Towards Amharic Semantic Search Engine

Fekade Getahun
Department of Computer Science
Addis Ababa University, Ethiopia
Email: fekade.getahun@aau.edu.et

Genet Asefa
Department of Computer Science and IT
Addis Ababa Science and Technology University, Ethiopia
Email: deb_liban@yahoo.com

ABSTRACT
Recently the amount of documents written in Amharic language has been dramatically increasing. Searching such content using localized and regional version of general search engine such as google.com.et returns documents containing search key terms while excluding specific characteristics of Amharic Language.
In this paper, we present the design and implementation of Semantic Search Engine for Amharic documents. The search engine has Crawler, Ontology/Knowledge base, Indexer and Query Processor that consider characteristics of Amharic language. The ontology provides shared concepts Sport. This ontology is built manually by language and sport domain experts and it is used in building semantic indexer, ranker and query engine. In addition, we show how the system facilitate meaning based searching, document relevant and popularity based documents ranking.

Categories and Subject Descriptors
Information Systems: Information retrieval- Document representation - Ontologies

Information Systems: World Wide Web - Web searching and information discovery - Web search engines – Web indexing

General Terms
Algorithms, Performance

Keywords
Search engine, Semantic information retrieval, Semantic indexing, Football domain ontology, Query processor, Document annotation.

1. INTRODUCTION

Noticing the growing number of non-English language web documents, google has provided a localized and region based search engine. In addition, there are a number of research works dedicated to searching non-Amharic documents [1,2,3]. However their approach do not consider the basic characteristics of the Amharic language.

In Amharic lexical variations are very common [4], a word may have more than one meaning also called Polysemy. For example, the word "ለጋ"/lega has two different meanings; i.e. "kicking a ball" or "very young"; different words may have similar meaning. For instance, "መምታት"/memtat and "መለጋት"/melegat have similar meaning "kicking". A search engine should have the capability of dealing with these characteristics of the language.

The number of Information retrieval systems/ researches associated with Amharic language is very limited [5,6,7]. In [5,6], text based searching approach is adopted whereas in [7] latent semantic indexing is used to identify concepts. However, both approaches fail to consider the semantic relationship between concepts and hence the results are restricted to mere occurrence of terms/ concept. Thus both are incapable in responding queries represented indirectly or with paraphrasing.

2. RELATED WORK

The size of news documents written in Amharic language has been increasing dramatically and yet there is a need to share them to the public. To make this happen, we need to have a search engine which is responsive to Amharic language.

The prominent researches conducted in the area of Amharic document retrieval search engines [5,6] utilize a keyword based indexing method which is incapable of representing the semantics of a document. In particular, the works are incapable in addressing two vital properties of Amharic language (Synonymy and Polysemy) as discussed in Section 1.

In [7], the researcher applied Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD) method to construct a semantic indexer. The LSI extracts concepts from a given corpus by looking for words that co-occur frequently without giving emphasis to the relationship between concepts. Therefore, this approach is incapable in responding to queries writing indirectly. In addition, as the size of documents increases, the performance of the indexer degrades.

In order to show the realm of this research let us consider a user query to get all sport documents about Lucy in 2013 African National Cup using the query i.e. "በ2013 ሴቶች የአፍሪካ ዋንጫ ሉሲ የተካፈሉችባቸዉ ግጥሚያዎች". In query the term "ሉሲ"/Lucy has two different meanings: the "Ethiopian Women's National Football Team" and the name of a female player. However, in this particular query, the term "ሉሲ" ("Lucy") is intended to mean Ethiopian Women's National Football Team rather than the female player.

Responding such query using existing approaches i.e. keyword based or LSI has difficulties. The keyword based approach in [5,6]

MEDES '15, October 25-29, 2015, Caraguatatuba, Brazil
 ISBN 978-1-4503-3480-8/15/10…$15.00
http://dx.doi.org/10.1145/2857218.2857235

uses the mere existence of words in the query in the document and does not address the semantic information embedded in the document. This may lead to irrelevant results.

In [7] , LSI that takes a query as a pseudo document and looks for documents very similar to it. However, this approach can't retrieve documents about "ኢትዮጵያ ሴት ብሄራዊ ቡድን"/*Ethiopian Women's national team* which is synonym to "Lucy" or "የሴቶች የአፍሪካ ዋንጫ"/ *African Women's Championship* which is related to Lucy as LSI based concept indexer ignore the relationship between concepts.

Researchers have been working on semantic search engine that a dedicated knowledge bases organized systematically enters of concepts, and associated relationships and documents. The semantic search engine can be domain specific or generic depending on the referenced knowledge base- VikiTron – uses mathematics, chemistry and geography knowledge base, Firmily – business search engine, Symbolab – scientific search engine, Evi – answer engine, Swoogle – Ontology search engine, Falcons – Full semantic search engine). However, these engines are restricted to foreign language and hence do not address basic characteristics of Amharic language.

In this paper, we provide a semantic search engine, AmhS2Eng, which returns list of Amharic documents very similar to the user query either directly or indirectly. In our approach, documents extracted from the web and user queries are annotated with semantic concepts extracted from football domain.

Table 2-1 shows summary of works related to the realm of this research being categorized based on language.

Table 2-1: Summary of related work

Language	Research title	Approach	Drawbacks
Amharic	[5]Amharic Search Engine	Keyword based	- Not concept or meantime based
	[6]Enhanced Amharic Search Engine		
	[7]Amharic Text retrieval	LSI	- Cannot infer new result - is time and memory intensive
English	Localized and region based engine [Google]	Keyword based	- Does take into account characteristic of the Amharic language
	[8] Context based Indexing in Search Engines using Ontology	Users must explicitly add context	- Incapability of handling indirect queries
	[2] Ontology based text indexing and querying for the semantic web	Ontology having uniform relationship is used	- Incapability of handling indirect queries - - Each concept in the ontology has equal weight - Dedicated to English language
	[3] An Ontology-Based Retrieval System Using Semantic Indexing	Knowledge based approach	- The similarity between concepts is not considered - Language specific

3. Architecture of Amharic Semantic Search Engine

Figure 1 shows the architecture of the proposed semantic search engine that takes into account specific features of the Amharic language. It is composed of four major components: Crawler, Query Interface, Ontology, Concept Indexer, and Query Processor.

The Ontology is built manually with the help of domain and language experts. It contains concepts, relationship between concepts and instances of each concept and similarity between knowledge entries.

The crawler extracts URLs and content of Amharic document and store them in a document repository.

The indexing component is responsible to annotate the crawled document with semantic information (Concept, Instance, Words) and associate semantic information with weight. This component is responsible to identify specific characteristic of the Amharic language – such as stemming.

The query processor responsible to annotate the original user query with semantic information, conduct document retrieval, rank the resulting documents based on relevance as explained in the next sub-sections.

3.1 Ontology management

The ontology development process is composed of three main activities:

1. Building the high-level ontological schema
2. Populating the ontology with instance value
3. Computing similarity

a) Building the ontology schema

In this work, ontology O is defined as collection of related concepts and it is represented as follows:

$$O = \{C, E, R\}$$

Where

C: Collection of concepts each represent words having similar words similar to SynSet in WordNet [9]

E: Collection of edges that connect related concepts and instances to concepts relationship

R: Collection of semantic relations such as IsK, partOf,

The ontology is built following Uschold and King's stated in [10]. The ontology provides a conceptual knowledge model of football domain which is used later in document and query annotation, and document ranking. Glossaries of terms are identified and further validated by domain experts to provide the list of domain terms

with their full descriptions and synonyms. Figure 2 shows extract of the domain ontology depicting concepts and relationship between concepts.

In this work, the ontology is represented in triplet data format having a subject, predicate and an object (i.e. value) and stored in Microsoft SQL server database.

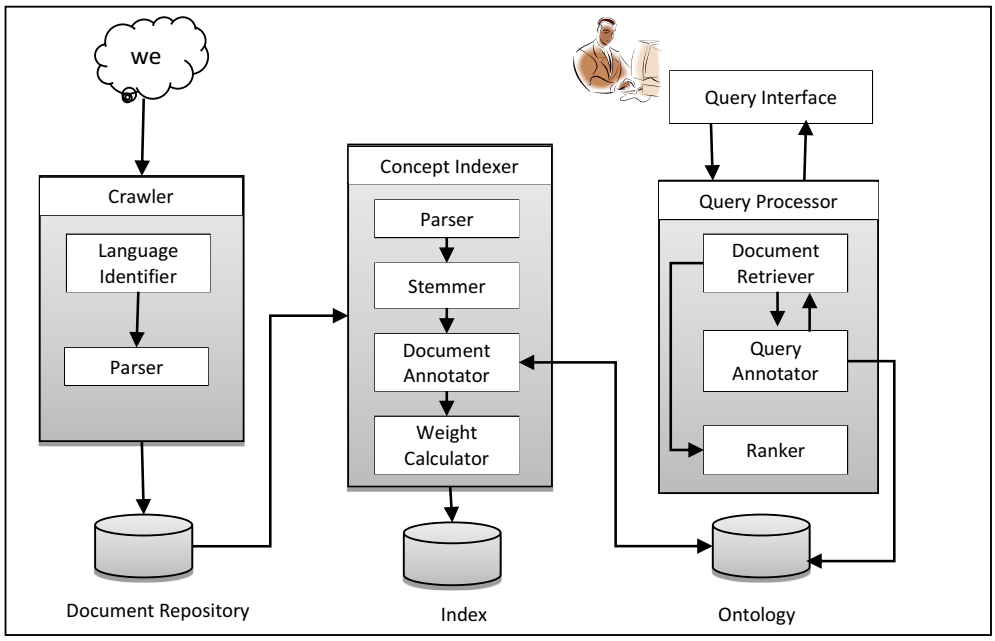

Figure 1: General Framework of AmhS2Eng

b) Populating the ontology

The ontology is populated with instances or terms of football domain. Sport domain experts identify and approve terms to be populated into the defined ontology. The ontology is populated and accessed using jenatoolkit[1] and SPARQL[2] queries.

c) Calculating semantic similarity

Semantic query processing demands the use of concepts or similar concepts that encompass each query term. One of the known approaches in measuring the semantic similarity between ontological terms or instances in the ontology depends on the amount of information common to concepts. In this work, the similarity between concepts is measured using the popular distance based approach proposed by Wu and palmer [11] and denoted as:

$$SIM(o_1, o_2) = \frac{2D}{2D + len(o_1, C) + len(o_2, C)}$$

Where

C: the least common ancestor of o_1 and o_2

D: the distance between the root element and C

$len(o_1, C)$ returns the distance between o1 and C,

3.2 Crawler

The crawler component is multithreaded and is responsible to identify Amharic web pages and associated content. Generally it performs:

- Extracts URL from the given page,
- Fetches a web page pointed by URL,
- Extracts set of links from the fetched web page,
- Ignores the extracted link that has been already fetched

Generally, the crawler uses http protocol and multiple threads that process concurrently to fetch pages. The fetched pages are then passed to Amharic language identification and stores in Document repository.

3.3 Concept Indexer

Concept indexer is responsible to represent concepts extracted from the unstructured Amharic documents and associates each concept with weight that shows its importance. It has the Document Annotator and Indexer as main components as presented in the next sub-sections.

[1] https://jena.apache.org/ - A Java system for RDF manipulation, parsing RDF/XML and N3, persistent storage with SQL

[2] www.w3.org/TR/rdf-sparql-query/ - SPARQL used to express queries and data stored natively as RDF.

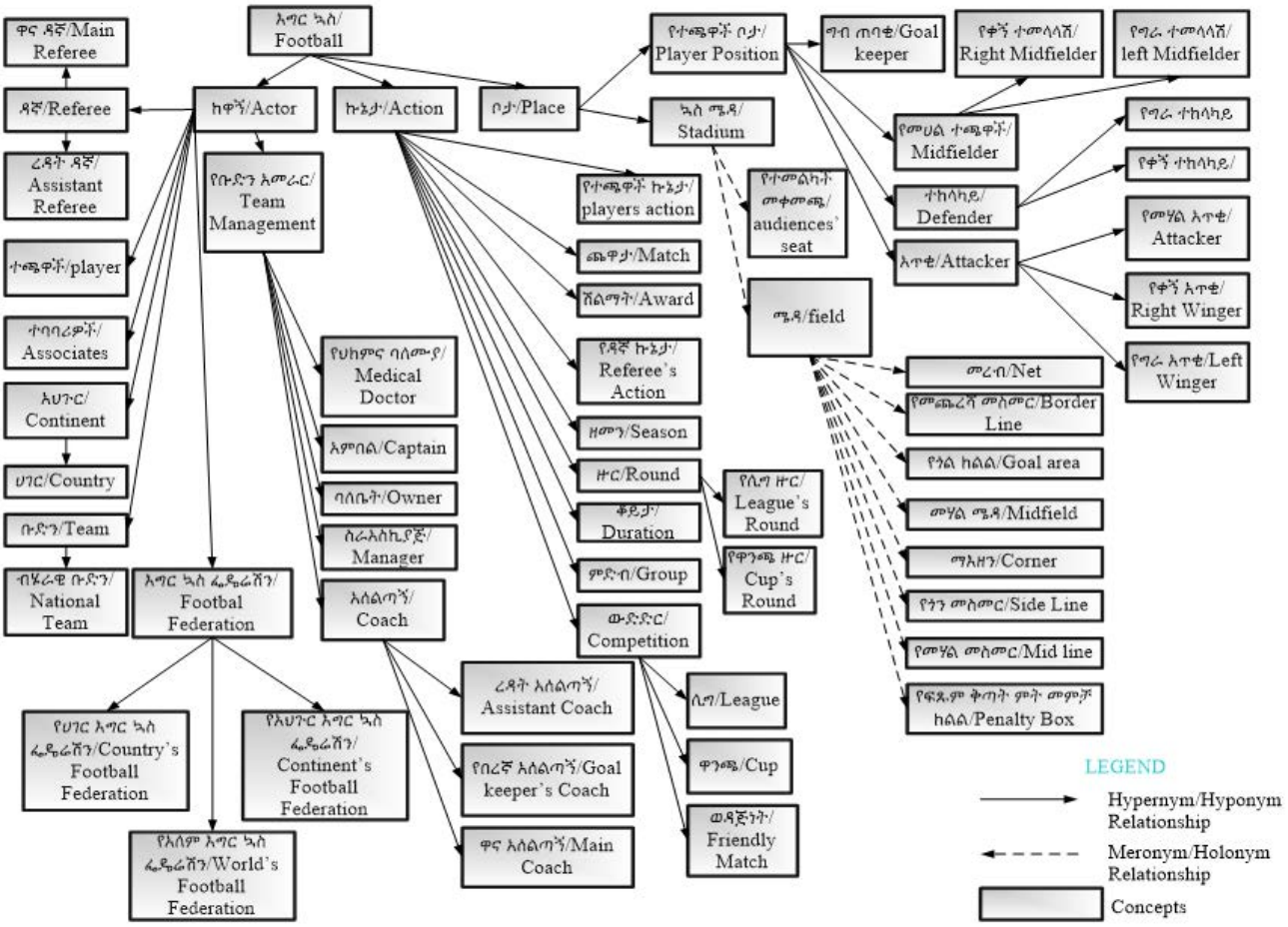

Figure 2: The concept taxonomy

3.3.1 Document Annotator (DA)

DA accepts a document as input, segment it into words, and map each word to semantic concept stored in the ontology.

In this work, word to concept mapping is done using the combination of predefine rules and looking for the best concept in the ontology approaches.

a) Rule based concept identification

Rules are regular expressions/ patterns, having numerals and literal values, defined by knowledge experts to represent concepts. Table 3-1 shows sample rules that extract the notion of "ነጥብ" / point, "ዉጤት"/ result, "ዙር" / round, "ደረጃ" / rank from textual corpus.

Table 3-1: Sample regular expressions for concept extraction

Concepts	Pattern
"ነጥብ" // point	"[0-9]+( )*ጎል"// //"[0-9]+()*Goal"
"ዉጤት" // result	"[0-9]+( )*ለ( )*[0-9]+" //"[0-9]+( )*to( )*[0-9]+"
"ዙር"//round	[0-9]+ኛ(ሳምንት|ዙር) //[0-9]th week*
"ደረጃ" // rank	[0-9]+( )*ነጥብ//[0-9]+( )*point

b) knowledge based concept identification

Even if the rule based approach is capable in identifying concepts, it is difficult to formulate patterns that detect all concepts of a given domain. For instance, considering the concept "ተጫዋች"/*players*, it is not easy to have a pattern that represents its instances which are person name (i.e. "አዳነ ግርማ", "ሳልሃዲን ሰይድ", "ጌታነህ ከበደ" and etc). Thus, consulting a knowledge base that contains such information is crucial.

Let us consider the following example to demonstrate the two approaches.

Example1. Concept identification

Consider the title of a sport news article

"የ16ኛው ሳምንት የኢትዮጵያ ፕሪሚየር ሊግን የደደቢት እና የሉሲ አጥቂ የሆነዉ ጌታነህ ከበደ በ13 ጎል እየመራ ነዉ።"

The news is about "Dedebit's and Lucy's striker Getaneh Kebde who is leading the 16th week Ethiopian primer league with 13 points"

The content is preprocessed, stemmed and stored in the Document repository. The indexer uses the stemmed version of the article:

“16ኛ ሳምንት ኢትዮጵይ ፕሪሚየር ሊግ ደደቢት ሉስ አጥቅ ሆን ጌታን ከበድ 13 ጎል መር ነው”

Applying the two concept identification approaches on the above stemmed text, we get the following results:

- Using rules: the following concepts are identified:

  [16ኛ ሳምንት]<ዙር>//[16th Week]<Round>

  [13 ጎል]<ነጥብ>//[13 Goal]<Point>

- Using knowledge base the following concepts are extracted:

  [ሉስ]<አሰልጣኝ>//[Lucy]<Coach>

  [ሉስ=ኢትዮጵይ ሴት ብሄራው ቡድን]<ብሄራዊ ቡድን> //[lucy=Ethiopian Women National Team]<National Team>

  [ሉስ=ብርቱክ በቀል]<ተጫዋች>//[Lucy=Birtuan Bekele]< Player>

  [ደደቢት]<ቡድን>//[Dedebit]<Club>

  [ኢትዮጵይ]<ሀገር>//[Ethiopia]<Country>

  [ፕሪሚየር ሊግ]<ሊግ>//[Premiere League]<League>

  [ጌታን ከበድ]<ተጫዋች>//[Getaneh Kebede] <Player>

Notice that all terms have exactly one meaning except for “ሉሲ”/“Lucy” which has 3 different interpretations (i.e. An instance of the concept *አሰልጣኝ/ Coach*, a nick name (synonym) for the instance *ብርቱክ በቀል/ Birtukan Bekele*, and a synonym for the instance *ኢትዮጵይ ሴት ብሄራው ቡድን/ Ethiopian Women National Team*). So, the sentence will have 3 different forms of annotations as;

1. [16ኛ ሳምንት/*16th Week*]<ዙር/*Round*> [ኢትዮጵይ/ *Ethiopia*] <ሀገር/ *Country* > [ፕሪሚየር ሊግ/ *Premiere League*] <ሊግ/ *League*> [ደደቢት/ *Dedebit*] <ቡድን> **[ሉስ/ *Lucy*] <አሰልጣኝ/ *Coach* >** አጥቅ ሆን/ *striker* [ጌታን ከበድ/ *Getaneh Kebede*] <ተጫዋች/ *Player* > [13 ጎል/Goal] <ነጥብ/ Point>መር ነው/ *leads*
2. [16ኛ ሳምንት/ *16th Week*] <ዙር/ *Round*> [ኢትዮጵይ/ *Ethiopia*] <ሀገር/ *Country* > [ፕሪሚየር ሊግ/ *Premiere League*] <ሊግ/ *League*> [ደደቢት/ *Dedebit*] <ቡድን> **[ሉስ / *Lucy*= ኢትዮጵይ ሴት ብሄራውቡድን/ *Ethiopian Women National Team*] <ብሄራዊ ቡድን/ *National Team*>** አጥቅ ሆን/ *striker* [ጌታን ከበድ/ *Getaneh Kebede*] <ተጫዋች/ *Player* > [13 ጎል/ Goal] <ነጥብ/ Point> መር ነው/ *leads*
3. [16ኛ ሳምንት/ *16th Week*] <ዙር/ *Round* > [ኢትዮጵይ/ *Ethiopia*] <ሀገር/*Country* > [ፕሪሚየር ሊግ/ *Premiere League*] <ሊግ/ *League* > [ደደቢት/ *Dedebit*] <ቡድን> **[ሉስ / *Lucy*= ብርቱክ በቀል/ *Birtuan Bekele*] <ተጫዋች/** *Player*> አጥቅ ሆን/ *striker* [ጌታን ከበድ/ *Getaneh Kebede*] <ተጫዋች/ *Player* > [13 ጎል/ Goal] <ነጥብ/ Point> መር ነው/ *leads*

3.3.2 Weighting

For each term, instance and concept extracted from the document weight determining its importance is computed. In this work, modified version of TF-IDF is used to compute weight and formalized as follows.

$$\boldsymbol{TFIDF_{t,d} = CF_{t,d} \times log(IDF_t)}$$

where:

- $\boldsymbol{CF_{t,d}}$ is the number of times a term t occurs in the document D as is or with its instances *i in t* and it is computed as follows:

$$\boldsymbol{CF_{t,d} = Count(t) + \sum_{i \epsilon t} Count(i)}$$

- $\mathbf{IDF_t}$ is the inverse document frequency – the number of documents divided by the number of documents in which the term *t* occurs.

3.3.3 Ranking

Ranking is the process of determining which annotated text has more sense than the other. Each annotated text is ranked according to the interrelationship it has with concepts/instances. The correlation among instances, concepts, and instances and concepts are defined taking into account the level of similarity denoted as Sim_{Inst}, Sim_{Conc} and $Sim_{InstCon}$ respectively and formalized as follows:

$$\boldsymbol{sim_{Inst}(T) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n} sim(I_i, I_k)}{n^2}}$$

Where:

- *I is the set of instances identified from the text T*
- *n is the number of instances in I.*
- *Sim(I_i, I_k) is the similarity between Instances I_i and I_k* in *I*

$$\boldsymbol{sim_{Conc}(T) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{n} sim(C_i, C_k)}{n^2}}$$

Where:

- *C is the set of concepts identified from the text T*
- *n is the number of concepts in C.*
- *Sim(C_i, C_k) is the similarity between concepts C_i and C_k* in C.

$$\boldsymbol{sim_{InstConc}(T) = \frac{\sum_{i=1}^{n}\sum_{k=1}^{m} sim(C_i, I_k)}{n * m}}$$

Where:

- ***C** is the set of concepts identified from the text T*
- ***I** is the set of instances identified from the text T*
- ***M** and **N*** are *the number of instances in **I** and the number of concepts in **C** respectively.*
- *Sim(C_i, I_k) is the similarity between concept C_i and Instance I_k stored in the Ontology.*

-

$$\mathbf{Rank}(\boldsymbol{T}) = \frac{\boldsymbol{Sim_{inst}(T) + Sim_{Cons}(T) + Sim_{InstConc}(T)}}{\mathbf{3}}$$

Total similarity is computed for each annotated text and the texts will be ranked according to their similarity values. The one

ranked at the top will be returned as a result of the document annotation process.

3.4 Query Processor

The Query Processor accepts user query, annotate it using semantic knowledge, retrieve documents and rank them.

a) Query annotator (QA)

QA is used to annotate a user query using predefined rules and concepts in the ontology. Give a specific user query Q, QA annotates it with several semantically complete and meaningful queries. QA component works using the same principles as DA as the user query Q is Query document. The annotated queries are ranked and the top most important query is evaluated automatically by the Query processor component; and the remaining queries will be presented to the user.

b) Document Retrieval (DR)

The role of the DR is to find list relevant documents to the semantically annotated user query. Algorithm 1 demonstrates the detail of this component.

The annotated query, returned by QA, has three category of terms: *instances*, *concepts* and *words* which are parsed in Line 2, 3 and 4 respectively. Instances and concepts have references in the knowledge base, KB sport ontology.

The DR looks for documents containing instances, concepts and words from the index along with weight as shown in Lines 5-19.

	Algorithm 1: Document Retrieval
	Input:
	oQ: Query // Original user query
	KB: Ontology
	Intermediate:
	D: Integer // Knowledge depth
	Q : query // Annotated Query
	InSet: Set // set of instances
	ConSet: Set // Concept Set
	Words: Set // Set of Words
	Output:
	DocSet: Set // Set of documents
	Begin
1.	Q:= QA (oQ); // Semantically annotate Query oQ
2.	InSet:= ParseInstances(Q);
3.	ConSet:= ParseConcepts(Q);
4.	Words:= ParseUnAnnotatedWords(Q);
5.	Foreach wd in (InSet U ConSet U Words)
6.	For i=0 to D
7.	If (i==0)
8.	W:=1;
9.	TFIDF:=TFIDF(Wd);
10.	DocSet:=Documents (Wd);
11.	Else
12.	W:= $Sim_{wu\&Palmer}$(Wd, Ci);
13.	TFIDF := TFIDF(Ci);
14.	DocSet:=Documents (Ci);
15.	END IF
16.	Rank := Rank + W*TFIDF
17.	DocSetR := DocSetR.Add(DocSet, Rank)
18.	Next
19.	Next
20.	Return DocSetR
21.	End

For each instance or concept,its weight that depends on the semantic similarity it has with related concept and TFIDF of the related concept as shown in line 6-16 is used.

4. EXPERIMENT AND EVALUATION

In order to validate the performance of our approach we have developed a semantic search engine, AmhS2Eng; and we compared its result against the result of classical keyword-based Amharic search engine [5]. The evaluation is using the popular information retrieval base relevance measures Recall, Precision, and F-value [12] are used. Both AmhS2Eng and the classical IR results are compared against the list of documents returned by experts. In most researches, expert judgments are considered to be correct and absolute. However, in this research, we argue that the initial expert judgments are very limited and we advised experts to refine their initial query result judgment seeing the result of the two search engine (following the relevance feedback approach in IR).

In the refinement process experts go through documents retrieved by both systems to determine whether the documents are relevant to the posed queries or not. Due to the refinement process the proposed search engine captured 8 relevant documents for 6 queries and the classical search engine returned 5 documents for 3 queries. This indicates that the relevance judgment made by experts is limited and also the proposed system has returned more relevant documents than the classical IR.

The Precision, Recall and F-values computed for 25 different queries before and after the refinement and are presented in Figure 3, Figure 4 and Figure 5 respectively.

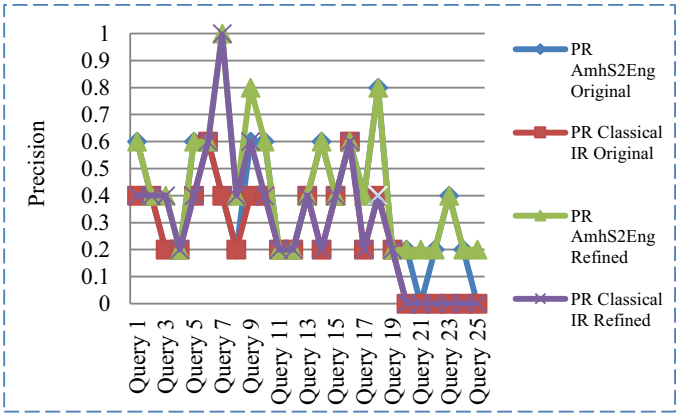

Figure 3: Precision graph

Considering Precision and Recall graphs shown in Figure 3 and Figure 4 respectively the precision has increased for "Query 3", "Query 7", "Query 8", "Query 9", "Query 21", and "Query 25" and the Recall increased for "Query 21" and "Query 25" for the case of refined expert judgment. For example, when we look at "Query 25", initially it doesn't have any relevant document. However, the AmhS2Eng returns one relevant document and the precision for this particular query increased from 0 to 0.2 and recall from 0 to 1. Like AmhS2Eng, the classical IR system captured missed documents for the direct queries ("Query 3", "Query 7", and "Query 9") as shown in **Error! Reference source not found.**. Thus, the precision for these 3 queries increased for the classical IR as well. However, no change has been observed on recall for any of the queries even though expert judgment refinement has been done. Besides, the classical IR did not capture any of the missed documents for the indirect queries. This shows that, unlike the classical IR, the proposed search engine method is based on semantics rather than simple key words.

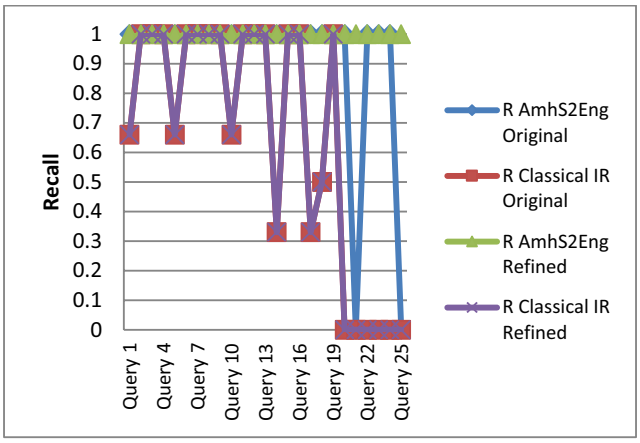

Figure 4. Recall value graph

Figure 4 shows that AmhS2Eng has a maximum recall value, 1, for all the queries even though it has returned some irrelevant documents. On the contrary, the classical IR system has missed some of the relevant documents of some queries and missed all for "Query 20" up to "Query 25". For example, "Query 23" – "የባፋናባፋናአሰልጣኝ"/*the coach of bafana bafana* has 2 relevant documents according to the refined relevance information but the classical IR returned none of them as the term "ባፋናባፋና"/*bafana bafana* does not exist in the documents. In contrast, AmhS2Eng captured the relevant document as "ባፋናባፋና"/*bafana bafana* has the same meaning as the term "የደቡብአፍሪካብሄራዊቡድን"/*South African national team*.

In order to show a better view of the capability of both systems F-value is computed and graph is shown in Figure 5. The F-values are computed with the intention of evaluating the result of the systems independent of recall and precision values. The F-values are computed for both systems and the refined expert judgment as shown in Figure 5. In addition, Figure 5 shows that F- values of AmhS2Eng for 14 queries are greater than that of the classical IR and the same for the remaining queries. This indicates that for more than half of the queries, the result of AmhS2Eng is much closer to expert judgment compared to the classical IR. Besides, the f-values computed for AmhS2Eng using the refined relevance information are greater than the values computed using the original/initial relevance information for some of the queries i.e. the "F-value AmhS2Eng Refined" line is much closer to the "Refined expert judgment" when compared to the "F-value AmhS2Eng Original".

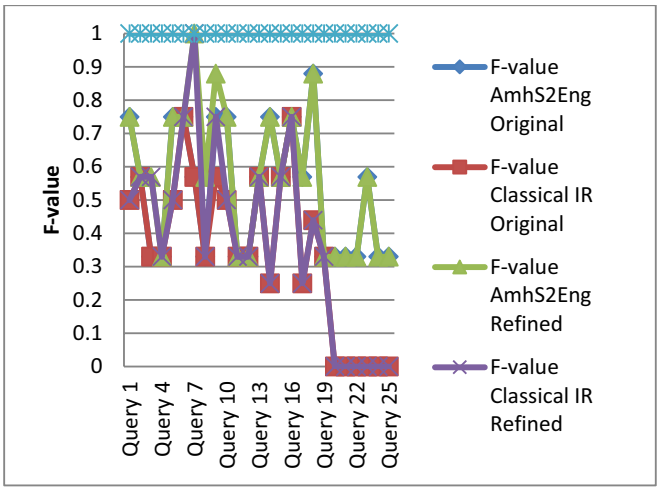

Figure 5. F-measure graph

For the 25 user queries and set of documents retrieved using each system, average recall, precision, and F-measure values are computed and presented graphically in Figure 6. The average values are calculated over the number of queries.

As it is illustrated in Figure 6, the average values for all the three evaluation techniques AmhS2Eng has a better f-value than the respective values of the classical IR system. When we put the average values in a percentage, the proposed system has 100% recall, 43.2% precision, and 53.68% F-measure whereas classical IR has 64.56% recall, 29.6% precision, and 36.04% F-measure.

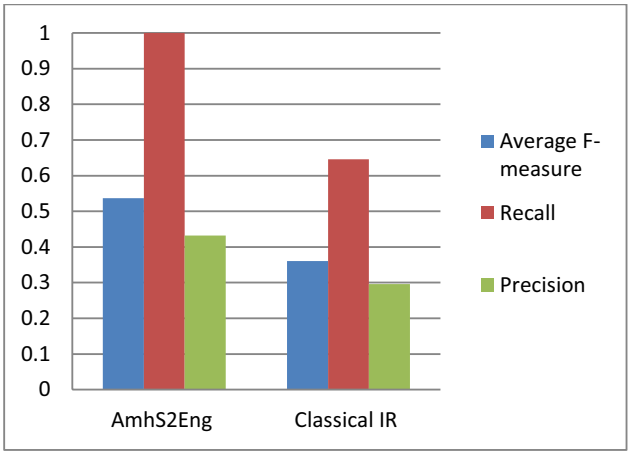

Figure 6. A comparison of the two systems

5. Discussion

Even though the recall of AmhS2Eng is 100%, the precision is 43.2% which is less than 50%. This happened due to the nature of the documents i.e. the content of almost all news documents are different and yet published by the same publisher, ERTA (Ethiopian Radio and Television Agency). Thus, the extent to which these news items will be similar is very rare. Therefore, for most of the queries the number of relevant documents in the

refined relevance information set is either 1 or 2. However, for query, “Query 7”, 5 relevant documents are returned as the top 5 documents returned by AmhS2Eng and the classical IR systems are taken into consideration. As a result, for queries that have limited number of relevant documents, 1 or 2, the precision goes down because of the false negative 4 or 3 irrelevant documents. This indicates that if the document collection was somehow different, the precision would be much better.

In addition, considering Figure 4, the recall values for both AmhS2Eng and the classical IR goes up and down when we go from the first query to the second query and up to the last. This happened because of the nature of the queries i.e. all the queries are completely different from one another. The recall would have increased linearly if the queries are related to one another i.e. if “Query 2” contains “Query 1” and “Query 3” contains “Query 2” and the same with the rest of the queries.

As it is mentioned in Section 4, the harmonic F-measure technique which has equal weight for recall and precision was used to evaluate both AmhS2Engand the classical IR system. The harmonic F-measure was chosen over the balanced F-measure because we cannot tell which one recall or precision is very important to users. In fact, it is likely that many people may prefer recall to precision. In such case, the balanced F-measure which takes recall as twice as precision is used for evaluation. If this particular balanced F-measure had been used for this study instead of the harmonic F-measure technique, the averaged F-measure would have been much higher than the one we have in this work.

6. CONCLUSION

In this paper, a semantic search engine based on domain ontology for Amharic text documents is presented. The goal of this study is to explore the advantages of ontology to build a semantic search engine. The concepts and individuals in the document collections are identified using sport knowledge based constructed manually by domain experts. All these concepts and individuals are used as index terms to represent documents.

The proposed search engine is tested based on the relevance information provided by domain experts. Besides, the proposed system is compared with the classical IR system developed by Tessema [5]. In order to test these two systems, 138 football news articles and 25 queries are used. The precision, recall, and F-measure techniques are used to evaluate the performance of the systems. AmhS2Eng has better average recall, precision, and F-measure values compared to the classical IR system.

7. ACKNOWLEDGMENTS

This research is sponsored by the Ministry of Information Communication and Technology, MCIT, Ethiopia. Our deepest gratitude goes to Mr. Tesfaye, Health and Physical Education expert in Ministry of Education, for facilitating access to resources from Ethiopian Football Federation and for his very helpful expert advises on the main parts of the work. Our appreciation also goes to Mr. Getachew Abebe and Zewudinesh Yirdaw, staff members of Ethiopian Football Federation, for spending their valuable time to respond to our questions and for providing the necessary data.

REFERENCES

[1] Parul Gupta and A.Sharma, "Context based Indexing in Search Engines using Ontology," *International Journal of Computer Applications*, vol. 1, no. 14, pp. 53-56, 2010.

[2] Jacob Kohler, Stephan Philippi, Michael Specht, and Alexander Ruegg, "Ontology based text indexing and querying for the semantic web," *Knowledge Based Systems - KBS*, vol. 19, no. 8, pp. 744-754, 2006.

[3] Soner Kara, "An ontology-based retrieval system using semantic indexing," 2010.

[4] Wolf Leslau, *Amharic Reference Grammar*.: ERIC Clearinghouse for Linguistics, Center for Applied Linguistics,1717 Massachusetts Ave. N.W., Washington, D.C. 20036, 1969.

[5] Tessema Mindaye Mengistu and Solomon Atnafu, "Design and Implementation of Amharic Search Engine," in *International IEEE Conference on Signal-Image Technologies and Internet-Based System - SITIS*, 2009, pp. 318 - 325.

[6] Hassen Redwan, Tessema Mindaye, and Solomon Atnafu, "Enhanced Design of Amharic Search Engine (An Amharic Search Engine with Alias and Multi-character Set Support)," in *AFRICON '09*, Addis Ababa, 2009, pp. 1-6.

[7] Tewodros Hailemeskel Gebermariam, "Amharic Text Retrieval: An Experiment Using Latent Semantic Indexing (LSI) with Singular Value Decomposition (SVD)," Masters Thesis, School of Information Studies for Africa, Addis Ababa University, Addis Ababa, Ethiopia, Unpublished 2003.

[8] Parul Gupta and A.Sharma, "Context based Indexing in Search Engines using Ontology," *International Journal of Computer Applications* , vol. 1, no. 14, pp. 53-56, 2010.

[9] George A. Miller, "WordNet: a lexical database for English.," *Commun. ACM* , vol. 38, no. 11, pp. 39-41, November 1995.

[10] Oscar Corcho, Mariano Fernandez-Lopez, and Asuncion Gomez-Perez, "Methodologies, tools and languages for building ontologies. Where is their meeting point?," *Data & Knowledge Engineering*, vol. 46, no. 1, pp. 41-64, 2003.

[11] Palmer and Wu., "Verbs Semantics and Lexical Selection," in *ACL.*, New Mexico, 1994., p. p. 133.

[12] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

.