Journal of Computer and Communications, 2015, 3, 67-76
Published Online December 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.312006

Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique

Abey Bruck, Tulu Tilahun

Department of Computer Science & IT, AMiT, Arba Minch University, Arba Minch, Ethiopia
Email: bruckabey@gmail.com, tuttilacs@gmail.com

Received 26 September 2015; accepted 13 December 2015; published 16 December 2015

Abstract

Information retrieval (IR) systems are designed to help information seekers retrieving relevant information from vast document. The need for relevant information from a vast amount of document gave birth to IR systems. Even though different IR systems exist, they cannot meet all users' expectations. A different level of users' knowledge makes queries to be expressed in different ways. As a result, the system may miss the core meaning of users query and retrieve dissatisfactory results. This happens mainly because of the ambiguities of words involved in the natural languages and expression mismatch among users and authors. The existing ambiguities in Amharic language have negative impacts on the performance of Amharic IR system. Some of the ambiguities for this type of problem are: spelling variants of the same word, polysemous and synonymous terms. If users are not fully knowledgeable about the information domain area, they will mostly formulate weak queries to retrieve documents. Thus, they end up frustrated with the results found from an IR system. This research has been conducted, aiming at augmenting the recall of previous work. Statistical co-occurrence technique has been used in order to expand query terms. The main reason for performing query expansion is to provide relevant documents as per users' query that can satisfy their information need. Statistical co-occurrence method considers, frequently appearing terms with the query term, regardless of their position. The efficiency of proposed technique has been tested on the prototype system and the result found compared with the result of previous study. Accordingly, 6% recall and 2% f-measure improvement has been made. Hence, the statistical co-occurrence method outperformed the bi-gram based IR system.

Keywords

Statistical, Co-Occurrence, Information Retrieval, Query Expansion, Amharic

How to cite this paper: Bruck, A. and Tilahun, T. (2015) Enhancing Amharic Information Retrieval System Based on Statistical Co-Occurrence Technique. *Journal of Computer and Communications*, **3**, 67-76.
http://dx.doi.org/10.4236/jcc.2015.312006

1. Introduction

As early as 1945 [1], Bush talked about a device in which an individual stores all his books, records, and communications, which has been mechanized so that it may be consulted with exceeding speed and flexibility. Bush's ideal device has become a reality, that the term "Information Retrieval" was coined as its name in 1952 [2]. IR is one of the ingenious solutions mankind invented to solve the obvious problem of searching for information. The idea of using computers to search for relevant pieces of information was popularized by Vannevar Bush in 1945 [1]. Bush [1] stated, "Instruments are at hand which, if properly developed, will give man access to and command over the inherited knowledge of the ages". This system predefined by Bush, is an IR system, which is used in all aspects of information retrieval now a days. The need for relevant information from a vast amount of document gave birth to information retrieval (IR). IR is the art and science of retrieving relevant documents by searching and measuring similarity between users query and documents in a certain corpora.

Information seekers thus, should implement an IR system to satisfy their information need [3]. The main goal of an IR system is, to support and nourish information-seeking behaviors of users [4]. Even though the task of information retrieval systems is retrieving relevant information, there is no one system, which is capable of retrieving relevant and only relevant documents as per users' query.

To investigate the underlying principles/theories, philosophies, techniques and tools of the various methods that can be adopted, manipulated or simply selected we have reviewed literatures and then collected data for testing purposes. And then, queries that have polysemous and synonymous terms are formulated in order to challenge the Amharic IR system and measure the extent to which it can discriminate their various meanings.

Nowadays, the IR research focuses on, the problem of making existing IR systems retrieve better results. Regarding this problem, different mechanisms have been introduced by many researches, in order to achieve a solution [5]. Some of these proposed techniques, as listed in [6], are: User-oriented search result organization; Incorporating user negative feedback; supporting effective browsing and Effective query reformulation. Among the listed mechanisms, effective query reformulation or query expansion based on semantic thesaurus has been found to be better [6].

After receiving a query, an IR system goes through a series of steps, in order to provide seemingly relevant information to users. First, it clusters documents as relevant and non-relevant as per the query, then ranks the relevant documents and displays them according to a certain similarity measurement [7]. Other than implementing such steps, a good IR system's designer should need to have knowledge of the relationship between a query, its user and the environment Stubinz and Whighli [8]. According to [8], before considering the experiments in designing a good information retrieval, impact of language and text needs to be considered. Researches introduced various methods in information retrieval to satisfy users' information need. Since the initiation of automated information retrieval, Boolean queries were an accurate reflection of user's information needs [9]. In time, language and needs outmoded this approach, leading to the need for approximated similarity measures, such as the inner product, the cosine similarity measure and probabilistic retrieval [10]. For many years these were the methods of choice, but were based on studies that neglected the user, and now the possibilities have been completely explored [8]. Therefore, the frontier of information retrieval research is, to exploit new methods which can take the exact users information need in to consideration. In this research, it is suggested to achieve this objective, through pre-calculated query operation.

IR system designers may face challenges on query reformulation issues, besides challenges that occur from the user's perspective. For example, when synonyms or polysemous terms, in queries are encountered, it can be difficult to relate and identify other terms for expansion. That is because, when such terms are used in queries, they seem to convey a certain meaning, while the users' intention was another [7] [11] [12]. Thus, this problem apparently becomes a major challenge, for IR designers to specify a solution to. Altogether, ambiguity that arises between users and authors, context similarity in written documents and existence of polysemous words in the natural language, exaggerates the problem further. Therefore, IR systems should comprise a more sophisticated technique and strategy that enables it, to cope up with the problem and have more of user-centered approach. Such techniques and strategies would be, ontology based query expansion method, meaning oriented thesaurus usage, expanding the whole query with common expanding terms and strategies which seek meaning out of users' queries.

Different methods can be used for query expansion [7] [13] [14]; such as word space nearest-neighbor and thesauruses method, cluster-based query expansion, latent semantic indexing, concept-based query expansion.

Even though many kinds of query expansion methods exist, their goal is the same; making users search task easy as much as possible, aiming at delivering information that satisfies users’ information need.

The purpose of this research is to design statistical co-occurrence based query expansion technique that can overcome the limitations of the previous research aiming at delivering relevant documents as close as professionals’ do. In turn, research result introduces a technique to develop a fully-fledged IR system that can behave like professionals so that users can be satisfied with results retrieved. Further, to enhances the precision of the system so that more relevant documents can be found among the retrieved ones.

2. Amharic Writing System and Ambiguities

Amharic is the official working language of Federal Democratic Republic of Ethiopia. Its alphabet is grouped into either consonantal or syllabary writing system. A consonantal system represents symbols separately as consonants and vowels, while the syllabary writing system has individual signs for syllables (*i.e.* consonant + vowel combinations). There is a dispute as to whether or not the Amharic writing system is a syllabary [15]. It could be argued that Amharic to some extent resembles other Semitic scripts such as the Arabic and Hebrew consonantal systems, which basically indicate consonants but, for teaching purposes, etc, have developed optional diacritics to signify vowels [15]. Unlike the Arabic and Hebrew, the Amharic writing system is written from left to right.

Amharic writing system has no upper and lower case latter variations and has no conventional cursive (*i.e.* written in a connected letters) form. Unlike Hebrew and Arabic, there are no systematic variations in the form of the symbol according to its position in the word [15].

When Ge’ez became the spoken and written language in common use in northern Ethiopia, it took only 24 of the 29 Sabaean symbols, modify most of them and add two new symbols to represent sounds of Greek and Latin loanwords not found in Ge’ez [7]. The two extra symbols are added to cope with two new sounds (/p/ as in “ፐ” and /p'/ as in “ጰ”), originally required for use in ecclesiastical Greek and Latin borrowings and names (e.g. Poulos, Police) [15]. Amharic inherited symbols from Ge’ez. It took all of the symbols and added eight new ones (“ቸ”, “ጨ”, “ጀ”, “ኘ”, “ቨ”, “ሸ”, “ኸ”, and “ዠ”) that represent sounds not found in Ge’ez [7]. Alphabetic system of the Amharic writing system consists of 34 base symbols with seven orders, representing seven vowels. For example the consonant /b/ has seven different symbols (*i.e.* በ, ቡ, ቢ, ባ, ቤ, ብ, ቦ read as bə, bu, bi, ba, be, bɨ, bo, respectively) representing seven vowels combined with the consonant /b/. Taking another consonant like /z/, the seven symbols ዘ, ዙ, ዚ, ዛ, ዜ, ዝ and ዞ read as zə, zu, zi, za, ze, zɨ and zo respectively can be found.

From the construction of the seven consonant-vowel combinations of the /b/ and /z/ consonants, one can see that there is a consistent pattern in the shapes, except the 1st symbols pronounced as (consonant + “ə”) and the 6th order which are less systematic. Considering all the 34 consonant symbols, there is considerable regularity of letter shapes, but some orders are more regular than others. The shapes are most consistent in the 5th order (e.g. ሄ, ሌ, ሜ, ሴ, ሬ having a ring on their right legs except ጬ), slightly less in the 3rd (e.g. ሂ, ሊ, ሚ, ሲ, ቢ having a small hyphen like extension on the bottom of their right legs, except ሪ and ዪ), slightly less again in the 2nd (e.g. ሁ, ሉ, ሱ, ቡ having a small hyphen like extension on their right legs except ሩ, ፉ and ው), still less in the 4th (e.g. ላ, ሳሻ, ባ, ካ, ዛ, ዣ, ዳ, ጃ, ጳ have their left leg shortened, on the other hand ጣ and ጫ has there two left legs shortened, others like ቃ, ታ, ቻ, ጋ, ፓ have bow like legs) and even less in the 7th (e.g. ሶ, ሾ, ቦ, ኮ, ዞ, ዦ, ጶ have their right legs shortened, on the other hand ጮ and ጦ have their two right legs shortened, others like ሆ, ሎ, ኖ, ኞ, ኖ, ሮ have a ring somewhere on their base letter or have some kind of modification like ሞ and ሦ) and the 6th (e.g. ህ, ል, ም, ስ, ድ, ቅ, ት, ፕ, ጭ having completely distinctive and pattern less structures) order, with the greatest number of patterns, [15]. Therefore the system is composed of largely unpredictable patterns. For example the set of syllables with /g/ starts off regularly enough except the 4th “ጋ”, 6th “ግ” and 7th “ጎ” orders. The /w/ set is even ambiguous having the 2nd “ው” read as “wu” and 6th “ዉ” read as “wɨ” symbols highly alike. The system has regularity in writing which makes it easy for a person to learn the language. For example, the symbol “ሰ” read as “sə” and “ሸ” read as “shə” have relatively the same kind of structure as their accent. Therefore, a person experiences any difficulty to learn a symbol, given that he/she knows the other (for some of the symbols). The Amharic writing system has 34 basic characters and their seven orders give 238 distinct symbols. In addition, there are forty others that contain a special feature usually representing labialization e.g ቿ, ቋ [7].

Amharic writing system adopted all the symbols in Ge’ez and added 8 other symbols and the other 44 symbols. The result is that there is a considerable systemic redundancy of several consonant sounds which lacks in the phonology of Ge’ez [15]. Ambiguities in Amharic writing system arise mainly due to symbol redundancy [7]. Thus, 4 distinct sets of 7 can represent the sound /h/ + vowel: (“ሀ”, “ሐ”, “ኃ”, “ኸ”), 2 sets represent /s/: (“ሰ”, “ሠ”) and 2 /s'/ (“ጸ”, “ፀ”) [15]. A similar problem is observed in usage of some letters interchangeably, like “ቆ” vs “ቃ” [7]. In addition to the symbolic redundancy of characters, Amharic writing system suffers slightly from visual similarity or different character, such as ፐ and ኘ, ፖ and ፓ, ዪ and ዩ, ጉ and ጐ [7]. The different forms of spelling variants of the same word are shown in **Table 1**. Because of those symbols having the same accent in speaking, they can be used interchangeably in the various words of Amharic language, thus, forming different forms of spelling for the same word. Using these symbols interchangeably in words doesn’t make reading, or forwarding ideas difficult for human beings. But unlike humans it is difficult for systems to consider them having the same meaning. Because IR systems only match the symbols in words to check weather a word from a document has the same meaning as in the query (*i.e.* if the words are a match then they are the same and have same meaning) encountering the different interchangeable symbols in words forces it to consider them as different *i.e.* the system considers ፀሀይ and ጸሃይ, as different words with different meaning.

The other challenge Amharic IR systems face is the combination of two words. There is no convention as to which words should be combined as one word or separately during writing. For example, the word “megneta bet” which means “bed room” can possibly be written as መኝታቤት (without space) and መኝታ ቤት (with space) and also the word “betemekides” which means “temple” can be written as “ቤተመቅደስ” (without space) and “ቤተ መቅደስ” (with space) which makes it difficult for the IR system to differentiate between them [7]. These kinds of situations makes IR systems’ task difficult.

According to 1998 statistical census [16], Amharic language has 17.4 million speakers as a mother tongue and 5.1 million speakers as their second tongue. Many recent researches have indicated that, electronic documents in Amharic keep on growing every year [7] [17]. This apparently makes it difficult for IR systems, to find relevant document from the vast amount of electronic documents available nowadays.

There are information retrieval systems designed for Amharic language [17]-[22] that attempts to retrieve relevant documents as per information need of users. Even though such IR systems exist, they cannot meet all users’ expectations. In order to satisfy the need of extracting relevant information, data mining and text processing techniques have been applied so far [23] [24].

Different levels of users’ knowledge makes queries to be expressed in different ways, as different authors express ideas using different terms. As a result, systems most of the time, misses the core meaning of users query and retrieve dissatisfactory results. This happens mainly because of the ambiguities of words involved in the natural languages and expression mismatch among users and authors.

The prevalence of synonyms query terms tends to decrease precision at higher recall levels [12]. One of the solutions suggested to solve such a problem is query operation. A recent research paper done by Alemayehu [7] is worth mentioning here, which attempts to apply query expansion to control synonyms words using thesaurus. Alemayehu tried to enhance the system’s recall at the expense of its precision. The performance analyses show that there is an enhancement of recall from an average of 0.29% to 0.65%. On the other hand, because of a tradeoff between recall and precision and because of polysemous query terms existence, his proposed system decreased the overall precision from 0.91% to 0.57% on average. This happens because some synonymous query terms can also be polysemous.

Words in the natural language can be regarded as polysemous or synonymous according to the context they are used in [25]. A word is polysemous if it has different contextual meaning. On the other hand, if a couple or more words refer to one meaning, they are termed as synonymous words. For example in these two sentence

Table 1. Spelling variants of the same word.

Canonical Amharic	Common Amharic	Possible but improbable Amharic
ዓለም	አለም	ዐለም ፣ አለም
ፀሐይ	ጸሃይ ፣ ፀሃይ ፣ ጸሐይ	ጸሀይ ፣ ጸሓይ ፣ ጸኅይ ፣ ጸኃይ ፣ጸኻይ ፣ ፀሀይ ፣ ፀሓይ ፣ ፀኅይ ፣ ፀኃይ ፣ ፀኻይ
ኃይለ ሥላሴ	ኀይለ ሥላሴ ፣ ኃይለ ስላሴ ፣ ኀይለ ስላሴ ፣ ኀይለ ስላሤ ፣ ሀይለ ስላሤ ፣ ሀይለ ሥላሴ ፣ ሃይለ ስላሴ ፣ ሃይለ ሥላሴ ፣ ሃይለ ስላሤ ፣ ኀይለ ሥላሤ ፣ ሀይለ ሥላ ሤ ፣ ሃይለ ሥላ ሤ ፣ ሐይለ ስላሴ ፣ ሐይለ ስላሤ ፣ ሐይለ ሥላሴ	ሐይለ ሥላሤ ፣ ሓይለ ሥላሤ ፣ ሓይለ ሥላሴ ፣ ሓይለ ስላሤ ፣ ሓይለ ስላሴ ፣ ኻኀይለ ሥላሴ ፣ ኻይለ ሥላሴ ፣ ኻይለ ስላሤ ፣ ኻይለ ስላሴ

“አበበ ካራ እየሳለ ነው” and “አበበ ቢላ እየሳላ ነው”, which their equivalent meaning in English is “Abebe is sharpening a knife”, the two words “ካራ” and “ቢላ” refers to a “knife” and thus, they are synonyms to each other. In another two sentences “አበበ ካራ እየሳለ ነው” and “አበበ ስእል እየሳለ ነው” the word “እየሳለ” refers to “sharpening a knife” in the first sentence and “drawing a picture” in the later. Therefore, this word “እየሳለ” changes its meaning according to the context it is used and thus it is polysemous. In a sentence “አበበ ጠላ ይወዳል” which means “Abebe likes tela”, “tela” or “ጠላ” refers to an Ethiopian traditional drink. In this context, the term “ጠላ” is a polysemous term which means “a traditional drink” or “hatred”, which can also be a synonym for other terms that bare one of its meanings. If the term “ጠላ” is used in a query, an Amharic query expansion technique may try to expand it using terms such as, “መጠጥ” and “ባህላዊ” for the meaning “traditional drink” or “መጥላት” and “አለመዋደድ” for “hatred”. Thus, expanding the query with all of the expansion terms mentioned above, may make the system to retrieve all the relevant documents, but with many non-relevant ones in between. In addition to the example given, there are many polysemous words in Amharic language. To mention some of them; “በቅሎ” means “a mule” or “to grow”, “ዋና” means “main” or “to swim”, “አንቀላፉ” means “to sleep” or “to die”, “ሳለ” means “to cough”, “sharpen a knife”, “to draw a picture” or it can even express a “period” in time, according to a context it is used. There are also phrases which can be equally ambiguous. The phrase “አበባው በቀለ” can be understood as a person’s name or “the flower has grown”. Therefore, it is necessary to consider polysemous nature of terms in addition to synonym terms for query expansion. It is therefore the purpose of the current research to investigate the possibility of building statistical co-occurrence based query expansion that can control polysemous and synonymous words to enhance the performance of the system by considering users’ information need.

3. Query Expansion Based on Statistical Co-Occurrence

An IR system takes a string query, and retrieves documents based on a certain similarity measurement technique. In addition, an IR system has its own performance level as measured in terms of recall and precision. As any other systems’ performance measurement, it is most unlikely to score 100% on both precision and recall in the case of information retrieval systems. But good systems are designed to enhance both precision and recall to the possible limit. Thus the aim of this research is, to design a good system, which can enhance the recall of the system without affecting its precision. This can be achieved by integrating a query expansion model with the system, so that it finds words having similar meanings with users’ query and retrieves relevant documents that satisfies users information need.

This method generates synonymous expanding terms for a query term based on index terms co-occurrence information. It analyzes the presence of a term with another term to decide whether they have same or different meaning. The overall architecture of process and data flows involved for the statistical co-occurrence technique is presented in **Figure 1**.

The set of top 10 documents retrieved using the refined query is the pseudo relevance feedback for statistical co-occurrence method. These 10 relevant retrieved documents are indexed in a separate inverted index to extract expansion terms easily. This index is over written every time a query is given to the system. That is because; different relevant documents are retrieved for each different users query. The reformulation process then, based on the statistical co-occurrence technique, selects expansion terms for each query term, and finally selects common expansion terms suitable for the whole query. The reformulation process involves query expansion and term selection sub-processes. Finally, the reformulated query is fed to the original IR system to retrieve re-ranked documents.

On expansion terms selection, co-occurrence frequency values of terms with the query term are analyzed, among the refined retrieved document set index. These found expansion terms for each query term is then saved separately, so that common terms found in all of the query terms can be extracted. Expanding users query using common expanding terms for the whole query has been hypothesized in this research. **Figure 2** shows a java written code for expanding term selection for each query terms.

As shown in **Figure 2**, a term is selected as an expansion term if it co-occurred with a query term even once, among the ten top retrieved documents set. Then these terms are ordered in descending order. The final task is to select those terms which are common for the whole query in order to eliminate polysemous terms ambiguity characteristics.

For statistical co-occurrence technique, the query expansion task is carried out based on the refined document set. For relevance feedback the first k out of the refined documents are selected. That is because; it is assumed

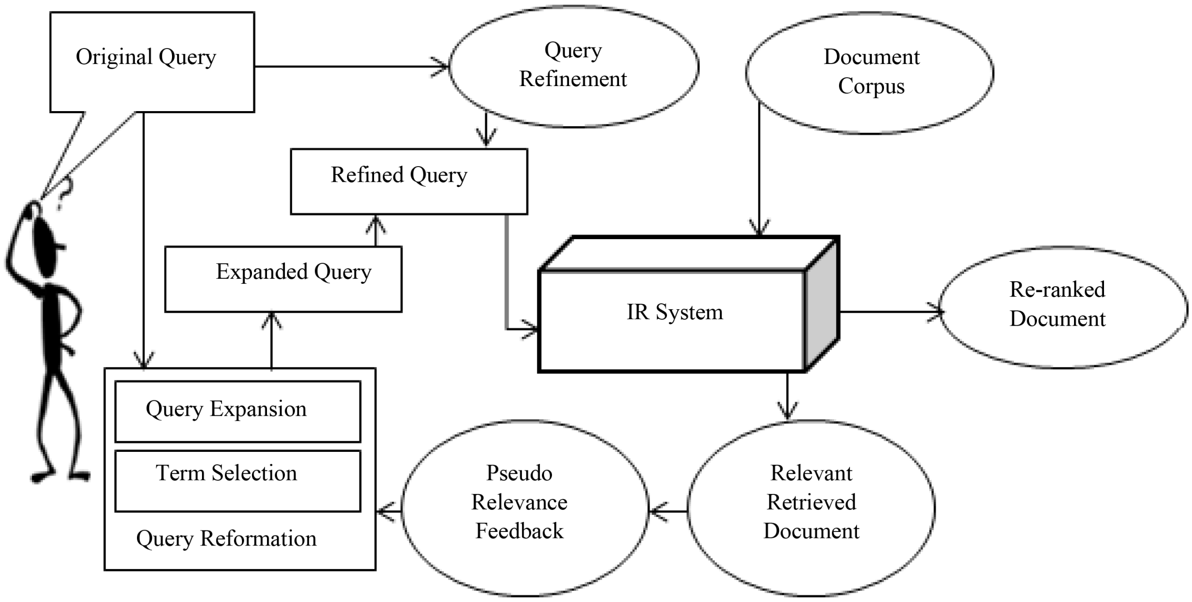

Figure 1. Statistical co-occurrence based query expansion architecture.

```
for(int i = 0; i<termsToSelectFromArray.length;i++)
  if(!termsToSelectFromArray[i].isEmpty())
  int docCount = hitCoutnt(queryTerm + “AND” + termsToSelectFromArray[i]);
  if(docCount >= 1)
  expandingTerms = expandingTerms + termsToSelectFromArray[i] + “-” + docCount
return expandingTerms
```

Figure 2. Statistical co-occurrence based query expansion.

that query refinement process ought to populate more relevant documents at the top retrieved documents set. Hence, these k documents are responsible for generating the expansion words. To overcome the polysemous characteristics of query terms, expansion terms are selected which represent every query terms. This is done by selecting those terms which appear in every expansion terms set of every query term. This technique assumes each query term as a polysemous term or having the potential of possessing polysemous characteristic as per the document collection. The second assumption in this technique is that, query terms other than the polysemous one have the potential to indicate the meaning the user intended. This means if a query term has two different meanings, other query terms aside from that polysemous query term has the potential to select from the two meanings.

For example let there be three query terms, q_1, q_2 and q_3. Let x, y and z be three meanings or collection of expansion terms for q_1, q_2 and q_3 respectively. A certain set of terms C which have common meaning with x, y and z is selected for expansion. **Figure 3** presents a pictorial representation this example given.

Here is a brief discussion of this technique. First the pseudo-relevance feedback of k top ranked relevant retrieved documents are taken and indexed in a separate inverted index. Then all the terms in this index are taken and each term is searched in that same index with each query term. Next the top n terms are selected, which frequently co-occurred with for each query term. The frequency is calculated by simply adding one every time a query term and a potential expansion term are found together in one of the k documents. Given q and t as query term and expansion term respectively and k amount of documents as pseudo-relevance feedback in which Equation (1) is used for frequency calculation.

$$Sim(q,t) = \sum_{j=1}^{k} x = \begin{cases} x = 1, & (q \in d_j \text{ AND } t \in d_j) \\ x = 0, & (q \notin d_j \text{ OR } t \notin d_j) \end{cases} \tag{1}$$

At last common expansion terms are selected from all of the query terms, which are also good expansion terms for the overall query rather than for each query term. Given query terms $q_1, q_2, \cdots, q_n$, index terms $t_1, t_2, \cdots, t_m$ and a constant v. **Figure 4** shows, how technique1 selects expansion terms.

E is a special variable and deliverable of the program which holds the total expansion terms of each query term. The $search$ (q_k, t_j) function compares the co-occurrence value for q_k and t_j, and returns the term t_j if it has a hit count greater than a certain value v, where v is less than k. k is the number of relevant retrieved documents gained from the pseudo-relevance feedback. The process left the variable E with the total expansion terms of the query. E is passed to the function SCT which Selects Common Terms from E and assigns it on itself. The last step returns common expansion terms for the whole query.

Given the query Q and the total expansion terms E **Figure 5** shows the algorithm; how the set of common expansion terms C is selected.

L is the number of query terms. An expansion term that found L times in E is finally taken as a common term.

4. Performance Evaluation

The experimentation phase holds implementation, testing and discussion of the challenges and findings that are recorded for each proposed techniques. Data preparation and selection, testing procedure through empirical testing and threshold selection are also discussed. A prototype has been built using Java NetBean.

In order to evaluate performance of this prototype system we have used recall, precision and F-measure. Determination of greater value for F-measure can be interpreted as an attempt to find the best possible compromise between recall and precision.

Precision and recall are the basic measures used in evaluating relevant document retrieval strategies. Recall is the ratio of the number of relevant records retrieved (*Ra*) to the total number of relevant records in the corpus (*R*), Equation (2). Precision is the ratio of the number of relevant records retrieved (*Ra*) to the total number of

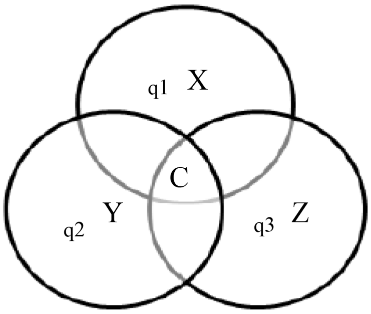

Figure 3. Pictorial representation of common meaning assumption.

```
for k = 0 to n do
     for  j=0 to  doc
    search($q_k$, $t_j$)
          if(hitcount > V) then
      add($t_j$, E)
    E= SCT(E)
    Return E
```

Figure 4. Pseudo code for statistical co-occurrence method.

```
SCT(Q, E)
L = length (Q)
for i=0 to length (E) do
      if count ($E_i$, E) = L
           C = C + $E_i$
return C
```

Figure 5. Pseudo code for common terms selection.

irrelevant and relevant records retrieved (*A*), Equation (3). Unless either recall or precision is needed it is good to show the result in F-measure (*F*) which is calculated from recall and precision. These evaluation metrics are usually expressed as a percentage, Equation (4).

Prototype system centric testing is carried out to evaluate the statistical co-occurrence technique's performance. The statistical co-occurrence's performance is recorded in terms of recall, precision and f-measure as shown on **Table 2**.

$$\text{Recall} = Ra/R \tag{2}$$

$$\text{Precision} = Ra/A \tag{3}$$

F-measure (*F*) is a harmonic mean evaluation measurement, which combines both recall (*Re*) and precision (*Pr*), Equation (4).

$$F = 2 * Re * Pr/Re + Pr \tag{4}$$

The intension of this study is to improve the recall of previous study which has been done based on bi-gram method [26]. Result difference shown in **Table 3** and **Figure 6**. Even though, precision of retrieving relevant

Table 2. Statistical co-occurrence's performance.

Query	REL	Using statistical co-occurrence method				
		RET	RETREL	R	P	F
ቀና እና ቅን ስራዎች	140	652	129	0.19	0.82	0.32
ሰው የተባለ አለቀ	10	28	7	0.7	0.25	0.36
የጌታዬ ባሪያ አለቀ	100	226	72	0.72	0.31	0.44
ዘይት ቀባው	69	121	45	0.65	0.37	0.47
መጥፎ ስራውን ገሰጸ	18	6	6	0.33	1.0	0.5
እጥፍ ሰጠው	26	9	9	0.34	1.0	0.51
አመድ እና ማቅ ለብሶ	35	60	27	0.77	0.45	0.56
አመታት ተቀመጠ	326	244	209	0.64	0.85	0.73
ሲጨልም አንቀላፋ	10	14	10	1	0.71	0.83
			Average	0.73	0.53	0.63

Table 3. Comparing performance of bi-gram based with statistical co-occurrence based.

	Average Recall	Average Precision	Average F-measure
Original query	0.83	0.39	0.53
The Bi-gram method	0.67	0.57	0.61
Statistical co-occurrence	0.73	0.53	0.63

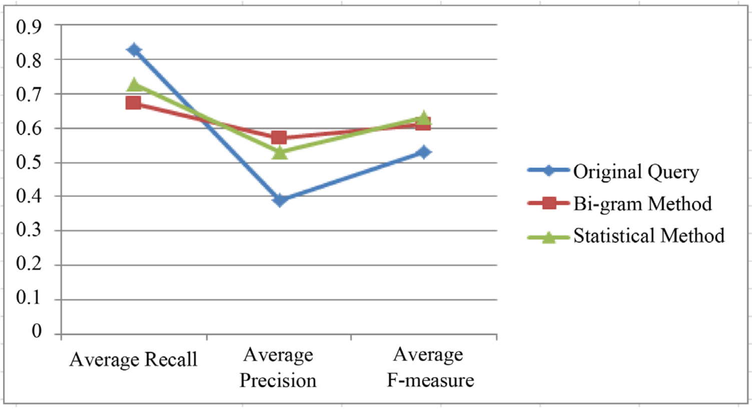

Figure 6. Performance graph of original query, bi-gram and statistical methods.

document down by 4%, the recall of retrieving relevant document improved by 6%. However, F-measure increased by 2%. Accordingly, statistical based information retrieval system performs more than bi-gram based information retrieval system.

5. Conclusion

Based on the frequency, statistical co-occurrence method selects common words which are expected to be synonyms for the polysemous word or words as per the user's intention. Then, these common words are expected to be synonyms for the polysemous word or words as per the user's intention. It is finally concluded that the statistical methods outperformed bi-gram based information retrieval system and scored 6% recall and 2% F-measure. This is an encouraging result to design an applicable search engine for Amharic language information retrieval system. The performance of the system can further be improved by designing hybrid, bi-gram and statistical co-occurrence, based query expansion.

References

[1] Bush, V. (1945) As We May Think. *The Atlantic Monthly*, **176**, 101-108.

[2] Baeza-Yates, R. and Ribeiro-Neto, B. (1999) Modern Information Retrieval. 2nd Edition, Addison-Wesley-Longman Publishers, England.

[3] Spink, A. and Wilson, T.D. (2002) Towards a Theoretical Framework for Information Retrieval (IR) Evaluation in an Information Seeking Context. *Journal of the American Society for Information Science*, **51**, 841-857.

[4] Belkin, N.J. (1993) Interaction with Texts: Information Retrieval as Information Seeking Behavior. Universität Regensburg and Universitätsverlag Konstanz, SchriftenzurInformationswissenschaft Band 12, 55-66.

[5] Wang, X. (2009) Improving Web Search for Difficult Queries. Unpublished paper available at University of Illinois at Urbana-Champaign.

[6] Greenberg, J. (2001) Optimal Query Expansion (QE) Processing Methods with Semantically Encoded Structured Thesauri Terminology. *Journal of the American Society for Information Science and Technology*, **52**, 487-498. http://dx.doi.org/10.1002/asi.1093

[7] Alemayehu, N. (2002) Application of Query Expansion for Amharic Information Retrieval System. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[8] Stubinz, J. and Whighli, S. (1998) Information Retrieval System Design for Very High Effectiveness. Division of Computer Science, Endeavour Research and Development (BVI), 12 October 1998, 1-7.

[9] Ashington, B. (1956) An Automatic System for Retrieval of Electronic Documents. *Third British Colloquium on Electronic Computing*, Manchester, 118-121.

[10] Dagnachew, A. and Worku, A. (1986) Yeamaregna Felitoch. Kuraz Asatami Dereget.

[11] Germann, D.C., Villavicencio, A. and Siqueira M. (2010) An Investigation on Polysemy and Lexical Organization of verbs. *Proceedings of the NAACL HLT* 2010 *First Workshop on Computational Neurolinguistics*, Los Angeles, June 2010, 52-60.

[12] Billhardt, H., Borrajo, D. and Maojo, V. (2002) A Context Vector Model for Information Retrieval. *Journal of American Society for Information Science and Technology*, **53**, 236-249. http://dx.doi.org/10.1002/asi.10032

[13] Kalmanovich, I.G. and Kurland, O. (2009) Cluster-Based Query Expansion. In: *Proceedings of the* 32nd *International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, 646-647. http://dx.doi.org/10.1145/1571941.1572058

[14] Li, L. (2007) A Query Expansion Method Based on Semantic Element. *Eighth ACIS International Conference on Software Engineering*, *Artificial Intelligence*, *Networking*, *and Parallel*/*Distributed Computing*, Qingdao, 30 July-1 August 2007, 587-590. http://dx.doi.org/10.1109/SNPD.2007.249

[15] Bloor, T. (1995) The Ethiopic Writing System: A Profile. *Journal of the Simplified Spelling Society*, **19**, 30-36.

[16] Gizaw, S. (2009) Multiple Pronunciation Model for Amharic Speech Recognition System. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[17] Redwan, H. and Atnafu, S, (1995) Design and Implementation-Algorithms of Amharic Search Engine System for Amharic Web Contents. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[18] Mengistu, T.M. (2007) Design and Implementation of Amharic Search Engine. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[19] Hailemeskel, T. (2003) Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD). M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[20] Saba, A. (2001) The Application of Information Retrieval Techniques to Amharic Documents on the Web. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[21] Gezmu, A.M. (2009) Automatic Thesaurus Construction for Amharic Text Retrieval. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[22] Bethlehem, M.A. (2002) The Application n-Gram-Based Indexing in Amharic Text Retrieval. M.Sc. Thesis, Addis Ababa University, Addis Ababa.

[23] Tilahun, T. (2014) Linguistic Localization of Opinion Mining from Amharic Blogs. *International Journal of Information Technology & Computer Sciences Perspectives*, **3**, 890.

[24] Tilahun, T. and Sharma, D. (2015) Design and Development of E-Governance Model for Service Quality Enhancement. *Journal of Data Analysis and Information Processing*, **3**, 55-62. http://dx.doi.org/10.4236/jdaip.2015.33007

[25] Jing, H and Tzoukermann, E. (1999) Information Retrieval Based on Context Distance and Morphology. In: *Proceedings of the* 22*nd Annual International Conference on Research and Development in Information Retrieval* (*SIGIR* "99), ACM, New York, 90-96. http://dx.doi.org/10.1145/312624.312661

[26] Bruck, A. and Tilahun, T. (2015) Bi-Gram Based Query Expansion Technique for Amharic Information Retrieval System. *IJIEEB*, **7**, 1-7.