2AIRTC: The Amharic Adhoc Information Retrieval Test Collection

Tilahun Yeshambel, Josiane Mothe, Yaregal Assabie

2AIRTC: The Amharic Adhoc Information Retrieval Test Collection

Tilahun Yeshambel[1], Josiane Mothe[2],Yaregal Assabie[3]

[1] Addis Ababa University, IT PhD program, Ethiopia
tilahun.yeshambel@gmail.com
[2] INSPE, Univ. de Toulouse, IRIT, UMR5505 CNRS, France
josiane.mothe@irit.fr
[3] Addis Ababa University, College of Natural and Computational Science, Ethiopia
yaregal.assabie@yahoo.com

Abstract. Evaluation is highly important for designing, developing, and maintaining information retrieval (IR) systems. The IR community has developed shared tasks where evaluation framework, evaluation measures and test collections have been developed for different languages. Although Amharic is the official language of Ethiopia currently having an estimated population of over 110 million, it is one of the under-resourced languages and there is not yet Amharic adhoc IR test collection. In this paper, we promote the monolingual Amharic IR test collection that we built for the IR community. Following the framework of Cranfield project and TREC, the collection that we named 2AIRTC consists of 12,583 documents, 240 topics and the corresponding relevance judgments.

Keywords: Information Retrieval, Amharic test collection, Adhoc retrieval, Evaluation, Data collection, Corpus, Under-resourced language.

1. Introduction

With the Internet technology, and the growth of online text information and globalization, information retrieval (IR) has gained more attention especially in commonly used languages. Both the research community and the industry have been very active in this field for more than 50 years [Sanderson and Croft, 2012]. IR also has an old history of evaluation. After the early framework in the Crandfield project, TREC has standardized adhoc retrieval evaluation [Buckley & Voorhees, 2005]. Performance evaluation of a system is indeed very crucial for scientific progress [Ferro, 2014]. There are many system evaluation criteria such as effectiveness, efficiency, usability, accessibility, utility, portability, and maintainability to complete the task. One of the major focuses of IR research evaluation is to measure the IR system effectiveness. In adhoc retrieval, where the task for the system is to retrieve the

relevant documents for a given query, effectiveness looks at the ability of the system to retrieve only relevant documents for a given user's query. This implies an evaluation framework consisting of a test collection as well as metrics. The standard adhoc retrieval test collection consists of three components: a corpus of documents to be searched in, a set of users' information needs or topics, and the associated relevance judgments indicating which documents are relevant for which topics. Test collections facilitate reproducibility of results and meaningful effectiveness comparison among different retrieval techniques.

In adhoc retrieval evaluation, a corpus is usually a systematic collection of naturally occurring fixed size documents in machine readable form. Building text collections for adhoc IR is a common task although it is resource demanding. A large number of shared tasks rely on such collections. Some of the well known text collections are Cranfield project [Cleverdon, 1959], Text REtrieval Conference[1] (TREC) and more specifically TREC adhoc [Harman, 1995], Cross-Language Evaluation Forum[2] (CLEF) datasets [Ferro, 2014], and NACSIS test Collection for Information Retrieval[3] (NTCIR) [Kando et al., 1999]. Indeed, the IR international conferences such as TREC, CLEF, NTCIR, INEX[4,] and FIRE[5] are held based on their own test collections.

One of the common techniques to build a test collection for adhoc retrieval is pooling, where the document pool to be judged by humans is built by putting together the top N retrieved results from a set of systems [Soboroff, 2007]. In this technique, documents outside the pool are considered as non-relevant. Pooling is a standard technique to create relevance judgment in TREC, CLEF, and NTCIR test collections. The second technique is exhaustive relevance judgment where every and each document in a dataset is judged according to each query [Kagolovsky & Moehr, 2003]. Cranfield and CACM test collections are built using exhaustive relevance judgment. A third way of preparing relevance judgment is crowdsourcing in which huge and heterogeneous mass of potential workers are assigned to a relevance task in the form of an open call through Internet [Samimi & Ravana, 2014]. Whatever the way they are built, test collections are one of the pillars for testing and comparison of retrieval system performance.

While there are a lot of existing collections in different languages such as English, French, Arabic, and Asian languages, to our best knowledge, there is no test collection for Amharic IR. The lack of reference collection is a major impediment to the development of Amharic IR as well as natural language processing (NLP) tools. Indeed, Amharic is one of the under-resourced languages: computational resources such as training and test data, electronic bilingual dictionary, stemmer, tagger, morphological analyzer, etc. do not exist or the existing ones are not fully functional and the number of studies reported on Amharic

[1] Text REtrieval Conference http://trec.nist.gov/

[2] Cross Language Evaluation Forum http://www.clef-initiative.eu/

[3] NII Test Collection for Information Retrieval http://research.nii.ac.jp/ntcir/

[4] Initiative for the Evaluation of XML Retrieval http://inex.mmci.uni-saarland.de.

[5] Forum for Information Retrieval Evaluation http://fire.irsi.res.in

is considerably limited compared to what is done for other languages such as English. Tools and resources that have been developed for other languages cannot be directly applied to Amharic because of its very specificities. We do believe that an Amharic reference collection would help to carry out more research and development works on Amharic IR and NLP. This paper describes the construction process and characteristics of the resulting Amharic IR test collection we deliver to the IR community.

2. Amharic Language

Amharic is an Afro-Asiatic language belonging to the South-west Semitic group [Hetzron, 1972; Argaw et al., 2005]. It is an official language of the government of Ethiopia currently having an estimated population of over 110 million [countrymeters, 2020]. It is the second-most commonly spoken Semitic language in the world next to Arabic [Abate & Assabie, 2014]. Although many languages are spoken in Ethiopia, Amharic is the lingua franca and the most literary language serving as a medium of instruction in the education system of the country for a long period. It uses Ethiopic alphabet for writing and has 34 base characters along with modifications on the respective base characters. The alphabet is conveniently written in a tabular format of seven columns. The first column represents the basic form with vowel ä /ä/ and the other six orders represent modifications with vowels in the order of ኡ/u/, ኢ/i/, ኣ/a/, ኤ/e/, እ/ə/, and ኦ /o/. For example, the base character መ/mä/ has the following modifications: ሙ/mu/, ሚ/mi/, ማ/ma/, ሜ/me/, ም/mə/, and ሞ/mo/. The language also uses punctuation marks such as ። (full stop), ፣ (comma), ፤ (semicolon), ፥ (colon), etc. It also adopts some other punctuation marks such as question and exclamation marks from English.

Amharic is known to have a complex morphology. A large number of words can be formed from a base form and word formation is complex involving affixation, reduplication, and Semitic stem inter-digitation. Thousands of surface words can be generated from an Amharic root and its stems by changing the shape of alphabets in a stem or root, and by adding affixes on stems [Abate & Assabie, 2014]. For example, the verb አሰበሸ/ässäbbärä/ is derived from the verbal stem ሰበር-/säbbär-/ which is itself derived from the verbal root ስ-ብ-ር /s-b-r/. Furthermore, a verb can be marked for a combination of person, case, gender, number, tense, aspect, and mood. Accordingly, the following verbs can be generated from the verbal stem ሰበር-/säbbär-/: ሰበርኩ /*säbbärku* 'I broke'/, ሰበርኩህ/*säbbärkuh* 'I broke you'/, ሰበርን/*säbbärn* 'we broke'/, ተሰበርኩ/*täsäbbärku* 'I was broken'/, ሰበረች/*säbbäräc* 'she broke'/, አይሰብርም/*ʔäysäbrəm* 'he will not break'/, etc. Thousands of words can be generated from a single verbal root making analysis, annotation and tagging of Amharic text a non-trivial task. This level of morphological complexity has significantly contributed to the difficulty of producing linguistic resources for Amharic.

3. Related Work

3.1 Test collections and evaluation standards

To investigate the performance of a given retrieval technique, IR research community uses reference collections that have been built for different languages and many of them are publicly available and freely accessible. They are now commonly used in IR studies and helped in promoting research in IR. In this section, we present some of the text test collections.

Cranfield test collection is the first IR test collection that also grounded the evaluation framework used nowadays in IR. It was created in late 1960's and contains the abstract of 1,400 documents, 225 queries and the corresponding relevance judgment [Cleverdon, 1967; Harman, 1995]. The Cranfield test collection is the base for the success of different conferences like Text Retrieval Conference (TREC). However, in the 1990's, the size of the Cranfield collection was considered as too small to generalize a given finding on it.

TREC was established in 1992 in order to support IR researches, to provide larger and more realistic collections, as well as to promote a standard for IR evaluation [Harman, 1995]. Since then, the TREC conference creates series of evaluation resources specifically for adhoc retrieval. What is now considered as a standard adhoc IR TREC collection consists of documents, topics that correspond to users' needs and that can be structured in various fields, and relevance judgment. While TREC initially focused on English, it had also considered other languages as Spanish, Chinese and Arabic that went later to other conferences.

The Cross-Language Evaluation Forum (CLEF) conference is one of the known conferences that have their own large-scale evaluation test collection for European languages. With the initiative of CLEF, large test collections for languages such as English, French, German, Bulgarian, and Hungarian are now available [Peter, 2001; Ferro, 2014]. One of the CLEF aims was to evaluate both monolingual IR and cross-lingual IR systems while it is now oriented more to other tasks such as image retrieval, health systems, etc.

NII-NACSIS Test Collection for Information Retrieval (NTCIR) focuses on cross-language search for Asian languages such as Japanese, Chinese, and Korean [Kando et al., 1999]. Since 1997, it promotes researches in IR, text summarization, information extraction, and question-answering with the aims of offering research infrastructure, forming research community and developing evaluation methodology.

3.2 Amharic corpora, resources and tools

The growth of Amharic digital data accelerates the demand for technologies and NLP tools for online data processing. Nonetheless, for experimental evaluation of applications and tools very few corpora and resources have been built. Some of the Amharic text corpora

which are available digitally and utilized for the development of Amharic NLP tools, IR or other text-centered tasks are presented in what follows.

Corpora

Walta Information Center (WIC) Corpus: This corpus has been built by linguists from Addis Ababa University and it is available both in Amharic characters and Romanized form. The corpus contains 1,065 Amharic news articles with 4,035 sentences. The domain of the corpus is much diversified including topics like politics, economics, science, sport, religion, business, etc. [Demeke & Getachew, 2006].

Ethiopian Language Research Center (ELRC) Corpus: This is the annotated version of WIC corpus. It has been annotated with part-of-speech (POS) tags manually by ELRC at Addis Ababa University. The corpus is tagged with 30 different POS tags [Demeke and Getachew, 2006].

Addis Ababa University NLP Task Force Corpus: This corpus is prepared by Language Technology staff members from Information Technology Doctoral Program at Addis Ababa University. The corpus is prepared to create parallel corpus for computational linguistics and includes Amharic, Afaan-Oromo, Tigrigna, and English languages with diverse content from historical documents and newspapers. The project is still on-going and the corpus is continuously being updated [Abate et al., 2018].

Amharic Corpus for Machine Learning (ACML): This corpus has been prepared by Gamback [Gamback, 2012]. The data set consists of free texts collected from Ethiopian News Headlines (ENH), WIC and Amharic fiction "Fikir Iske Meqabir" (FIM). It is a set of 10,000 ENH articles with a total of 3.1 million words, 1,503 words from the WIC corpus, and 470 words from FIM book.

These corpora are simply collections of documents and as such are not appropriate to evaluate adhoc retrieval since there is no query set nor the associated relevance judgment. They are mainly collections of documents from which formatting tags are removed and additional semantic annotations are provided. Moreover, most of the corpora are not publicly accessible.

It is also worth to mention the non-European languages multi-lingual IR track at CLEF where queries in Amharic language were used; although document collections were in European languages [Di Nunzio et al., 2005; Di Nunzio et al., 2007]. For Amharic-English IR, one of the most successful approaches was a dictionary-based one [Argaw et al., 2005].

Resources

Amharic Machine Readable Dictionaries: Some of the commonly used dictionaries involving Amharic are Amharic–English dictionary containing 15,000 Amharic words [Amsalu, 1987], Amharic-French dictionary containing 12,000 Amharic entries [Berhanu, 2004], and Amharic-Amharic dictionary containing 56,000 entries [Kesatie, 1993]. Entries of the Amharic machine-readable dictionaries are represented by their citation forms.

Stopword lists: Few researches were conducted on building Amharic stopword list. Researchers who conduct studies on Amharic IR usually build their own list of stopwords. For example, Mindaye et al. [2010] built the stopword list with 77 entries while [Eyassu & Gambäck [2005] created the stopword list with 745 entries. They collected stopwords from different sources but these lists have not been evaluated by linguists. Recently, Yeshambel et al. [2020] built morpheme-based stopword list by systematically analyzing the morphology of the language and their distribution in the corpus. The list consists of 222 stopwords.

Tools: Amharic faces challenges in the development of NLP tools and applications. The major obstacles that hinder the progress on the development of Amharic NLP applications are complex morphology of the language, lack of sufficient corpora, and lack of standards in resource construction and application development. As a result, only few Amharic NLP tools have been developed thus far using rule-based and machine learning approaches. Among the available tools are morphological analyzers [Argaw & Asker, 2006; Gasser, 2011; Abate & Assabie, 2014], stemmer [Alemayehu & Willett, 2003], and parser [Argaw & Asker, 2006; Tachbelie et al., 2011]. However, the development of these tools is at prototype stages. They are also limited in scope.

4. Amharic IR Test Collection

4.1. Methodology to build the 2AIRTC collection

The document collection creation has been carried out in two steps: initial corpus and Web corpus. We initially collected documents from various sources that produce documents in Amharic without considering any specific topics. We then complement this collection with web documents retrieved considering our target topics. To do this, we run the query part of our topics on a Web search engine and gather the retrieved documents. The topic set was created by considering both current issues but also considering the topics that were likely to be treated in our initial sources. The document relevance judgments were manually done using a precise guideline. For each topic, we run Lemur (http://www.lemurproject.org/) on the initial document set and fuse these results with the ones from the web search engine. The fused lists were then manually assessed with binary relevance for each topic. These steps are described in more details in the next sub-sections.

4.2. Document set

Since we wanted documents to be diverse enough, we initially collected documents from different sources. More precisely, we collected 777 documents from news agencies sites

(Walta Media and Communication Corporate[6], Fana Broadcasting Corporate[7], Amhara Mass Media Agency[8]) and social media (Facebook), 701 historical documents from blogger (Daniel Kibret[9]), and 15,000 documents from Amharic Wikipedia[10]. In addition, 8,522 news articles were collected from Walta Information Center and we also collected 1,189 religious documents, 1,773 news articles, and 772 documents (letters, opinions and reports) from various sources. Accordingly, the total number of documents collected is 28,734. The document collection represents various topics about business, sport, entertainment, education, religion, politics, technology, health and culture.

After the topics were created, to select the documents to be assessed, we ran the title fields of these topics on our initial corpus using Lemur toolkit. We also ran the same queries (topic title) using Google on the Web. Here our idea was to use not only the documents from our initial collection but also to enrich the collection with documents that were retrieved on the Web. Our aim was to complement the document collection and to avoid topics with either no or a very few relevant documents. For the search engine documents, we considered a maximum of 50 documents per topic and we collected 2,880 documents in that way on the Web. For relevance assessment, we fused both retrieved document list (Lemur on our initial collection and Web documents for the complementary documents). Each document from the fused list was then judged for relevance.

Finally, the document collection consists in two sub-collections: documents that have been assessed for at least one topic (either relevant or non-relevant) and the entire document collection.

Assessed document sub-collection: This collection is created only from the judged documents in the initial corpus and Web documents. The top retrieved documents from both Lemur and Google of each query are fused and organized in separate files and then assessed independently. This sub-collection consists of 12,583 assessed documents though the collected documents were more than this. Out of these, 6,960 documents have been assessed as relevant for at least one topic and the remaining 5,623 documents have been judged as non-relevant. These documents are full length, processed to remove unnecessary parts such as tags and English alphabets, and plain text form. All the documents are stored in a single text file using TREC-like format. Each document has a unique document identification number. The content of each document is enclosed with <TEXT> and </TEXT> tags. One document is delimited from the other by "DOC" and "</DOC>" tags. As shown in Table 1, the relevant document set contains various document lengths from small to very large. The same holds for documents that have been judged as non-relevant.

---

[6] http://www.waltainfo.com/

[7] https://www.fanabc.com/

[8] https://www.facebook.com/AmharaMassMediaAgencyAMMA/

[9] http://www.danielkibret.com/

[10] https://am.wikipedia.org/wiki/

Table 1: Statistics of the 2AIRTC relevant judged document corpus

Parameter	**Size**
Number of documents	6,960
Number of sentences	63,081
Total number of words in the documents	2,243,372
Number of unique words	6,446
Minimum number of words per document	43
Average number of words per document	1,357
Median of words per document	219
Maximum number of words per document	74,804
Size of the relevant judged document	28.8MB

Entire document collection: While the previous sub-collection contains the documents that have been assessed for at least one topic, the entire collection consists of our initial documents plus the ones retrieved from the Web making a total of 31,614 documents. The documents from this collection are not formatted.

4.3. Topic set

<top>
<num>2</num>
<title_A> የኢትዮጵያዊያን የዘመን አቆጣጠር </title_A>
<title_E> Ethiopian calendar </title_E>
<desc_A> ስለኢትዮጵያ ዘመን አቆጣጠር ሥርአት የሚያትቱ ሰነዶችን መለየት። </desc_A>
<desc_E> Identifying documents discussing on Ethiopian calendar system. </desc_E>
<narr_A> ስለ ኢትዮጵያ የዘመን አቆጣጠር ታሪክና አመሰራረትት የሚያትቱ ሰነዶች ጥሩ የመረጃ ምንጮች ናቸው፡ ከዚህ በተጨማሪ የበአላት ቀናት እና የአቆጣጠር ስሌት የሚያትቱ ሰነዶች ጠቃሚ የመረጃ ምንጮች ናቸው። ይሁን እነጅ ስለአውሮጳውያን የዘመን አቆጣጠር ወይም ሌሎች ሀገሮች የቀን አቆጣጠር የሚገልጹ ሰነዶች ጠቃሚዎች አይደሉም። እንዲሁም ስለአዲስ አመት የሚያትቱ ሰነዶች ጠቃሚ የመረጃ ምንጮች አይደሉም።</narr_A>
<narr_E> Documents discussing the origin and history of Ethiopian calendar are good sources of information. In addition, documents explaining about holidays and methods for finding the dates and day in each year are relevant. However, documents discussing on Gregorian calendar or other calendars are not relevant. Moreover, documents discussing on new year are not relevant. </narr_E>
</top>

Fig. 1: 2AIRTC topic number 2

We created the topic set using Amharic language statements from our search experience. The topics were built in such a way to reflect real word information need and cover diverse issues. The assumption considered during topic creation was that words in the topic titles are expected in document collection. The 2AIRTC contains a set of topics which prescribe

information needs to be met. We manually created 240 topics for the adhoc IR task where the topics are about specific entities (e.g., people, places or events). Many of these topics were created by skimming some documents in the initial corpus, few of them were created by considering current issues, and the other were simply made up. The topic set is written both in Amharic and its translated version of English. Both the corpus and topic set are coded in UTF-8 (see Figure 1). Each topic has a unique identification integer number. The title field contains fewer search words which describe a topic and could be a typical query to be submitted to a retrieval system. Topic titles vary in terms of length and types. The topic titles include short topics, medium topics, and collocation. Since Amharic is a morphologically rich and complex language, the topic titles were designed to reflect real operational environment. The base of Amharic words might be stem or root. For Amharic retrieval, on top of stems, the roots of words which are derived from verbs are important rather than stems. Therefore, various types of words are included in the titles. Some of them consist of primary words and the others are from derived words.

The description field contains the description of the topic area in one or two sentences. It is the description of the user's information need. Conceptually, it is consistent with the topic title and states its purpose in a sentence form. The narrative field provides further explanation about each title to decide which types of documents are relevant and which are not. It consists of more than two sentences. Assessors judge document relevance based on this field. Table 2 presents the detailed information on the topic set.

Table 2: Statistics of the topic set

Parameter	**Size**
Number of topics	240
Minimum number of words per topic title	1
Average number of topic title's words per topic	4
Median of topic title's words per topic	3
Maximum number of topic title's words per topic	7

4.4. Relevance judgments

The relevance judgment is the third element of an adhoc IR test collection. It indicates the set of relevant documents to each topic. With regard to the document list to be reviewed by assessors, as mentioned previously, we ran the title field of the topics on our initial corpus using Lemur toolkit to get the first top 50 retrieved documents list per topic. We also ran the same topic using Google and considered a maximum of 50 documents. The topics and the associated retrieved results of both were distributed to assessors who judged them. Duplicated documents were removed. The relevant assessment was made manually by reading documents and using the narrative part of each topic. In addition, exhaustive relevance judgment was used on some topics in the initial document collection to get a larger number of relevant documents.

A document is marked as relevant based on the narrative information in the topic; thus it should not simply contain words from the query but rather fulfill the information need. Document relevance assessment has been done by students taking the IR course at University of Rift Valley. The students formed groups in which each group consisting of five students was given 20 topics and the top 50 retrieved documents for each topic by Google and Lemur. The students judged the relevance of each document for the given topic based on its narrative information and their satisfaction as users. There was one assessor group per topic. However, the five students in the group needed to agree to decide the relevance of each document. Therefore, a judgment represents the shared information needs of a group of students.

While assessing the top 50 documents retrieved by Lemur, sometimes we could not get any relevant document from our initial corpus of 28,734 documents. Therefore, some documents which were not retrieved for any of the topics using Lemur were assessed carefully in exhaustive relevance judgment during the second phase. For those documents, students had read each of them and judged as relevant or non-relevant to each topic.

As a result of manual relevant assessment, some topics have many relevant documents, while other topics have fewer relevant documents. Each topic has at least 10 relevant documents. This indicates that for some topics it will not be possible to measure effectiveness above rank ten; this is also the case in TREC or alike test collections.

Finally, the 2AIRTC relevance file was produced using the TREC format, as follows: topic ID, 0, document ID and relevance fields. Topic and document IDs are unique identification numbers of topics and documents, respectively. The number zero (0) is common to all topics and documents. The relevance indicates the relevance value of the considered document/topic pair and is 1 if the document is relevant to a topic, 0 otherwise.

Table 3: 2AIRTC relevance judgment statistics

Parameter	**Size**
Total number of topics	240
Average number of relevant documents per topic	22
Minimum number of relevant documents per topic	10
Maximum number of relevant documents per topic	172

5. Conclusion

Using standard test collections in IR is a common experimental practice. Various test collections for different languages have been built and are used by many research groups. Amharic IR test collection is an under explored research area. In comparison with other resourceful languages, few resources and tools have been built for Amharic. Even the existing Amharic corpora do not have any associated Amharic topics and relevance judgments. Furthermore, most of the existing corpora and resources are small in size and

not publicly accessible. However, the importance of building and sharing test collections is well acknowledged. We built the first reusable test collection for IR system benchmarking. This collection will be made available to research communities. It will be accessible on request at tilahun.yeshambel@gmail.com. Our Amharic test collection is reproducible and contains representative documents and topics. This collection, named 2AIRTC[11], can serve as a reliable resource for the evaluation and comparison of various Amharic IR systems. We do believe this collection will help in enhancing new research on Amharic IR.

Acknowledgement

We would like to thank students who participated in the creation of the test collection. We also thank the owners and sources of documents (Walta Media and Communication Corporate, Fana Broadcasting Corporate, Amhara Mass Media Agency and Daniel Kibret) who provide and permit their documents to carryout academic research experiments and annotate, use and share them for research communities.

References

Abate, M., & Assabie, Y. (2014). Development of amharic morphological analyzer using memory-based learning. In *Proc. of the 9th Int. Conf. on Natural Language Processing*, pp. 1-13, Warsaw.

Abate, S. T., Melese, M., Tachbelie, M. Y., Meshesha, M., Atinafu, S., Mulugeta, W., Assabie, Y., Abera, H., Seyoum, B. E. Abebe, T., Tsegaye, W., Lemma, A., Andargie, T. & Shifaw, S. (2018) Parallel Corpora for Bi-Lingual English-Ethiopian Languages Statistical Machine Translation, In *Proc. of the 27th Int. Conf. on Computational Linguistics*, pp. 3102-3111, New Mexico, USA.

Alemayehu, N., & Willett, P. (2003). The effectiveness of stemming for information retrieval in Amharic. *Program: electronic library and information systems*, 37(4), 254–259.

Amsalu, A. (1987). *Amharic-English Dictionary*. Kuraz Printing Press, Addis Ababa, Ethiopia.

Argaw, A. A., & Asker, L. (2006). Amharic-English information retrieval. In *Workshop of the CLEF* (pp. 43-50). Springer, Berlin, Heidelberg.

Argaw A.A., Asker L., Cöster R., Karlgren J., Sahlgren M. (2005). Dictionary-Based Amharic-French information retrieval. In *Workshop of the CLEF* (pp. 83-92). Springer, Heidelberg.

Berhanu, A. (2004). *Amharic-Français Dictionnaire*. Shama Books, Addis Ababa, Ethiopia.

Buckley, C. & Voorhees, E. (2005). Retrieval system evaluation, In *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press, Cambridge, MA, Chapter 3, 53–75.

Cleverdon, C. W. (1959). The evaluation of systems used in information retrieval. In: *Proc. of the Int. Conf. on Scientific Information*. pp.687-698, Washington, DC.

Cleverdon, C. (1967). The Cranfield tests on index language devices. In *Aslib proceedings*. MCB UP Ltd.

Countrymeters. (2020). Ethiopian Population. URL http://countrymeters.info/en/ethiopia. Accessed: 05-05-2020.

[11] The test collection will be made publicly available upon publication of this paper.

Di Nunzio, G. M., Ferro, N., Jones, G. J., & Peters, C. (2005). CLEF 2005: Ad hoc track overview. In Workshop of the CLEF. pp. 11-36, Springer, Berlin, Heidelberg.

Di Nunzio, G. M., Ferro, N., Mandl, T., & Peters, C. (2005). Clef 2007: Ad hoc track overview. In Workshop of the CLEF. pp. 13-32, Springer, Berlin, Heidelberg.

Demeke, G. A., & Getachew, M. (2006). Manual annotation of Amharic news items with part-of-speech tags and its challenges. Ethiopian Languages Research Center, 2:1–16.

Harman D. (1995).Overview of the second text retrieval conference (TREC-2). Information Processing & Management, 31(3):271-289.

Ferro, Nicola (2014). CLEF 15th Birthday: Past, Present, and Future. ACM SIGIR Forum, 48(2): 31–55.

Eyassu, S., & Gambäck, B. [2005]. Classifying Amharic news text using self-organizing maps. 71. https://doi.org/10.3115/1621787.1621801

Gamback B. (2012). Tagging and Verifying an Amharic News Corpus. Proceedings of the Eighth International Conference on Language Resources and Evaluation, pp. 79–84.

Gasser, M. (2011). HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. Conference on Human Language Technology for Development, pp. 94–99.

Hetzron,R.(1972). *Ethiopian Semitic: Studies in Classification*. Manchester Univesity Press.

Kagolovsky, Y., & Moehr, J. (2003). Current Status of the Evaluation of Information Retrieval. *Journal of Medical Systems*, 27(5): 409–424.

Kando, N., Kuriyama, K., Nozue, T., Eguchi, K., Kato, H., & Adachi, J. (1999). The NTCIR Workshop : the First Evaluation Workshop on Japanese Text Retrieval and Cross-Lingual Information Retrieval. *4thW. on Information Retrieval with Asian Languages*, (1), INV-1-INV-7.

Kesatie B. (1993). YeAmarinja Mezgebe Qalat. *Ethiopian Languages Research Center*, Artistic Publisher, Addis Abeba, Ethiopia.

Mindaye, T., & Atnafu, S. (2009). Design and implementation of Amharic search engine. In *Proc. of the 5th Int. Conf. on Signal Image Technology and Internet Based Systems*, pp. 318–325.

Peters, C., & Braschler, M. (2001). European research letter: Cross-language system evaluation: The CLEF campaigns. *J. of the Am. Society for Inf. Science and Technology, 52*(12), 1067-1072.

Samimi, P., & Ravana, S. (2014). Creation of Reliable Relevance Judgments in Information Retrieval Systems Evaluation Experimentation through Crowdsourcing :A Review. *The Scientific World Journal*, Volume 2014.

Sanderson, M., & Croft, W. (2012). The history of information retrieval research. *Proceedings of the IEEE*, *Special Centennial Issue*, pp.1444-1451.

Soboroff,I.(2007).A comparasion of pooled and sampled relevance judgments in the TREC 2006 Terabyte Track.The first international workshop on evaluation information access,Tokyo,Japan.

Tachbelie, M. Y., Abate, S. T., & Besacier, L. (2011). Part-of-speech tagging for underresourced and morphologically rich languages—the case of Amharic. *HLTD (2011)*, 50-55.

Yeshambel, T., Josiane, M. and Assabie,Y. (2020). Construction of Morpheme-Based Amharic Stopword List for Information Retrieval System. Accepted In *the 8th EAI Int. Conf. on Advancements of Science and Technology*, Bahir Dar, Ethiopia.