Temporal Information Extraction from Unstructured Amharic Text

Anonymous authors
Paper under double-blind review

Abstract

In information extraction, temporal information extraction is one of the types that extract the specific knowledge of certain incidents and it's temporal arguments from texts. Temporal information extraction has been done on different languages texts but not on one of the Semitic language Amharic.In this study, we present a system that extracts temporal information from unstructured Amharic text.The system has designed by standalone supervised machine learning followed by rule-based approach.The model from the supervised machine learning detects events from the text, then, handcrafted rules extract events and it's temporal flags(mentions) from the text.The study has shown that the supervised learning technique has outperformed the standalone rule-based method. For the temporal information extraction, we have been extracting event arguments and mentions.Event arguments identify event triggering words or phrases that clearly express the occurrence of the event.The event argument attributes can be verbs, nouns, occasionally adjectives such as አቤል በ1995 ዓ.ም ተወለደ ። /(Abel was born in 1995.) and time as well. Event mentions or temporal flags that has gained attention in this study indicates happening or long lasting of an event.

1 Introduction

Amharic is a Semitic language, related to Hebrew, Arabic, and Syriac. Next to Arabic, it has been the second most spoken Semitic language with around 27 million speakers Mulugeta and Gasser (2012) primarily in Ethiopia. It is currently the official language of government in Ethiopia, and has been since the 13th century. It has been the medium of instruction in primary and secondary schools as well as the source language for a large body of historical text. 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 problems of under-represented languages; There has been lack of resources Sohail and Elahi (2018) for understanding and extracting relevant information from unstructured text. Leading to fewer online resources available to people in their everyday lives and a lack of access to Amharic source texts for scholars and other interested group of people; Now a days, researchers in linguistic and computing disciplines face difficulties because of Amharic presents sophisticated language-specific issues. Thus, any information extraction systems developed for Hebrew, Arabic, or other languages cannot address Amharic problems. Event in Amharic are expressed predominantly through verbs and nominal, but, the linguistic structure and morphological richness highly matter to apply models used for other languages. Beside this, in Amharic written text sources temporal event mentions possibly have various linguistic and syntactic representations.For instance, Geez numeral, Arabic and alphanumeric representation of Amharic temporal information could be found in the same text of Amharic document.

Because of its prominent significance of extracting events from unstructured Amharic text for high level Natural Language Processing (NLP) tasks such as textual entailment, question answering, character identification, semantic role labeling and others we are interested to tackle this problem. In this study we present a comprehensive technique for extracting events and it's temporal flags from Amharic unstructured text.

2 RELATED WORK

There are some progressive works that has been done so far on Amharic natural language analysis 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. The challenge is that temporal expressions in Amharic 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 and a very limited works in event extraction task. However, relative to other Semitic languages there are studies as in Al-Smadi and Qawasmeh (2016) which has done for Arabic language on automatic event extraction using knowledge driven approach which concentrates on tagging the event trigger instances and related entities. There has been one great contribution in Al-Smadi and Qawasmeh (2016) which links event to 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, which has more common words with Arabic. In Ramrakhiyani and Majumder (2015) solely focused on Temporal Expression Recognition in Hindi using interactive handcrafted rules. Ramrakhiyani and Majumder (2015) aims to carry out two basic goals, 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) proposed a state-of-the-art supervised machine learning approach for extracting events out of Arabic tweets. This paper 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. Smadi and Qawasmeh (2018) claim that the third task is relatively better than previous works done using similar techniques.

Another work proposed by Arnulphy et al. (2015) detects French and English Time Markup Language(ML) Events by using a combination of 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, which are event agent, event location, event trigger, event target, and event product and event time. The tools and dataset 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. Al-Smadi and Qawasmeh (2016) 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 English, there has been much less research in other languages. 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 where state of the art Event detection 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

3.1 DATA SET PREPARATION AND PREPROCESSING

Unlike other languages, Amharic language does not have any standardized annotated publically available corpora like Treebank and propbank for English. The news domain has been preferable data source because of its publically availability and rich source of information for any NLP applications such as entity extraction, event and temporal information extraction and co-reference resolution. In this study we build our own data set by scraping top websites Zehabesha[1], Satenaw[2] , Ezega[3] ,and BBC Amharic[4] which contains relevant Amharic unstructured text contents. A Python Beautiful Soup library [5]has been used for pulling data out of HTML and XML files. The scraped texts are from all domains such as economy, social, politics, technology and sport. A total of 659,848 words have extracted. Along with our own dataset we have used Amharic corpora which have been prepared by the Ethiopian Languages Research Center of Addis Ababa University in a project called *the annotation of Amharic news documents*.The corpus has 210,000 words collected from 1065 Amharic news documents of Walta Information Center Demeke and Getachew (2006), a private news and information service located in Addis Ababa. Because of Amharic language has different characters with the same meaning and pronunciation with different symbols. For Example:- (ሀ፡ሐ፡ኀ) , (ሰ፡ ሠ) and the rests have the same meaning Gasser (2011). As a result, we develop a character normalizer which enables to normalize those characters to an ordinary conceivable form. In this study well known preliminary NLP preprocessing tasks have performed include stop word removal,Tokenizer, part-of-speech tagging and morphological analyzer. Another crucial step in our preprocessing module is normalizing Amharic temporal arguments. There have been various representations of date time expressions in Amharic including Arabic, Geez and using alphanumeric characters.

3.2 EVENT DETECTION USING SUPERVISED MACHINE LEANING

In this study, supervised machine learning technique has been employed. Supervised machine learning classifiers typically predict new events, based on the given labeled training sets. Such learning algorithms deduce event properties and characteristics from training data and use these to generalize the unseen situations.

In this study, the datasets are unstructured text and documents. However, these unstructured text sequences must be converted into a structured feature space using mathematical modeling. Feature extraction for classification can be seen as a search among all possible transformations of the feature set for the best one. This preserves class reparability as much as possible in the space with the lowest possible dimensionality. Features have properties of a text that have used to provide necessary information associated to a given events and increase the confidence level of predicting a token as an event. Thus, in this study the feature extractor component 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 represents the

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

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

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

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

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

instance which contains event trigger keywords ; Whereas the off-event class refers the instance which is do not infer the event trigger keywords.

From the machine learning algorithms Naive Bayes,decision and SVM algorithms have been selected based on their widely use in text classification task such as Pranckevicius and Marcinkevicius (2017); Bilal and Israr (2016); Sarkar and Chatterjee (2015)

The models have trained on above algorithms using the labeled data-set as input. The prediction phase gets new input and detects an event as on-event and off-event classes. The best feature that has been recommended by the system which is more powerful than other feature sets to predict the event classes. The POS has found as the best syntactic feature to detect the events based on the feature selection recommendation.

3.3 EVENT AND TEMPORAL INFORMATION EXTRACTION USING RULE BASED APPROACH

A standalone rule-based approach has proposed to extract event arguments and it's temporal flags. Unlike other languages, Amharic has a subject-object verb agreement and other morphological features makes cumbersome the construction of rules. As Yunita Sari and Zamin (2010) has mentioned, construction of extraction pattern is based on syntactic or semantic constraint and delimiter based 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 , to disambiguate nominal events we need morphological features of the instances. To do so, morphological analyzer has employed to get the morphological features of event mention instances. e.g (ምሳሌ)፦(የኢትዮጵያ ህዝቦች ከዚህ በኋላ ፈፅሞ አምባገነናዊ ስርአት አያስተናግዱም ።) 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 help of a carefully constructed list of gazetteers. The word class (POS tag) and lemma of the word itself have been used as a abasement for the handcrafted linguistic rules. Different components have used to get syntactic features of words using Tree Tagger for Java (T4J), Hornmorpho for Amharic, Tigrigna and Affan Oromo. The pattern extractor has been developed based on those syntactic features. Simple rules have been applied to extract detected events. e.g / (ምሳሌ)፦ (አበበ) N (ትላንት) ADV (የገዛው) VN (በሬ) N (ሞተ) VP (።) Here the snippets of handcrafted rules have been tackled based on the POS tagger results. In addition, the formal structures are not always regular to develop stable rules. Whereas the morphological analyzer had been very helpful, because of the existence of deverbal events which have been act as ambiguous.

Some of the rules those have been applied in this 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,cit,object, grammar and preposition
3. Usually events expressed using verbs and nouns. Check the neighboring words using bi-gram language models. Because not all nouns have been events and sometimes nouns come at the beginning are the subjects or participant of the event not exactly the event.
4. Identifying the nominal events; To do so, the morphological analyzer has great role it indicates the citation of the respective nouns; i.e words which have been exactly nominal can be deverbal or non deverbal nouns, but, deverbal nouns has a citation of verbs.
5. Words which has been categorized as verbs and verb group word classes as part-of-speech and it's infinitive forms have selected as primary candidates.
6. words and phrases which indicate temporal expression have been identified by simple regular expressions
7. 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.

8. Identifying words which contains temporal keywords; Those temporal indicator keywords have been carefully built 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.
9. Example የአበበ ሰርግ ነገ ነው ። / "Abebe's wedding is torrow." from the above sentence the word **ሰርግ** is a deverbal nouns which 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**

4 Experimental Results

In this study, a total of five experiments have been conducted. The results of the experiments have been evaluated using standard information extraction performance metrics including precision, recall-measure, and ROC (Receiver operating characteristics curves).

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	**OF-Event**
	0.915	0.916	0.915	**Weighted Ave.**
LIBSVM	0.895	0.395	0.548	**ON-Event**
	0.825	0.984	0.897	**OF-Event**
	0.843	0.833	0.808	**Weighted Ave.**
J48	0.891	0.698	0.783	**ON-Event**
	0.903	0.971	0.935	**OF-Event**
	0.9	0.9	0.896	**Weighted Ave.**

The following table shows experimental results of rule based event and temporal information extraction system.

Table 2: Rule based temporal information and event argument extraction experiemntal result

Techniques	**Standard measures**		
	Precision	***Recall***	***F-measure***
Event Argument	0.976	0.952	0.959
Temporal information	0.846	0.897	0.87

5 Conclusion and future work

In this study we presented an event detection and event argument extraction with it's mentions from unstructured Amharic text. Supervised machine learning classifiers such naïve bayes, support vector machine and decision tree have used to detect events. Stand alone rule based approach have also employed to extract event arguments and it's mentions. Our system has been evaluated on our own data-set using standard evaluation metrics precision, recall and F-measure. From the study we have showed that the supervised learning technique have been informative in coverage.Where the standalone rule based approach is more accurate in sensitivity to extract event arguments and temporal mentions of those events.

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

REFERENCES

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

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

B. Arnulphy, V. Claveau, X. Tannier, and A. Vilnat. Supervised Machine Learning Techniques to Detect TimeML Events in French and English. In C. Beimann, S. Handschuch, A. Freitas, F. Meziane, and E. Métais, editors, *20thInternational Conference on Applications of Natural Language to Information Systems, NLDB 2015*, volume 9103 of *Proceedings of the NLDB conference*, Passau, Germany, June 2015. Springer. doi: 10.1007/978-3-319-19581-0_2. URL `https://hal.archives-ouvertes.fr/hal-01226541`.

M. Bilal and H. Israr. Sentiment classification of roman-urdu opinionsusing naı ve bayesian, decision tree and knnclassification techniques. *Journal of King Saud University –Computer and Information Sciences*, 28(333), July 2016.

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

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

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

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

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

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

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

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

A. Sarkar and S. Chatterjee. Text classification using support vector machine. *International Journal of Engineering Science Invention*, 4(33), November 2015.

U. Sikdar and B. Gambäck. *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 2018. ISBN 978-3-319-77112-0. doi: 10.1007/978-3-319-77113-7_22.

M. Smadi and O. Qawasmeh. 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, 2018. doi: 10.1109/SNAMS.2018.8554560. URL `https://doi.org/10.1109/SNAMS.2018.8554560`.

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

J. Tourille, O. Ferret, X. Tannier, and A. Neveol. 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, 2017.

M. F. H. Yunita Sari and N. Zamin. Rule based pattern extractor and named entity recognition: A hybrid approach. *IEEE*, 2010. doi: 10.1109/ITSIM.2010.5561392.