Available online at http://www.ajol.info/index.php/njbas/index
Nigerian Journal of Basic and Applied Science (2011), 19 (2): 269-276 ISSN 0794-5698

Query Expansion: Is It Necessary In Textual Case-Based Reasoning?

***Aminu Bui Muhammad and Ahmed Tambuwal Yusuf**
Computer Science Unit, Department of Mathematics, Usmanu Danfodiyo University, Sokoto, Nigeria
[*Corresponding Author: aminuuu@udusok.edu.ng]

ABSTRACT: Query expansion (QE) is the process of transforming a seed query to improve retrieval performance in information retrieval operations. It is often intended to overcome a vocabulary mismatch between the query and the document collection. Query expansion is known to improve retrieval effectiveness of some information retrieval systems, however, its effect in Textual Case-based reasoning (TCBR) which is closely related to the field of Information Retrieval has not been well studied. In this research, a TCBR System intended for storage and retrieval of Frequently Asked Questions (FAQs) named FAQCase was developed. Experiments were conducted to examine the effect of synonym-based query expansion on the system. The result has shown that there is significant retrieval improvement in FAQCase with query expansion over FAQCase without query expansion, in a situation where vocabulary mismatch between new questions and the stored FAQs is high.
Keywords: Query expansion, Textual case-based reasoning, Word sense disambiguation, WordNet

INTRODUCTION

In most text collections, the same concept may be referred to using different words/phrases. This issue known as *synonymy* has an impact on the *recall* of most Information Retrieval systems. In traditional Information Retrieval (Van Rijsbergen, 1979), a relevant document to a query may not be retrieved if it does not contain a word that exactly matches any word in the query.

Synonym-based query expansion is the process of transforming a seed query to include synonyms of the original query terms. It is one of the popular techniques used in addressing synonymy problem in *best match* text retrieval techniques such as Information retrieval (IR) and Textual Case-based Reasoning (TCBR) (Burke *et al*., 1997). Query expansion has been well researched in Information Retrieval (White and Marchionini, 2007) and its behaviour is fairly well understood. In particular, it is generally known to improve *recall* (Salton and McGill, 1983).

TCBR systems (Leake, 1996) solve new problems by reusing previous similar problem-solving experiences documented as text. The basic principle is that similar problems have similar solutions and it is therefore easier to modify a past solution in solving a new problem than solving the problem from scratch.

In this work, we investigated the effect of synonym-based query expansion in TCBR and developed a TCBR system called *FAQCase* intended for retrieval of Frequently Asked Questions (FAQs).

FAQs are documents of question and answer pairs used to provide basic information to users in an organization. A recurring question can be solved once and stored, and for its subsequent recurrence, its already stored answer can be reused instead of treating the question as a new problem. The Case Based Reasoning (CBR) problem-solving approach (Kolodner, 1993) is naturally suited to this task as it advocates the reuse of previous cases to solve new problems. Here the previous cases are the stored FAQs while the problem would be a new question.
An experiment to investigate the effect of query expansion on FAQCase was set-up where synonyms of each query word were retrieved from *WordNet* (Fellbaum, 1998) after word sense disambiguation and part-of-speech tagging.

An adapted *Lesk Algorithm* (Lesk, 1986) was used for the word sense disambiguation while a part-of-speech tagger called *QTAG* was used for part-of-speech tagging.

WordNet as the Source of Synonyms (synsets)
WordNet is a lexical database created at the Cognitive Science Laboratory of the Princeton University (Fellbaum, 1998; Sourceforge, 2010). It can be seen as a machine readable dictionary, but, unlike most dictionaries (Fellbaum, 1998), WordNet contains only open-class words (i.e. nouns, verbs adjectives and adverbs), it does not contain any other word which is not in this category. This means that all *w*ords must be part-of-speech (PoS) tagged before being passed to WordNet. In this work, WordNet was used as the source of the synsets for query expansion.

The basic relationship between words in WordNet is the Synonym relation called *Synset*. Words in the same synset are synonymous in a particular *sense*. Word sense is the meaning a word can take depending how it is used. For example the word "*bank*" could mean a financial institution in one sense and a river bank in another sense.

Each synset of a word contains one or more words including the word itself and has a *gloss* associated with it. A gloss for a word sense is the definition of the word in that particular sense and typically includes example sentence(s). For instance, one of the synsets of 'bank' is {depository financial institution, bank, banking concern, banking company} and its gloss is (a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home").

In this work, WordNet version 2.0 was used in the implementation of Word Sense Disambiguation component of FAQCase. This version contains about 152,000 different words which have a total of about 203,000 different senses. It has over 115,000 synsets, of these synsets about 80,000 are noun synsets, 13,500 are verb synsets, 18,500 are adjective synsets and 3,700 are adverb synsets(Fellbaum, 1998). Similarly, we used the *Lesk Algorithm* (Lesk, 1986) and extended it to work with WordNet for automatic word sense disambiguation.

Part-of-speech (PoS) Tagging
WordNet requires words to be tagged with their part-of-speech before being passed to it. One of the commonly used tools for automatic PoS tagging of words is a probabilistic PoS tagger called *QTag* (Softpedia, 2010). It is a program that reads text and for each token in the text, it returns the PoS (e.g. noun, verb, punctuation, etc). It is probabilistic because it works using statistical methods. Thus, errors may arise from its tagging just like every PoS tagger, but it is fairly robust and from informal evaluation, it tags texts with good accuracy (Softpedia, 2010).

In this research, Qtag was used to tag every query word with its PoS and map the tags to WordNet recognizable PoS. Senses of a query word are then extracted from WordNet given the word and its PoS. The Lesk algorithm is then used to disambiguate the word sense. The synsets of the disambiguated sense are then extracted and used for query expansion. For example, the question "*when does the semester ends*?" was tagged by Qtag as [WRB, DOZ, IN, NN, VBZ, ?] corresponding to the tokens [when, does, the, semester, ends, ?]. These tags were then converted to their WordNet recognizable equivalents i.e. [s, s, s, n, v, s], 's' was used to denote words whose PoS cannot be mapped to the four WordNet PoS, so that, they will not be looked-up in WordNet.

Table 1 below shows complete Qtag PoS, their description and how they are mapped to WordNet PoS in this work.

Dataset Characteristics
The Dataset consists of **22** FAQs from the Robert Gordon University which are obtained from staff members who attend to students' complaints and questions.

A total number of **143** different terms were used in forming the test FAQs out of which **127** were found in WordNet, among those found in WordNet, **114** have the sense to be disambiguated and **13** do not have. The detailed dataset statistics is shown in Table 2.

The definition of Table 2 parameters are as follows:

Number of Different Words: This is the index vocabulary of the system.

Terms Found in WordNet: These are the index terms that were found in WordNet database. Although a term/word may exist in WordNet, it will not be in the list of terms found in WordNet if it is not tagged as one of the four WordNet recognisable PoS. For example, for the question *"I want to suspend my studies"* list of terms found in WordNet is [I, want, suspend], we can see that the term "studies" is not in the list, this is not because it cannot be found in WordNet, but, because it was not tagged as noun by the PoS tagger (QTAG).

Number of Terms with Targeted Sense in WordNet: Not all terms found in WordNet have the sense to be disambiguated in WordNet, some terms may be in WordNet, but, the sense to be disambiguated will not be there.
For example, in the question: "*my school email account wasn't working"* the term "account" was among the list of words found in WordNet, but then, WordNet does not have "email account" sense in the list of senses for "account". When measuring the performance of the algorithm, terms that do not have the sense to be disambiguated in WordNet were not considered. This is because there is no way the system can disambiguate their sense when they do not exist in WordNet.

Terms Successfully Disambiguated: This is the number of terms that have the sense to be disambiguated in WordNet and were successfully disambiguated by the system.

Proportion of Success: this is the proportion of query terms that are in WordNet (with the senses to be disambiguated also in WordNet) and are successfully disambiguated by the system. This is given by the equation

$$success = \frac{No.\ Terms\ Successfully\ Disambiguated}{No.\ of\ \ Terms\ with\ Right\ Sense\ in\ WordNet}$$

Table 1: QTag PoS and their WordNet equivalents

QTag PoS	Description	WordNet PoS
NN	noun, common singular (action)	
NNS	noun, common plural (actions)	
NP	noun, proper singular (Thailand, Thatcher)	n
NPS	noun, proper plural (Americas, Atwells)	
VB	verb, base (believe)	
VBD	verb, past tense (believed)	
VBG	verb, -ing (believing)	v
VBN	verb, past participle (believed)	
VBZ	verb, -s (believes)	
JJ	adjective, general (near)	
JJR	adjective, comparative (nearer)	a
JJS	adjective, superlative (nearest)	
RB	adverb, general (chronically, deep)	
RBR	adverb, comparative (easier, sooner)	
RBS	adverb, superlative (easiest, soonest)	r
RP	adverbial particle (back, up)	
others	Stopwords and punctuations etc	s

Table 2: Dataset statistics

No. of FAQs	No. of Different words	Words found in WordNet	Words with targeted sense in WordNet	Words with out targeted sense in WordNet	Words successfully disambiguated	Proportion of success in disambiguation
22	143	127	114	13	104	0.913 (91.3%)

Evaluation Metric

Performance evaluation in TCBR is still an open issue because it often comprises of multiple aspects (e.g. coverage similarity component, semantic similarity component etc) and obtaining user feedback in the form of ranked lists is time consuming. Typically evaluation is designed to measure the performance of the component which and according to Brüninghaus and Ashley (1998), there are three major issues to be considered when evaluating a TCBR system, i.e., *What performance measure should be chosen? What should the performance be compared to?* and *What to focus the evaluation on?.*

Considering the above mentioned issues, the performance measure chosen in this research was *Normalised Precision* (Van Rijsbergen, 1979). This compares the precision performance of systems between worst case and best case which ranges between 1 and 0. Worse case (0) is when the system did not retrieve the targeted answer, best case (1) is when the answer is retrieved and ranked the top and anything between 0 and 1 is when the answer is retrieved but not ranked the top. This performance measure has been chosen because there is only one FAQ (if it exists) in the case base which is expected to answer user question/query. Thus, it measures how well the system retrieved and ranked this answer between the systems in comparison.

The ability of the word sense disambiguation algorithm to disambiguate senses was also evaluated.

Experimental Design

A total of 22 FAQs questions of the stored FAQs were rephrased by three users, the users were drawn from the Robert Gordon University, one of the users is staff member, one is a research student and the other one is a masters student. This sample represents categories of the real life users of the developed system (FAQCase). Some of the rephrased FAQs contain combination of terms in their original FAQs and other different words while others were formed with completely different words from their original FAQs. Example, for the original FAQ: "*I was unable to attend exam due to illness – what should I do?"* the rephrased FAQ is: "*I missed examination because of sickness - any help?"*
Performances of two systems were measured. The description of the two systems is as follows:

System A: This is the version of the developed application that transforms user query following word sense disambiguation of the query terms (i.e. FAQCase with word sense disambiguation and query expansion).

System B: This is the same as System A except it does not perform word sense disambiguation and query expansion (i.e. FAQCase without word sense disambiguation and query expansion).

The three rephrased FAQs from the case-base FAQs formed the test samples/queries; the ability of the two systems to recover from the paraphrasing of the FAQs questions was measured using Normalised Precision.

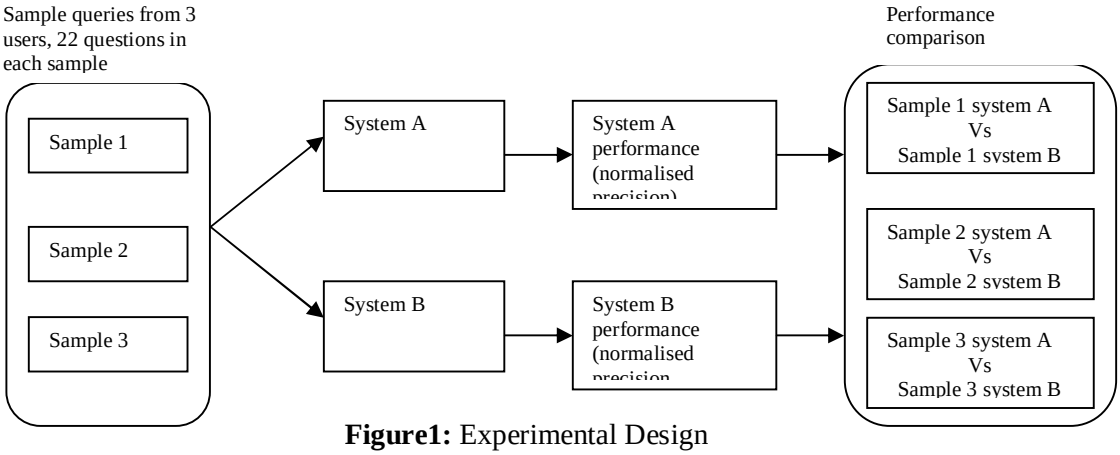

Figure1: Experimental Design

RESULTS AND DISCUSSION

Table 3 shows results of testing System A (system with query expansion) against the three sets of test FAQs. Table 4 shows the results of testing System B (System with no query expansion) against the three sets of test FAQs. Each table contains question number, rank of the answer FAQ retrieved by the system and normalized precision for each question. When the answer is not retrieved by the system, it is given the last rank (i.e. 22) so that its normalized precision is zero.

Table 3: System A Results

Question No	System A (Query Expansion) Ranks			Normalised Precisions		
	Sample 1	Sample 2	Sample 3	P-Sample1	P-Sample2	P-Sample3
1	1	1	1	1	1	1
2	3	1	1	0.9	1	1
3	1	1	1	1	1	1
4	5	1	1	0.81	1	1
5	1	1	1	1	1	1
6	1	1	1	1	1	1
7	1	1	1	1	1	1
8	7	1	1	0.71	1	1
9	2	1	1	0.95	1	1
10	1	1	1	1	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	22	1	1	0
14	4	1	1	0.86	1	1
15	1	1	1	1	1	1
16	1	1	1	1	1	1
17	1	1	1	1	1	1
18	1	1	2	1	1	0.95
19	1	1	5	1	1	0.81
20	1	1	2	1	1	0.95
21	1	1	1	1	1	1
22	1	1	1	1	1	1

Note: *Sample 1, 2 and 3 are the rephrased test FAQs questions from user 1, 2 and 3 respectively.*
P-Sample 1, 2 and 3 are the Normalised precision value for FAQs questions from Sample 1, 2 and 3 respectively.

Table 4: System B Results

Question No	System B (No Query Expansion) Ranks			Normalised Precisions		
	Sample	Sample	Sample 3	P-Sample1	P-Sample2	P-Sample3
1	1	1	1	1	1	1
2	22	1	1	0	1	1
3	1	1	1	1	1	1
4	22	1	1	0	1	1
5	1	1	1	1	1	1
6	1	1	22	1	1	0
7	1	22	1	1	0	1
8	22	1	1	0	1	1
9	22	1	1	0	1	1
10	1	1	1	1	1	1
11	1	1	1	1	1	1
12	1	1	1	1	1	1
13	1	1	22	1	1	0
14	22	1	1	0	1	1
15	1	1	1	1	1	1
16	1	1	1	1	1	1
17	1	1	1	1	1	1
18	1	1	2	1	1	0.95
19	1	1	3	1	1	0.9
20	1	1	2	1	1	0.95
21	1	1	1	1	1	1
22	1	1	1	1	1	1

The charts (Figures 2-4) show graph comparison of the two Systems on the three Samples: Figure 2 shows performance of both System A and System B on sample 1. It can be observed from the chart that at some instance there is great disparity between the two bars whereas at some other instance the bars are of equal heights. However, there is no instance in which System B bar is higher than that of System A. This means that there is no instance in which System outperforms System B.

Figure 3 shows performance of System A and System B on Sample 2. In most cases the two different bars representing System A and System B are of the same height. This means performance of both systems on sample 2 set of test FAQs were mostly the same. This can be attributed to the fact that most of the test FAQs in this sample are not much different from equivalents in the case base. Similarly, the two systems did not exhibit much difference on sample 3 as can be seen from Fig. 4.

The two systems were statistically compared using Wilcoxon Signed Rank Test on the three samples and the result has shown that System A is significantly better than System B on Sample 1 and there is no significant difference between the two systems on Samples 2 and 3. Table 5 shows the result of the test.

The difference between the two systems manifested only on sample 1 set questions. This is because most of these questions were formed with completely different words with their equivalents in the case-base. Unlike sample 1, most of sample 2 and 3 questions have some common words with their equivalent questions in the case-base.

The result has, therefore, shown that, when user questions are likely to be completely of different words from a FAQ question that is likely to answer them then query expansion is desirable otherwise it is not necessary.

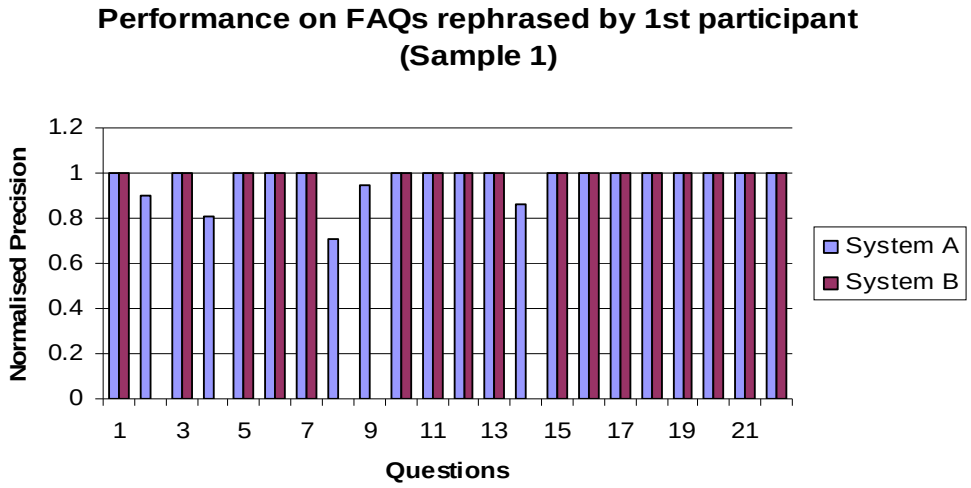

Figure 2: System A Vs System B on Sample

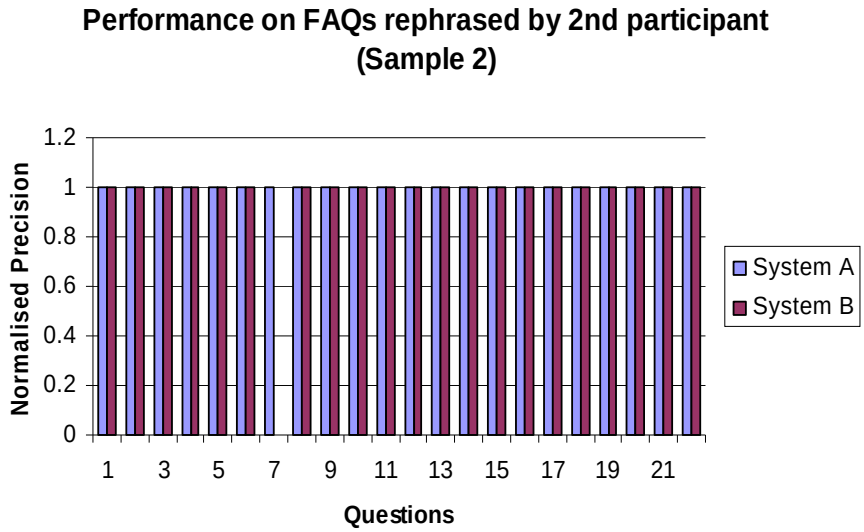

Figure 3: System A Vs System B on Sample 2

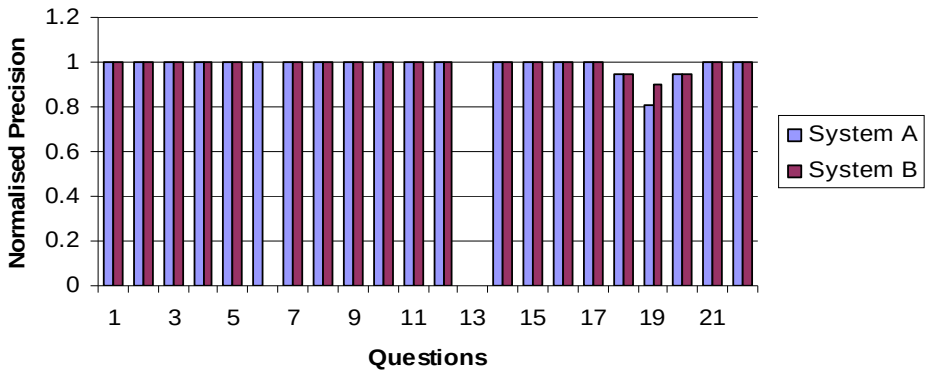

Figure 4: System A Vs System B on Sample 3

Table 5: Wilcoxon Signed Rank Test result

Test	N	Test statistics	p	Median
System A Vs System B on sample 1	22	15.0	0.03	0.00
System A Vs System B on sample 2	22	1.0	0.50	0.00
System A Vs System B on sample 3	22	2.0	0.50	0.00

CONCLUSION

In this research, a TCBR system (FAQCase) has been presented which uses WordNet lexical database to transform user queries/questions. The query transformation was achieved by using WordNet's glosses and the Lesk Algorithm to disambiguate query terms after which, appropriate synonyms of the query terms are added to the query.

Experiment was conducted to observe the effect of query expansion on FAQCase and the result has shown significant retrieval improvement on FAQCase with query expansion over FAQCase without query expansion where new problems (questions) are made of highly different words from the experiences (FAQs) capable of providing solution to them.

Although WordNet is a general English database, it lacks some domain specific terms. There are some terms that cannot be found in WordNet such as the term resit. Similarly, a term may be found in WordNet but a particular usage of the term may not be found. For instance the term account could be found in WordNet, but, account in the sense of 'email account' could not be found in WordNet. Therefore, future work related to this research should be in the use of more promising knowledge sources in word sense disambiguation such as web resources (e.g. wikipedia and google search history)

REFERENCE

Brüninghaus, S. and Ashley, K.D. (1998). *Evaluation of Textual CBR Approaches*. In: Proceedings of the AAAI-98 Workshop on Textual Case-Based Reasoning, Madison, WI.: USA

Burke, R.D., Hammond, K.J., Kulyukin, V., Lytinen, S.L., Tomuro, N. and Schonberg, S. (1997). *Question-Answering from Frequently-Asked Question Files: Experiences with the FAQ-Finder System*. AI Magazine, 18(2): pp.57-66 **NOT CITED**

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database.* London; The MIT press.

Kolodner, J. (1993). *Case-Based Reasoning.* San Mateo, California: Morgan Kaufmann;

Leake, D. (1996). *Case-Based Reasoning: Experiences, Lessons, and Future Directions.* Menlo Park: AAAI Press/MIT Press.

Lesk, M. (1986). *Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to tell a pine cone from an ice cream cone*, Proceedings of the 5th annual international conference on Systems documentation, pp. 24 – 26.

Salton, G. and McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Softpedia (2009). *QTAG Tagger*. [Online] Available from http://www.softpedia.com/get/System/File-Management/OM-Tag.shtml [Accessed June 11 2009]

SourceForge (2009). Java WordNet Library (JWNL) [Online] Available from http://sourceforge.net/projects/jwordnet [Accessed August 2, 2009]

Van Rijsbergen, C.J. (1979). *Information Retrieval*. 2nd ed. London: Butterworths.

White, R.W. and Marchionini, G. (2007). *Examining the Effectiveness of Real-time Query Expansion*. *Journal of Information Processing and Management* **43(3):** 685-704.