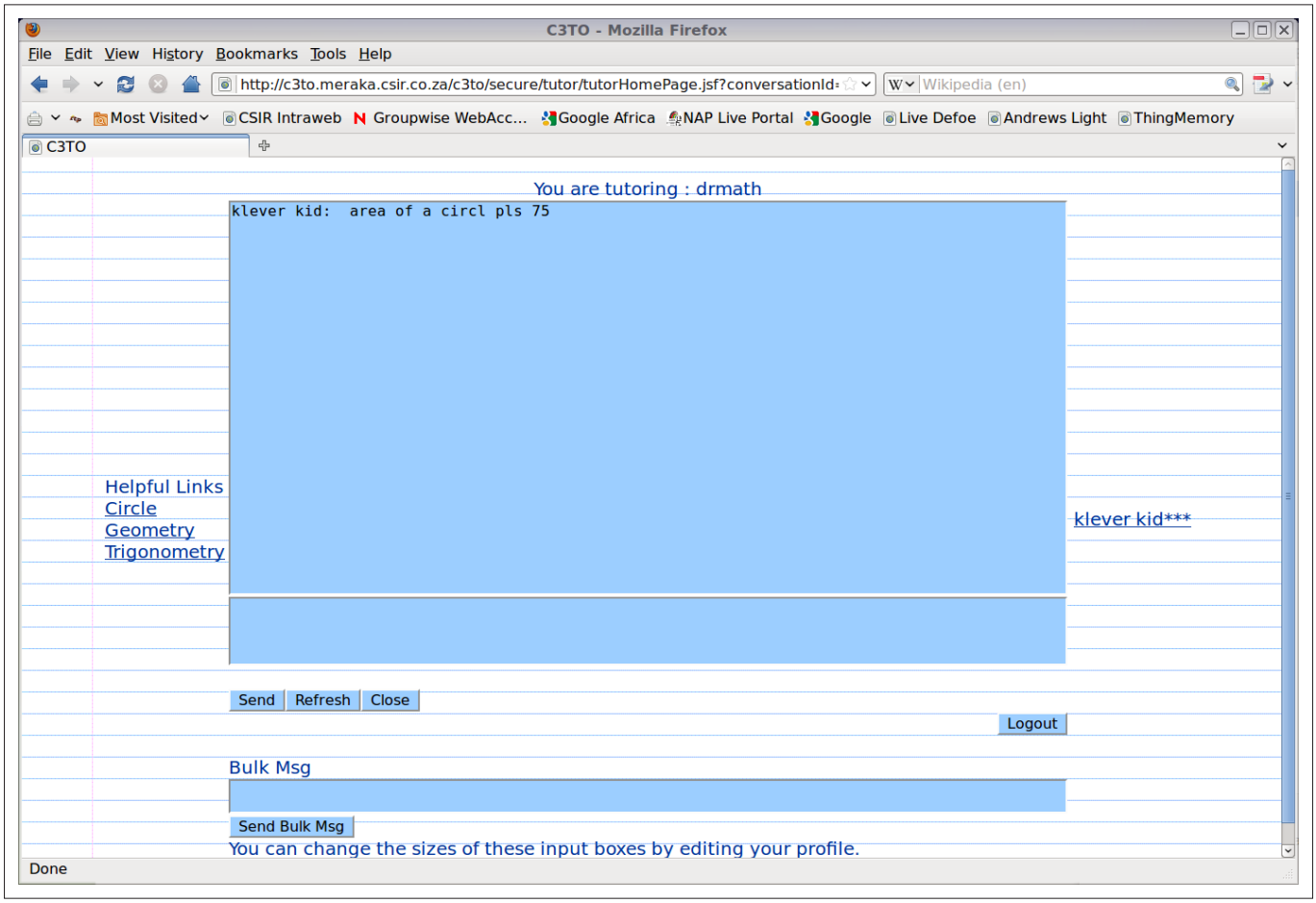

FIGURE 6: Example tutor screen.

in the environment provides a high degree of confidence in the correctness of the classification. Nevertheless, we are embarking on a project to build a tagged and validated corpus of Dr Math conversations to be used by this and related projects.

Of the 1399 conversations, 805 did not cover any topics in the mathematics curriculum. These 805 conversations included general discussions about examinations, information about how the Dr Math project works, and requests for non-mathematical help, and were removed from the sample.

The remaining 594 conversations varied in length as well as in mathematical content. Similar to during the initial identification of mathematics terms, conversations could be assigned to more than one topic and more than one subtopic. For example, the following conversation would be classified as falling under the topics of both trigonometry and graphs:

> hi how can i help?
> **hey dr maths i hv some prblms**
> **i ned 2 knw abt trig graphs coz i hv a crus prblm wif dat graph**
> well lets start with a sin graph it starts at 0, and goes between 1 and -1 over 360 degrees

The μ topic spotter then processed the 594 conversations and the topics determined by the μ topic spotter were compared with the topics which had been manually assigned. In order to be considered to be correct, at least one of the topics suggested by the μ topic spotter had to match at least one of the manually assigned topics. The results can be seen in Table 3. The μ topic spotter spotted the correct topic of discussion in more than 75% of the conversations, and was thus able to supply the tutor with fast, relevant supporting documentation three-quarters of the time.

In addition, some characteristics of the successful and unsuccessful topic selections were generated. The average number of words and number of characters in each of the 594 conversations were calculated. The average number of lines in the conversations was ignored because the concept of a line and the length of a line varied greatly between learners using a small screen on a mobile phone and tutors using a normal Internet-based workstation.

As can be seen in Table 4, messages which were classified correctly had 24% more words in the message and more than double the number of characters. The longer the conversation, the better the μ topic spotter is at correctly determining the topic under discussion.

TABLE 3: Results of topic-spotting tests.

Classification	Count	Percent
Correct	456	77
Wrong	138	23

TABLE 4: Message sizes categorised by correct or wrong classification.

Classification	Average word count	Average character count
Correct	26	538
Wrong	21	259

Discussion

This article presented the μ model that aims to identify mathematics topics in MXit conversations and discussed an instantiation, the μ topic spotter, in the Dr Math tutoring environment. The instantiation demonstrated the feasibility of implementing the μ model. Initial evaluation of the results showed the μ topic spotter to provide appropriate supporting documentation in more than 75% of cases. Considering the idiosyncratic nature of MXit lingo, this definitely represents a useful result for the purposes of the model.

In this study a corpus of conversations coded by a single person was used. Work has already started on providing a tagged corpus of Dr Math conversations for use in this and other projects. Such a validated corpus will allow more formal evaluation of the model and enable several other research areas.

During the implementation of the μ topic spotter several lessons were learned and observations made that will allow further refinement of the μ model and thus of the μ topic spotter.

One shortcoming that was identified is grounded in the observation that MXit spelling changes often occur within a word; for example, in this article 'trasnform' was equated to 'transform', since the similarity value was just slightly above the cut-off point. However, just adapting the cut-off point introduces many new false-positives. This demonstrates the limitations of N-gram processing to cater for misspelled words. To complicate matters, transform could just as easily been spelled as 'tr@ns4m' in MXit lingo. Currently μ model cannot equate such cases to the word transform.

To cater for these strange spellings, better tokenising of MXit lingo is needed. Numerals often appear in MXit words in normal conversation. For example, 'n0t' is a common MXit spelling of the English word 'not'. However, in view of the fact that this implementation of the μ topic spotter was for a mathematics tutoring environment, numerals were used as word delimiters to cater for mathematical expressions. For example, learners often typed expressions such as 'x2plus5xplus6' without any spaces, and it was necessary for the Dr Math project that the word 'plus' could be extracted from that string of characters.

As a better understanding of the MXit lingo specifically as it relates to mathematics tutoring is achieved, it may be possible to do even more domain-specific tweaking. One such suggestion is to add weighted values to the mathematical terms or, possibly, even for specific strings. For example, while searching the log files for this research one conversation was found where a learner asked about 'that python thing'. Being in a mathematics tutoring environment and not a discussion forum about snakes or programming languages, the likelihood of the question being about Pythagoras theorem is high. Perhaps future research could indicate whether just receiving the string 'py' at the beginning of a word would be sufficient to present the tutors with supporting documentation about Pythagoras theorem.

Conclusion

The Dr Math project is an important on-going project with tens of thousands of learners having used the system since its inception. Volunteer tutors are often swamped with questions and could use assistance in dealing with them. The μ model aimed to provide such assistance to tutors by providing timeous access to supporting documentation by automatically identifying the mathematics topic being discussed.

Although there are still many opportunities for improvement, the model as it stands achieved the research objective, in that a topic spotter that can timeously identify mathematics topics in conversations between Dr Math tutors and learners was created. We also believe that we have met the seven requirements for good design science, as stipulated by Hevner et al. (2004) and summarised in Table 5. Firstly, two clear, purposeful artefacts were produced, thereby meeting the basic requirement of the first guideline. Guideline 2 is implicitly matched, as a relevant problem from an operational tutoring environment was selected as a problem area.

Providing appropriate supporting documentation in just over 75% of the conversations about mathematics clearly demonstrates the functionality and usefulness of the model, which addresses guideline 3. Given the nature of MXit lingo, a model yielding a greater than 75% match can be argued to be sufficient to make a clear contribution, as required by guideline 4. The methods used in construction of the model are all well-known and accepted practice in the area, meeting the requirement for rigorous methods as stipulated by guideline 5. However, it could be argued that a corpus of Dr Math conversations using multiple coders would provide additional strength and rigour. Guideline 6, seeing design as a search process, was met through the process of circumscription, while publication of this model contributes to the communication requirement set by guideline 7.

The Dr Math tutoring service provides help to many learners where other assistance is not accessible. We believe that the μ model will aid tutors in helping learners more effectively, and we will therefore continue to refine the model and produce further initiatives around the model and the Dr Math platform.

Acknowledgements

We acknowledge the assistance of Michelle van den Heever, who holds a BA Hons in Applied Language Studies with a major in Theory of Second Language Acquisition. She reviewed many of the MXit-based conversations from a linguistic point of view and provided insight into the spelling conventions used in MXit lingo.

R.B. thanks the National Research Foundation for partially supporting his research, while L.B. thanks the Rupert Family Trust for the Rupert Gesinstigting Award that she received to support her PhD studies.

Competing interests

The Dr Math™ project is hosted at the Meraka Institute at the Council for Scientific and Industrial Research (CSIR). The term 'Dr Math' is a trademark of the Meraka Institute. L.B. is an employee of the Meraka Institute, CSIR, and, as such is eligible for various employee benefit-sharing programmes with regard to intellectual property rights.

Authors' contributions

L.B. is a PhD student and R.B. is her supervisor. Both L.B. and R.B. contributed to the conceptualisation of the research method and the resultant model in the study. L.B. did the programming and coding of conversations. L.B. wrote the manuscript with R.B. reviewing drafts thereof.

References

Barinaga, M. (1997). New insights into how babies learn language. *Science, 277*(5326), 641. http://dx.doi.org/ 10.1126/science.277.5326.641

Baskerville, R. (2008). What design science is not. *European Journal of Information Systems, 17*(5), 441–443. http://dx.doi.org/10.1057/ejis.2008.45

Butgereit, L. (2011). *C³TO: A scalable architecture for mobile chat based tutoring.* Unpublished Master's dissertation. Nelson Mandela Metropolitan University, Port Elizabeth, South Africa. Available from http://www.nmmu.ac.za/documents/theses/Laura%20Lee%20Butgereit.pdf

Butgereit, L., & Botha, R.A. (2011a, September). *A Lucene stemmer for MXit lingo.* Paper presented at the Annual Conference on World Wide Web Applications (ZA-WWW2011), Johannesburg.

Butgereit, L., & Botha, R.A. (2011b). Stop words for "Dr Math". In P. Cunningham, & M. Cunningham (Eds.), *Proceedings of the IST-Africa 2011 Conference, May 2011* (pp. 1–9). Gabarone, Botswana: IIMC International Information Management Corporation. Available from http://researchspace.csir.co.za/dspace/bitstream/10204/5043/1/Butgereit_2011.pdf

Cavnar, W.B., & Trenkle, J.M. (1994). N-gram-based text categorization. In *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR-94), 11–13 April* (pp. 161–175). Las Vegas, NV: UNLV Publications/Reprographics.

Chigona, W., Chigona, A., Ngqokelela, B., & Mpofu, S. (2009). MXit: Uses, perceptions and self-justifications. *Journal of Information, Information Technology, and Organizations*, *4*, 1–16.

Güran, A., Akyokuş, S., Bayazıt, N.G., & Gürbüz, M.Z. (2009). Turkish text categorization using N-gram words. In *Proceedings of the International Symposium on Innovations in Intelligent Systems and Applications (INISTA 2009), 29 June 2009* (pp. 369–373). Trabzon, Turkey. Available from http://www.zahidgurbuz.com/yayinlar/Turkish%20Text%20Categorization%20Using%20N-Gram%20Words-2009.pdf

Hatcher, E., & Gospodnetic, O. (2004). *Lucene in action*. Stamford, CT: Manning Publications.

TABLE 5: Guidelines for good design science.

No.	Guideline	Description
1	Design as an artefact	The research should produce a purposeful artefact which addresses an important problem
2	Problem relevance	The problem should be relevant in the research community
3	Design evaluation	The functionality, completeness and usability of the research output should be demonstrated
4	Research contributions	Effective research must provide clear contributions in the research area
5	Research rigour	Rigorous methods should be applied in both the construction and evaluation of the research output
6	Design as a search process	An iterative search process should be used
7	Communication of research	Research should be presented to a wide audience

Source: Hevner, A.R., March, S.T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly, 28*(1), 75–105

Hevner, A.R., March, S.T., Park, J., & Ram, S. (2004). Design science in information systems research. *Management Information Systems Quarterly, 28*(1), 75–105.

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In V. Kešelj, & T. Endo (Eds.), *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING'03), August 2003* (pp. 255–264). Nova Scotia, Canada: Dalhousie University, Halifax.

Khreisat, L. (2006). Arabic text classification using N-gram frequency statistics: A comparative study. In S.F. Krone, S. Lessmann, & R. Stahlbock (Eds.), *Proceedings of the 2006 International Conference on Data Mining (DMIN'06), 26–29 June 2006* (pp. 78–82). Las Vegas, NV: CSREA Press. Available from http://ww1.ucmss.com/books/LFS/CSREA2006/DMI5552.pdf

March, S.T., & Smith, G.F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. http://dx.doi.org/10.1016/0167-9236(94)00041-2

Vaishnavi, V.K., & Kuechler. W. (2007). *Design science research methods and patterns*: *Innovating information and communication technology.* Boston, MA: Auerbach Publications. http://dx.doi.org/10.1201/9781420059335

Wilbur, W.J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science, 18*(1), 45–55. http://dx.doi.org/10.1177/016555159201800106