Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2103–2109
Marseille, 11–16 May 2020

Event Extraction from Unstructured Amharic Text

Ephrem Tadesse, Rosa Tsegaye Aga, Kuulaa Qaqqabaa
Jimma University, Armauer Hansen Research Institute, Addis Ababa Science and Technology University
Jimma, Addis Ababa, Akaki Kality Sub-City
ephe11ta@gmail.com, rosatsegaye@gmail.com, kuulaa@gmail.com

Abstract

In information extraction, event extraction is one of the types that extract the specific knowledge of certain incidents from texts. Event extraction has been done on different languages text but not on one of the Semitic language, Amharic. In this study, we present a system that extracts an event from unstructured Amharic text. The system has designed by the integration of supervised machine learning and rule-based approaches. We call this system a hybrid system. The system uses the supervised machine learning to detect events from the text and the handcrafted and the rule-based rules to extract the event from the text. For the event extraction, event arguments have been used. Event arguments identify event triggering words or phrases that clearly express the occurrence of the event. The event argument attributes can be verbs, nouns, sometimes adjectives (such as ሰርግ/wedding) and time as well. The hybrid system has compared with the standalone rule-based method that is well known for event extraction. The study has shown that the hybrid system has outperformed the standalone rule-based method.

Keywords: Event extraction, under-resourced language, Machine learning algorithms, Nominal events.

1. Introduction

Amharic is a Semitic language, related to Hebrew, Arabic, and Syriac. It has been the second most spoken Semitic language by around 27 million speakers (Mulugeta and Gasser, 2012) primarily in Ethiopia next to Arabic language. It is currently the official language of government in Ethiopia, and has been since the 13^{th} century. In addition, it is the medium of instruction in primary and secondary schools as well as the source language for a large body of historical texts. As a result, most documents in the country have been produced in Amharic and there has been an enormous production of electronic and online accessible Amharic documents.

The predominant problem of underrepresented languages is the lack of resources (Sohail and Elahi, 2018). Most recently on the web fewer online Amharic textual resources are available for people in their everyday lives. However, researchers and other interested group of people in linguistic and computing disciplines face difficulties because of Amharic presents sophisticated language-specific issues.

The existing information extraction systems that have developed for Hebrew, Arabic, or other languages have not represented the linguistic structure and morphological richness of the languages. But events in Amharic text are predominantly expressed through verbs and nouns. Therefore, these systems can not be used directly for Amharic texts.

For example, consider the following sentence "ሰኞ መስከረም 1965 ኢትዮጵያ በውጥረት ነግሳ ነበር." /"Ethiopia was in turmoil in Monday, September , 1965". In this sentence "በውጥረት" and "ነግሳ" refers to an event, whereas the phrase "ሰኞ መስከረም 1965" is a time argument which indicates when the event happened. The word "ኢትዮጵያ" refers the named entity or participant of the event.

Because of this prominent significance of extracting events from unstructured Amharic text for high level Natural Language Processing (NLP) tasks we are interested to tackle this problem. In this study we present a comprehensive technique for extracting events from Amharic unstructured text.

The rest of the paper is organized as follows. Section 2. discusses the related works of this study. Section 3. explains the methodology of the study. It motivates and elaborates the event extraction models and algorithms that have used in the study. Section 5. presents the experimental results of the study, and discussion and comparison of the different result of the models that have proposed in the study. The study has concluded in Section 6. by conclusion and future work.

2. Related Work

Recently event extraction has gained popularity due to its wide applicability for various NLP applications. Most event extraction systems support English and European language texts from different domains using a variety of techniques. Now a days, Semitic languages are typically a topic of interest for researchers. Event extraction for Amharic has not been done yet; therefore this study is the first in this particular information extraction (IE) application. Due to the variation of the language structure the existing techniques and tools applied to other languages can't be directly used for this particular task.

There are some progressive work that have been done so far on Amharic NLP tasks with promising results including part of speech tagging, morphological analyzer, named entity recognition, base phrase chunking and text classification as in (Adafre, 2005; Ibrahim and Assabie, 2014; Sikdar and Gambäck, 2018; lasker et al., 2007). Various techniques have been widely employed for each task to enhance the accuracy and handling linguistic exceptions. However, there have not been ready-made pre-components and well organized datasets. Besides these limitations there has not been any undergoing research on event extraction from unstructured Amharic text due to difficulties in syntactic and semantic status of class of functional verbs. The other challenges are identifying event arguments. In our case temporal event arguments have considered. However, it has a challenge in Amharic texts. Amharic texts have represented in various forms such as; sequence of words, Arabic and Geez'e script numerals. As such it needs extra normalization and syntactic analyzing scheme to tackle temporal argument.

Semitic languages like Arabic, Hebrew and Amharic have much more complex morphology than English. The morphological variation limits the research progress on Natural language processing, in general. However, there are studies relative to other Semitic languages. For example, (Al-Smadi and Qawasmeh, 2016) have done their study on automatic event extraction for Arabic language using knowledge driven approach which concentrates on tagging the event trigger instances and related entities. One of their main contribution is to link event with the entity mention. However, in our case we mainly concentrate on extracting events and its arguments with the advantage of hand crafted rules and machine learning classifiers.

Hindi is another under-resourced an indo European language that has more common words with Arabic. In (Ramrakhiyani and Majumder, 2015) solely has focused on Temporal Expression Recognition in Hindi using interactive handcrafted rules. They aim to carry out two basic goals that are identification of the temporal expressions in plain text and classifying the identified temporal expression. However, extracting events along with the corresponding arguments gains more advantage for the ease of chronological ordering of events in their occurrences. In addition it can be extended for event argument relationship extraction tasks.

(Smadi and Qawasmeh, 2018) has proposed a supervised machine learning approach for extracting events from Arabic tweets. The study mainly focuses on four main tasks: Event Trigger Extraction, Event Time Expression Extraction, Event Type Identification, and Temporal Resolution for ontology population. Significant scores have resulted for each task covered under this paper includes; T1: event trigger extraction F-1= 92.6, and T2: event time expression extraction F-1= 92.8 in T3: event type identification Accuracy= 80.1. They have claimed that the third task is relatively better than the previous works done using similar techniques like document-term matrix or bag-of-words. (Arnulphy et al., 2015) has also proposed supervised machine learning approach but to detect French and English Time Markup Language (ML) Events. The study has suggested the approach to be used by combining different supervised machine learning algorithms such as conditional random field, decision tree and k-nearest neighbor including language models.

(Al-Smadi and Qawasmeh, 2016) has proposed knowledge-based approach for event extraction from Arabic Tweets. There are three subtasks covered under their study such as event trigger extraction, event time extraction, and event type identification. The event expression includes important event arguments that are event agent, event location, event trigger, event target, and event product and event time. The tools and dataset that have used in their study have utilized twitter streaming API and preprocessed through AraNLP Java-based package. Moreover, after the visualization services event extraction like calendar, timeline supplied through the help of ontological knowledge bases. In their study the experimental results show that the approach has an accuracy of 75.9 for T1: event trigger extraction, 87.5 for T2: Event time extraction and 97.7 for T3: event type identification. Their study claims that applying this kind of domain dependent approach to extract events from tweets scores significant results.

In general there has been a lot of work in event extraction such as (Arnulphy et al., 2015; Tourille et al., 2017) in European languages, predominantly in English; But, much less research in other languages. An approach or technique that has used in one language to extract events might be used in languages as well if they have a similar grammar and character set. However, If languages have very different grammar, or a very different written representation, it will be difficult to use related approaches or techniques to extract events.

There has been research in part-of-speech tagging on Amharic text (Adafre, 2005) and on Amharic morphology (Mulugeta and Gasser, 2012) which are helpful for event detection, but not directly related to the actual event extraction task from Amharic text. For this particular task the state of art Event detection system typically uses a robust machine-learning techniques. Examples of such systems are (Arnulphy et al., 2015). Because of the lack of sufficient labeled training data for Amharic, we bootstrap an event extractor using a rule-based algorithm.

3. Methodology

According to (Frederik Hogenboom and Kaymak, 2016) event extraction techniques have been evaluated based on the works on a set of qualitative dimensions that are the amount of required data, knowledge, expertise, interpretability of the results and the required development and execution times. In this study, supervised machine learning techniques, handcrafted rules and hybrid techniques have employed to detect and extract events and its arguments from unstructured text. Our focus of interest has been extracting events and event arguments from unstructured Amharic text. Event arguments include identification of event trigger words; where in Amharic unstructured text nominal events become ambiguous. Such events can be arguments of other events, and they often have been hard to be identified.

3.1. Dataset preparation

Unlike other languages, Amharic language does not have any standardized annotated publically available corpora like Treebank[1] and PropBank[2] for English. The news domain is more preferable data source. Because its publicly available and contains rich source of information that helps

[1] https://catalog.ldc.upenn.edu/LDC99T42

[2] http://www.nltk.org/howto/propbank.html

for any NLP applications such as entity extraction, event and temporal information extraction and co-reference resolution. In this study, we build our own dataset by scraping top local websites. These are Zehabesha[3], Satenaw[4] , Ezega[5], and one international website BBC Amharic[6] that contains relevant Amharic unstructured text. A Python Beautiful Soup library [7] has been used for for scraping the sites. The scraped texts are from all domains such as economy, politics, technology and sport. Simple regular expressions have been used to retrieve only relevant text contents. A total of 659,848 words have extracted. Along with our own dataset, we have used Amharic corpora that have been prepared by the Ethiopian Languages Research Center of Addis Ababa University in a project called *the annotation of Amharic news documents* (Demeke and Getachew, 2006). The project has been tagging manually each Amharic word in its context with the most appropriate parts-of speech tag. The corpus contains 210,000 words that has collected from 1065 Amharic news (documents of Walta Information Center (Demeke and Getachew, 2006)). Walta Information Center is a private news and information service provider located in Addis Ababa, Ethiopia.

3.2. Data preprocessing

In this step, data has converted to the appropriate format required for the respective information extraction process. In this study the scraped texts have many junks such as markup tags and other special characters. The first step in our study is raw text preprocessing. This step contains cleaning unwanted junks, sentence splitting, tokenizing, word stemming, character normalization, stop word removal and Part Of Speech tagging (POS). Unlike other languages, Amharic is a morphologically rich language that posses complicated syntactic features. This makes cumbersome the preprocessing task to analyze the morphological features of representative tokens. The sentence splitter splits using Amharic sentence demarcations (። ፧ ? !).

Amharic language has different characters with the same meaning and pronunciation. Those different characters should be treated equally because there is no change in meaning regardless of the linguistic view of orientation among the characters. For example:- (ሀ፤ሐ፤ኀ), (ሰ፤ ሠ), (ዐ፤ ዓ፤ አ) and (ጸ፤ ፀ), each group has the same meaning (Gasser, 2011). As a result, we develop a character normalizer that enables to normalize those characters to an ordinary conceivable form This task helps the performance of our system. The other preprocessing task is stop word removal. Like other language, Amharic has its own list of stop words such as conjunctions, articles and prepositions. In our case we have adopted stop word lists that has used in (Tsedalu, 2010) study. In addition, we have built our own stop word lists, as well with the help of linguistic experts. Then a total of 235 stop word have identified.

The other important preprocessing task is analyzing Amharic verb morphology to identify lemma of words and their derivation. The lemma of a word is very crucial feature for the classifier. We have applied hornmorpho [8] that is a system to process the morphology of Amharic. The system works for the other Ethiopian local languages such as Amharic, Affan Oromo, and Tigrinya languages. However, the system misses some unique and compound words. Thus, we have developed our own unique exceptional dictionary (Gazetteer) to handle exceptional keywords. Finding a pattern to get only the lemma of the hornmorpho result has also other difficulties; because sometimes the co responding word doesn't contain full information. In that case the Hornmorpho skips subject, object, grammar, or word classes of a specific words. For Amharic Language, Hornmorpho has evaluated using 200 randomly selected verbs and nouns/adjectives in (Gasser, 2011) study. The output has compared with manually identified Amharic verbs and nouns. 99%; Amharic nouns: 95.5%. Although, we prefer to use this tool in our study, because of the lack of other ready-made NLP components for Amharic language. The Jython library[9] has been used to integrate the python based morphological analyzer for Amharic to get morphological features of words.

Besides analyzing the verb morphology, annotating the exact word class of the instance is also the required preprocessing task in this study. To do so, we have been using the publically available language independent part-of-speech tagger, which is TreeTagger[10]. TreeTagger is a tool for annotating text with part-of-speech and lemma information. It has been successfully used to tag German, English, French, Italian, Danish, Swedish, Norwegian, Dutch, Spanish, Bulgarian, Coptic and Spanish texts. It is adaptable to other languages as well if a lexicon and a manually tagged training corpus are available (Schmid, 1994). It consists of two programs: the training program that creates a parameter file from a full-form lexicon and the lexicon generator along with a hand tagged corpus. The tagger program reads the parameter file and annotates the text with part of speech and lemma information. To prepare a parameter file for TreeTagger we used a total of 217 000 Amharic manually tagged corpora with 9 distinguished word classes and corresponding lemmas. We have conducted evaluation of TreeTager using 92,456 randomly untagged tokens. The output of TreeTager results 99.9% accuracy compared with manually tagged Amharic words.

The other crucial step in our preprocessing module is normalizing Amharic temporal arguments. There are various representations of date time expressions in Amharic such as Arabic, Geez and using alphanumeric characters.

For example, the following sentences show the different date time representation:

(አቤል በ1995 ዓ.ም ተወለደ ።) using Arabic characters
(አቤል በሽ ዘጠኝ መቶ ዘጠና አምስት ዓ.ም ተወለደ ።) using alphanumeric characters
(አቤል በ ፡፲፱፻፱፺፭፡ ዓ.ም ተወለደ ።) using geez characters

[3] http://www.zehabesha.com/amharic/

[4] https://www.satenaw.com/amharic/

[5] https://www.ezega.com/News/am/

[6] https://www.bbc.com/amharic

[7] https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[8] https://www.cs.indiana.edu/ gasser/HLTD11/

[9] https://www.jython.org/

[10] https://reckart.github.io/tt4j/

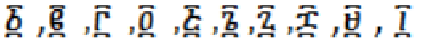

Figure 1: Geez numerals

The above sentences refer logically similar meaning with various syntactic representation. In order to handle temporal arguments of the event, a normalization and conversion scheme to convert temporal representations into one form. The conversion of Ge'ez numerals to uniform Arabic number system is not straight forward as other normalization tasks because of the irregularities of Unicode values for Ge'ez numerals. Some of the Ge'ez numerals are presented in Figure 1.

3.3. Event detection using supervised machine leaning

In this study, supervised machine learning approach has been employed. Supervised machine learning classifiers predict new events based on the given labeled training sets. It uses event properties and characteristics from training data and generalize the unseen situations to predict events. In this study, the supervised learning approach has been used to detect the events from a text.

In this study, the datasets are unstructured text and documents. Therefore, the unstructured text sequences have converted into a structured feature space using mathematical modeling. For classification, feature extraction can be seen as a search among all possible transformations of the feature set for the best one. This preserves class maintainability as much as possible in the space with the lowest possible dimension. In this study, the features contain information of the text that have used to provide necessary information associated to a given events. These features increase the confidence level of predicting a token as an event. Thus, the feature extractor component that has used in this study is responsible for extracting candidate attributes for the classifier. The features that have used in this study are the following:-

- Words of the instance
- POS of the corresponding word
- Lemma of the corresponding word
- List of lexicons for exceptional events

A binary classifier has been used to detect events from Amharic text. The classifier detects events from the text and classify the text as *on-event* and *off-event*. The on-event class refers the instance that contains event trigger keywords; Whereas the off-event class refers the instance that does not infer the event trigger keywords. From the machine learning algorithms, Naive Bayes, decision tree and SVM algorithms have proposed based on their widely use in text classification tasks (Pranckevicius and Marcinkevicius, 2017).

Naive Bayes classifier is linear classifier that is known for being simple and very efficient for text classification tasks. The probabilistic model of naive Bayes classifiers is based on Bayes' theorem. This algorithm works on the assumption that the features in a dataset are independent from each other.

LIBSVM is a library for Support Vector Machines (SVM). It has gained wide popularity in machine learning and many other areas. SVM finds an optimal solution and maximizes the distance between the hyperplane and the difficult points close to decision boundary. As (Chang and Lin, 2011) stated, if there are no points near the decision surface, then there are no very uncertain classification decisions.

The other classifier algorithm that has used in this study is decision tree. Decision tree is a Tree-based classifier for instances that have represented as feature-vectors. There is one branch for each value of the feature, and leaves specify the category. It represents arbitrary classification function over discrete feature vectors. For the decision tree, J48 algorithm have used. J48 is an algorithm used to generate a decision tree. The decision trees generated by J48 can be used for classification, and for this reason, J48 is often referred to as a statistical classifier.

The above algorithms have been used to train the models using the labeled dataset as an input. Then the models have detected the instances even on a test set as on-event and off-event classes. The POS tag feature has showed good performance as the best syntactic feature to detect the events based on the feature selection recommendation.

3.4. Event Extraction using Rule Based Approach

Rule based learning is one of the information extraction method that utilizes the extraction pattern to retrieve information from a text document. In this study, a standalone rule-based approach has proposed to enhance the accuracy of event extraction system. Unlike other languages, Amharic has a subject-object-verb agreement and other morphological features that makes cumbersome the rules construction. As (Yunita Sari and Zamin, 2010) has mentioned, construction of extraction pattern is based on syntactic or semantic constraint and delimiter or combination of both syntactic and semantic constraint. Events dominantly exist as nominal and verbs (Ramesh and Kumar, 2016). The nominal events are ambiguous, in which they can appear in deverbal or non-deverbal nouns form. Thus, we need morphological features of the instances to disambiguate nominal events. To do so, morphological analyzer has employed to get the morphological features of the event that have mentioned in the instances. For example:- (የኢትዮጵያ ህዝቦች ከዚህ በኋላ ፈፅሞ መዘናጋት አይፈልጉም ።) In this sentence, the underline word (ፈፅሞ) is derived from the verb (ፍጽም). It seems an adjective, but, it's a deverbal entity we call it a *nominal event*. The rules have been developed based on syntactic features of words with the help of a carefully constructed list of gazetteers. The POS tag and lemma of the word have been used as an abasement for the handcrafted linguistic rules. Different components have been used to get syntactic features of words using Tree Tagger and Hornmorpho. The pattern extractor has been developed based on the syntactic features. Simple rules have been applied to extract detected events.

For example:- (አበበ) N (ትላንት) ADV (የገዛው) VN (በሬ) N (ሞተ) VP (።) . In this example, the snippets of handcrafted rules have tackled based on the POS tagger results. The formal structures are not always regular to develop stable rules. In contrast, the morphological analyzer is very helpful, because of the existence of deverbal events that have been act as ambiguous.

Some of the rules that have applied in this study for the hybrid system includes the following:-

1. Automatically label preprocessed texts with their corresponding word classes or parts-of-speeches.

2. Get the morphological features of words including word, subject, root, lemma , object, grammar and preposition

3. Usually events are expressed using verbs and nouns. Check the neighboring words using bigram language models. Because, not all nouns have been events and sometimes nouns come at the beginning, then they are the subjects or participant of the event not exactly the event.

4. Identifying the nominal events; To do so, the morphological analyzer has main role on indicating the citation of the respective nouns; i.e words that have exactly nominal can be deverbal or non deverbal nouns. But, deverbal nouns has a citation of verbs.

5. Words that has categorized as verbs and verb group word classes as part-of-speech and it's infinitive forms have selected as primary candidates.

6. Check non deverbal nouns (usually acts as events) from carefully built gazetteers (list of non deverbal noun lexical). Because of our limited dictionary a ternary search tree algorithm has been applied to enhance the efficiency.

7. Identifying words that contain temporal keywords. The temporal indicator keywords have carefully built the list of commonly used temporal expressions in Amharic. In addition, regular expressions have been constructed to tackle regular date-time expressions. Bi-gram language models have been applied to find temporal arguments.
   For example:- የአበበ ሰርግ ነገ ነው ። / "Abebe's wedding is tomorrow."
   From this sentence, the word **ሰርግ** is a deverbal nouns that has been extracted as an event and it's actually an event, where the word **ነገ** is an event argument extracted as temporal event argument of the major event **ሰርግ**

3.5. Event Extraction using hybrid Approach

Unlike knowledge driven systems; in hybrid event extraction systems the amount of required data increases, due to the usage of supervised machine learning techniques, yet typically remains less than the case with purely data-driven techniques. Where complexity and hence required expertise is generally high due to the combination of multiple techniques compared to pure knowledge driven techniques.Moreover, the interpretability of a system benefits to some extent from the use of semantics as in knowledge-based techniques(Baradaran and Mineai-Bidgoli, 2015).

The other technique that has employed in this study is combining both supervised machine learning and rule-based techniques to extract events from Amharic unstructured text. The machine learning approach mainly focuses on coverage (Recall) apart from sensitivity (precision) while, the handcrafted rules approach is on achieving the highest potential of precision value based on the incorporated rules. In our case, the machine learning classifiers ignores nominal events in comparison with the verbal events. Therefore, we incorporate some rules to tackle the missed events from the machine learning classifiers result. Deverbal nouns exhibit both nominal and verbal syntactic representations They serve as concrete nouns, but also participate in verbal constructions where they require arguments and accept the aspectual modification. Nominal events sometimes appear as deverbal and non-deverbal, in which deverbal entities have been derived from verbs in-contrary the non-deverbal entities have not derived from verbs. e.g. (ፊጽሞ) is a deverbal entities that is an event derived from verb (ፍጽም). An event is a situation that lasts for a moment. By this definition, nominal can be an event e.g (ሰርግ)/ wedding is a non-deverbal nominal event. Another example (የአበበ ሰርግ ሃምሌ 16፣ 2010 ዓ.ም ነው.) . Simply knowing the morphological variation of words and having a common non-deverbal nominal list from the gazetteers (list of exceptional non deverbal events) help to get rid of event ambiguity. We also get those deverbal events from the morphological analyzer and non-deverbal events from the gazetteers. Applying such disambiguation scheme improves accuracy of our system in proportion to the standalone rule based approach.

4. Model Evaluation

Among the standard information extraction evaluation metrics precision, recall and F-measure have been used to evaluate the performance of models. In this study a 10-fold cross-validation technique has used to split the dataset. In this case by shuffling the dataset randomly, 80% of the data has used for training and 20% has used for test.

5. Experimental Results

In this study, a total of five experiments have been conducted. Three experiments are on supervised learning algorithms (Naive Bayes, Decision Tree and SVM) to detect events from the unstructured Amharic text. One experiment is on the rule-based approach to extract the event from the unstructured Amharic text. The other experiment is combining the supervised learning and the rule-based approaches.

The first three experiments are training a model using the three selected supervised learning algorithms. All features have used to see the effect of each attribute on the event detection. Each algorithm has been experimented on the full features.

As the result shows in Table 1, among the three algorithms, the Naïve Bayes (NB) classifier has outperformed the other classifiers to detect events. It has showed F-score of 0.915% on the weighted average, 0.831 on the On-Event class, and 0.944 on the Off-Event class. This experimental result confirms the advantage of Naïve Bayes classifier for event detection task. We get encouraging result using a machine learning classifier for event detection task. The problem resides on deverbal entities ambiguousness.

Table 1: Experimental results for machine learning Algorithms to detect events

Algorithms	**Measures**			**Classes**
	Precision	***Recall***	***F-measure***	
NB	0.866	0.798	0.831	**On-Event**
	0.932	0.957	0.944	**Off-Event**
	0.915	0.916	0.915	**Weighted Ave.**
LIBSVM	0.895	0.395	0.548	**On-Event**
	0.825	0.984	0.897	**Off-Event**
	0.843	0.833	0.808	**Weighted Ave.**
J48	0.891	0.698	0.783	**On-Event**
	0.903	0.971	0.935	**Off-Event**
	0.9	0.9	0.896	**Weighted Ave.**

From the machine learning event detection system, it has observed that due to linguistic features verb triggered events have equal weight by the classifier with the non-event class. This has been the reason that motivates the study to come up with developing hand crafted rules to get rid of the ambiguities. In this particular technique, in order to make a clear comparison with the hybrid based event extraction system, similar dataset have been used.

The other experiment is on the rule-based approach. As the result Table 2 shows, the F-score of this approach model is 0.959. This shows that it has outperformed the supervised machine learning three models.

The last experiment that has been conducted in this study is on the hybrid event extraction technique. The performance of this method relay on the power of having the advantage of the rule based and supervised machine learning methods in conjunction. The machine learning classifiers have labeled the instances as on-event and off-event binary classes by assigning different weights. An instance that has assigned high probability value by the classifier is categorized under on-event class; which is actually an event. On the other hand an instance which has assigned low probability value than the on-event class instance has been mostly non-event or categorized as off-event class. Thus positive predicated values accepted as it's i.e. instances categorized as on-event with highest weighted value. Because, it is predicted exactly as an event, while instances getting equal weight by the classifier in both class are going to be the target instances for the heuristics. Equal weighted instances are considered as ambiguous. Using the help of syntactic features, ambiguous instances have been handled. As a result, the number of on-event instances correctly extracted increases when heuristics has applied. In order to get the false negative and the false positive values we have used a manual scanning of the result to be accurate.

Table 2 shows the hybrid technique experimental result as well and compare with the experimental result of rule based technique. As the table shows, the combination of both rule based and supervised machine learning classifiers bring significant result to extract events from unstructured Amharic text.

Table 2: Over all event extraction evaluation of experimental result comparison

Techniques	**Standard measures**		
	Precision	***Recall***	***F-measure***
Rule based Approach	0.976	0.952	0.959
Hybrid Approach	0.979	0.962	0.971

6. Conclusion and future work

In this study we have presented a system that extract events from unstructured Amharic text. The system has built by combining supervised learning and the standalone rule-based techniques. The supervised machine learning have used to detect events and the standalone rule-based technique to extract the even from the unstructured Amharic text. For the supervised machine learning, the three algorithms (naïve bayes, support vector machine and decision tree) have proposed. Then Naïve bayes has outperformed to detect events from the unstructured Amharic texts.

The standalone rule based approach has evaluated independently to extract events from unstructured Amharic text. However, the proposed hybrid system has outperformed using the Naïve bayes algorithm to detect the event.

In the future we need to address other relevant event extraction tasks such as building larger events and temporally annotated corpus, employing powerful deep learning techniques to extract relation between event and time, extracting relation between events and document creation time.

7. Bibliographical References

Adafre, S. F. (2005). Part of speech tagging for amharic using conditional random fields. In *Proceedings of the ACL workshop on computational approaches to semitic languages*.

Al-Smadi, M. and Qawasmeh, O. (2016). Knowledge-based approach for event extraction from arabic tweets. *International Journal of Advanced Computer Science and Applications*, 7(6).

Arnulphy, B., Claveau, V., Tannier, X., and Vilnat, A. (2015). Supervised Machine Learning Techniques to Detect TimeML Events in French and English. In Chris Beimann, et al., editors, *20thInternational Conference on Applications of Natural Language to Information Systems, NLDB 2015*, volume 9103 of *Proceedings of the NLDB conference*, Passau, Germany, June. Springer.

Baradaran, R. and Mineai-Bidgoli, B. (2015). Event extraction from classical arabic texts. *International Arab Journal of Information Technology"*, 12(5).

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May.

Demeke, G. and Getachew, M. (2006). Manual annotation of amharic news items with part-of-speech tags and its challenges. 01.

Frederik Hogenboom, F. F. and Kaymak, U. (2016). A survey of event extraction methods from text for decision support systems. *Elsevier*.

Gasser, M. (2011). Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development*.

Ibrahim, A. and Assabie, Y. (2014). Amharic sentence parsing using base phrase chunking. In *COLING 2014*.

lasker, L., Argaw, A. A., and Gamback, B. (2007). Applying machine learning to amharic text classification. In *Proceedings of the 5th World Congress of African Linguistics*.

Mulugeta, W. and Gasser, M. (2012). Learning morphological rules for amharic verbs using inductive logic programming.

Pranckevicius, T. and Marcinkevicius, V. (2017). Comparison of naïve bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. In *Baltic J. Modern Computing*.

Ramesh, D. and Kumar, S. S. (2016). Event extraction from natural language text. *International Journal of Engineering Sciences and Research Technology (IJESRT)*, 5(7).

Ramrakhiyani, N. and Majumder. (2015). Approaches to temporal expression recognition in hindi. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 14(1).

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Proceeding*.

Sikdar, U. and Gambäck, B., (2018). *Named Entity Recognition for Amharic Using Stack-Based Deep Learning: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part I*, pages 276–287. 01.

Smadi, M. and Qawasmeh, O. (2018). A supervised machine learning approach for events extraction out of arabic tweets. In *Fifth International Conference on Social Networks Analysis, Management and Security, SNAMS 2018, Valencia, Spain, October 15-18, 2018*, pages 114–119.

Sohail, O. and Elahi, I. (2018). Text classification in an under-resourced language via lexical normalization and feature pooling. In *Twenty-Second Pacific Asia Conference on Information Systems*.

Tourille, J., Ferret, O., Tannier, X., and Neveol, A. (2017). Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 of *EACL*, page 739–745.

Tsedalu, G. (2010). Information extraction model from amharic news texts. Master's thesis, Addis Ababa University.

Yunita Sari, M. F. H. and Zamin, N. (2010). Rule based pattern extractor and named entity recognition: A hybrid approach. *IEEE*.