Jimma University

Jimma Institute of Technology

Faculty of Computing

Character Identification in Multiparty Dialogues; Identifying Mentions of the Characters from Amharic TV show Transcripts

---

By: Dawod Yimer

A Thesis Submitted to Faculty of Computing

Jimma Institute of Technology

in Partial Fulfillment of the requirements for the Degree

of Master of Science in Information Technology

Jimma University

Jimma Institute of Technology

Faculty of Computing

==========

Character Identification in Multiparty Dialogues; Identifying Mentions of the Characters from Amharic TV show Transcripts

---

By: Dawod Yimer

Advisors:

==========

1. Dr. Wondwesen Mulugeta(PhD)

2. Mr. Efrem Tadese (MSC)

June 26, 2019

Acknowledgment

All Praise to ALLAH S.W.T the Almighty, for giving me the blessing, the strength, the chance and endurance to complete this study. I would like to express my sincere gratitude to my supervisor Dr. Wondwesen Mulugeta for his time, generous guidance, patience and encouragement throughout the whole dissertation project, from which I have learned a lot regarding my title. I would also like to acknowledge Mr. Efrem Taddese for his suggestive and constructive comments which strengthen this research work. He has been working day in day out with me in the all over work. I extend my thanks to the colleague for providing me the necessary data that is important for training and testing of the prototype developed. I could not have finished this study without full support of my beloved mother, my sisters, and brothers with their blessed Du'aa. Their love, encouragements and continuous pray have make me more strong each and every day on completing this study. Finally my gratitude goes to Selomon, Getamesay and all my classmates for the discussion we have and for the ideas we share which was very helpful for the successful completion of this work.

Contents

List of Figures

List of Tables

List of Acronyms

CIMD - Character Identification on Multiparty Dialogue.

NLP Natural Language Processing.

IR- Information Retrieval.

IE- Information Extraction.

CONLL- Conference On Computational Natural Language Learning.

ACNN - agglomerative Convolutional Neural Network.

CEAF - Constrained Entity-Aligned F-measure.

MUC- Message Understanding Conference.

B^3- Bcube.

BLANC - Bilateral Assessment of Noun-Phrase Conference.

GATE - GATE: General Architecture for Text Engineering.

NER - Named Entity Recognition.

POS - Part of Speech Tagger.

ROC - Receiver Operating Characteristic.

EL - Entity Linking.

ACE - Automatic Content Extraction.

RMSProp - Root Mean Square Propagation.

IPA - International Phonetic Alphabet.

TnT - Tri grams'n' tags.

Abstract

Character Identification is an entity linking task that finds the global entity of each personal mention on multiparty dialogue. In this work, we combined coreference resolution and entity linking to accomplish a more complicated task, which is identifying the characters in multiparty dialogue. The personal mentions are detected from nominals referring to certain characters in a show, and the entities are collected from the list of all characters in those series of the show. To tackle this task, we introduce a novel coreference resolution algorithm that selectively create clusters to handle both singular and plural mentions, and also a convolutional neural network based entity linking model that jointly handles both types of mentions through multitask learning.

Our approach for tackling this problem has been to model this task as co-reference resolution followed by entity linking for assigning character labels to clusters of named entity mentions. Using an agglomerative convolutional neural network that takes groups of features and learns mention and mention-pair embeddings vastly improved the cluster purity scores for coreference resolution. By integrating the two basic tasks deep learning model was designed to identify the global personal mentions that refers a human characters.

Adjusted evaluation metrics are proposed for these tasks as well to handle the uniqueness of mentions. Three basic evaluation metrics such as Bcube, BLANC and Ceafe are practiced and each experiment shows that the new coreference resolution and entity linking models significantly outperform on the model developed. To the best of our knowledge, this is the first time that dialogue mentions are thoroughly analyzed for resolution tasks. Transcripts of TV shows are collected as corpus and manually annotated with mentions by linguistically motivated rules. These mentions are manually linked to their referents. The dataset used in this work is based on [10] and [15] format, and consists of dialogue from Two Amharic TV shows: Gemena and Sewlesew in text (transcribed) form. So that, 25 episodes of the shows are annotated, which comprises a total of 164 dialogues, 155 scenes, 1840 mentions, and 146 entities. We use common evaluation metrics to evaluated our models using those transcribed dataset, and achieve a character identification accuracy of 80.65% and an F1-score of 77.2% on the held-out episodes of the annotated test datasets, and Accuracy of 87.2% and F1-score of 63.2% on the overall dataset used in this research work.

[17]Key words: Character Identification from Amharic multiparty dialogue, Coreference resolution, Entity Linking, Deep learning approach for entity linking, Convolutional neural network approach for character Identifica-

Chapter One

1 Introduction

Character Identification is an entity linking task that finds the global entity of each personal mention in multiparty dialogue. A mention is a nominal referring to humans (e.g., **አሷ**, **እናት**), where as an entity is an an actual character in the show(e.g., **ማህሌት**). We introduce a new entity linking task, called character identification that links mentions in multi-party dialogue to their referent entities. Mentions in this task are nominal's implying humans and entities are certain characters in the TV show. Example: mentions **አንተ**, **እናት** and **ብሩክ** are linked to specific characters in the show if a referent entity is applicable.

Character identification is a preliminary task for character mining which is a task of linking mentions with the referent entities. In this research work, we introduce an entity linking task, called character identification, that maps each mention in multiparty conversation to its referent character(s). Extracting characters in dialogues is particularly hard because speakers take turns to form a conversation such that it often requires connecting mentions from multiple utterances together to derive meaningful inferences. Coreference resolution is a common choice for making connections between these mentions[1, 2, 3]. Furthermore, linking mentions to one another may not be good enough for certain tasks such as question answering, which requires to know what specific entities that mentions refer to. This implies that the task needs to be approached from the side of entity linking, which maps each mention to one or more pre-determined characters.

Amharic, the second most spoken Semitic language, is a language that poses its own challenges in natural language processing. At the moment, there is no Amharic dialogue corpus available to train deep learning models for entity linking using such mentions. Thus, a new corpus is developed by collecting transcripts of TV shows and mentions are annotated with their referent characters. It is worth pointing out that character identification is just the first step to a bigger task called character mining. Character mining will be an extended task that utilizes the results of character identification. It focuses on extracting information and constructing knowledge base associated with particular characters or any personal entities in contexts. The target entities are primarily participants, either spoken or mentioned, in dialogues. The task

tion.

can be subdivided into three sequential tasks, character identification, attribute extraction, and knowledge base construction. In this thesis work character identification is explored which is steppingstone of character mining.

The context can be drawn from any kind of document where characters are present (e.g., dialogues,narratives, novels). This research work focuses on contexts extracted from multiparty conversation, especially from transcripts of Amharic series TV shows. So that entities are mainly the characters in the shows or the speakers in conversations, that are predetermined due to the nature of the dialogue data. The study introduces a working model which aims to create a large scale dataset for character identification. This is a work to establish a robust framework architecture for annotating referent information of characters with a focus on TV show transcripts.

Character identification is distinguished from coreference resolution because mentions are linked to global entities in character identification whereas they are linked to others without considering any global entities in coreference resolution.

The original transcripts collected were formatted in plain text; we converted them into JSON so that it could be easily processed. And this structured data were then manually checked for potential errors.

The expected goal is to assign each mention to an entity, who may or may not appear as a speaker in the dialogue. In figure 1, the mention አባት is not one of the speakers in the dialogue; nonetheless, it clearly refers to a real person that may appear in some other dialogues. Identifying such mentions as actual characters requires cross-document entity resolution, which makes this work challenging.

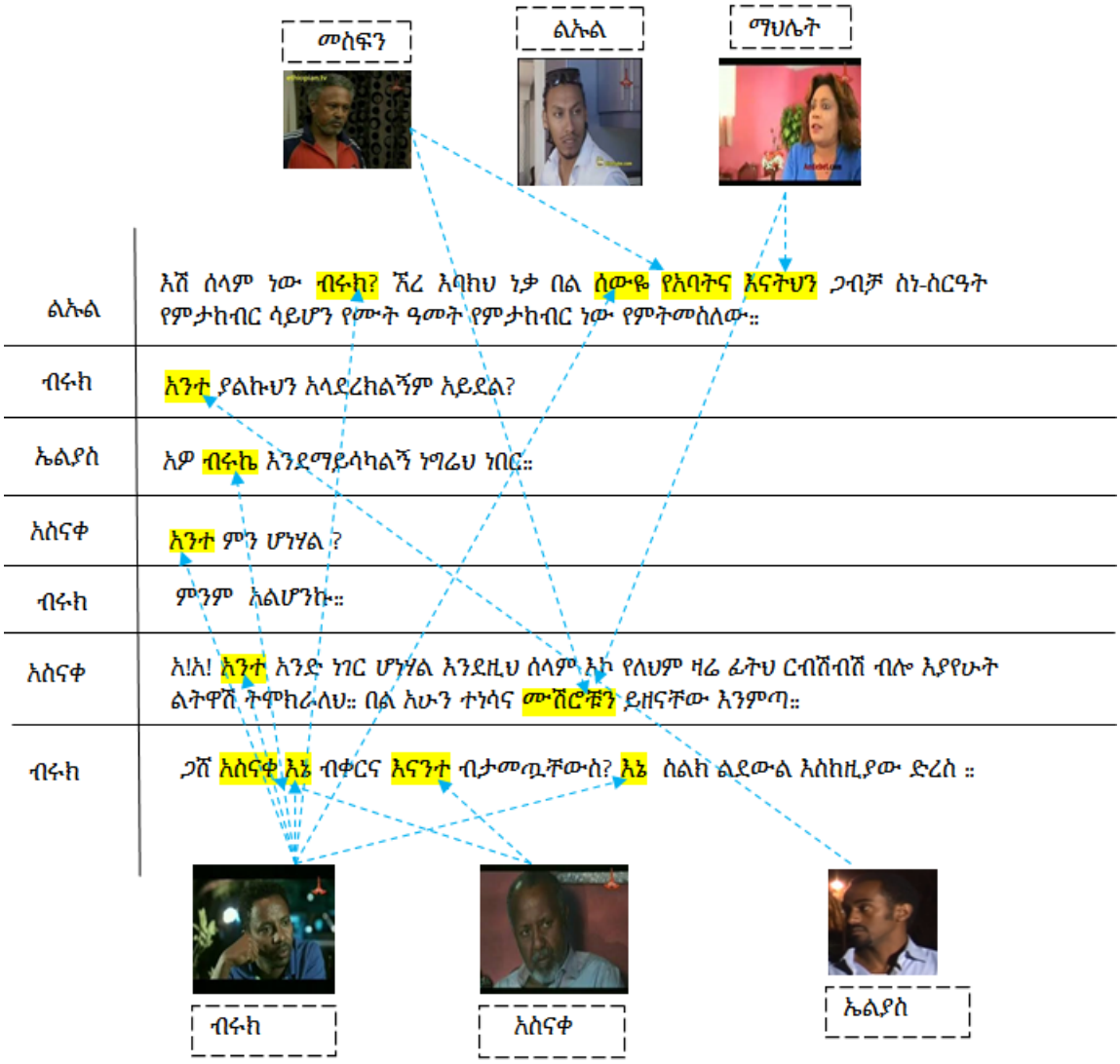

Figure 1: An example of character identification

All four speakers are introduced as characters before the conversation (**ብሩክ**, **አስናቀ**, **ልኡል** and **ኤልያስ**), and two more characters are introduced during the conversation(**መስፍን** and **ማህሌት**). The goal of this task is to identify each mention as one or more of these characters.

A mention be a nominal that refers to a singular or a collective entity(e.g., **እኔ**, **እናት**, **ሙሽሮቹ**), and an entity be the actual person that the mention refers to. Given a dialogue transcribed in text where all mentions are detected, the objective is to find the entity for each mention, who can be either active or passive in the dialogue. In Figure 1, entities such as **ብሩክ**, **አስናቀ**, **ኤልያስ** and **ልኡል** are the active speakers of the dialogue, whereas **መስፍን** and **ማህሌት** are not although they are passively mentioned as **አባት** and **እናት** in this context. Linking such mentions to their global entities demands inferred knowledge about the kinship from other dialogues.

1.1 Motivation

The motivation behind introducing and conducting research work on the task of character identification is the smart feature of machine learning that aims to provide syntactic and semantic rich information for the better understanding of Amharic as natural language text. And, character identification will serve as a stepping stone to a bigger task called Character Mining, and dialogue generation. Those will be an extended task that utilize the results of character identification [10]. It focuses on extracting information and constructing knowledge base associated with particular characters or any personal entities in contexts. And there is no any research work done in the field of mention detection or speaker identification in general character identification on Amharic multiparty dialogue (conversations). This motivates us to work on such area and will invest a drop of knowledge to contribute some features in the natural language processing discipline.

1.2 Statement of The Problem

Different researchers propose different methods that make natural language processing and information extraction practical with tremendous growth. System developed for English or any other language on a specific domain can work for other languages of the same domain with some modification. This is due to the reason that natural language processing and information extraction system has to be trained about the different nature of the language and the domain for which they are developed.

Amharic is one of widely used language in Ethiopia which has its own phonetic and grammar. In this regard, building efficient character identification system from Amharic multiparty dialogue is an essential task for introducing Amharic to NLP. It could be easy for humans to identify which character say what dialogue in a TV show or identify characters from transcripts, but it is hard for machines to correctly do so. When it comes to Amharic language, it becomes even harder because Amharic is a morphologically rich language and detection of ambiguous mention such as general and collective mention, is also much harder. Due to the nature of multiparty dialogues where speakers take turns to complete a context, character identification becomes a critical step to adapt higher-level NLP tasks (e.g., dialogue generation, emotion detection, character mining, opinion mining) to this domain. Perhaps the most challenging aspect comes from colloquial writing that consists of ironies,metaphors, or hidden pronouns.

Despite all the challenges, we believe that the output of this task will enhance inference on dialogue contexts by providing finer-grained information about individuals. But the use of character identification is road map of character mining that improves many NLP tasks. Currently, NLP for Amharic language is one of the major research areas in Ethiopia but there is nothing so far done in the area of character identification from conversational dialogue as to the best knowledge of the researcher.

By nature entities in dialogue make conversation in different scene with different names, so that linking the nominals to the referent is difficult. When we investigate Amharic pronoun there are challenges especially detecting 3rd person and 2nd person pronouns. The orthography of those pronouns have exception for detecting human mentions. Some will write **አንች** or **አንቺ** to 2^{rd} person pronoun you and **እሷ** or **እርሷ** and **እሱ** or **እርሱ** to indicate 3^{rd} person pronoun she and he respectively. The possessive pronoun **የአንቺ** or **ያንቺ**, **የአንተ** or **ያንተ** also indicate the same mention with different layout in singular 2^{nd} person female and male respectively. There are addition pronouns to indicate respect for elder peoples, such as, **እርሰዎ** and **እርሳቸው**. So that identifying those pronouns and link to corresponding referent is another challenge. Those are some of the exception that make character identification tough. Based on this research gap; we list some research questions to dig out the best way and contribute a feasible improvement to the science.
Research Questions:

- How to extract noun phrases which are actors in the dialogue without extracting every common noun?
- How to correctly identify a mention from each turns that appear in the conversation?
- How to link each mention to the corresponding referent entity?
- what types of tools are used to annotate conversational data to create dialogue corpus?
- How to disambiguate & adjudicate the errors and unknown characters in the dataset?

1.3 Objectives

1.3.1 General Objective

The general objective of this research work is to design a model for Character Identification on Amharic Multiparty Dialogues.

1.3.2 Specific Objectives

Since character identification on Amharic multiparty dialogue is a brand new task we proposed, there are some specific objectives of our work in order to tackle the task in a systematic fashion.

- Developing a corpus for character identification with the help of linguistics experts.
- Conduct literature review on coreference resolution and entity linking to better understanding of the state of-the-art character identification.
- Explore different techniques and algorithms that have been used in the area of character identification.
- Propose a model for Character identification from Amharic dialogue text.
- Develop a prototype system based on the model
- Evaluating our model using the test data.
- Infer conclusion and recommendation based on experimental results.

To the best of my knowledge, this is the first time that character identification on multiparty dialogue is formally proposed as a research work on Amharic conversational transcript corpus.

1.4 Research Methodology

Although Amharic is a language rich in drama, film, novel and narration transcripts, it does not contribute in the field of Information technology especially in the Natural Language Processing (NLP) and information extraction (IE). By taking this into consideration and reading different supportive articles, journals and books, etc, a model is designed to identify characters from multiparty dialogue using TV show transcript as a corpus. Transcripts of TV shows are collected and passed through data pre-processing techniques followed by corpus annotation task to differentiate mentions exist in the dialogue. Mentions indicating humans are annotated by rule to utilize dependency relations, named entities, and a personal noun dictionary provided by the open-source toolkit, such as NLTK, TnT tagger and SpaCy dependency parser. Rules are set to detect mentions like this; a word sequence is considered a mention if it is a person named entity, it is a pronoun or possessive pronoun, it is proper noun that indicate entities, or it is in the personal noun dictionary. To achieve the general and specific objectives of the study, the following methods are employed;

1.4.1 Literature review

Literature review is one of the most important tasks or methods needed for the successful completion of research papers. So, the researcher reviewed different kind of speaker Identification, Coreference Resolution and Entity Linking related papers, articles, conference proceeding papers, thesis documents, and books to get clear understanding on the subject area. Different kind of existing approaches to solve the character mining for English and Germany languages are also reviewed. The works done in the reviewed literature evaluated on a set of qualitative and quantitative dimensions, i.e., the amount of required data, knowledge, and expertise, as well as the interpretation of the results and the required development and execution times.

The different facts about Amharic language like the grammatical structure, the character representation and other language specific issues that are important for the research work have been reviewed and presented. This helps us to understand the nature of the language with regard to natural language processing and information extraction.

1.4.2 Design and Implementation of CIMD System

A deep learning approach is used to develop CIMD (Character Identification on Multiparty dialogue) system. It contains the document preprocessing,and post processing such as mention detection and identification, mention clustering, co-referencing, linking as a main components.

1.4.3 Data Collection

Since most of the state-of-art researches are conducted using data driven approach which are highly dependent on large amount of corpus, it is obvious why data is crucial. We choose Sewlesew and Gemena TV shows as our sources of multiparty conversation data. In this research work TV shows are selected, because the dialogue represents everyday conversation well, even if it can very well be domain specific depending on the plots and settings. The content and exchanges between characters are written for ease of inclusion. Moreover, prior knowledge regarding characters is usually not required and can be learned as show proceeds. So that we use the first 11 episodes of Sewlesew and 14 episodes of Gemena drama transcripts as our training and testing datasets.

We use data preprocessing which is a necessary and time consuming step in most information extraction and retrieval systems. In this step, data is converted to appropriate format required

for character identification process. Since we deal on dialogue data, there are irrelevant information in a document that will hinder the performance to achieve the main goal. To get ride off those undesirable or irrelevant information, we have to do some preprocessing task like Stop word removal, tokenization, character normalization, and sentence segmenting.

1.4.4 Modeling tools and Techniques

In order to develop a prototype for this research work, different appropriate tools have been selected and used. The Python programming language is used to implement the different language specific algorithms developed. Different module of python such as Tensor flow, Keras, Fast-text and Numpy are used to facilitate the completion of this research work. A statistical TnT part-of-speech tagger is used as a component for automatically assigning words with appropriate word classes and a central component for higher level NLP tools such as parser, noun phrase Chunker, named entity recognizer etc.

1.4.5 Annotation Task

Annotation consists of pieces of information added to the language data. The data may have various forms, it can be audio, video, or textual data. The added information can be external, such as the author's name, the date of recording/writing, or the type of font, this type of annotation is often called metadata. In this work we are more interested in linguistic information, such as part of speech, syntactic analysis, mention and entity annotation, adjudication and disambiguation of the textual data. Example as given in Figure 2 below, we identify mentions(አንተ ፣ እኔ ፣ አባት ፣ ሙሽሮቹ)with referent entities(ብሩክ ፣ አስናቀ ፣ ኤልያስ). Entities are speakers in the conversation(e.g., ብሩክ) or mentioned in the dialogues like መስፍን. The task also touches three linguistic annotation phases. Those are - a morphosyntactic layer, only dealing with morphosyntactic ambiguity, part of speech, inflectional and morphological annotation, - a layer dealing with syntactic relations of different degree of depth (oriented toward constituency or dependency annotations), and - a layer focused on different aspects of semantic and discourse relations such as word sense disambiguation, anaphoric relations, etc. For our system we use the morphosyntatic layer to annotate the corpus used. In this stage the core task of character identification is done especially detection of mentions are get desirable annotation format. The dataset used in the research work is annotated using Json file for easy processing of the datasets.

Example-

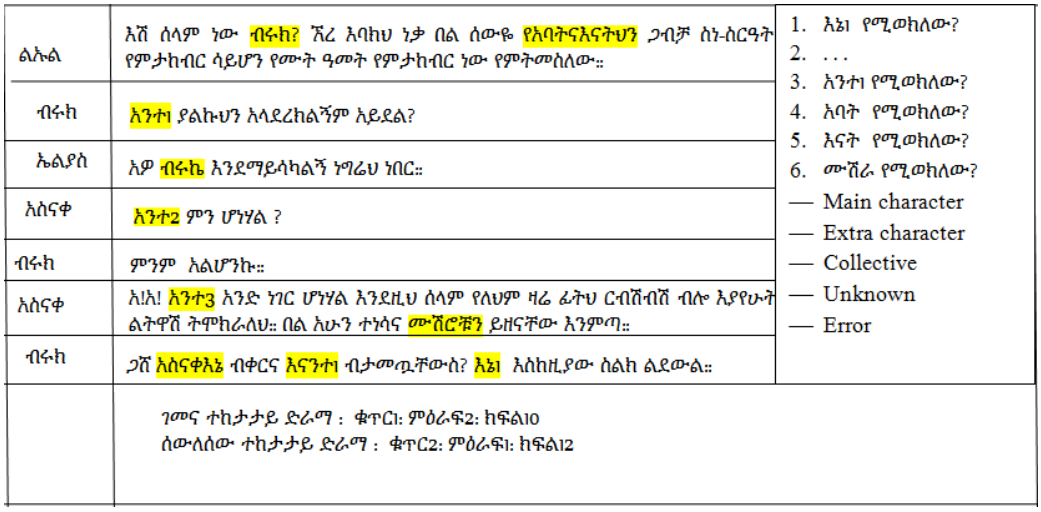

Figure 2: Annotation task

Mention is annotated with main character, scondary character, extra character, or as collective, unknown or other. Collective indicates the use of **እናንተ** or **እኛ**, that the referent entities are not identifiable. Example **እኛ ተዋንያን**. Error indicates an incorrectly identified mention that does not refer a human like **አንተ ጨረቃ**. Here **አንተ** refers the object moon, which is not human entity. Unkown refers an unknown character which is not listed as an option.(e.g., **እሱ ይርዳኝ**).

The work provides transcripts from the TV show Sewlesew and Gemena consisting of 25 episodes annotated in JSON file to preserve relevant information of nominals that refer to human entity. A subset of the first one seasons of this show was annotated for the task of character identification following the annotation guidelines suggested by [15]. As a result, a subset of the first season of Sewlesew and Gemena drama is completely annotated for character identification in our corpus. The annotated dataset are used to train models which represent individual film characters or groups of characters. Finally we use the CoNLL'12 shared task format to get advantage of speaker information, statement and utterance of each mention agreement. The converted format preserves all necessary annotation for our task.

1.4.6 Adjudication

If a transcript data contains at least one annotation disagreement we redact it into adjudication. The same task as that for the annotation is used for the adjudication, except that options for

the mentions are modified to display options selected by the previous annotators. This task is done in the annotation phase to avoid ambiguous nominals. Example one character may play in the show with two different names in different episodes such as ብሌን and ጥሩወርቅ. In such case we annotate the character as one character ብሌን.

1.5 Analysis and Evaluation

Analysis of the results and evaluation of the performance is crucial and the target for the conclusion in any system. The popular evaluation metrics used in the Natural Language processing environment are precision, recall, and F-measure. The output and performance of any NLP application relays on the quality of the input data and the techniques used. Since character identification is one of NLP application for our work we need to apply the same metrics in terms of BLANC, Bcube, and Ceafe.

1.6 Scope and Limitations

The scope of our work is developing a system used for Identifying Characters from Amharic TV show Transcripts on Multiparty Dialogues. The overall task of character identification in a multiparty dialogue setting could be divided into two major sub-tasks; those core tasks are coreference resolution and entity linking. By integrating the two modules a system is proposed which map mentions to their referent characters(entities) introduced or may not introduced during the dialogue. There are a lot of subtasks that should be done such as:

- Mention Detections
- Coreference Resolution
- Entity Linking and Character extraction

Through our investigation in character identification on multiparty conversations, we hope to assess the feasibility of the task and tackle an unexplored yet crucial branch of machine learning. Due to time and complexity; our work will not include all dialogue where mentions are present like that of novels and narratives; So our system is specific on TV show transcripts where multiple turns exist by incorporating different topics.

1.7 Application of Results

By exploring a relatively new task, character identification in multiparty dialogues, and introduce a novel perspective on approaching the task with coreference resolution and entity linking. Character identification is a preliminary task for character mining. It is roadmap to facilitate and provide entity-specific knowledge for systems like question answering and dialogue generation, Emotion Detection, Reading Comprehension. It also used to tackle tasks like speaker or speech recognition and next utterance generation. It is a stepping stone to design models to interpret implicit and explicit contexts such as interpersonal feelings or personal identification.

1.8 Structure of the Thesis

In summary this work is organized as follows. Section one presented introduction to the research background of character Identification, statement of the problem, objectives, research methodology, scope and limitation, and finally application areas of the result. Section two talks all about the theoretical background of the character Identification system as Literature including coreference resolution and entity linking tasks, components and existing approaches with their category and the necessary evaluation metrics for character identification system. The third Section is about related works done in character identification using different approaches under different languages. Section four discuss about the structure of Amharic language including common linguistic characteristics such as punctuation marks, major word classes, Amharic verb morphology, some normalization schemes etc. The fifth Section concerns all about the main components of this work. It is comprise of design of the prototype and implementation issues along with their functional operation of components and sub components. The approaches used and the algorithms developed are briefly described in this Section. Section six contains the experimentation and evaluation of this work. In the Final Section we discuss about the contribution of this work, conclusion, along with recommendation to extend the system as a future work.

Chapter Two

2 Literature Review

Character identification is a task that has been proposed and worked on before. In their work, [17] proposes different approaches to identify speakers at the turn levels for film dialogue scripts. Main focus of research work in mean time was making systems intelligent enough to accept instructions from humans in natural language processing for simple tasks like Pronoun identification, coreference resolution and entity linking. To achieve the objective of this study, incorporating works related to mention detection, Coreference resolution, entity linking and speaker Identification was necessary.

In order to understand the problem domain from the literature background and to identify clear boundary of our works from the current state of arts different books, and research works which are related to Character Identification, Coreference Resolution and Entity Linking have been reviewed. This chapter discuses about the different forms of coreference resolution , mention detection and character Identification and other related issues. Moreover, it gives brief introduction on approaches of the state-of- art used in coreference resolution and entity linking.

2.1 Character Identification

Character entities verbalize their thoughts in different ways through dialogues. The differences in their expressions, be they striking or subtle, can serve as clues to the speaker's identity as well as the nominal mention referents when they are recouped. This research work investigates the possibility of identifying characters in anonymous multiparty dialogues. Research in this area, however, has mostly been focused on acoustic features, which are absent in many situations (e.g.,online chats, discussion forums, text messages)[83]. In addition, it is commonly acknowledged that natural language texts themselves reflect the personalities of characters, in addition to their semantic content [84]. Various experiments have demonstrated significant differences in the linguistic patterns generated by different participants, suggesting the possibility to perform speaker identification with text-based data. An increasing number of large unstructured dialogue datasets are becoming available, although they comprise only the dialogue transcripts without speaker labels [85]. This work attempts to identify the 4 main characters in the dia-

logues occurring in the first seasons of the TV show, Sewlesew. The minor characters in the show are to be identified collectively as Other.

2.2 Preprocessing of Input Texts

Text corpus often consists of a raw natural language texts. A big part of the relevant information can be distinguished by some regularity found in the linguistic properties of texts. In this phase the language property i.e. its structure, the position where most relevant information in the text are located, how the co-reference between sentences is presented in the text and other language and domain specific information will be studied and implemented into different linguistic components as part of the extraction system. The following linguistic components are proved to be useful for developing an IE or NLP system as they are described in [59].

2.2.1 Morphological analysis

Before applying any learning methods to a dialogue corpus, it is common practice to perform some form of pre-processing. The aim of pre-processing is to standardize a dataset with minimal loss of information. This can reduce data scarcity, and eventually make it easier for models to learn from the dataset. In natural language processing, it is commonly acknowledged that pre-processing can have a significant effect on the results of the natural language processing system, the same observation holds for dialogue.

Although the specific procedure for pre-processing is task- and data-dependent, in this section we highlight a few common approaches, in order to give a general idea of where pre-processing can be effective for dialogue systems. Pre-processing is often used to remove anomalies in the data. For text-based corpora, this can include removing acronyms, slang, misspellings and phonemicization (e.g. where words are written according to their pronunciation instead of their correct spelling). For some models, such as the generative dialogue models discussed later, tokenization (e.g. defining the smallest unit of input) is also critical. In datasets collected from mobile text, forum, microblog and chat-based settings, it is common to observe a significant number of acronyms, abbreviations, and phonemics that are specific to the topic and userbase [53].

There are several operations which usually compose the primary step of the character identification process. The first of them is splitting a text into the fragments which are defined differently

throughout the papers from different researchers like sentences, segments or tokens. This procedure can be performed by the components named as tokenizers, segmenters or splitters.

As stated by [33], tokenization is a quite straightforward task for the texts in any European language, where the blank space between characters and punctuation indicate the boundaries of a word and a sentence respectively. But,texts like Chinese or Japanese language, where the boundaries are not so obvious this operation is not simple and requires much more effort to fulfill it. The next task within the initial processing stage is usually the morphological analysis which includes part-of-speech tagging and phrasal units (noun or verb phrases) identification. Part-of-speech tagging might be helpful to the next step which is the lexical analysis. It handles unknown words and resolves ambiguities, some of them by identifying part-of-speech of the words which cause those ambiguities.

In addition, the lexical analysis involves working with the specialized dictionaries and gazetteers, which are composed of different types of names: titles, countries, cities, companies and their suffixes, positions in a company, etc. If a word in a document is found in a gazetteer it is tagged with the semantic class the word belongs to.

After passing the preprocessing step we must identify the proper names which is one of the most important operations in the chain of information extraction, and used for the identification of various classes of proper names, such as names of people or organisations, dates, currency amounts, locations, addresses, etc. They can be encountered in almost all types of texts and usually they constitute the part of the extraction scenario. These names are recognised using a number of patterns which are called regular expressions [34]. However, usually most authors do not classify this operation as a separate task within the whole information extraction process.

2.2.2 Part-of-speech Tagger

Amharic is one of the morphological rich languages. It is a major language spoken mainly in Ethiopia and belongs to the Semitic branch of the Afro-Asiatic super family. Amharic is related to Hebrew, Arabic and Syrian. Like other Semitic languages such as Arabic,Amharic exhibits a root-pattern morphological phenomenon. A root is a set of consonants (called radicals) which has a basic lexical meaning. A pattern consists of a set of vowels which are inserted among the consonants of a root to form a stem. So we need a tool that used to tag the word class of Amharic Texts.

Amharic part of speech taggers developed by [40] for factored language modeling is one of the work designed. They use Hidden Markov Model (HMM) and Support Vector Machine (SVM) and got good performance. Another POS is developed by [42] extracts totally 23 POS tags from 300 words which is also used for training and testing the POS tagger. [41] attempted to develop POS tagger for Amharic using Conditional Random Fields which consists 10 tags by collapsing some of the categories stated by Getachew cited as [42].

The POS tagger in our work adopts a multilingual freely available tree tagger which is annotated with the corresponding word class. The Horn Morpho[48] developed by Gassar will be adapted as tagger. This will be used as input for the extraction component with morphological analyzer and Gazetteers. It is one of the important features for the machine classifier component.

2.2.3 Syntactic analysis

In contrast to POS tagging, syntax analysis, also called syntax parsing, looks beyond the scope of single words. Syntax analysis identifies syntactical parts of a sentence (verb group, noun group and prepositional phrases) and their functions (subject, direct and indirect object, modifiers and determiners). Simple sentences, consisting, for instance, of a main clause only, can be parsed using a finite state grammar. Simple finite state grammars are often not sufficient to parse more complex sentences, consisting of one or more subordinate clauses in addition to the main clause, or containing syntax structures, such as prepositional phrases, adverbial phrases, conjunction, personal and relative pronouns and genitives (possession) in noun phrases. Basically syntactic analysis is used to parse a sentence when it is needed for higher level NLP applications.

2.3 Information Extraction

As it is defined by [39] Information extraction is the task of locating specific pieces of data from a natural language document, particularly useful sub-area of natural language processing (NLP). In IE, the data to be extracted from a natural language text is given by a template may be either one of a set of specified values or strings taken directly from the document. [38] also defines IE as a form of shallow text processing that locates a specified set of relevant items in a natural language document, transforming unstructured text into a structured database. Systems for this task require significant domain-specific knowledge. So generally, IE is the process of extracting relevant and factual data from unstructured or free text.

IE usually uses NLP tools, lexical resources and semantic constraints for better efficiency. The General Architecture for Text Engineering (GATE) which is the widely known open source software system for computations related to natural language [60] defines IE as a system which analyses unstructured text in order to extract information about pre-specified types of events, entities or relationships. Information extraction involves the processing of natural language text to produce structured knowledge, suitable for storage in a database for later retrieval or automated reasoning. An active area of research for over twenty years, the community has developed several core information extraction tasks that comprise an extraction pipeline. The number of Amharic documents on the Web is increasing as many newspaper publishers started providing their services electronically. To increase the performance of extracting and exploiting the valuable information from Amharic text tools are designed.

As indicated in chapter one character identification is sub goal of information extraction, as a result many papers are used the same approach for both of them to met their goals. To achieve the state of art of character identification using TV show transcript we must pass the major task of Information extraction. we take Transcript as a raw text and, apply data preprocessing, Such as Tokenization, Stop word removal(words like ወደ ፣ እና ፣ ስለ ፣ ስለዚህ ፣ ነው ፣ ነበር ፣ ወይም ፣ እዚያ ፣ ሌሎች ፣ ከ ፣ ሁሉ ፣ ናቸው ፣ ብቻ ፣ ከላይ, . . .), Character and number normalization to feed it for the POS tagger. POS tagger classifies the word class of the preprocessing text to identify the mentions or pronouns indicating actors who is participant or mentioned in the show. In order to remove errors we apply annotation task followed by adjudication process.

2.3.1 Components of information extraction

Different authors divide the process of information extraction in different steps of different granularity, combining them into bigger stages and assigning the components of the information extraction systems to accomplish the tasks involved [35, 33, 34]. However, analysing those different approaches indicate the summarization of the general pipeline for information extraction process.

2.4 Speaker Identification

Speaker identification is a task that has been proposed and worked on before. In their work, [17] proposes different approaches to identify speakers at the turn levels for film dialogue scripts.

For each line in the transcripts, the system makes prediction of a possible speaker groups, and there are three main categories of the speaker groups, primary, secondary, and others using K-NN clustering model and Conditional Random Field(CRF) model. The work has helped to filter speaker candidates of individual utterances, however, the scope and the applications of the proposed systems are limited for character identification. Speaker identification does not concern the mentions within utterances and serves more of a documentation classification task. It does not identify the exact speakers of utterances but only the groupings of the speakers.

[25] proposed a text-based speaker identification approach using Logistic Regression and Recurrent Neural Network (RNN) to learn the turn changes in movie dialogues. Their task is fundamentally different from the task of character identification, as their main focus is on the turn changes of dialogues instead of identifying the referent mentions. The task is beneficial to us, given the scenarios where the speaker of conversations is known, thus marking it a valuable task in the study.

Considering this the deep conversation model we proposed may therefore inadvertently learn responses that remain within the same dialogue turn instead of starting a new turn. Furthermore, these dialogues contain multiple references to named entities (in particular, person names such as fictional characters) that are specific to the dialogue in question. These named entities should ideally not be part of the conversation model, since they often draw on an external context that is absent from the inputs provided to the conversation model. For instance, the mention of character names in a movie is associated with a visual context (for instance, the characters appearing in a given scene) that is not captured in the training data.

2.5 Mention Detection

A mention is a reference or representation of an entity or an object that appeared in texts. Mentions can have different mention types and the entity a mention is referencing can have different entity types. For example, in the sentence "He is Obama, the president.", "He", "Obama" and "the president" are referencing to the same object (Barack Obama himself), so they are all mentions to the entity Barack Obama[36]. This is task of detecting mentions by finding the maximal projection of every noun and pronoun. Different researchers use varieties of standard approaches to detect mentions by considering each word span is an NP or the word is a pronoun.

One of the crucial steps toward understanding natural languages is mention detection, whose

goal is to identify a reference to the mentions, whether named (አቢይ), nominal (ጠቅላይ ሚኒስተር), pronominal (እሱ, እሷ).

For Example in the sentence እሱ ጠቅላይ ሚኒስተር አቢይ አህመድ ነው::, እሱ, አቢይ and ጠቅላይ ሚኒስተር are referencing to the same object አቢይ አህመድ himself, so they are all mentions to the entity አቢይ አህመድ. This is an extension of the named entity recognition task which only aims to extract entity names.

Mention detection is application of information extraction basically person's names and Pronouns. It is necessary for many higher-level applications such as relation extraction, knowledge population, information retrieval, question answering and so on. Detecting any nominal indicating characters (singular/plural/collective) are mentions which are needed for our system. Mentions indicating humans are annotated by rule based mention detector by using dependency relations, named entities and personal noun dictionary.

It is common to divide the coreference resolution task into two main subtasks: mention detection and resolution of references [16]. Mention detection is concerned with identifying potential mentions of entities in the text and resolution of references involves determining which mentions refer to the same entity. Although Mention Detection has close ties to named-entity recognition (NER hence forth), it is more general and complex task than NER because besides named mentions, nominal and pronominal textual references also have to be identified.

Our goal is to develop a robust mention detector as the basis for developing an end-to-end coreference resolution system to accomplish the task of character Identification from Amharic multiparty dialogues. As Amharic is a less resourced language in the area of character Identification, there is a considerable lack of linguistic resources, which makes it particularly challenging to develop highly accurate tools for the chore of mention detection.

2.5.1 Linguistic Analysis of Mentions

With reference to the subtask of mention detection, in this section we establish what mentions we regard as potential ones to be included in a coreference chain. In general, we take into account noun phrases (NP), focusing on the largest span of the NP. In the case of nouns complemented by subordinate clauses and coordination, we also extract the embedded NPs of larger NPs as possible candidates for a coreference chain. The mention detection module aims to annotate all mentions in given texts. It has an annotator that accepts a TextAnnotation and give it a new

view which contains all the mentions as Constituent.

We propose the following mention classification:

1. Proper names -These are the name of the person Such as ዳውድ ፣ሳራ ፣ሰሎሞን etc.

2. Pronouns - such as he/ እሱ or she እሷ, we/ እኛ, I/ እኔ, You/ አንተ ፣ አንች ፣ አናንተ and they/ እነሱ.

   Example መልእክተኞቹ ወደ ንጉሱ ተመለሱ :: እርሱ ግን ለምን እንደተመለሱ ጠየቃቸው:: In this example the pronoun እርሱ stands by it self without embedding in other word classes called independant pronoun. Sometimes pronouns will exists by embedded with verbs.

   For example – አስናቀ የገዛው በግ ታረደ::

   In this example the personal pronoun እርሱ found in the verb ታረደ, In general the total number of independent personal pronouns in Amharic is 8. In addition to this, there are two personal pronouns (እርስዎ and እሳቸው) to show respect or politeness.

3. Possessives - For each personal pronouns there are corresponding possessive pronouns Such as የእሱ/him, የእሷ/her የእነሱ/your, የእኛ/our etc..

4. Nominal - i.e. noun phrases that have a noun as ahead Such as a man/ አንድ ሰው/ ወንድ , a woman/ አንድ ሴት, መምህር ፣ተማሪ, etc....

   For example ትላንት አንድ ወንድ እና ሴት ወደ ሱቁ መጡ::

5. Verbal nouns - This are embedded pronouns discussed above, There are Verbs that have been nominalised and function as the head of the mention, with the corresponding case marking suffix. The whole clause governed by the verbal noun has to be annotated.

2.6 Coreference Resolution

Coreference resolution is concerned with identifying mentions of entities in text and determining which mentions are referring to the same entity. Coreference resolution is the process of determining whether two expressions in natural language refer to the same entity in the scope. It is an important subtask in natural language processing systems. Coreference resolution is linking mentions to the correct coreferents [16]. It is a technique or phenomena of pointing back to an entity that has been introduced with more descriptive phrase in the text than the entity or expression which is referring back [6].

Coreference resolution process includes tasks like interpretation of pronouns, definite descriptions and others whose correct interpretation contributes greatly to the effectiveness of resolution process. In the domain of Coreference Resolution two main research paradigms have gained prominence - Knowledge based [37] employ large sets of linguistic rules to deterministically classify pairs of mentions and Data-driven methods [16] on the other hand require access to annotated data. As the data-driven approaches have successfully been applied to a number of Natural Language Processing (NLP) tasks, the availability of corpora marked with coreference information has made them favorable candidates.

Once an entity has been introduced in a text or a conversation, it may be referred to multiple times later or never again. Specially in Dialogue Linking the utterance of a character to referent mentions in conversation is particularly hard because speakers take turns to form a conversation such that it often requires connecting mentions from multiple utterances together to derive meaningful inferences. Coreference resolution is a common choice for making connections between these mentions. Our system detects mentions by finding the maximal projection of every noun and pronoun. In coreference resolution, an entity is an object or set of objects in the world, while a mention is the textual reference to an entity [23].

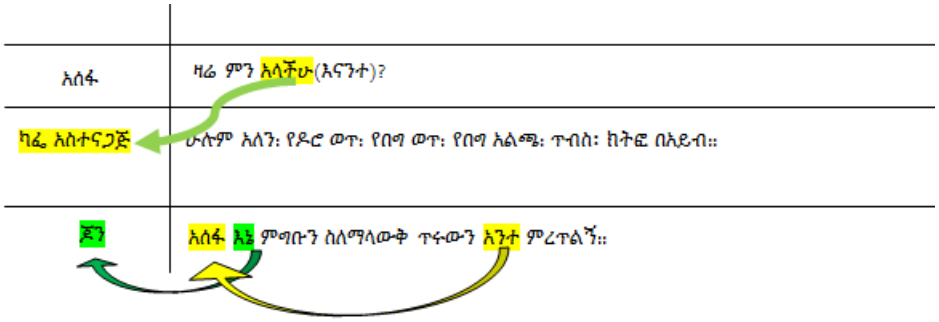

Figure 3: An example of Linking mentions to the correct antecedents (anaphoras)

Here we understand that a coreference resolution is system of linking mentions participated or mentioned in the conversation to the exact coreferents, whether they are exist within sentence(intrasentential) or cross sentence (Intersentential). For example– The verb noun **እናንተ** references the antecedent **አስተናጋጅ** which is cross sentential corefering. And the first person pronoun **እኔ** references the antecedent **ጆን**, which is intrasentential coreference.

Consider the following sentences:

ፍቃዱ ፣ እሱን አታንሽው ምንም ማለት አይደለም ብሌን የእኔ እና የአንቺ ዋናው ክፍያችን ይሔ ጥናት ነው።.

አንቺ, which refers to **ብሌን**, and **እኔ**, which refers to **ፍቃዱ** are all coreferent mentions because they refer to other entities within this document. But the mention **እሱን** is not referred to again. Such a mention is called a singleton. It is a challenging and important task to accurately separate out coreferent mentions from singleton mentions.

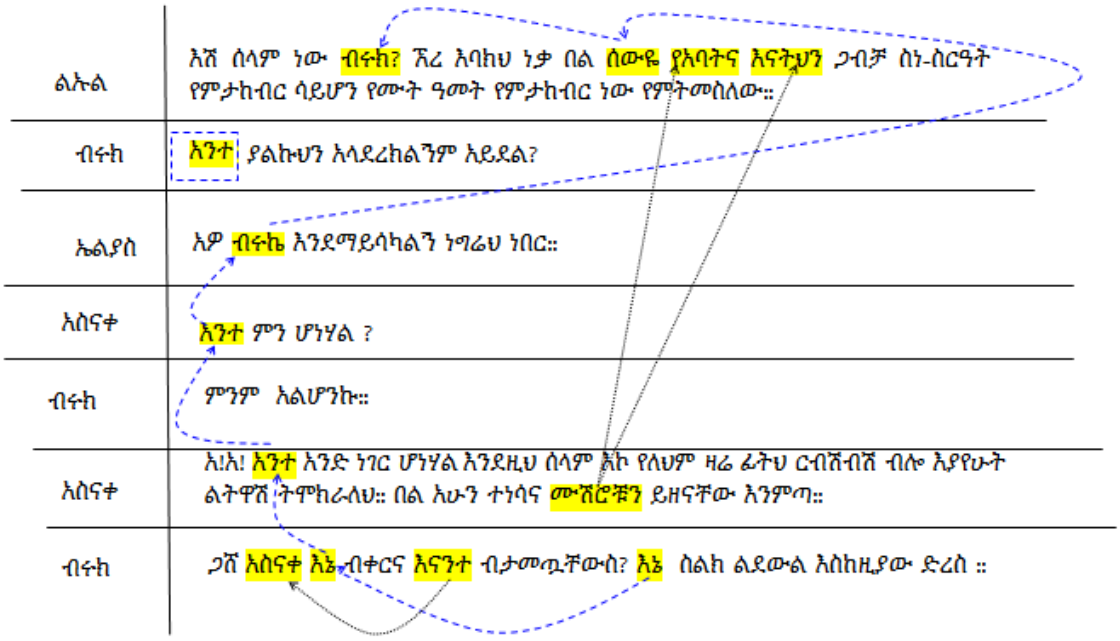

Figure 4: Example of Coreference resolution

A number of NLP tasks use mention detection as first step, for example: the core NLP problem of coreference resolution is identifying which entity a mention refers to. Identifying and filtering out singleton mentions reduces the search space and hence can improve the accuracy of downstream coreference resolution.

Coreference resolution, the task of identifying which mentions in a text refer to the same real world entity, is fundamentally a clustering problem. However, many recent state-of-the-art coreference system operate solely by linking pairs of mentions together[2, 28].

An alternative approach is to use agglomerative clustering, treating each mention as a singleton cluster at the outset and then repeatedly merging clusters of mentions deemed to be referring to the same entity. Such systems can take advantage of entity-level information, i.e., features between clusters of mentions instead of between just two mentions. [29] this work, train a deep

neural network to build distributed representations of pairs of coreference clusters. This captures entity-level information with a large number of learned, continuous features instead of a small number of hand-crafted categorical ones.

Character identification is distinguished from coreference resolution because mentions are linked to global entities in character identification where as they are linked to one another without considering global entities in coreference resolution [10]. In this paper, writers augment and create corpus for character identification from multiparty dialogue, and propose an end-to-end deep-learning system that combines rule based models for coreference resolution and entity linking to tackle the task of character identification. The ability to link coreferring noun phrases both within and across sentences is critical to discourse analysis and language understanding in general.

2.6.1 Forms of anaphora

Anaphora can be divided into pronominal anaphora, lexical noun phrase anaphora, noun anaphora, verb anaphora, adverb anaphora, or zero anaphora based on the form of the anaphora or syntatic category of the anaphora[6].

- Pronominal anaphora: is the type of anaphora that is used in many research papers and as a result it is the most common and known type of anaphora. Such type of anaphora are pronouns[27].

  Example,

  Today I met Sara and her friend.

  In this example, "her" is the anaphor and "Sara" is its antecedent.

  Personal, possessive and demonstrative pronouns both singular and plural are categorized under pronominal anaphora indicated in[6].

- Lexical noun phrase anaphora: This type of anaphora occurs when the anaphor is categorized as definite noun phrase while the antecedent is proper name.

  Example,

  አቢይ አህመድ የኢምሬቱን አልጋ ወራሽ ተቀብለው በቢሯቸው አነጋገሩ። ጠቅላይ ሚኒስትሩ ኢትዮጵያ ለኢንቨስትመንት አመች የሆነ አየር ንብረት እንዳላትና ከኢምሬት ባለሃብቶች ጋር የመስራት ፍላጎት እንዳላትም ለልዑሉ ገልጸውላቸዋል።

  In the above example the definite noun phrase **ጠቅላይ ሚኒስትሩ** is the anaphor whereas the proper name **አቢይ አህመድ** is its antecedent.

In addition to these there are also verb and adverbs that corefer in different situations, in our work the entity extracted is a human mentions we need all the nominal and noun phrases indicating human beings.

2.7 Entity linking

Entity Linking (EL) is a central task in information extraction given a textual data, identify entity mentions and link them to the corresponding entry in a given Knowledge Base (KB, e.g. Wikipedia or Freebase). Recent research on entity linking (EL) has introduced an over-plus of promising techniques, ranging from deep neural networks to joint inference. But despite numerous papers there is surprisingly little understanding of the state of the art in entity linking. In this work entity linking can seen in the side of linking each personal mentions to the global referent entities rather than linking with predetermined knowledge bases. So that the task involves aligning a textual nominals of a personal mention to the referent named entity that represents the mentioned character, if it is present.

Entity linking is a natural language processing task of determining entities and connecting related information in context to the entities [4]. In the previous stage, coreference resolution groups mentions into clusters, but it does not assign character labels to the clusters, which is required for character identification. It is distinguished from named entity recognition, because named entity recognition is identifying and classifying tokens into six predefined classes person, organization, location, time, title, and others(non-named entity tokens), whereas Entity Linking is recognizing entity mentions in text and linking them to the corresponding entries in the knowledge base.

Potential applications include information extraction, information retrieval, and knowledge base population. However, this task is challenging due to name variations and entity ambiguity even very harder when we practice it in Amharic. Entity linking is limited to recognizing entities for which a target entry exists in the reference knowledge base; each entry is a candidate. It is assumed that the document provides sufficient context for disambiguating entities. Thus, an entity linking model is required that takes the mention embeddings and the mention-pair embeddings generated by the machine learning algorithm and classifies each mention to one of the character labels.

Entity Linking aims to associate concepts to their corresponding Wikipedia pages. The task of

entity linking, such as Wikification, primarily emphasizes on disambiguating the referred entity of mentions in discourses [7, 8]. For instances, given a sentence:

“ሚኒሊክ ሆስፒታል በአገሪቱ የህክምና አገልግሎት ታሪክ ውስጥ ፈር ቀዳጅ ሲሆን የግንባታ ሥራው ከ1890 ዓ. ም እስከ 1891 ዓ. ም ተከናውኖ የተወሰኑ ክፍሎቹ ለአገልግሎት በቅተዋል”, a system should identify “ሚኒሊክ” as ሚኒሊክ ሆስፒታል rather than the Ethiopian emperer, ዳግማዊ ሚኒሊክ.

Such branch of entity linking takes the advantage of massive Wikipedia corpus for Entity Disambiguation. It is distingushed from entity linking that finds the distinct one-to-one or one-to-many relations between mentions to concepts. Entity disambiguation aim to clarify the connections to concepts when the constructs are confusing due to their similar names or traits[11, 24]. A critical step to achieve this goal is to link named entity mentions appearing in text with their corresponding entities in a knowledge base, which is called entity linking.

Figure 5: An example of Linking mentions to their referent entity

Other than linking mentions to their Wikipedia concepts, entity linking have also done on domain-specific information using local context [15]. Instead of using Wikipedia which serves as a universal knowledge base that contains factual information of various domains, entity linking systems can be trained on corpus of specific domains, such as law and medicine, to create in more versatile systems that help with computations and automation's in other fields.

In the case of character identification, a system will be trained on genre specific corpus, transcripts of multiparty conversations, rather than a domain-specific corpus. However; it can easily be trained on TV show conversations that occur around a particular character or topic, and thus making it domain-specific to the character. In the menagerie of tasks for information extraction, entity linking is a new beast that has drawn a lot of attention from NLP practitioners and researchers recently. Entity Linking, also referred to as record linkage or entity resolution, involves aligning a textual mention of a named-entity to an appropriate entry in a knowledge base, which may or may not contain the entity.

Unlike English language where capitalization is the major clue for recognizing Named Entities, Amharic lacks this feature and in addition to it, there are so many other problems faced by researchers while designing entity linker system for Amharic language such as ambiguity in names, lack of standardization and spelling, non-availability of large gazetteer, scarcity of resources and tools etc. A named entity may also have multiple surface forms, such as its full name, partial names, aliases, abbreviations, and alternate spellings.

Example– ኃይለስላሴ or ሃይለስላሴ, An entity linking system has to identify the correct mapping entities for entity mentions of various surface forms. Such type of ambiguities must be normalized in the preprocessing step of morphological analysis. On the other hand, an entity mention could possibly denote different named entities. For instance, the entity mention " ጸሀይ" can refer to the star at the center of the Solar System, or a name of person named ጸሀይ in the transcript. So our system must identify the person ጸሀይ from the star ጸሀይ during the analysis of morphological semantics.

2.8 Approaches to Character Identification

The dominant category of Character Identification approaches can be seen in terms of coreference resolution and entity linking and include data-driven or statistical, Hidden Markov Model, knowledge based and hybrid approaches, Machine Learning, and Rule based approaches.

Hidden Markov Model is a generative model which incorporates double stochastic process. First stochastic process generates the sequence of states where as second stochastic process is responsible for generating the sequence of observations from the sequence of states.
Rule based is another approach which has been used for designing named entity recognition system [30, 31]. In rule based approach, rules are crafted for every entity and named entities are recognized using these rules. The detailed descriptions of each approach are explained below and in chapter 3 of this document.

2.8.1 Rule Based Approach

This approach is used to identify mentions indicating humans by annotating using rule-based mention detector, which utilizes dependency relations, named entities, and a personal noun dictionary provided by the open-source toolkit, NLP4J. Indicated in [10] mentions extracted from text applying a rule as a word sequence is considered as a mention if it is a person named entity, it is a pronoun or possessive pronoun excluding it, or it is in the personal noun dictionary. A mention is not identified by the detector, the approach considered it as a "miss". If a detected mention does not refer human character(s), it is considered an "error". Such approach is the state-of-the art of Character identification to accomplish the goals in terms of coreference resolution and entity linking, and scores good performance.

2.8.2 Machine Learning Approach

The first step in the machine learning phase was to somehow convert these mentions into a feature vector. Then we could feed these features to a variety of machine learning algorithms and see how they perform. We started by trying to build our own word encodings using one-hot vectors and a skip-gram model with a window size of 2. This caused our feature vectors to be completely massive and extremely sparse. This would have taken ages to train so we ended up utilizing word2vec which was more efficient.

Naive Bayes- This classifier is statistical classifiers based on Bayes theorem. Naive Bayes is one of the best text classification techniques with various applications in personal email sorting, document categorization, email spam detection, sexually explicit content detection, language detection and sentiment detection etc. Bayesian performed the absolute worst of all of the machine learning algorithms tried, though this is typical, another factor could be that all of the

word2vec attributes are not independent of each other, meaning the independence assumption on which Naive Bayes relies doesn't hold.

SVM - though cited in the literature as being one of the most promising algorithms for this task still performed rather poorly, this may be due to the class not being linearly separable with the given features and the kernel function not being sufficient to separate them.

2.9 Evaluation metrics

In any information extraction tasks the evaluation are expressed in terms of precision, recall and F-measure. The computation of those evaluation metrics are based on the notion of false positives, true positives, false negatives and true negatives. The values which is extracted correctly are true positives whereas false positives are wrongly extracted values. On the other hand true negatives refer values, which is relevant but not extracted and false negatives (false drop) refers the values which is not important and not extracted. In order to evaluate the performance of mention detection and coreference resolution systems, there are three mainstream evaluation metrics used MUC, B^3, and $CEAF_e$.

MUC [44] concerns the number of pairwise links needed to be inserted or removed to map system responses to gold keys. The number of links the system and gold shared and minimum numbers of links needed to describe coreference chains of the system and gold are computed. Precision is calculated by dividing the former with the latter that describes the system chains, and recall is calculated by dividing the former with the later that describes the gold chains.

In simple terms, Precision (P) is the proportion of correctly extracted entities ($N_{correct}$) to the total number of extracted entities ($N_{response}$) (the ratio between number of needles in a hand and number of needles and straws in the hand). Recall (R) is the proportion of correctly extracted entities ($N_{correct}$) to the total number of entities which are extracted manually (N_{key}) (the ratio between number of needles in the hand and total number of them in the haystack). Thus,

$$\mathrm{P}=\frac{N_{correct}}{N_{response}}, \qquad \mathrm{R}=\frac{N_{correct}}{N_{key}}$$

In order to combine precision and recall, the F measure was introduced in one of the MUCs. Thus,

$$\mathrm{F}=\frac{(\beta^2+1)PR}{\beta^2 P+R}$$

B^3 [45] metric computes precision and recall on a mention level, instead of evaluating the coreference chains solely on their links. System performance is evaluated by the average of all mention

scores. Given a set M that contains mentions denoted as m_i. Coreference chains S_{mi} and G_{mi} represent the chains containing mention mi in system and gold responses. Precision(P) and recall(R) are calculated as below:

P(mi) = $\frac{|S_{mi}| \cap |G_{mi}|}{|S_{mi}|}$, R(mi) = $\frac{|S_{mi}| \cap |G_{mi}|}{|G_{mi}|}$

$CEAF_e$ [47] metric further points out the drawback of B^3, in which entities can be used more than once during evaluation. As result, both multiple Coreference chains of the same entity and chains with mentions of multiple entities are not penalized. To cope with this problem, CEAF evaluates only on the best one-to-one mapping between the system's and gold's entities. Given a system entity S_i and gold entity G_j. An entity-based similarity metric $\varphi(S_i, G_j)$ gives the count of common mentions that refer to both S_i and G_j. The alignment with the best total similarity is denoted as $\Phi(g^*)$. Thus precision(P) and recall(R) are measured as below.

P=$\frac{\Phi(g^*)}{\sum_i \varphi(S_i,S_i)}$, R = $\frac{\Phi(g^*)}{\sum_i \varphi(G_i,G_i)}$

Chapter Three

3 Related Work

Research on dialogue systems has also involved considerable efforts on speaker identification includes both rule-based and machine-learning approaches. Most Speaker Identification systems support English and European language texts from different domains using variety of techniques. For many Semitic languages, there is a research gap particularly in the area of dialogues. Character Identification on multiparty dialogue from Amharic TV series has not been done still, therefore our work is the first in this particular IE application. Due to the variation of the language structure adapting the existing techniques done on Latin and Spanish language might not be applied for Semitic languages like Amharic.

This thesis introduces a new model that uses mention detection, coreference resolution, and entity linking techniques, to achieve the task of character Identification. Our task is different from general and traditional entity linking because we are working in a limited domain, namely, TV show transcripts, and we are including pronouns as entities, whereas in the past, pronouns were never considered as entities in entity linking in order to accomplish the task of character Identification.

3.1 Data Driven Approaches

During the past decade, several areas of speech and language understanding have witnessed substantial discoveries from the use of data-driven models. In the area of dialogue systems, the trend is less obvious, and most practical systems are still built through significant engineering and expert knowledge. Nevertheless, several recent results suggest that data-driven approaches are feasible and quite promising.

Data Driven based Natural Language Processing (NLP) approach includes the statistical and all forms of machine learning models. In machine learning model, there are lot of different algorithms are trained on various task of Information Extraction such as Artificial Neural Network Algorithms, Deep Learning Algorithms, Bayesian Algorithms, Clustering Algorithms, and many other algorithms categorized as supervised, unsupervised and semi supervised machine learning models. In Data Driven approach a Machine learning models are homogeneous to functions

that will predict some output for a particular given input. Generally, in order to generate a machine learning model we need Sample Data with target attribute given and Machine Learning Algorithm chosen according to the nature of target attribute.

The process of training a Machine Learning (ML) model involves providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The models are used to detect pattern or formula from the dataset, that we give to actually apply to new data. However, statistical methods have a long-standing focus on inference, which is achieved through the creation and fitting of a project-specific probability model. The model allows us to compute a quantitative measure of confidence that a discovered relationship describes a true effect that is unlikely to result from noise. Furthermore, if enough data are available, we can explicitly verify assumptions (e.g., equal variance) and refine the specified model, if needed. Below we define some basic algorithms of machine Learning models related to character Identification.

3.1.1 Stanford Entity-Centric System

A machine learning system used in character Identification is the Stanford entity-centric system [1]. The system takes an ensemble-like statistical approach that utilizes global entity-level features to create feature clusters, and it is stacked with two models. The first model, mention pair model, consists of two tasks, classification and ranking. Logistic classifiers are trained for both tasks to assign probabilities to a mention. The former task considers the likelihood of two mentions are linked.The latter task estimates the potential antecedent of a given mention. The second model, entity centric coreference model, aims to produce a final set of coreference clusters through learning from the features and scores of mentions pairs. It operates between pairs of clusters unlike the previous model. Iteratively, it builds up entity-specific mention clusters using agglomerative clustering and imitation learning. This approach is particularly in alignment with my task, which finds groups of mentions referring to a centralized character. Furthermore, it allows new models to be trained with the developed corpus. This would give insight on whether a task can be learned by machines and whether a generalized model can be trained to distinguish speakers in all contexts. The system evaluated on the English portion of the 2012 CoNLL Shared Task dataset and show that it improves over the current state of the art.

3.1.2 Hobbs Naive Syntax-Based Algorithm

In this algorithm the morpho syntactic information plays an important role both in filtering certain types of interpretation (gender, binding constraints) and in determining preferred interpretations (subject assignment, parallelism). Several algorithms have been developed that incorporate these types of syntactic knowledge for anaphora resolution, in particular for the resolution of pronouns.

The earliest and best-known of these syntax-based algorithms is the pronoun resolution algorithm proposed by [49]. This algorithm, still often used as a baseline, traverses the surface parse tree breadth-first, left-to-right, and then going backwards one sentence at a time, looking for an antecedent matching the pronoun in gender and number. Although [49] developed his algorithm already in the 1970s, it can in fact be seen as a very early indicator of a research paradigm shift, moving towards shallow processing strategies that are chiefly based on less knowledge-rich sources of evidence in the field of speaker identification.

3.1.3 Amharic Anaphora Resolution Model using Knowledge-poor Approaches

[6] have proposed Amharic anaphora resolution model using knowledge poor anaphora resolution approach. His approach uses low levels of linguistic knowledge like morphology to build anaphora resolution systems avoiding the need of complex knowledge like semantic analysis, world knowledge and others. The proposed model takes Amharic texts as input and preprocesses to tag the texts with word classes and various chunks. The researchers used constraint and preference rules and other techniques like WordNet ontology that helped them solve anaphora resolution problem. Some of the constraint rules they used were language dependent while others were language independent. The main contribution of this work is resolving hidden anaphors. Identification of verbs and then extracting their morphological properties is needed to extract hidden anaphors inside verbs. This component identifies anaphors which are personal pronouns hidden inside verbs. The constraint rules used in this work are gender, number and person agreements. They are presented as follows:

1. Anaphors and antecedents should agree in gender information
2. Anaphors and antecedents should agree in number information
3. Anaphors and antecedents should agree in person information.

Application of preference rules

Preference rules are rules that give more preference to antecedents that satisfy the rule while give nothing or negative value that don't satisfy the the rule. Preference rules used in Temesgen's work are discussed as follows.

- Subject place preference
- Definiteness
- Recency
- Boost pronoun
- Mention frequency

3.1.4 Machine Learning models for character identification

Artificial neural network was first proposed in the 1940s as a learning model that simulates neuron activities in human brain [21]. The model is made possible for learning tasks, like regression, classification, and prediction, with the later advancements in computer technology and the introduction of back-propagation [22]. Artificial neural network has shown successes in various applications, particularly in learning non-linear and complex features. Models tend to outperform traditional statistical models when trained on large datasets. Different architectures of the neural network have since been introduced as the concept gains its popularity.

One particular architecture of neural network, used in Character identification system is the Convolutional Neural Network (CNN). This model takes groups of features incorporating several dialogue aspects, feeds them into deep convolution layers. Convolutional neural network model dynamically generates mention embedding and mention-pair embedding, which are used to create the cluster embedding that significantly improve the performance of the entity linking model.

In the work [15] proposes a deep learning approach to coreference resolution and entity linking for character identification. The paper introduces the agglomerative convolutional neural network that takes groups of features and learns mention and mention-pair embeddings for coreference resolution. They also propose another neural model that employs the embeddings learned and creates cluster embeddings for entity linking.

Entity links are then mapped to each referent separately by cluster embeddings. This method takes into consideration, 20 labels viz. Top 9 characters and an unknown label. It emphasizes on the intuition that the coreference resolution accuracy depends upon the size of the clusters. The combined implementation of cluster and mention embeddings bettered the singular use of mention embeddings in terms of accuracy. Although, agglomerative CNN being an incremental feature approach has its fair share of advantages in terms of word embeddings, even if the existing approach still lacks the handling of plurals and collective nouns. They use CNN, because It has the advantage of inexpensive computation compared to other neural network architectures. To remove the ambiguity of unknown mentions writers introduce three labels of differentiation as a constraint. Those are;

1. General: Mention used in reference to a general case.
2. Generic: Mention referring to an unidentifiable entity.
3. Other: Mention referred to insignificant singleton entity.

The work is relatively new task, character identification on multiparty dialogues, and introduce a novel perspective on approaching the task with coreference resolution and entity linking.

To evaluate performance of the system Episode and scene-level models are developed separately for all three systems using the same dataset for training (TRN), development (DEV), and evaluation (TST) sets. All system outputs are evaluated with the MUC[44], B^3[45] and $CEAF_e$ [47]. The coreference resolution model shows comparable results to other state-of-the-art systems. The Stanford and the Harvard systems reported μ scores of 65.73% and 64.21% on the CoNLL'12 dataset, respectively. Their entity linking model significantly outperforms the previous work, showing the F1 score of 86.76% and the accuracy of 95.30% for character identification.

[17] used the k-nearest neighbor (k-NN), Naïve Bayes' (NB) and Conditional Random Field (CRF) classifiers to identify the speakers from the dialogue scripts at turn levels. A considerable research on speaker identification from speech dialogues has been conducted. Even though, the work reports speaker identification task from film dialogue which includes textual features only. They have used a text corpus of film scripts, extracted from the Internet Movie Script Database (IMSDb) archive and annotated with speakers.

Researchers have attempted to recognize the speakers based on their style of speaking as a preference rules. Those are numbers of words per sentence as well as part of speech (POS).

They use a supervised learning technique for the task of speaker identification or character identification from film scripts. Each script has been dissected into 9:1 ratio where the first portion is used as training data and the latter is used as test data. In a film script, there are typically a few frontline characters who account for most of the turns in it. Therefore, to achieve this feature, they apply a heuristic approach of calculating the average number of turns per speaker. The researchers have been adopted to identify the major and minor speakers in a given script. The speakers that have turns higher than the above computed average are termed as major whereas the values lower than the above computed average are labeled as OTHERS. Therefore in their classification problem, the set of class labels is comprised of the names of all the characters that are identified as major speakers by the above heuristic and the label OTHERS. This is used to avoid overfitting by clubbing all the minor characters from the training set.

To evaluate the system they use 135 scripts from IMSDB. The average accuracy for the baseline system was found to be 16.76%. NB performs better than the baseline with an average accuracy of 23.59%. For the k-NN method, different values of k were selected and performs better than the other two in terms of average accuracy. In research paper the scheme only classifies the speakers as primary character, secondary character and other. It doesn't show the mentions that are not speaker but they are participant of the conversation. And also the system doesn't show any mechanism of linking the turn by turn utterance of the speakers.

Serban and Pineau [25] proposed their text-based speaker identification approach using Logistic Regression and Recurrent Neural Network (RNN) to learn the turn changes in movie dialogues. This work used scripted dialogues to identify turn-taking and differences in speakers, where the actual identities of the speakers are irrelevant. They define two classification tasks. The first is a turn taking binary classification task, and the second task is a 6-way classification task, which refer to as the speaker identification task. This work automatically infers turns and speakers from scripted dialogues. Using Movie-Scriptolog corpus as dataset, writers got feasible result to estimate turn taking and speaker identity with high accuracy. The result suggest that recurrent neural networks offer a rich paradigm for building models for speaker characterization. The paper is good and the model used are promising, which outperforms an accuracy of 69.47% in the combination of RNN and Word2Vec models. Shortcoming we identified from this work is that a simple RNN model is unable to perform speaker identification based on textual data.

The work proposed by [74] uses a convolutional neural network model for text-based speaker

identification on multiparty dialogues extracted from the TV show, Friends. In this research work they attempt to identify speakers in dialogues using their speech patterns as captured by transcriptions and idiolect styles to the TV show. The improvements produced by the consideration of neighboring utterances in the CNN's prediction process indicate that contextual information is essential to the performance of text-based speaker identification. The paper enables identification of speaker, where the names of the speakers are associated with their own utterances, a novel attempt in text-based speaker identification. This paper investigates the possibility of identifying speakers in anonymous multiparty dialogues using multi-document convolutional neural network, and shows an accuracy of 31.06% and a macro average F1 score of 29.72, exhibiting promising performance on the text-based speaker identification task. The experimental results confirm the value of textual features in speaker identification on multiparty dialogues is promising. One of the challenges, that we try to identify from this research work, is the large number of misspellings and colloquialisms in the dataset as a result of the mistakes in the human transcription process and the nature of human dialogues affects the over all performance of the system.

3.1.5 Statistical Entity Mention Detection

A number of EMD approaches are based on machine learning. Almost all of them recast the task as a sequence labeling problem. Different algorithms have been proposed recently: Hidden Markov Models [55], Maximum Entropy classifiers [56], Support Vector Machines [61], and Conditional Random Fields [58]. They do not use sets of hand-written rules, but need annotated corpora to derive them. These systems appeal to a trainable classifier to decide for each word whether or not it is part of a mention. State-of-the-art EMD systems include a variety of features, representing different sources of information. The most commonly used features such as, lexical features derived directly from tokens, e.g. the word itself, prefixes and suffixes (the first n characters at the start/end of the word), capitalization information, etc. Features derived by using other NLP techniques, e.g. sentence splitting, POS tagging, lemmatization, text chunking, parser, etc. External information given by gazetteers (lists of proper names of persons, locations, organizations), WordNet, Wikipedia, etc.

Clark and Manning [1] introduces an entity centric system using mention pair models as features. Agglomerative clustering is used to build coreference chains formed by merging pairs of clusters at each step. A key aspect of any incremental coreference system is its local deci-

sions. Using this to full advantage, costs are assigned to each action which are in turn trained using a cost-sensitive classification. For the ranking model, the current mention is matched to the candidate antecedents simultaneously competing with each other. The resultant prediction model depends upon the previous actions which violate the IID(independent and identically distributed) assumption of statistical learning.Hence, imitation learning is used to classify whether a particular action is the one the expert policy would take at the current state.

The work in[5] proposes a new approach to coreference resolution using distributed word representations. An incremental coreference system is defined which acts as a feed forward neural network for mention clusters rather than mention pairs. The usage of mentin pairs do not enforce transitivity and therefore rely only on local pairwise information to make coreference decisions Mention clusters on the other hand facilitate previous coreference decisions to inform the latest ones. This is an extension to Intelligent Tutoring Systems where features are created between mention clusters using the pairwise probabilities of the mention pair model. This is extended by consideration of all features from vector representations of mention pairs to produce cluster level features. The actual benefits of deep learning on coreference are the lack of hand engineered features. This is leveraged by Clarke in creating a simple feature set which outperforms state- ofthe- art approaches.

Sam Wiseman et al [2] presents a mention ranking model for coreference resolution. It emphasizes on anaphoricity detection and antecedent ranking with respect to learning feature representations. The training model using backpropagation is preceded by a pre-training segment comprising of two tasks viz. anaphoricity detection and antecedent ranking. The mention ranking model is trained with the slack-rescaled max-margin training objective which facilitates separation between highest scoring true and false antecedents of the current mention.

A major challenge of coreference systems is resolving an anaphoric mention that has no previous head term. This paper intuitively evaluates the possibility of overcoming this challenge by means of non local decision making. It provides a conclusion that pronouns may not be the only coreferent mentions causing these errors and therefore a local model can also be tweaked with respect to a loss function to achieve this.

3.2 Knowledge Driven Approaches

Also known as a rule based approach. Generally, rule-based means manually designed sequence of words, or part-of-speech, or another way of representing words in a sentence, and match these sequences with the text. we could call this an "expert system". Rule based approach is always based on a large set of handcrafted rules and a small lexicon to handle the exceptions.

3.2.1 Stanford Multi-Sieve System

The modern heuristic approach to the development of coreference systems is grow vigorously. Many such systems are still being developed; indeed, the Stanford Deterministic Coreference Resolution System based on the so called Stanford Sieve approach[51, 52] a version of the precision first approach. The Stanford DCR was the best performing system at the CoNLL 2011 coreference shared task.

The architecture of the Stanford Deterministic Coreference Resolution (DCR) proposed by [52] is articulated around two main stages: a high recall (and highly precise) mention detection component based on Stanford CoreNLP, a high quality NLP pipeline; and a coreference resolution stage consisting of 10 components called sieves and also ordered from the highest precision to lowest precision. The operation of the coreference resolution stage is based on the following principles:

- The system keeps track of entities (i.e., the discourse entities of systems such LaSIE(Large Scale Information Extraction) sets of mentions that have already been determined to belong together), while keeping track of properties such as number, gender, and named entity type.

- Each sieve operates on entities rather than mentions, and on the whole discourse, rather than on a sentence or a paragraph at a time.

- For sieves that compare two mentions, the system keeps track of a representative mention in each cluster (typically the first one, as it is usually the longest, whereas subsequent mentions are shortened or only expressed as pronouns). In addition to this, the Stanford DCR has ten Sieves component start from Speaker Identification sieve, which used to identify speakers and the last sieve Pronoun resolution, that resolve the pronouns indicating referent mentions.

In this paper they have covered in detail most of the best-known non-statistical approaches to anaphora resolution. As seen discussing the Stanford Deterministic Coreference System, such approaches still achieve state-of-the-art performance, and very few new ideas about the linguistic features playing a role in anaphora resolution have been introduced in more recent systems. The Stanford Deterministic Coreference System achieved the highest score of 59.5 at the CONLL2011 coreference shared task and, the approach has been extensively evaluated on a variety of other datasets, always achieving state-of-the-art results.

Another Stanford multi-pass sieve system incorporated by [10] is used to provide a baseline of how a coreference resolution system performs on character identification task.
It is deterministic rule-based coreference resolution system based on entity-centric, precision-ranked rules. The system is composed of multiple sieves of linguistic rules that are in the orders of high-to-low precision and low-to-high recall. Information regarding mentions, such as plurality, gender, and parse tree,is extracted during mention detection and used as global features. Intra-document coreference resolution clusters together textual mentions within a single document based on the underlying referent entity. Mentions are usually noun phrases (NPs) headed by nominal or pronominal terminals. Pairwise links between mentions are formed based on defined linguistic rules at each sieve in order to construct coreference chains and mention clusters.

3.2.2 Rule-based mention detector

Rule-based approaches aim at creating a large number of hand-crafted rules typically, regular expressions exploiting the context around the mention to identify both the boundary and type. For datasets with Automatic Content Extraction(ACE)[12] style annotation guide lines, focusing on specific semantic types, POS tags, syntactic features, and orthographic features, such as capitalization, are some of the most common features. Such systems often have a simple structure, they are generally easy to understand, yet difficult to design. This is because rules must be correctly written and optimized by human experts, and to port on different domains new rules need to be written. Rule based systems show some examples or regular expressions for PERSON mentions that are used by a typical Mention Detection system.

If the annotation guidelines assume that all the noun phrases in a document (and sometimes even other chunks) are considered mentions, the mention extraction modules rely on a combi-

nation of parsing trees and lists of Named Entities. Unfortunately, not many studies go into details on this technique.

A work done by [54] proposes a parsing-based mention detection algorithm that used to identify a human mentions from entities other than human beings. The work sets the rules as a constraint to consider certain entity is a mention as (1) Named entities are considered mentions if and only if they correspond to sequences of parsing constituents by prohibited any partial overlap between Named Entities. (2) Possessive pronouns are mentions if they are not parts of named entities. (3) Noun phrases (including non-possessive pronouns) are candidate mentions if they are not parts of named entities. The set of candidate mentions is filtered to eliminate pairs of NPs with the same head noun embedding NPs are discarded. The remaining NPs are added to the set of mentions. For example,
guitarist Daniel Ash and Daniel Ash, get aligned and become one mention.

The model also uses dependency relations, named entities, and a personal noun dictionary,which is provided by the open-source toolkit,NLP4J. In this model the general rules are set down as follows: a word sequence is considered a mention if it is a person named entity,it is a pronoun or possessive pronoun excluding it or it is in the personal noun dictionary chosen from Freebase and DBpedia. Plural (e.g., we, them, boys) and collective (e.g., family, people) nouns are discarded. The system evaluated for their mention extraction component alone from the SemEval-2010 Task 1 and CoNLL-2011/2012 evaluation campaigns. At SemEval-2010, the algorithm achieved an F-score of 78.1% on the English data for the mention detection subtask. At CoNLL-2011, most systems relied on parsing-based mention detection techniques, showing EMD F-scores of 65–75%.

The key characteristic of the rule-based system developed by [50] for MUC-6 is that it builds on the finite-state grammar developed for the FASTUS system (Finite State Automaton Text Understanding System), versions of which participated in several editions of MUC. The anaphora resolution system described in [50], therefore, approximates appositional/copular constructions and (originally syntactic) salience within the pattern-based approach in FASTUS. The work points out that these approximations lead to a loss of precision with respect to perfect or good parses used in other systems; however, the loss due to this approximation approach is not as large as one could imagine, and the most obvious cases where a syntactic analysis would help (reflexives and disjoint reference filtering) are relatively infrequent. In this system Mention de-

tection takes mentions (template entities) as input which, besides their span, already have some linguistic features that are useful for subsequent processing. These features are the preference rule of the model, and those are;

- The determiner or pronoun type (definite, indefinite, or pronominal);
- Grammatical number (singular or plural, or a modifying cardinal expression);
- Head string and modifiers of the mention's noun chunk;
- A semantic class that is assigned based on the head, and comes from a shallow hierarchy;
- Sentence and paragraph positions;
- Information about the enclosing text region (headline or main text);

Information about the enclosing text region is used to model the assumptions for text-region accessibility that were employed in the MUC-6 annotation, namely that a mention in the headline region can be coreferring with a mention from the text, whereas mentions in the text region can be resolved to any preceding mention within the text region. The system is scored 59% recall and 72 % precision (F=0.65) in the official MUC-6 evaluation, which was the best overall performance.

3.2.3 Joint Mention Extraction and Coreference Resolution

In character identification most state-of-the-art systems combine mention detection and Coreference resolution in a pipeline architecture. First, a set of mentions is created for a document and each item of this set is assigned various linguistic properties; second, the set is partitioned into equivalence classes, or entities.

The work done by [62] propose a joint model for entity detection and coreference resolution. Their approach involves joint inference at both testing and training steps. Unlike other work, the paper do not restrict the scope of the study to mention classification, but aim at the full-scale mention detection task. They design framework called Learning as Search Optimization that extends a standard search algorithm to incorporate learning. This work involves finding text spans that correspond to an entity, identifying what type of entity it is (person, location, etc.), identifying what type of mention it is (name, nominal, pronoun, etc.) and finally identifying which other mentions in the document it corefers with.

The difficulty lies in the fact that there are often many ambiguous ways to refer to the same entity. Consider the following example:
Bill Clinton$^{NAM}_{per-1}$ gave a speech today to the Senate$^{NAM}_{org-2}$. The president$^{NOM}_{per-1}$ outlined his$^{PRO}_{per-1}$plan for budget reforms to them$^{PRO}_{org-2}$.
There are five entity mentions in these two sentences, each of which is underlined (the corresponding mention type and entity type appear as superscripts and subscripts, respectively, with coreference chains marked in the subscripts), but only two entities: (Bill Clinton, The president, his) and (the Senate, them). The mention detection task is to identify the entity mentions and their types, without regard for the underlying entity sets, while coreference resolution groups a given mentions into sets.

In this work, the Entity Mention Detection(EMD) component has identified, among others, two mentions, the Senate (name, organization) and them (pronoun, person). The coreference component in a pipeline system has a low chance of resolving them to the Senate, due to the semantic class disagreement. This happens because the EMD component has prematurely assigned an incorrect label to an ambiguous mention. The error could have been avoided if a system was able to postpone such decisions, making use of the information provided by the coreference component at the later stage. Using the (ACE-2004) dataset for evaluation purpose, the approach outperforms a pipeline system with a scores of 89.1. So that the paper is informative to identify the mentions and entities especially when there are many mentions corefering each other. I learn many things to design my system.

3.3 Hybrid Approach

Hybrid Approach combines the feature of two or more models together in order to accomplish specific task in Natural Language Processing, Information Extraction,or any other field of studies.

Henry (Yu-Hsin) Chen & Jinho D. Choi in their paper [10] introduces a subtask of entity linking, called character identification, that maps mentions in multiparty conversation to their referent characters. Transcripts of TV shows are collected as the sources of corpus and automatically annotated with mentions by linguistically-motivated rules. These mentions are manually linked to their referents through crowdsourcing. They show the distinguishing feature of different concepts as; Character identification is distinguished from coreference resolution because mentions

are linked to global entities in character identification whereas they are linked to one another without considering global entities in coreference resolution. Furthermore, this task is harder than typical entity linking because contexts switch of topics more rapidly in dialogues.

In this work, mentions that are either plural or collective nouns are discarded, and the knowledge base does not get populated from the context dynamically. Mentions indicating humans are annotated by rule-based mention detector, which utilizes dependency relations, named entities, and a personal noun dictionary provided by the open-source toolkit, NLP4J. In this paper Character identification is tackled as a coreference resolution task, which takes advantage of utilizing existing state-of-the-art systems even if it may not result the best for their task since it is more similar to entity linking. Most of the current entity linking systems are accustomed to find entities in Wikipedia which is not the case in character identification from multiparty dialogue, that is task of determining entities and connecting related information in context to the mentions in the knowledge base.

The paper evaluated in three evaluation metrics such as B^3, $CEAF_e$ and MUC with CONLL12 dataset format using Stanford multi-pass sieve and Stanford entity-centric models, and yield close performance when run out-of-box. It is interesting because both rule-based and statistical models give similar baseline results. Stanford multi-pass sieve is a rule-based system whereas Stanford entity-centric uses its pre-trained model.

The approach proposed by [15] and [10] use an Agglomerative CNN and a rule based model for the task of character Identification respectively. Both of them use a dialogue data as a corpus, and fulfill the state-of- art of character Identification from multiparty conversations. Here [10] propose a rule based mention detection to detect all the human mentions found in the conversation by setting some set of instruction as a preference rule, a Stanford Multi sieve system to tackle the task of coreference resolution, and Stanford entity centric system to link the human mentions to the correct referents found in the knowledge base.

However, [15] trains their system on the same corpus with some modification using machine learning approach performs better comparing to [10]. The system is evaluated in terms of Train (TRN), Development(DEV) and Test(TST) dataset and fulfills the state-of-art of coreference resolution and entity linking. Both of the systems are designed to train on TV show transcripts written in English which is completely different from Amharic. So adapting some of the best feature of the systems especially from the papers proposed by[15], we design system used to

identify characters from Amharic multiparty dialogues using deep learning models. To the best of my knowledge, this is the first deep learning model that performs character identification from Amharic multiparty dialogues.

Chapter Four

4 The Amharic Language

This chapter discusses the different issues about the Amharic language that is needed in the development of character identification system for Amharic TV show transcripts. It begins by introducing the Amharic writing system, the numerals and punctuation marks in Amharic language and Amharic pronouns, nominal and possessives are also described.

4.1 Introduction

Amharic is a Semitic language, a descendant language of the original Semitic (Proto-Semitic) language, spoken some 6,000 years ago[87]. In spite of its number of speakers, Amharic has remained less known than other languages of the Semitic family, such as Arabic, Hebrew, Aramaic, and Akkadian. Regardless of, the importance of Ethiopia in the ancient Middle Eastern world as a juncture for trade between Asia Minor, Egypt, India, and the rest of Africa, the Semitic languages of Ethiopia have, relatively speaking, been little studied.

Amharic language, which is categorized under Semitic languages family, is a national language of Ethiopia (i.e. it is official working language of the federal democratic republic of Ethiopia). It is the second most spoken language among Semitic language families in the world, next to Arabic. Amharic is one of the Ethiopian Semitic languages, which are a sub-grouping within the Semitic branch of the Afro Asiatic languages.

The actual size of the population of speakers of different languages in Ethiopia must be based on estimates [66] analysed the Ethiopian census from 1994 and indicated that more than 40% of the population then understood Amharic. It is spoken in many parts of Ethiopia, a country of more than 73.92 million people by the as reported in the 2007 census from Ethiopia central statistics agency [67]. Then, Amharic become the official or working language of several of the states/regions within the federal system, including Amhara and the multi-ethnic Southern Nations, Nationalities and Peoples region. Outside Ethiopia, Amharic is the language of millions of emigrants (notably in Egypt, USA, Israel, and Sweden), and is spoken in Eritrea [68].

4.2 Amharic phonology

In this section, the consonant and vowel sounds attested in the present corpus are described. Amharic has its own typical phonological and morphological features that characterize it. Some of the features of Amharic phonology that gives the language its characteristic sound when one hears it spoken are the weak indeterminate stress; the presence of glottalic, palatal, and labialized consonants; the frequent gemination of consonants and central vowels; and the use of the automatic epenthetic vowel. Those descriptions of Amharic phonetics and phonology are to be found in [75, 76]. As regards its dialect situation, Amharic is in great difference from one place to another. The speech of Addis Ababa has emerged as the standard dialect and has wide currency across all Amharic speaking communities. The most divergent dialect is that of Gojjam province, though the Menz and Wollo varieties also show their own marked features, especially in phonology. Amharic has the 32 distinctive consonant sounds, or consonant phonemes and 7 distinctive vowel phonemes[43].

Manner of Articulation	**voicing**	**Place of Articulation**					
		Labials	**Dentals**	**Palatals**	**Velars**	**Labiovelar**	**Glottals**
Stops	Voiceless	ፐ [p]	ተ [t]	ቸ [ʧ]	ከ [k]	ኩ [kʷ]	አ [ʔ]
	Voiced	በ [b]	ደ [d]	ጀ [ʤ]	ገ [g]	ጉ [gʷ]	
	Glottalized	ጰ [p']	ጠ [t']	ጭ [tʃʼ]	ቀ [q]	ቈ [qʷ]	
Fricatives	Voiceless	ፍ [f]	ሰ [s]	ሸ [ʃ]			ሀ [h]
	Voiced	ቨ [v]	ዘ [z]	ዠ [ʒ]			
	Glottalized		ጸ [s']				
	Rounded						ኍ [hʷ]
Nasals	Voiced	ም [m]	ን [n]	ኘ [ɲ]			
Liquids	Voiced		ል [l] ር [r]				
Glides	Voiced	ው [w]			ይ [j]		

Figure 6: The 32 Amharic consonant phonemes [46]

Variants of Constants

Labiovelar consonants Amharic has labialized consonants such as $[b^*], [k^*], [t^*]$, in which lip-rounding anticipates the release of the consonant. These usually occur with the vowel a, and

may be considered sequences of a consonant and w: bw, kw, tw.

y and w insertion When vowels meet, one of the glides y or w may be inserted. Typically w is inserted if the first vowel is o or u, and y is inserted if the first vowel is i or e.

Long consonants All the consonants except z and h may be long, sustained single-articulations with the approximate duration of a two-consonant sequence. Thus the word አለ'(he is present) is different from አለ(he said), and ዋና'(principal) different from ዋና (swimming). Obstruent consonants (stops, affricates and fricatives) may exhibit a three-way contrast at the same point of articulation between voiceless, voiced and glottalized. The latter, called sometimes ejectives, produce a sharp sound and are analogous to the emphatic consonants of Arabic and other Semitic languages. Another distinctive trait of the consonantal system of Amharic is the existence of labialized gutturals. All consonants, except h and the glottal stop, may occur in a long or geminated form.

Vowels: Amharic has the seven vowels shown as below. The vowels are written here with phonetic symbols of the International Phonetic Association appropriate for their qualities[80]. The vowel system is remarkable by the occurrence of three central vowels and, also, by its symmetry. The table shows the articulation of the vowels as tongue position in terms of three degrees of height and three of frontness.

	front	central	back
High	ኢ [i]	ኧ [ɨ]	ኡ [u]
Mid	ኤ [e]	ኧ [ə]	ኦ [o]
low		ኣ [a]	

Figure 7: The seven Amharic vowels

Variants of Vowels

Vowel elision When Amharic words are constructed from their parts with the result that i or e are adjacent to another vowel, ኧ and ኧ are typically elided, or omitted.

Vowel insertion When Amharic words are assembled from their parts, the high central vowel ኧ often appears to separate resulting consonant sequences which are disallowed by the requirements of Amharic word structure. The vowel is termed Epenthetic, and typically appears when prefixes and suffixes combine with stems, for example to separate y-n and gr of Yngr –> yingir

'he tells'. If vowels precede and follow the sequence of consonants, the Epenthetic Vowel is unneeded and absent, as in አይናገሩም('they don't tell). Probably most occurrences of Amharic ኦ may be considered epenthetic.

Voiceless vowel The word-final vowel of the suffix of past tense verbs voiceless - pronounced approximately as if whispered. As noted above, Amharic script does not distinguish between consonants that are not followed by a vowel and consonants that are followed by the high central vowel ኦ. Amharic phonology that gives the language its characteristic sound when one hears it spoken. Moreover, vowels are a crucial for proper pronunciation in Amharic speech synthesis.

4.3 History of Amharic Writing

The writing system of Amharic is an adaptation of the writing system evolved some 2000 years ago for Ge'ez, the Semitic language of the ancient Ethiopian kingdom of Axum, also known as ETHIOPIC. Ethiopic writing has been adapted for use in writing a number of modern Semitic Ethiopian languages, notably Tigrinya and Amharic, and for Cushitic languages of Ethiopia as well.

Ethiopic- In Ethiopia, by contrast, writing underwent popular adaptation, with graphs became less angular and formal. Ethiopic writing, in forms regularly distinct from South Arabian, is well evidenced in a number of lengthy commemorative inscriptions attributed to Ezana, during his reign. Ethiopic writing began to include representation of vowels, not as separate symbols as in Greek but as extensions and other modifications of consonant symbols. Ethiopic writing began to be written from left-to-right, opposite that of most other Semitic writing systems. This may have been an influence of Greek, a left-to-right writing system also known in Axum and employed on inscriptions and coins.

It is not known whether the Ethiopian innovation of vowel-writing was a unique invention of the Ethiopians, or perhaps an inspiration from Indian Brahmi writing, which somewhat similarly represented vowels, and at a somewhat earlier time. Knowledge of vowel-writing could have come to Ethiopia with regular trade known to have existed between cities of Western India and Aksum ([69, 70, 71]). Such local evolution of the system might have been expected, however, certainly if Ethiopians had even superficial knowledge of the Brahmi system, as the natural result of a tendency for stylistically variant graphs of a single consonant to become associated with some vowel [73]. The record, however, seems to show a rather abrupt appearance of vowel

representation. The Ethiopic numbers seem to have been adapted from those of Greek, in which the numbers were letters in the sequence of the Greek alphabet.

The original and basic consonantal graphs were reanalyzed as the consonant plus the most common vowel, so 0 was reanalyzed as ba, and in Amharic a became g, The structure of the Ethiopic writing system in its adaptation as Amharic writing. Because of their shared history as adaptations of Sinaitic writing, Greek and Ethiopic (and Amharic, which is derived from Ethiopic) have similarities apparent in a number of comparisons of graphs in the Greek and Ethiopic columns.

Amharic- After the decline of Axum around 600 AD, Ethiopic writing is absent in the archaeological record, but reappears in manuscripts, on parchment, from about 1250 AD, in use for writing Christian religious literature and chronicles of the kings of Ethiopia. About 200 years later, a slightly modified form of Ethiopic began to be used for writing Amharic. This Amharic adaptation of Ethiopic consists largely in development of a regularized system of punctuation, and invention of a set of graphs for the series of palatalized consonants, which were not regularly used in Ge'ez. Subsequently, Ethiopic was adapted for use to write Tigrinya, and Amharic writing was adapted for use to write other Ethiopian languages. Amharic writing today fulfills all the needs of modern literate society, for letters, novels, poetry, legal decrees, newspapers, and magazines.

The Amharic word for the graphs or letters of Amharic writing is called Fidel. Restricting the term letter to graphs of the Greek or Latin based alphabets; here we shall refer to the Amharic graphs by their Amharic name, Fidel. In modern Ethiopic script each syllograph (syllable pattern) comes in seven different forms (called orders), reflecting the seven vowel sounds. The first order is the basic form; the others are derived from it by modifications indicating vowels. There are 33 basic forms, giving (7*33 = 231) syllographs, or Fidel (Fidel, or alphabet in Amharic, refers both to the characters and the entire script). Unlike Arabic and Hebrew, Amharic is written left-to-right in its own unique script (inherited from the clerical Ge'ez. Despite its large number of graphs in comparison to a European-language alphabet, the Ethiopic Amharic writing system is quite efficient and systematic.

Amharic characters were represented by computer using Unicode. Unicode provides a unique number for every character, no matter what the platform, program, or language. Ethiopic characters (fidel - ፊደል) have more than 380 Unicode representations including punctuation and

special characters (U+1200- U+137F). The problems of writing Amharic by typewriter, the large number of graphs and the inappropriately small size of typescript, have been solved by computer mediated writing.

4.4 Amharic Punctuation Marks and Numerals

Punctuation is the use of spacing, conventional signs, and certain typographical devices that helps to understand, or to read a handwritten or printed text. It is the practice, action, or system of inserting points or other small marks into texts, in order to aid interpretation; division of text into sentences, clauses, etc., by means of such marks.

In Amharic, there are different punctuation marks used for different purposes. In writing system of Amharic, there are nine punctuation marks. Those are; (1) word separator- Literally (two dots or two points ፡). This mark is used to separate words. Since the rise of digital publishing the mark is primarily applied today in a handwritten document. (2) Preface colon(አስረጂ ሰረዝ ፦) is used following clarification of a certain subject. It will preface validation statements and examples that support the clarification. Introduces speech from a descriptive prefix. In transcribed interviews, after the name of the speaker whose transcribed speech immediately follows, compare the colon in western text. (3) Colon or comma(ነጠላ ሰረዝ ፥ or ፣) often used to separate comparative and sequential list of names, phrases, or numbers as well as to separate parts of a sentence that are not complete by themselves. A special note of explanation is needed here. While the Unicode standard refers to ፥ as "ETHIOPIC COLON" the correlation with "colon" from Western practices as the name implies can given the wrong impression over the functional role of the symbol in writing. (4) semicolon (ድርብ ሰረዝ ፤) is used to separate equivalent main phrases in one idea. Even though it is not placed at the end of a paragraph, it can be used to separate sentences with similar ideas in a paragraph. (5) Question mark(ሦስት ነጥብ ፧) used at the end of the questioning sentence. In modern writing "?" is preffered. (6) Full stop (period)(አራት ነጥብ ።)- This mark is placed at the end of the sentence that describes the completeness of an idea. (7) Section Mark used to divide sections or subsections; generally three or more used together on a line of their own. (8) Paragraph Separator(፠) is used to conclude the final paragraph of a section in Literature of Ethiopic. Adopted into Ethiopic writing practices are enclosing punctuation such as parenthesis, brackets, single and double quotation marks and guillemets. Expressive punctuation such as question mark, exclamation point, inverted excla-

mation mark, and ellipsis are also incorporated into Ethiopic practices.

In Amharic, numbers can be represented using either the symbols of Arabic number system or the symbols of the Ethiopic number system. The Amharic numbers are not used for mathematics, but often in dates, and page and chapter numbers of books. For mathematics, the Arabic numerals known in English writing (1, 2, 3, etc.) are used. There is no zero in Amharic number system. Amharic numbers are written as in Table 1.

Amharic Numerals				
1. ፩	6. ፮	20. ፳	50. ፶	100. ፻
2. ፪	7. ፯	21. ፳፩	60. ፷	146. ፻፵፮
3. ፫	8. ፰	22. ፳፪	70. ፸	210. ፪፻፲
4. ፬	9. ፱	30. ፴	80. ፹	1000. ፲፻
5. ፭	10. ፲	40. ፵	90. ፺	2007. ፳፻፯

Table 1: Examples of Amharic numbers Systems

In Amharic, fractions and ordinals have their own way of representation. Table 2 shows fraction and ordinal representations in Amharic. As numbers are one of the information that is extracted in this research work its representation in letters in Amharic text is important and it is presented in the following table.

Fraction	Amharic Representation	Ordinals	Representation
$\frac{1}{2}$	ግማሽ	1^{st}	አንደኛ/ ቀዳማዊ
$\frac{1}{3}$	ሲሶ	2^{nd}	ሁለተኛ/ ዳግማዊ
$\frac{1}{4}$	ሩብ	3^{rd}	ሶስተኛ/ ሳልስ
$\frac{2}{3}$	ሁለት ሶስተኛ	4^{th}	አራተኛ/ ራብዕ
$\frac{3}{4}$	ሶስት አራተኛ	−−	–
$\frac{1}{10}$	አስራት	8^{th}	ስምንተኛ
$2X$	እጥፍ	9^{th}	ዘጠነኛ

Table 2: Examples of Amharic fraction numbers Systems

4.4.1 Problems in Amharic Writing System

Homophonous Fidels - It is common many languages have some consonant sounds, which written with more than one alphabet. For example In English language k, c, and q may all be

read [k], or s and c both read [s] in different situations. Even if Amharic has less of this than English, but there are some fidels which have the same sound. Such as (አ ፣ ዐ, ሀ ፣ ሐ ፣ ኀ ፣ ኸ, ሰ ፣ ሠ, and ጸ ፣ ፀ) represented as .?, h, s and s' respectively in fidel. So that there is 2 ways to write .?, 4 ways to write h, 2 ways to write s, and 2 ways to write s'.

- **Two ways to write ?a** - In Ge'ez and presumably earlier in Amharic, there were phonetically similar consonants .?a, a glottal stop written አ, and ዐ. The two different ?'s are distinguished by name as follows: allefu ?a (አሊፉ አ) and aynu ?a (ዐይኑ ዐ).

2 ways to write .?							
	a	u	i	a	e	(i)	o
?a	አ	ኡ	ኢ	ኣ	ኤ	እ	ኦ
?'a	ዐ	ዑ	ዒ	ዓ	ዔ	ዕ	ዖ

Table 3: Two ways to write ?a

- **Four ways to write h** - In earlier Amharic there were four different h-like consonants: a voiceless glottal fricative, phonetic [h], ሀ, a voiceless pharyngeal fricative [h] (ሐ), a voiceless velar fricative ኀ[x], a second voiceless velar fricative ኸ [x], which arose as a weakened or lax pronunciation of voiceless velar stop ኽ [k]. Three of the h's are named as follows: haletaw ha (ሃሌታው ሀ), hameru ha(ሐመሩ ሐ), and bizuhanu ha (ብዙኃኑ ኀ). The newest h, ኸ, so far lacks a standard name. Most frequent is the set of ሀ.

4 ways to write h							
	a	u	i	a	e	(i)	o
h	ሀ	ሁ	ሂ	ሃ	ሄ	ህ	ሆ
h	ሐ	ሑ	ሒ	ሓ	ሔ	ሕ	ሖ
h	ኀ	ኁ	ኂ	ኃ	ኄ	ኅ	ኆ
h	ኸ	ኹ	ኺ	ኻ	ኼ	ኽ	ኾ

Table 4: Four ways to write.h

- **Two ways to write.s** - In the past Amharic, there were two different s like consonants: ሰ and a similar sound, perhaps a lateral fricative ሠ. The two different s's are distinguished by name as follows: nigusu se(ንጉሱ ሰ) and isatu se(እሳቱ ሠ). More frequent is the set of ሰ. Amharic spelling prefers one of the set of homophonous Fidel in particular words, and the spelling of the cognate word in Ge'ez, if there is one, which is authoritative. For example, sillase 'trinity' is written ሥላሴ and not ስላሴ, ስላሜ, or ሥላሜ

2 ways to write s							
	a	u	i	a	e	(i)	o
s	ሰ	ሱ	ሲ	ሳ	ሴ	ስ	ሶ
s	ሠ	ሡ	ሢ	ሣ	ሤ	ሥ	ሦ

Table 5: Two ways to write.s

- **Two ways to write.s'** - In Amharic prehistory there were different consonants s'. The two different s' fidel are distinguished by name as s'elotu s'e(ጸሎቱ ጸ) and s'ehayu s'e(ፀሐዩ ፀ). More frequent is the set of ጸ. Again Amharic spelling prefers one or the other in particular words. The spelling of the equivalent word in Ge'ez, if this is known, is preferred. For example, s'ehay 'sun' is written as ፀሐይ and not ጸሐይ.

2 ways to write s'							
	a	u	i	a	e	(i)	o
s'	ጸ	ጹ	ጺ	ጻ	ጼ	ጽ	ጾ
s'	ፀ	ፁ	ፂ	ፃ	ፄ	ፅ	ፆ

Table 6: Two ways to write.s'

Compound Words in Amharic Language- In the Amharic writing system, an ambiguity is often observed regarding the representation of compound words. Some compound words are used as a single word in some instances (either by fusing the two words or by inserting a hyphen between them) and as two separate words at other instances. For example- አድስ አበባ, is compound word representing a single entity when we consider it as single word, or አድስ and አበባ separately represents two different entities አድስ 'new' and አበባ 'flower' or the proper name of person አድስ and person አበባ. The same for, ገብረሚካኤል, ሃይለሥላሴ, etc.

4.5 Amharic Morphology

Amharic language is morphologically compex due to the Semitic languages nature. A significantly large part of the vocabulary consists of verbs, and like many other Semitic languages, Amharic has a rich verbal morphology based on triconsonantal roots with vowel variants describing modifications to, or supplementary detail and variants of the root form. Subject, gender, number, etc., are also indicated as bound morphemes on the verb, as well as objects and possession markers, mood and tense, beneficative, malfactive, transitive, dative, negative, etc. Amharic nouns (and adjectives) can be inflected for gender, number, definiteness, and case, although gender is usually neutral. The definite article attaches to the end of a noun, as do conjunctions, while prepositions are mostly prefixed.

4.5.1 Word Categorization in Amharic

The words in Amharic are categorized under five basic categories by Baye Yimam [65] that uses the morphology and position of the word in Amharic sentence as criteria. These five categories are ስም (noun), ግስ (verb), ቅፅል (adjective), ተውሳከ ግስ (Adverb) and መስተዋድድ (preposition).

A word will be categorized as a noun, if it can be pluralized by adding the suffix ኦች ፣ዎች and used as nominating something like person and animal. It is used as a subject in a sentence. Pronouns, which were considered as independent category in the previous works by the linguistics professionals is categorized under nouns after considering the unique nature of the language as the earlier linguists just adopt the English language structure for Amharic language. In this research work we only considers the personal pronouns, because In character identification the main task is identifying the human mentions that participate or mentioned as actor in the dialogue. In Amharic there are 8 independent personal pronouns. These are; እኔ ፣አንተ ፣አንቺ ፣እሱ ፣እሷ ፣እኛ ፣እናንተ ፣እነሱ, in addition to these there are two a personal pronouns indicating respect such as እርሰዎ ፣እርሳቸው.

Independent Personal Pronouns In Amharic	Personal Pronouns In English	Independent Personal Pronouns Romanized	Person Information	Number Information	Gender Information
እኔ	I	Ine	1st Person	Singular	Male/Female
አንተ	You	Ante	2nd Person		Male
አንቺ	You	Anci	2nd Person		Female
እሱ	He	Isu	3rd Person		Male
እሷ	She	Iswa	3rd Person		Female
እኛ	We	Ina	1st Person	Plural	Male/Female
እናንተ	You	Inante	2nd Person		Male/Female
እነሱ	They	Inesu	3rd Person		Male/Female
እርስዎ	You	Irswo	2nd Person	Singular	Male/Female
እሳቸው	He/She	Isacew	3rd Person		Male/Female

Figure 8: List of Amharic independent personal pronouns (Adapted from [57])

The polite pronouns are typically used for older persons and others, such as priests, to whom respect should be shown, and also for adults whom one does not know. Feminine forms may be used by boys and young men as affectionate reference to close friends.

A word which can be placed at the end of a sentence and which can accept suffixes and, which is used to indicate masculine, feminine, and plurality is classified as a verb. Amharic has a complex verbal morphology, using prefixes, suffixes and changes in the vowel pattern of the stem. A verb has different verb forms such as; perfect, simple imperfect, compound imperfect,imperative, gerunds. Many of the verb forms may be combined with auxiliaries.

Adjective is a word that comes before a noun and add some kind of qualification to the noun. But every word that comes before a noun is not an adjective. There are words which typically function as adjectives, modifiers of nouns, such as big, small, and pretty, and color words such as black. But these may also function as nouns.

Adverb is a word that qualifies the verb by adding extra idea from time, place and situations point of view. There are primary and derived adverbs. Primary adverbs are those not derived

from a verbal or a nominal form, derived adverbs are based on one form or another.

Preposition is a word that doesn't take any kind of suffix and prefix, that can't be used to create other words and which doesn't have meaning by itself but can represent different adverbial roles when used with nouns. The different propositions include ከ ፣ ለ ፣ ወደ ፣ ስለ ፣ እንደ... etc. The prepositions are attached to nouns, pronouns, and adjectives. A preposition consisting of one letter is never written as a separate word such as በቤት. Prepositions consisting of more than one letter may be written as separate words such as ወደቤት. All the prepositions(except ኧ) end in ኣ the rules for the meeting of two vowels will be applied: thus ለኔ for ለእኔ, ላንተ for ለአንተ.

4.5.2 Amharic Sentence

A sentence, is a group of words that follow with the grammatical arrangement of the language and capable of conveying meaningful message to the audience. A sentence in Amharic can be a statement which is used to declare, explain, or discuss an issue. When Amharic sentence is viewed from grammatical structure point of view it is a combination of noun phrase and verb phrase. The noun phrase comes first and then the verb phrase follows. Based on the number of phrases they contain sentences in Amharic are categorized under two basic categories simple sentence and complex sentence. Simple sentence only contains a single verb while complex sentence is constructed by combining more than one noun phrases and verb phrases.

Most of the sentences are declarative sentences. The multiparty dialogue use the declarative sentence for expressing different information on different issues. There are also Interrogative Sentence which is used to ask a question, Exclamatory Sentence which is used for emphasis and emotion, and feelings.

4.5.3 Amharic literature

The early years of the Derg era in Ethiopia (1975-91) released considerable pent-up creative energy, often expressed as political writings, from propaganda to poetry, but especially Amharic fiction. Since then, Addis Ababa bookstores include Amharic writings of all sorts of poetry, translation, newspapers, literary and news magazines, drama, novels, history, textbooks, etc[80]. Amharic language magazines are also published in the U.S. and Europe to serve the growing expatriate populations there.

The first manifestation of Amharic literature is songs of the 15th century court, such as in praise

of Emperor Yishaq (1406-21). Later are religious writings for and against Catholicism in the era of Portuguese influence, circa 1540-60, followed by a few works including a commentary on the Psalms during the reign of Iyasu (1730-55), but then nothing extant until the time of Emperor Tewodros (reigned 1855-66), who began to promote the language, including its use to write his chronicles[77].

The first publication of Amharic fiction is considered by Kane[77] to have been the 1908 Libb Weled Tarik "Fictional History" of Afewerq Gebre lyesus, published in Rome. A well respected novel of the Haile Sellassie era is Fiqir iska Meqabir 1965-1958 Eth. calendar) by Haddis Alemayehu.

Summary

In this research work we use a transcripts of multiparty dialogue written in Amharic. Our corpus are collection of conversational sentences in a form of utterance or statement as spoken by the participant mention or mentions adverted in the dialogue. So our dataset contains collection of declarative, interrogative, or exclamatory sentence from series of TV show. The corpus may contain ambiguous content, and must be normalized during the preprocessing step of our work.

Chapter Five

5 System Design and Implementation

In this chapter we talk about the proposed models of our work along with the design constraints and implementation issues. The main components of character Identification model along with sub components and the interaction between the cooperative components will be presented. As described in many literature's background, Amharic is a language rich in drama, films, novels and narrations. Unfortunately, those resources are not annotated. So that, It does not incorporated in the field of Information technology especially in the Natural Language Processing (NLP) and information extraction (IE). By taking this in consideration and reading different supportive articles, journals and books, we design a model that used to identify characters on multiparty dialogue using TV show transcript as a corpus.

5.1 Introduction

As discussed in Section three various models have been proposed to identify character from dialogue datasets of different languages. The models depend on the characteristics of the language's structure designed for. As a result, It is difficult to apply the models proposed for other languages to Amharic dialogues directly. As a result, we have proposed a new character Identification model for Amharic. This tasks requires us to build a system which can identify different mentions with corresponding referent character entities from TV show multiparty dialogues.

5.2 Datasets

There are no existing Amharic corpora of multiparty conversations, so that we do not have sufficient resource annotated which are specific to this research task. Thus it is necessary for us to generate a corpus for this study. TV shows are chosen as sources of multiparty conversational data.

TV shows are selected because it represents everyday conversation well, nonetheless TV series are very well be domain-specific depending on the plots and settings. The contents and exchanges between characters are written for ease of understanding. Moreover, prior knowledge regarding characters is usually not required and can be learned as show proceeds. TV shows

also cover a variety of topics and are carried on over a long period of time by focused groups of characters.

Transcripts of the TV show, Gemena and Sewlesew is selected. The show serves as an ideal candidate due to its casual and day-to-day conversations among their characters. Dialogues in this corpus Constitutes daily conversations that are more natural and various in topics than other Amharic dialogue novels.

In this work the CoNLL-2012 shared task data format is used for training and evaluating our models. This data has documents from TV show series. Each constituent sentence in a document has its words annotated with Parts of Speech (POS) tags, and Named entity (NE) tags. The CoNLL data format allows to preserve speaker information for each statement. The format preserves all necessary details of mentions and entities necessary to identify characters from multiparty dialogue.

Example dialogues-

አስናቀ	አ!አ! አንተ አንድ ነገር ሆነሃል እንደዚህ ሰላም እኮ የለህም ዛሬ ፊትህ ርብሽብሽ ብሎ እያየሁት ልትዋሽ ትሞክራለህ። በል አሁን ተነሳና ሙሽሮቹን ይዘናቸው እንምጣ።
ብሩክ	ጋሽ አስናቀ እኔ ብቀርና አናንተ ብታመጪቸውስ? እኔ ስልክ ልደውል አስከዚያው ድረስ ።

In addition, greedy mentions for each document are not available as a separate list. However, list of mentions including noun phrases, pronouns and named entities, animate, personal names, and all the mentions which are actually coreferent, i.e they are mentioned somewhere else in the document or not mentioned are available as a Gazetteer.

DataSet	Season	Episodes	Scene	Speakers	Utterance	Sentences	Tokens
Gemena	1	10	49	40	647	849	6023
Sewlesew	1	11	91	45	1107	1447	11557
Miscellaneous	1	4	15	14	146	320	1937
Total	3	25	155	99	1900	2616	19517

Table 7: Corpus Statistics used for training and evaluating character identification system

The original transcripts collected was plain text format; we convert the raw transcript into JSON file so that it can be easily processed. This structured data is then manually checked for potential errors. Table 7 shows the distributions from the subset of the character identification corpus used for this research task. The provided dataset is divided into seasons (ምዕራፍ), each season is divided into episodes(ክፍል), each episode is divided into scenes (ትዕይንት), each scene contains utterances, where each utterance indicates a turn of speech.

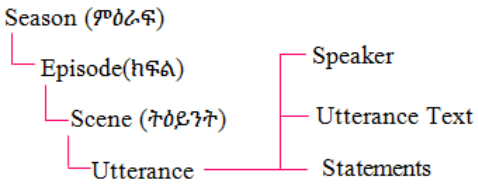

Figure 9: Structure of the corpus

In general the distribution of mentions and entities in the the overall corpus are summarized in the table 8 as follows. All columns show the counts of corresponding columns.

	Episodes	Scenes	Utterances	Tokens	Speakers	Mentions	Entities
Total	25	155	1900	19517	99	1840	137

Table 8: Summary of mentions and entities annotated for character identification task

The table contains list of mentions by counting the plural mention as one entity, and excludes generic mentions. The entities are also counted by considering entities that appear in different episodes with different name will counted as one. So that the number of mentions and entities will increase based on the document delimiter we use. Table below shows the statistics of our annotated corpus which includes 1965 singular and plural mentions. The Plural mentions alone compose around 6.31% of the entire dataset, which is significant enough to make a difference in resolution. Each cluster contains about 5 mentions on average when each scene is treated as an independent dialogue. All mentions were double-annotated by help of linguistic experts.
Table 10 below shows the distributions of entity types. Each column shows the number of mentions annotated with the corresponding entity type. Note that the total number of mentions here is different from the one in Table 9 (1965 vs. 2091) because each plural mention is counted more than once in this table.

Season	General				Mention			Entity	
	Episode	Scene	Utterance	Speaker	Singular	Plural	Total	Cluster	Type
1	25	155	1900	99	1841	124	1965	706	137

Table 9: The overall statistics of our corpus.

Season	Known Entities		Ambiguous entities				Total
	Primary	Secondary	General	Generic	Other	Collective	
1	1171	734	49	52	59	24	2091

Table 10: The distributions of entity types used in Character identification System

Plural (e.g., እኛ, እነሱ, ልጆች) and collective (e.g., ቤተሰብ, ሰው) nouns are annotated. The next table shows composition of the detected mentions as NE: named entities, PRP: pronouns, PNN(%): singular personal nouns and its ratio to all nouns.

	NE	PRP	PNN	All
	565	752	774	2091
Percentage(%)	27.02	35.96	37.01	

Table 11: Composition of the detected mentions

In addition to this we use 31768 common and singular personal nouns are prepared as Gazetteer. The singular personal nouns are collected from Addis Abeba Condominium list names. And, the system uses animate and inanimate names during features construction for the learning process.

5.3 Proposed Approach

Based on related works reviewed in Section 3, different researchers used different approaches to identify characters from multiparty dialogues. Researchers have their own point of view with evidence to apply a particular approach in such domain. To make it precise we don't need to go detail explanation of each class of approaches because we already state them with their pros and cons in Section 3. In this Section we need to highlight about the approaches used in this research work and the reasons behind it before we are going to the system architecture.

The overall task of character identification in a multiparty discourse setting, could be divided into two sub-tasks – coreference resolution and entity linking. By integrating the two modules a new system is proposed that identifies a global entity of each personal mentions from multiparty dialogues introduced during the discourse. The proposed model is biasing on the deep learning approach which involves learning mention and mention pair embeddings using convolutional neural networks. These embeddings would be used to get cluster embeddings for the subsequent stage of entity linking.

The first step in the deep learning phase was to somehow convert these mentions into a feature vector. Then we could feed these features to a variety of neural network layers and see how they perform. We started by trying to build our own word encoding using one-hot vectors exracted from facebook fasttext corpora [78] and train it using a continuous bag of word (CBOW) model with a window size of 2. And we use word2vec which is more efficient.

5.4 System Architecture

This work introduces a new task, called character identification, that is a subtask of coreference resolution followed by entity linking. This work is proposed for the first time in such domain, so that a new corpus is annotated, which comprises multiparty conversations from TV show transcripts for the training and evaluating the overall performance. The data will pass through the common preprocessing and annotation tasks. Our annotation scheme allows the creation of appropriate dataset with the personal mentions and their referent characters. We further disambiguate our corpus and introduce generic groupings of mentions with abstract referent entities. The nature of this corpus is analyzed with potential challenges and ambiguities identified for future investigation. To tackle the task of coreference resolution and entity linking, we propose a deep learning CNN mention-to mention ranking model that provides better mention and mention-pair representations learned from feature groupings of dialogue data. Our developed system consists of three major stages: mention detection, followed by coreference resolution, and finally, entity linking. In the first stage, mentions are extracted and relevant information about mentions, e.g., gender and number, is prepared for the next step.

In general the following are some of the general components of character identification system. Document preprocessing-this component allows the system to identify language specific aspects and to normalize the document in order to save the CPU cost , space, and to enhance the

system performance.

Corpus annotation component helps us to identify the group of entity types like primary, secondary entities, general and collective entities with their importance mention referents. We apply corpus disambiguation in order to remove disagreement occurred in the annotation step. The last component is character identification that focuses on the main features which are extracted to check the relevancy. Template generation: finally the extracted mentions should be prepared in considerable format for further analysis, application inputs, reasoning, and performance evaluation. The proposed system architecture of our character identification system is adopted from the general architecture of coreference resolution and entity linking system. Our own system specific components are incorporated along with the general architecture. The following figure renders our proposed system architecture.

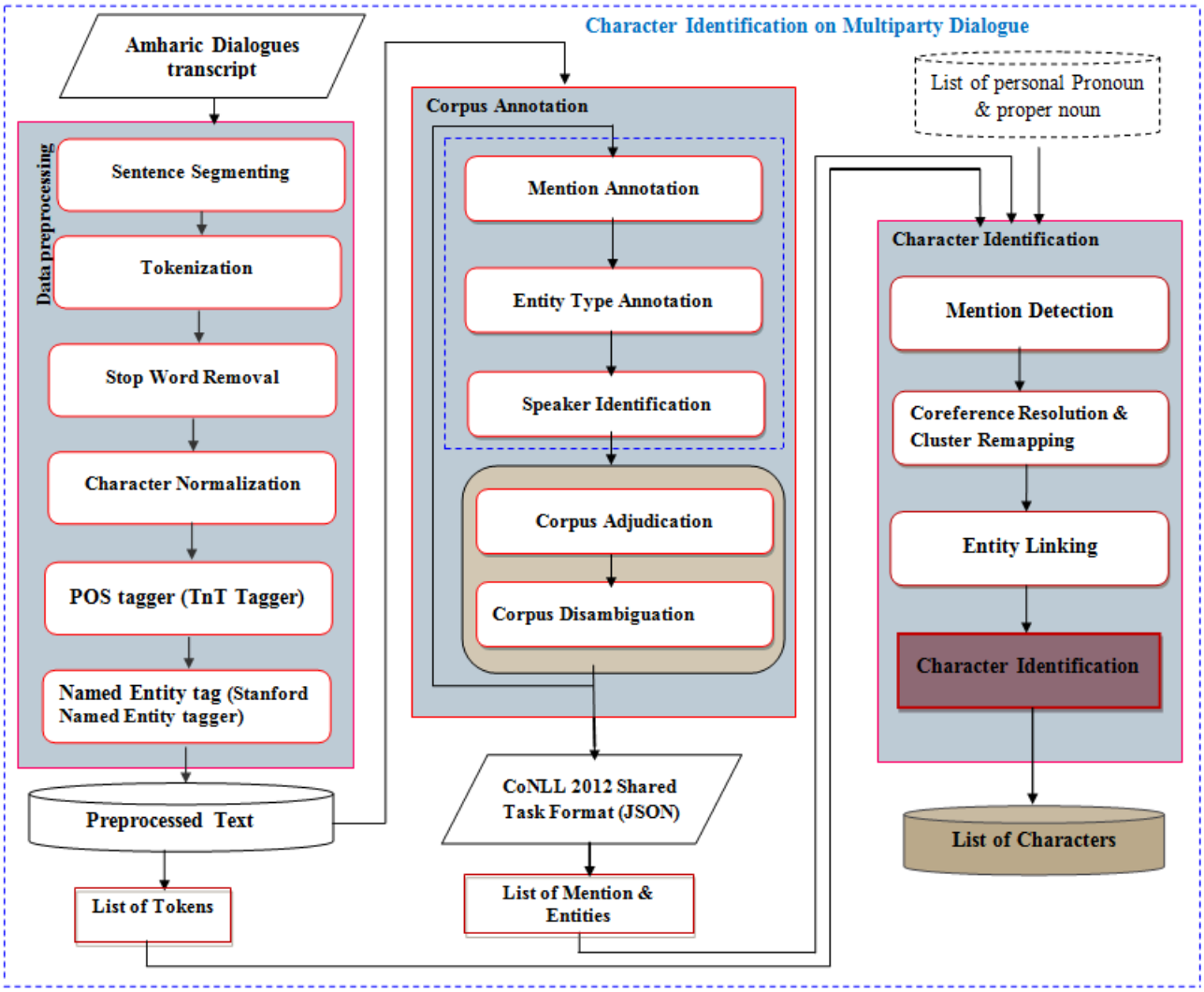

Figure 10: Architecture of character identification System

In the following subsections we give a more detailed description of the components along with their subcomponents of our proposed character identification system model.

5.5 Data Preprocessing

Amharic conversation corpus need different type of preprocessing before we made ready for Character Identification system, as which Amharic has different language specific features that should be normalized. The document preprocessing component such as, tokenization, sentence segmenting, character normalization, stop word removal etc.. are applied on the dataset to handle language specific issues.

5.5.1 Sentence Segmenting

A sentence segmenter divides a paragraph into sentences. We need sentence segmenter because in turn of conversation a speaker may state their idea with more than one sentence. Most of the time Amharic sentence is ended in full stop(::) or in question marks(?) and sometimes with exclamation marks(!). To segment a sentence we use nltk sentence tokenizer which is punctuation based sentence tokenizer. This instance is already trainable and works well for Amharic sentence.

5.5.2 Tokenization

This is the task of breaking texts in to piece of meaningful tokens. Sometimes it can be defined as a sequence of characters or a defined document unit. Tokenization is the task of chopping it up into pieces, perhaps at the same time throwing a way certain characters such as punctuation. A token is an instance of a sequence of characters in some particular document that are grouped together as a useful semantic unit for processing.

In Amharic sentence or phrase a single space would be added between the word and punctuation by the system. The tokenizer then tokenizes all the text segments which have space between each other as independent token. We use NLTK word tokenizer for this task.

5.5.3 Character Normalization

As described in Section four of this document Amharic language has different characters with the same meaning and pronunciation but with different symbols. The different symbols should

be treated equally because there is no change in meaning regardless of the linguistic view of orientation among characters. [ሀ ፣ኅ ፣ሐ ፣ሃ ፣ሓ ፣ኃ],[ሰ ፤ሠ],[ጸ ፣ፀ] and[አ ፣ዓ] all these different forms of characters (Fidels) represent the same sound ሀ, ሰ, ጸ and አ respectively. Therefore, these variants of characters (Fidels) need to be converted to the same or common character form (shape) in order to avoid representing the same words/phrase using different letters having the same sound which will increase the number of words representing the document without any relevance as a result the performance of the system will down a bit. Such characters should be normalized to a single character like ሀ ፤ሰ ፤ጸ ፤አ because there is no any change in their meaning.

5.5.4 Stop word removal

Like any other language Amharic language has its own list of stop words including conjunctions, articles and prepositions. It's obvious that stop words has a great significance to write the meaningful document from linguistic perspective, But when we design an NLP application like IE we need the relevant words to represent the document. As such stop words are frequently occur words in any document without a meaning to describe about the document. So that , in order to enhance the performance of the system as well as to save computing resources we need to get rid of those irrelevance before starting the learning and extraction component. Consequently, we remove stop words by preserving pronouns and possessive pronouns because pronouns indicate mention referents in case of character identification system.

5.5.5 Amharic POS tagging

Part-of-speech (POS) tagging is a classification task with the goal to assign word classes to the words in a text. POS tagging is the core task to achieve the goal of Natural Language processing. We used POS tagging because it helps to identify mentions and independent pronouns, and used as input for the next step. Because of the unavailability of free working POS tagger for Amharic, we train TnT tagger as Amharic POS tagging. Using the Walta labeled corpus to train the TnT tagger and got precision of 79% which is comparably good. Such tagger is used to tag our corpus.

5.5.6 Amharic named entity tagging

Considering recent increases in computing power and decreases in the costs of data storage, data scientists and developers can build large knowledge bases that contain millions of entities

and hundreds of millions of facts about them. These knowledge bases are key contributors to intelligent computer behavior. Not surprisingly, Named Entity Extraction operates at the core of several popular technologies such as deep interpretation of natural language. This task is used to generates a list of the mentions, identified by NER, found in each sentence in a document. Recognizes named entities (person and company names, Date, Event, etc.) in text. In this work, we need named entity tagger to identify person names participated or mentioned in the corpus used. We train Stanford named entity tagger to handle Amharic named entities.

5.6 General Corpus annotation

The character identification corpus was first developed by collecting transcripts from series TV show Sewlesew and Gemena. This is transcripts of dialogues and passes through different stages of preprocessing in the form of plain text. This is used as an input for the annotation module. This module contains the annotation, adjudication and disambiguation of corpus used in the training and evaluating the proposed system.

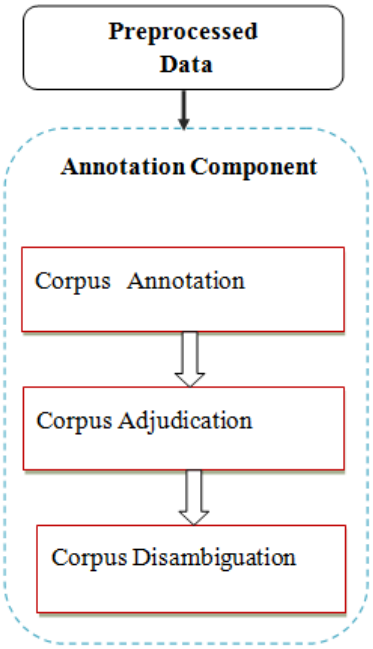

Figure 11: Annotation of Corpus module

5.6.1 Corpus annotation

In this work we introduce a systematic framework for annotating referent information of mentions in order to create a large scale dataset for character identification. In this stage the plain text changes to JSON file. The JSON file preserves all the relevant information about the entities, speakers and mentions found in the dialogue. Start from splitting utterances into sentences and personal mentions in every sentence are manually annotated with their entities. For example below, the second utterance consists of one sentence including four mentions. The mentions, **አስናቀ** and **እኔ**, are singular that refer to **አስናቀ** and **ብሩክ**, respectively. The last mention, **እናንተ** is plural mention that refers to both **አስናቀ** and **ልኡል** as shown in figure 13.

አስናቀ	አ!አ! አንተ አንድ ነገር ሆነሃል እንደዚህ ሰላም እኮ የለህም ዛሬ ፊትህ ርብሽብሽ ብሎ እያየሁት ልትዋሽ ትሞክራለህ። በል አሁን ተነሳና ሙሽሮቹን ይዘናቸው እንምጣ።
ብሩክ	ጋሽ አስናቀ እኔ ብቀርና እናንተ ብታመጯቸውስ? እኔ ስልክ ልደውል እስክዚያው ድረስ ።

Figure 12: Sample dialogues that compose Two utterances

In order to give the model enough context we use all the utterances in a particular scene as a single sample. Each scene is then represented by a concatenation of a sequence of tokens ([' **ጋሽ**',' **አስናቀ**',' **እኔ**',' **ብቀርና**',' **እናንተ**',' **ብታመጯቸውስ**', ' **እኔ**',' **ስልክ**',' **ልደውል**', ' **እስክዚያው**',' **ድረስ**', ' **።**']), a sequence of speakers ([**ብሩክ**]), a sequence of indicators whether there exists or not a mention is ([0, 1, 1,0, 1, 0, 1, 0, 0,0,0,0]). The JSON file contains the id of each episodes, scenes, utterances to identify each statements. The POS tag, Named entity tag of each token were included in file during this phase.

Mentions are annotated by linking with the corresponding global referent entities. Each mention is annotated by the following scheme: $[begin\_index, end\_index, entity(entity)*]$

All datasets follow the format of the data released for the CoNLL 2012 shared task [16]. We have listed all the columns that are typically in a co-reference resolution shared task. The following are the columns for every token in a utterance. (1) Document ID: (e.g., s01). (2) Scene ID: the ID of the scene within the episode. (3) Token ID: the ID of the token within the sentence. (4) Word form: the tokenized word. (5) Part-of-speech tag: the part-of-speech tag of the word (we use TnT tagger, which is a statistical pos tagger). (6) Speaker: the speaker of this sentence. (7) Named entity tag: the named entity tag of the word(we use Stanford named entity tagger). (8) begin and end index of the character entity of each mention that appear across all

```
{
 "utterance_id": "s01_e01_c04_u008",
 "speakers": ["ብሩክ"],
 "transcript": "ጋሸ አስናቀ እኔ ብቀርና እናንተ ብታመጪቸውስ እኔ ስልክ ልደውል እስከዚያው ድረስ ።",
 "transcript_with_note": null,
 "tokens": [
 ["ጋሸ","አስናቀ","እኔ","ብቀርና","እናንተ","ብታመጪቸውስ","እኔ","ስልክ","ልደውል","እስከዚያው","ድረስ","።"]],
 "tokens_with_note": null,
 "part_of_speech_tags": [
  ["VP", "N", "PRP", "NC", "PRP", "N", "PRP", "N", "VP", "PRON", "PREP", "PUNC"]],
 "named_entity_tags": [
  ["O", "PERSON", "O", "O", "O", "O", "O", "O", "O", "O", "O", "O"]],
 "character_entities": [
  [
   [1,2,"አስናቀ"],[2,3,"ብሩክ"],[4,5,"አስናቀ","ልኡል"],[6,7,"ብሩክ"]]
 ]
}
```


Figure 13: The label of the entity scheme

documents.


```
{ #begin document(s01_e01_c04)
   "utterance_id":"s01_e01_c04_u008",
   "speakers":["ብሩክ"],
   "transcript":"ጋሸ አስናቀ እኔ ብቀርና እናንተ ብታመጪቸውስ እኔ ስልክ ልደውል እስከዚያው ድረስ ።",
   "transcript_with_note":null,
   "tokens":[
      [
         "ጋሸ","አስናቀ","እኔ","ብቀርና","እናንተ","ብታመጪቸውስ","እኔ","ስልክ","ልደውል","እስከዚያው","ድረስ","።"
      ]
   ],
   "tokens_with_note":null,
   "part_of_speech_tags":[
      [
         "VP","N","PRP","NC","PRP","N","PRP","N","VP","PRON","PREP","PUNC"
      ]
   ],
  "named_entity_tags":[
      [
         "O","PERSON","O","O","O","O","O","O","O","O","O","O"
      ]
   ],
   "character_entities":[
      [
         [
            1, 2,"አስናቀ"
         ],
         [
            2,3,"ብሩክ"
         ],
         [
            4,5,"አስናቀ","ልኡል"
         ],
         [
            6,7,"ብሩክ"
         ]
      ]
   ]
} #end document
```


Figure 14: Annotation of Corpus in CONLL-2012 shared task format using JSON

5.6.2 Corpus Adjudication

A critical part of Annotation stage is adjudication where you take your annotators work and use it to create the gold standard corpus that you will use for machine learning. During the course of working on this research, we found quite an annotation errors in the dataset. We present a qualitative analysis of the kind of errors made by the Adjudication while assigning the character labels to mentions in the dialogue corpus. Most common error is the incorrect resolution of the full names of a given character and the boundary errors where stray punctuation symbols such as ellipsis, apostrophe are added to the named entity phrase and considered a new entity. These problems are probably very common when annotation is constructed. A subset of the disagreed annotation was adjudicated manually from which we found that taking the union of the entity sets annotated effectively give the correct set of entities for each of those disagreed plural mentions.

Any part of episode (scene) containing at least one annotation disagreement is put into adjudication. The same process as that for the annotation task is used for the adjudication, except that options for the mentions are modified to display options. Nonetheless, corpus adjudication still have the flexibility of choosing any option from the complete list in the annotation task.

5.6.3 Corpus Disambiguation

Three labels are introduced to disambiguate Unknown mentions like General, Generic, and Other in addition to primary and secondary entity labels. Generic provides abstract groupings for unidentifiable entities, and each group is assigned a unique number for differentiation.

Primary characters are main characters which covers most of the dialogues conversation, where as Secondary characters are supporting characters with names identified. And the remaining are ambiguous entities annotated in the form of Generic ወንድ/ ሴት/ ሰው entities, Collectives - collective use of እኛ, and እናንተ. General - reference to general cases, and Other annotated for irrelevant and singleton entities.

Here we use similar annotation guidelines for both singular and plural mentions with the only difference in annotation between these two types of mentions is the number of entities to which the mentions refer.

	Primary	Secondary	Generic	Collective	General	Other	Total
	1171	734	52	24	49	59	2091

Table 12: Count break down of mentions in our corpus after disambiguation.

Formally, each mention m is annotated with a set of entities E, where each element in E belongs to one of the following four groups:

1. Known entities : include all the primary and secondary characters recurring in the show. primary characters are list of characters which appear in train, development and test set in common, where as secondary characters are character which appear in train and development set or train and test set in common.

2. GENERIC : indicates actual characters in the show whose identities are unknown across the show.
   Example- በኋላ የሚባል ነገር የለም። አቶ አስናቀ ቶሎ እንዳቀርብ ስለሚፈልግ አሁኑኑ ተልኮ ጸሃፊዋ ጋር እንዳገኘው ይፈልጋል።

3. GENERAL : indicates mentions referring to a general case rather than a specific entity.
   e.g., የመጨረሻ አሪፍ የሆነች ልጅ ተዋውቂያለሁ እሷን አገባልሽና አለምሽን አሳይሻለሁ ።

4. OTHER : indicates actual characters in the show whose identities are unknown in this dialogue but revealed in some other dialogue.
   Example- የመጨረሻ አሪፍ የሆነች ልጅ ተዋውቂያለሁ እሷን አገባልሽና አለምሽን አሳይሻለሁ ።

In addition to the four entity groups, we annotate collective and non-entity types are annotated to handle mentions represented by group name and the use of the pronoun እሱ refering with non- human entities. Collective type is introduced to indicate mentions represented by group of entities.

Example-, ከዚያ ሰዎች ሲንጫጩ ሰምቸ ወጣሁ ።,

here the mention ሰዎች indicates collection of entities represented by one name but they are not represented by an identifiable entities, And we annotate mention እናንተ or እኛ when the entities referred by the mention are unknown.

Example -, እኛ ተዋንያን።.

Non-Entity indicates pronoun that does not represent human mention.

Example, እሱ ለጊዜው ስለማያስፈልግህ ነው ።

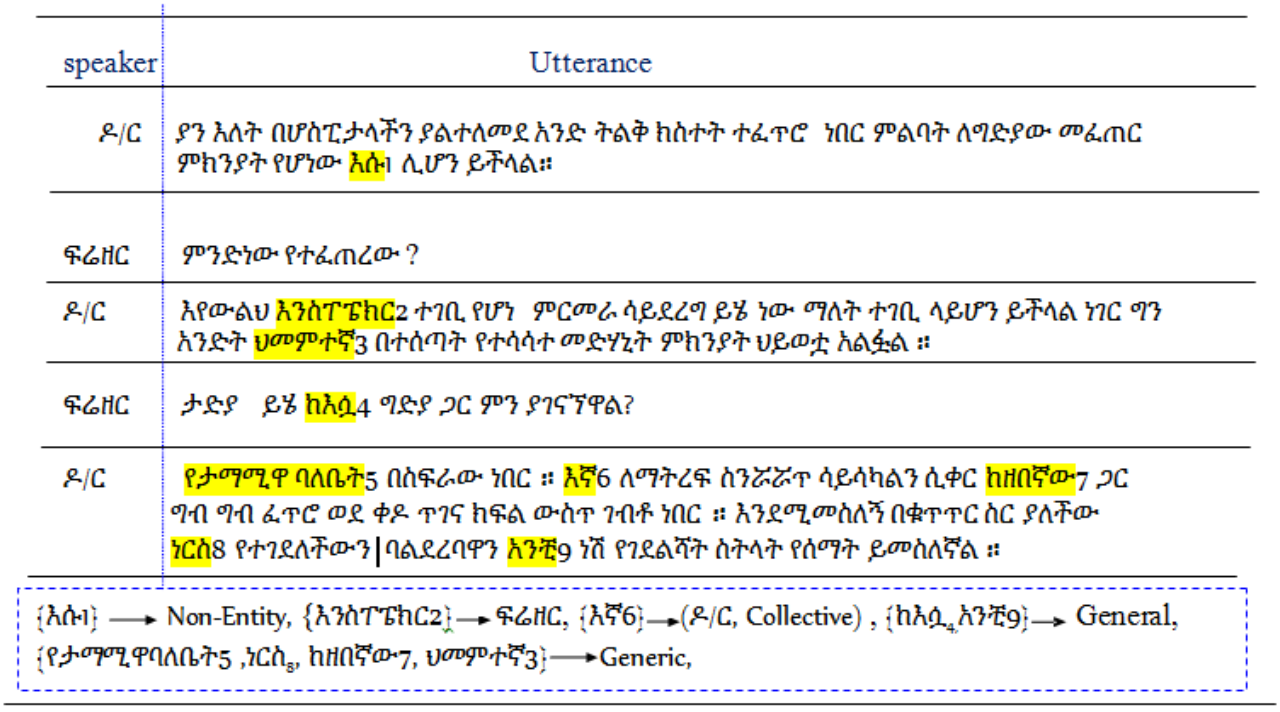

Figure 15: An example of a multiparty dialogue extracted from the corpus

- General: Mention used in reference to a general case (e.g., **አንቺ**).
- Generic: Mention referring to a unidentifiable entity (e.g., **የታማሚዋ ባለቤት**5 , **ነርስ**8, **ከዘበኛው**9, **ህመምተኛ**3).
- Other: Mention referred to insignificant singleton entity (e.g., **እሱ**1). We perform this disambiguation manually with two main guidelines: only mentions originally labeled Unknown are included, and the labels introduced above are provided to annotators in addition to the known entities. The result of the disambiguation is shown in Table with detailed break down of the counts of mentions in each group.

5.7 Mention Detection

A noun phrase is a mention if it is either

1. A PERSON named entity, or
2. A pronoun or possessive pronoun, or
3. One of the personal noun gazetteers that are common and singular personal nouns (e.g., **ማሚ**, **እህት**, **ወንድም**) and titles of nomination like **ኢንስፔክተር**, **መምህር** etc,.

The most likely annotation is a fairly simple and straight forward method to determine the entity being referenced. When we look at a models we see that we might want to encode the corpus tokens to better understand about the relationship between words. Additionally we might look at modeling the relationship between the speakers. Simply we need to capture the speaker making the reference, and the word the speaker uses to reference the entity.

5.8 Coreference Resolution

Coreference resolution is a task of finding all expressions that refer to the same entity in a text. It plays a crucial role in Natural Language Understanding tasks like document summarization, information extraction and question answering. We would be baising on the agglomerative Convolutional Neural network approach introduced by [10] for our coreference resolution. Convolutional neural network can learn mention and mention pair embedding. These embedding would be used to get cluster embeddings for the subsequent stage of entity linking. In this work, we have leveraged the benefits of convolutional neural networks to build a mention-to-mention ranking model. Thereby features which have common properties are segregated into groups and they are trained on separately. This model has resulted in efficient mention and mention pair embeddings.

We use a scoring function s_m for determining the likelihood of a link between the two mentions given its mention pair representation $r_m(m_i, m_j)$ between m_i and m_j.
$s_m(m_i, m_j) = \sigma(W_m r_m(m_i, m_j) + b_m)$, Where w_m and b_m are the weights and bias of the scoring function. Th scoring function is essentially a regression model used to train our model with a mean squared error loss function. Let $A(m_i)$ be the list of antecedents of each mention m_i and $C(m_j)$ be the cluster containing the mention. For each mention , the goal is to find the training instances up to the closes antecedent with a linking score of 1. This condition for the gold linking score p(a, m) is given as follows. $p(a,m) = \{^{1\ if\ m \in C(a)}_{0\ else}$. Through back-propagation of the loss function, the model learns mentions and mention pair representations which in turn optimizes the task of mention ranking.

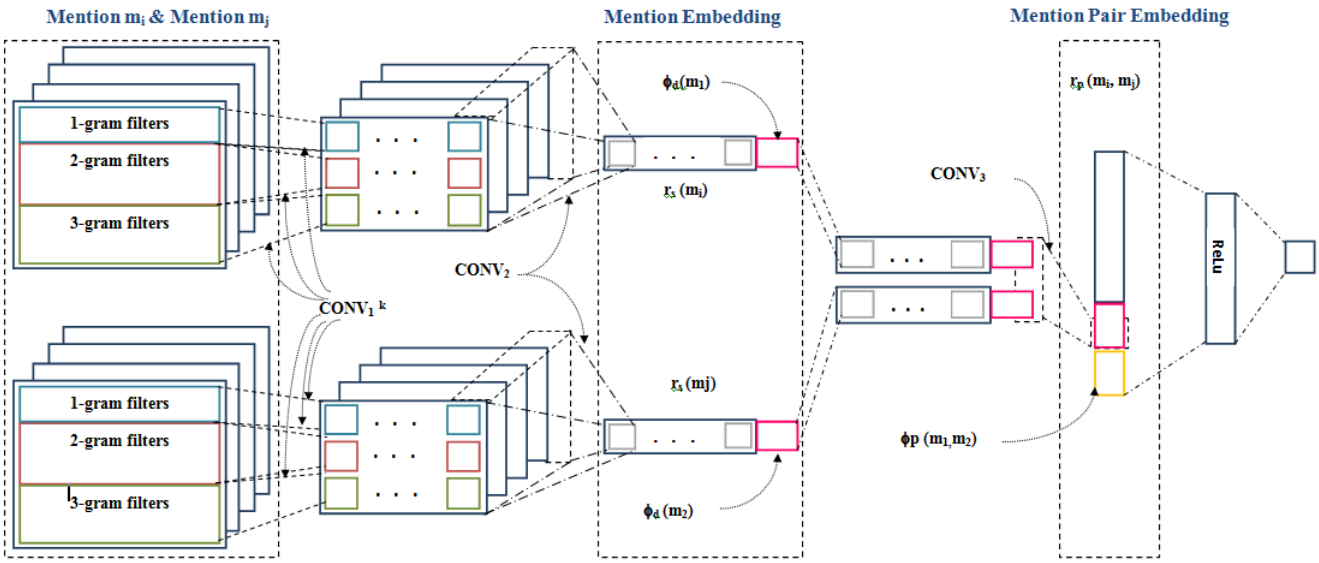

Figure 16: Agglomerative Convolutional Neural Networks adapted from CNN model[72]

During the training phase, a CNN automatically learns the values of its filters based on the task that to be performed. For example, NLP tasks are sentences or documents represented as a matrix. Each row of the matrix corresponds to one token, typically a word, but it could be a character. That is, each row is vector that represents a word. Typically, these vectors are word embeddings (low dimensional representations), but they could also be one-hot vectors that index the word into a vocabulary.

The second part of the neural network utilizes the learned mention embedding $r_s(m)$ to create the mention pair embedding. Tanh is used as an activation function with a dropout rate of 0.8. To control the number of features the CNN model is learning and to avoids over fitting Max pooling layer picks maximum values from the convoluted feature maps. Another feature map $\phi_p(m_i, m_j)$is defined to extract pairwise features between mentions m_i and m_j. The third convolution layer CONV3 is applied to the stacked mention embeddings, $r_s(m_i)$ and $r_s(m_j)$. The result is max-pooled and concatenated with the pairwise features extracted by $\phi_p(m_i, m_j)$ to form the mention-pair embedding $r_p(m_i, m_j)$, defined as follows:

This leads us to the next step to convert mention embeddings to mention pair embeddings.

$$r_s(m) = Conv_2\left(\begin{bmatrix} Conv_1^1(\phi_e^1(m)) \\ \\ Conv_k^1(\phi_e^k(m)) \end{bmatrix}\right) || \phi_d(m)$$

The other feature type in the feature space are the biword features $\phi op(m_i, m_j)$. The third convolutional layer CONV_3^k is applied on mention embeddings $\text{r}_s(m_i)$ and $\text{r}_s(m_j)$. This result is concatenated and max-pooled with the bi-word features to form $\text{r}_p(m_i, m_j)$ defined as follows:

$$\text{r}_p(\text{m}_\text{i}, \text{m}_\text{j}) = \text{Conv}_3\left(\begin{bmatrix} \text{r}_s(\text{m}_\text{i}) \\ \\ \text{r}_s(\text{m}_\text{j}) \end{bmatrix}\right) || \, \Phi_p(\text{m}_\text{i}, \text{m}_\text{j})$$

The learned mention pair embeddings are passed through a hidden layer with ReLu and sigmoid function $\sigma(m_i, m_j)$ to determine the coreferent relation between m_i and m_j defined as follows:
$\text{h(x)} = \text{ReLU}(\text{w}_h\text{x} + \text{b}_h)$
$\sigma(m_i, m_j) = \text{sigmoid}(\text{w}_s\text{h}(\text{r}_p(\text{m}_i, \text{m}_j)) + \text{b}_s)$
For each mention, $\sigma(m_i, m_j)$ performs binary classifications between m_i and m_j. It considers a halfway threshold and the model considers no coreferent relation between m_i and m_j if it's below the threshold and therefore creates a new cluster with it. The rules for mention clusters are defined as follows: Finally we regularize ovrfitting of parameters using RMSProp optimizer.

Algorithm 1 Mention Cluster Algorithm

1. If $\forall 1 \leq \text{j} \leq \text{i}.\max(\sigma(\text{m}_i, \text{m}_j)) < 0.5$

create a new cluster C_{m_i}

2. If $\exists 1 \leq \text{j} \leq \text{i}.\max(\sigma(\text{m}_i, \text{m}j)) \geq 0.5$

   1. $\text{C}_{mk} \leftarrow \text{C}_{m_k} \cup m_i$,

   2. $\text{m}_k = \arg_j\max(\sigma(\text{m}_i, \text{m}j))$.

This leaves the parameters of optimizer at their default values (except the learning rate, which can be freely tuned). This is used to create cluster embeddings and fed as input to our entity linking model.

5.8.1 Algorithm

This Section introduces our new coreference resolution algorithm that creates clusters with corresponding to different mention types. This algorithm ensures singular mentions representing different entities assigned to separate clusters. For example, let m_p be a plural mention and m_i

be a singular mention such that $m_p \longleftarrow m_i$. When the referent relation is found, the cluster C_i is created and both m_p and m_i are assigned to C_i. Let m_j be another singular mention such that $m_p \longleftarrow m_j$. Now, the algorithm must decide whether to assign m_j to C_i or create another cluster C_j for m_j. If $m_i <\longrightarrow> m_j$, m_j should be assigned to C*i*; otherwise to C_j.

Our algorithm allows a model to learn this decision during training so that the clusters can be created accordingly during decoding. For each mention m_j, our algorithm compares it against all of the preceding mentions m_i to determine whether or not they are referent, where i and j are the ordered indices such that $0 < i < j$. Additionally, two more mentions, m_g and m_o, are compared to m_j that represent the General and the Other types,respectively. For each mention pair (m_i, m_j), the algorithm assigns one of the following three labels for multi-classification:

1. N: m_i is not referent to m_j.

2. L: m_j gets assigned to the cluster that m_i belongs to. If m_i does not yet belong to any cluster, a new cluster C_i is created and both m_i and m_j are assigned to C_i.

3. R: m_i gets assigned to the cluster that m_j belongs to. If m_j does not yet belong to any cluster, a new cluster C_j is created and both m_i and m_j are assigned to C_j.

During training, labels are determined by consulting the oracle. L is labeled if mi is a singular mention. R is labeled if m_i is plural and m_j is singular. N is labeled for all the other cases.

5.8.2 Mention to mention pair ranking

In this paper, we have leveraged the benefits of convolutional neural networks to build a mention to-mention ranking model. Thereby features which have common properties are segregated into groups and they are trained on separately. This model has resulted in efficient mention and mention pair embeddings.

5.8.3 Feature Extraction

Three main categories of features have been used in this model namely mention embedding, singleton and bi-word features. The word embeddings are trained with Word2vec and Fast text. Utterance and sentence vectors are considered to be the weighted average word embeddings of all words in an utterance and a sentence.

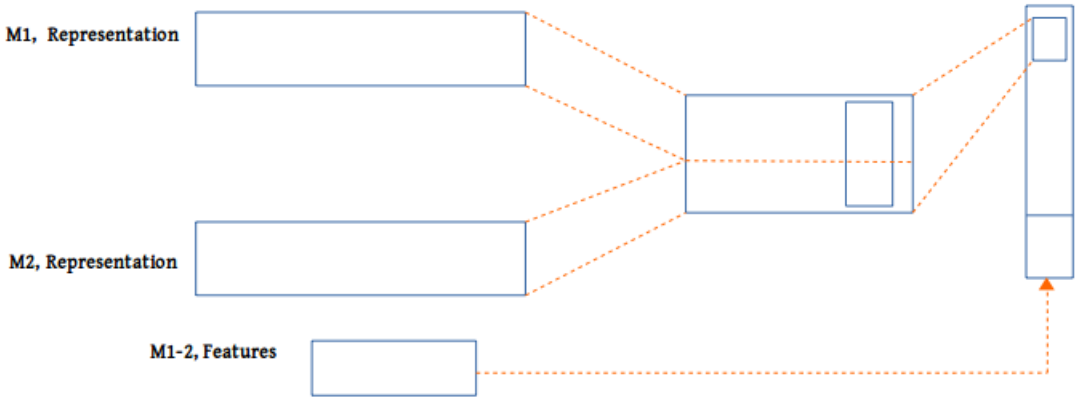

Figure 17: Mention-pair representation

Feature Group	List of Features
Discrete- Φ_d	Average plurality information of all words in a mention. Speaker embedding of current and previous utterance. Average gender information of all words in a mention. Average word animacy of all the words in m.
Pair-wise Features - Φ_p	Speaker information of the mention pair Longest Common Subsequence of words in mentions Common words between mentions. Distance and position metrics between mentions. Sentence and mention distance between m_i and m_j. Speaker match between m_i and m_j.
Mention Embedding - Φ_m	Utterance vectors of current,previous and successor utterances Sentence vectors of current,previous and successor mentions before utterances. Average word embeddings of all words in a mention. Words embeddings of n preceding and succeeding words in a mention.

Table 13: Mention Feature Template.

By taking dw is the dimension of the mention embedding feature., and f be the number of filters used in each layer. Pooling and convolution layers are next added into the system architecture. The window sizes of the convolutional layers are given the values 1xdw, 2xdw and 3xdw and are trained on each mention embedding feature. The output of each convolution

layer is max-pooled along the line of the columns. This provides us with a resultant vector of 1 x f by stacking both the convolution and pooling layers into a matrix of dimension 12xf. An additional max-pooling and convolution of sizes 1xf and 12x1 respectively are stacked on top of this existing architecture to output a 1 x f vector which will contain the categorized mention embedding features.

5.9 CNN Entity Linker

The task of character identification requires each mention to be identified by the names of actual characters (e.g., አስናቀ, መስፍን). Figure below gives the overview of our entity linking model, which adapts the underlying architecture from the entity linking model proposed by [15] and generalizes it to handle detected mentions. It assumes the output from CNN, such that for each mention m_i, the embedding of that mention and the set of clusters C_1. . . C_k that m_i belongs to are taken. For each cluster C_a, CNN gives the list of mention pair embeddings $m_{i,j}^{Ca}$, where m_i, $m_j \in C_a$. The CNN model creates multiple cluster and cluster pair embeddings when m is assigned to more than one cluster during coreference resolution so that the average vectors of those embeddings are generated, which get concatenated with the mention embedding of m_i and passed onto the fully-connected layers for prediction.

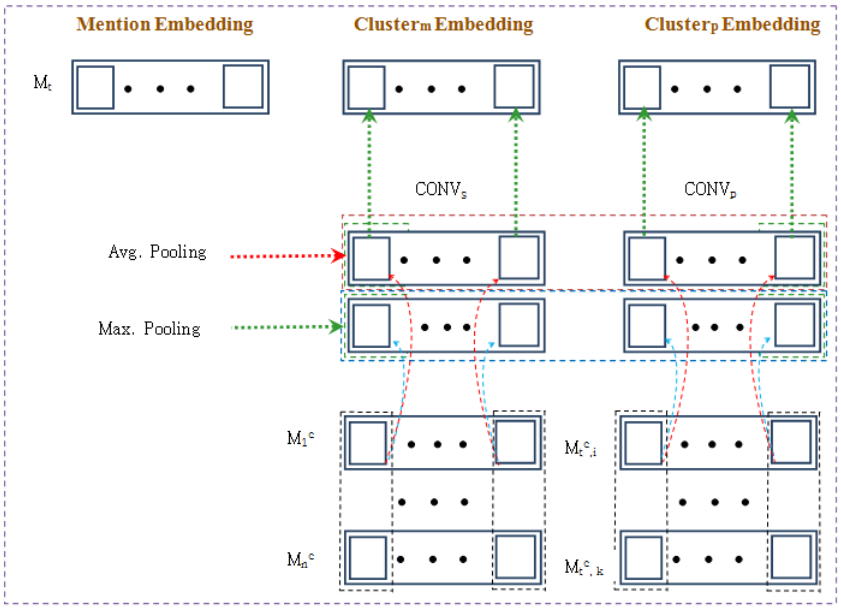

Figure 18: Neural Model for Entity Linking Embedding

In the previous stage, coreference resolution groups mentions into clusters, but it does not assign character labels to the clusters, which is required for character identification. Thus, an entity

linking model is required that takes the mention embeddings and the mention-pair embeddings generated by the CNN and classifies each mention to one of the character labels. This would involve training a deep learning approach to classify each of the mentions to an entity label. Once the coreference resolution system is built and trained, we can predict coreferences represented by clusters for each scene document. The next step should be making predictions from clusters to TV show character id. We model this process as an entity linking task.

We prepare three embeddings(Figure 18) generated by mentions which are predicted by previous co-reference system. The first embedding is mention embedding; the second is embedding of the cluster including the mention; the third is generated by mention-pair embedding, which pairs the mention with reaming mention in the same cluster.

Here we can derive a formula to mention cluster as:

$\mathrm{Rs(Cm)} = [\mathrm{r}_s(m_1), r_s(m_1), ..., r_s(m_{|C_m|})]$. . . . . (1)

$\mathrm{Rp(C}_m, m) = [\mathrm{r}_p(\mathrm{m}_i, m)|m_i \neq m]$ . . . . . (2)

In order to fix the input tensor size of both cluster embedding and mention-pair embedding, we perform avg-pooling and max-pooling for both embeddings. Then each of pooling layers is passed to a convolutional layer.

Finally, we concatenate the mention embedding, cluster CNN embedding and mention-pair CNN embedding. After concatenation, we feed them into a Convolutional neural network with two hidden layers.

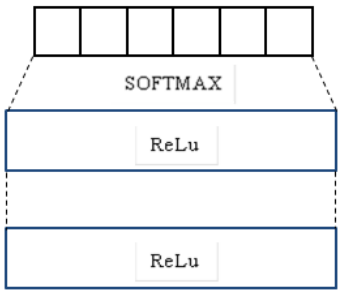

Figure 19: Entity Linking Activation Function

Literally, the character identification problem is tackled as a coreference resolution task with a further step on entity linking. In terms of this task, the baseline model generates mentions from a coreference system, and then each coreference chain is linked to a specific character identity. Both parts are implemented with convolutional neural network.

```
    hidden1 = Dense(dense_dim, activation='relu')(cm_vec)
    hidden2 = Dense(dense_dim, activation='relu')(hidden1)
    probs = Dense(nb_labels, activation='softmax')(hidden2)

self.linking_model = Model(inputs=[mrepr, crepr, cmmft], outputs=[probs])
self.linking_model.compile(optimizer=RMSprop(),
                           loss=['sparse_categorical_crossentropy'],
                           metrics=['sparse_categorical_accuracy'])
```

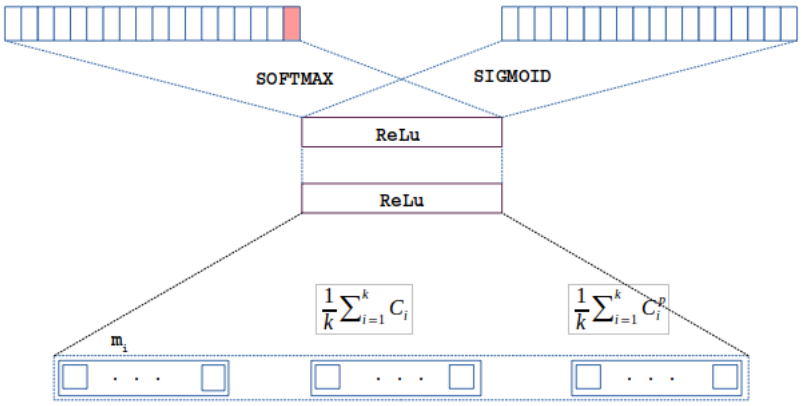

Figure 20: The overview of our entity linking model using multi-task learning.

We create our CNN with three hidden layers using a relu activation function with seven input tensors. We used the default RMSProp optimizer for our results. The final ReLu layer is fed to two output layers optimized by softmax and sigmoid functions, respectively. The dimension of the output layer from softmax is |E| + 1 where E is the set of all entities such that each cell represents an entity and the extra cell gives an indication of m being plural. When this extra cell is predicted, the output layer from sigmoid is used, whose dimension is |E|, to predict multiple entities for m. Since the sigmoid function optimizes each cell to be between 0 and 1, any entity whose score is greater than 0.5 is taken. These two output layers are optimized jointly, treating the resolution of singular and plural mentions as multi-task learning. The linking model creates multiple cluster and cluster pair embedding when m is assigned to more than one cluster during coreference resolution so that the average vectors of those embedding are generated, which get concatenated with the mention embedding of m_i and passed onto the fully-connected layers for prediction.

5.10 Implementation

Python 3.5 programming language with Pycharm editor is used to develop CIMD prototype. We also employed tools including TnT tagger, Hornmorpho, Tensorflow library, JSON, and many other open source libraries. The reason we choose the Python programming language is that deep learning algorithms can easily integrated and used in the environment. The publically available language independent part-of-speech tagger, which is Trigram'n'Tagger, is used to annotate Amharic texts with their proper part-of-speech tag.

Summary

In this chapter the CIMD model is presented and the main tasks of the different components are described. The CIMD model comprised of document preprocessing, co-reference resolution and entity linking. The document preprocessing handles the processing of language related issues, the co-reference resolution component creates a cluster that contain the referent mentions and the last component handles the linkage of mentions extracted with the corresponding entities annotated. The CIMD model that is proposed in this research work is a generic model which can be used for any other data domain in Amharic language.

Chapter Six

6 Experiment

The objectives of this work is to investigate the assessment of solving the task of character identification using deep learning models. In this section, we present the results of the experiment that was conducted to test the deep learning model in terms of co-reference resolution and entity linking.

6.1 Experimental procedures

In this section, we present the results of the different experiments that were conducted to test the deep learning model of character identification by integrating co-reference resolution and entity linking. Experiments are conducted on two tasks, coreference resolution and entity linking. Based on the distribution of characters which appear in all documents, we measure the label accuracy of our proposed system. Characters that appear in all documents are considered as main characters and characters that appear either in training and test set or training and development set in common are considered as extra characters.

The experiments were performed in terms of coreference resolution and entity linking. So that hyper parameters used in the experiment are defined. We used tanh and ReLU for the activation function in the convolutional layers. The hidden layers had 150 dimension, and the dropout of all layers was set to 0.8. The learning was done using RMSprop optimizer[86] to regularize the out-layers and the learning rate was reduced by 50% for every 5 epochs.

6.1.1 Data collection

The study uses the training and test data provided form TV drama series, which span the first season of the TV show Sewlesew and Gemena, divided into scenes (train: 125 scenes from 19 episodes; test: 18 scenes from 3 episodes and dev: 12 scenes from 3 episodes). In total, the training and test data contain 1524 and 274 nominal mentions (e.g., ብሩክ, እኔ; described in section one), respectively, which are annotated with the key of the entity to which they refer(e.g.,[3, 4," ማህሌት"]). The utterances are further annotated with the name of the speaker (e.g., መስፍን). Overall there are 120 entities in the training data and 86 entities test data.

	Episodes	Scenes	Utterances	Tokens	Speakers	Mentions	Entities
	25	155	1900	19517	135	1899	280

Table 14: Statistics of the character identification corpus used for this task.

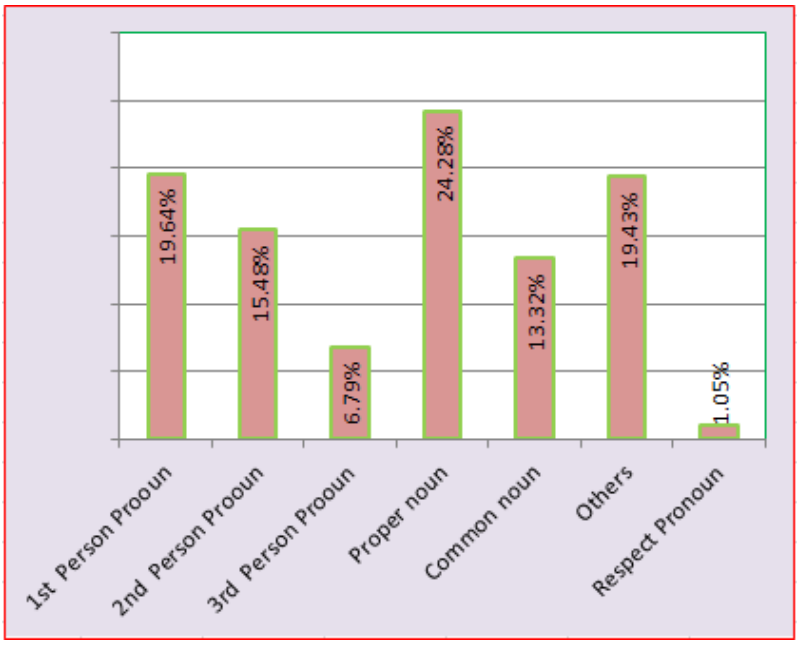

Figure 21: The distribution of mentions used in our system

Among this we take 3 episodes for evaluating the proposed system, that covers 12% of the total scenes used for overall corpus annotated. we summarize the test set as follows:

	Episodes	Scenes	Utterances	Tokens	Speakers	Mentions	Entities
	3	18	291	2775	36	274	86

Table 15: Test Set used for evaluating the character identification system.

The proposed models use some of the provided automatic linguistic annotations, such as PoS or named entity tags and some of them are annotated manually with the help of linguistic experts. In addition, we used the publicly available 300-dimensional word vectors that were pre-trained and provided by facebook fastText corpus with the word2vec continous bag of words model [78]. The word vectors are available in both binary and text formats.

Our annotated corpus can be formatted with Scene-delimiter documents which treat every scene

as a document. The final result of the annotations in our corpus are shown in Table 16 for scene-delimiter documents.

	D	M	C	S	S%	Avg|C|	M:D	C:D	S:D
	155	1965	706	1840	9.4	2.6	12.7	4.55	11.8

Table 16: Annotation statistics and constituent ratios for scene-delimiter documents.

D/M/C: Counts of documents/mentions/coreferent clusters. S/S%: Count of singletons and its composition ratio to all clusters. Avg(|C|): average count of mentions in a chain.
The conll format that was chosen does not have many usable parser. A standardized format such as JSON would have been much better to use. The data was not formatted correctly (tab delimited) which causes issues when we attempt to parse it, since the data is rich in feature there are things it tags as entities which makes no logical sense (punctuation, interjections).

6.1.2 Data Split

All results reported from these experiments are averages of three randomly initialized trials. The corpus in table 17 is split into training, development, and evaluation sets, where all models are tuned on the development set and the best models are tested on the evaluation set. Among the total 25 episodes, 76%, 12%, and the rest episodes of the season are used to generate the training, development, and evaluation sets, respectively. The distribution is based on the number of episodes, so that among the total 25 episodes used in the research work 19 episodes are used for training, 3 episodes are used as development and the remaining 3 episodes are used for testing. In terms of the number of scenes used in the study 80%, 8%, and 12% are used for training, development and testing respectively this research work.

Dataset	Episodes	Scenes	Utterances	Tokens	Speakers	Mentions	Entities
TRN	19	125	1482	15520	71	1524	114
DEV	3	12	127	1222	28	101	80
TST	3	18	291	2775	36	274	86
Total	25	155	1900	19517	135	1899	280

Table 17: Distributions from the subset of corpus used for the character identification.

From the total season used, episodes 1∼ 19 are used for training (TRN), 20∼ 22 for development (DEV), and 23 ∼ rest for evaluation (TST).

6.1.3 Test data set preparation

In this study we applied different methods for the character Identification tasks. The deep learning classifier based on Coreference resolution and entity linking is intensive and takes lot of time to prepare the training and testing data set. To evaluate the performance of the system we perform two separate tests using 86 entities and 274 mentions found in the test Set. Tests are done by evaluating the performance of coreference resolution and entity linking. And we use three different evaluation metrics which fit for the testing of system having mentions and referents. Evaluation is done below in Section 6.2.1 and 6.2.2.

Entity Creation

The mentions identified in the first phase were clustered, according to the output of the classifier, using a greedy clustering algorithm. Starting from the second mention in the document, each mention is compared to all previous mentions, which are collected in a global mentions table. If the pair-wise classifier assigns a probability greater than a given threshold when checking for coreference a new mention against a previously encountered mention, the new mention is assigned to the same entity of the previous one. In case more than one entity candidates, the new mention is assigned to the most likely entity, i.e. the entity including the mention which received the highest coreference score by the classifier. This strategy has been described as best-first clustering by [81]. In principle the process is not optimal since, once a mention is assigned to an entity, it cannot be later assigned to another entity to which it more likely refers.

6.2 Performance evaluation

The evalution metrics proposed by [16] for the CoNLL'12 shared task are B3, CEAFe, and BLANC,are used to evaluate the performance of our character identification models.

6.2.1 Coreference Evaluation Metrics

The study trained as part of the preliminary experiments, were all evaluated with the official CoNLL scorer on the three metrics for measuring coreference resolution: MUC, B^3, CEAFe. The results can be discussed as belows.

According to [45], instead of evaluating exclusively on the coreference chains, the B^3 metric calculates precision and recall values on a mention level basis. The average of all these mention scores is the performance of the system. Hence, for a set M containing mentions m_i, consider coreference chains S_{m_i} and G_{m_i} to be denoting to system and gold responses respectively. Precision - P and Recall - R are calculated as follows:

$$P = \frac{|S_{mi}| \cap |G_{mi}|}{|S_{mi}|},$$
$$R = \frac{|S_{mi}| \cap |G_{mi}|}{|G_{mi}|}$$

For Example -

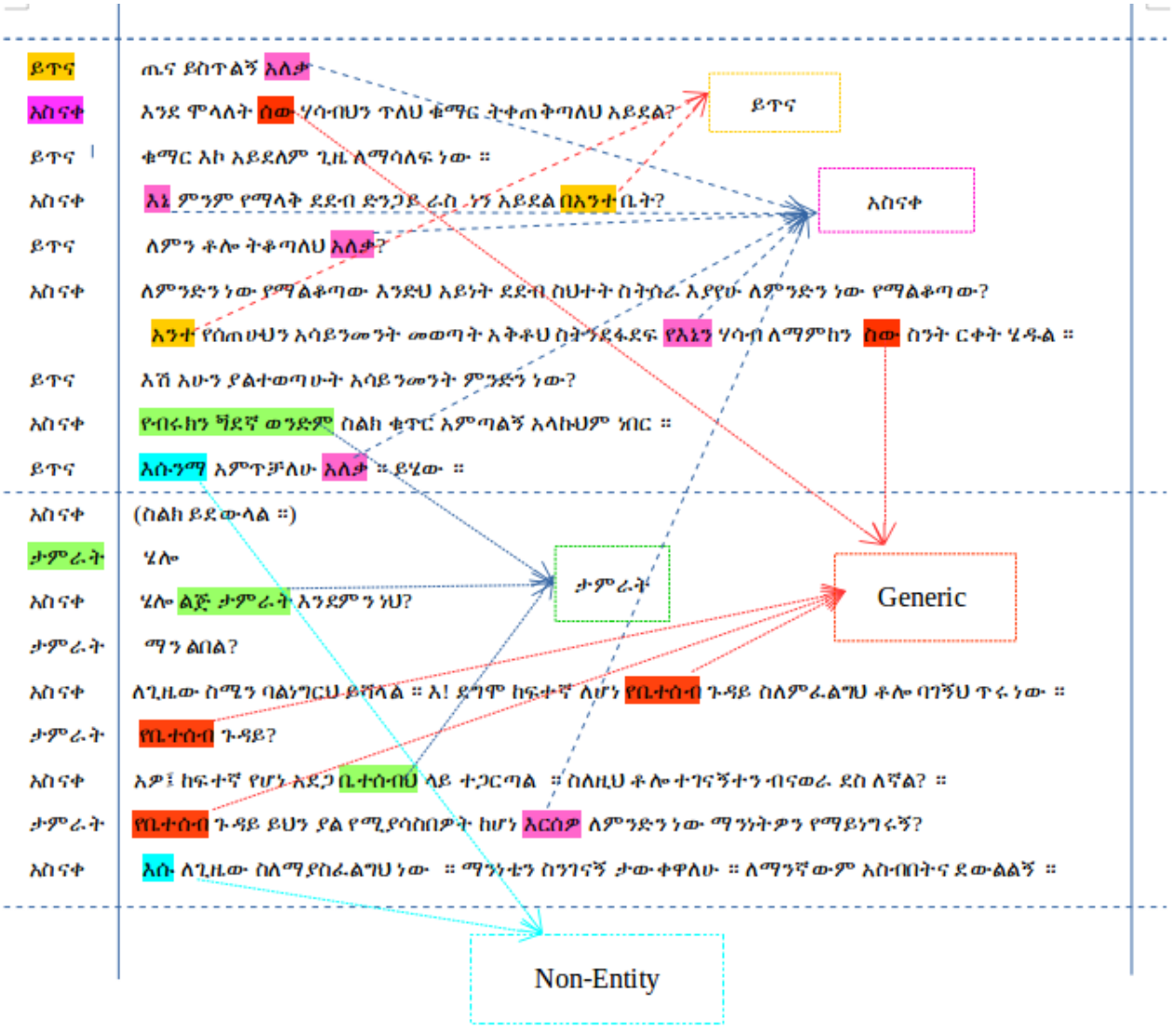

Figure 22: Sample Scene taken from test set.

Mentions identified from the given scene with expected clusters are the following;

```
[አለቃ, ሰው, እኔ, በአንተ, አለቃ, አንተ, የእኔን, ስው, የብሩክን ቫደኛ ወንድም, አለቃ, ልጅ ታምራት, ቤተሰብ, የቤተሰብ, እርሰዎ]

{የቤተሰብ: [ሰው, ስው, ቤተሰብ, የቤተሰብ],
በአንተ: [በአንተ, አንተ, የእኔን],
ልጅ ታምራት: [የብሩክን ቫደኛ ወንድም, ልጅ ታምራት],
ቤተሰብ: [ሰው, ስው, ቤተሰብ, የቤተሰብ],
እርሰዎ: [አለቃ, እኔ, አለቃ, አለቃ, እርሰዎ],
አንተ: [በአንተ, አንተ, የእኔን],
አለቃ: [አለቃ, እኔ, አለቃ, አለቃ, እርሰዎ],
ሰው: [ሰው, ስው, ቤተሰብ, የቤተሰብ],
የእኔን: [በአንተ, አንተ, የእኔን],
አለቃ: [አለቃ, እኔ, አለቃ, አለቃ, እርሰዎ],
እኔ: [አለቃ, እኔ, አለቃ, አለቃ, እርሰዎ],
የብሩክን ቫደኛ ወንድም: [የብሩክን ቫደኛ ወንድም, ልጅ ታምራት],
ስው: [ሰው, ስው, ቤተሰብ, የቤተሰብ],
አለቃ: [አለቃ, እኔ, አለቃ, አለቃ, እርሰዎ]}
```

Figure 23: Sample mentions taken from test set.

Table 18 shows manually annotated mention with the corresponding referent characters;

	አስናቀ	ይጥና	ታምራት	Generic
	አለቃ	አንተ	የብሩክ ጓደኛ ወንድም	ሰው
	እኔ	በአንተ	ልጅ ታምራት	ሰው
	አለቃ			የቤተሰብ
	አለቃ			የቤተሰብ
	የእኔን			ቤተሰብ
	እርሰዎ			

Table 18: Mentions with correspondent entity

The formula for Precision and recall is:

$$P = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{|C_m^s \cap C_m^o|}{|C_m^s|} \qquad R = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{|C_m^s \cap C_m^o|}{|C_m^o|}$$

,Where C_m^s is system mentions and C_m^o is oracle(Gold) mentions.

Precision $= \frac{1}{14} * (\frac{3}{5} + \frac{2}{5} + \frac{1}{5} + \frac{2}{2} + \frac{1}{2} + 1 + \frac{2}{2} + \frac{1}{2} + \frac{2}{2} + \frac{1}{2} + \frac{3}{3} + \frac{2}{3} + \frac{1}{3})$

Precision $= \frac{1}{14} * 10 = 0.7143$

This precision is calculated from mention exist in one scene. To calculate the overall precision we must calculate precision of the 18 Scenes used in our test set and finally we take the average which is 0.9574, and it true for the recall.

Recall = $\frac{1}{14} * (\frac{3}{6} + \frac{2}{6} + \frac{1}{6} + \frac{2}{6} + \frac{1}{6} + \frac{1}{6} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{2}{5} + \frac{1}{5} + \frac{3}{5} + \frac{2}{5} + \frac{1}{5})$

Recall = $\frac{1}{14} * 5.97$

Recall = 0.4264

In general, we were able to achieve a coreference resolution precision of 95.74%, a recall of 44.03% and F1-score of 60.32% on Bcube evaluation metric.

BLANC(BiLateral Assessment of Noun-phrase Coreference) as defined in [82] is best explained considering two kinds of decisions in Mention identification. This evaluation computes the correctness of the mentions that are being resolved, regardless of the structure of coreference links. Standard P and R are computed to compare the sets of mentions of Gold and System. P is defined as the number of common mentions between Gold and System divided by the number of system mentions; R is defined as the number of common mentions between Gold and System divided by the number of true mentions.

The coreference F-measure and non-coreference F-measure can be extended as follows. Coreference recall, precision and F-measure are adapted as:

$R_c = \frac{|C_k \cap C_r|}{|C_k \cap C_r| + |C_k \cap N_r| + |\frac{C_k}{T_r}|}$ . . . . . (1)

$P_c = \frac{|C_k \cap C_r|}{|C_r \cap C_k| + |C_r \cap N_k| + |\frac{C_r}{T_k}|}$ . . . . . (2)

$F_c = \frac{2P_cR_c}{P_c + R_c}$ . . . . . (3)

Non-coreference recall, precision and F-measure are as follows:

$R_n = \frac{|N_k \cap N_r|}{|N_k \cap N_r| + |N_k \cap N_r| + |\frac{N_k}{T_r}|}$ . . . . . (4)

$P_n = \frac{|N_k \cap N_r|}{|N_r \cap N_k| + |N_r \cap N_k| + |\frac{N_r}{T_k}|}$ . . . . . (5)

$F_n = \frac{2P_nR_n}{P_n + R_n}$ . . . . . (6)

The extended BLANC continues to be the arithmetic average of F_c and F_n:

BLANC = $\frac{F_n + F_c}{2}$

Since $T_k = T_r$ and $C_k \subset T_k$, we have $C_k \subset T_r$; thus $C_k \backslash T_r = \Phi$, and $|C_k \cap Tr| = 0$. This establishes that $R_c = R_c^g$. Indeed, since Ck is a union of three disjoint subsets:

$C_k = (C_k \cap C_r) \cup (C_k \cap N_r) \cup (C_k \setminus T_r)$, R_c^g and R_c can be unified as $\frac{|C_k \cap C_r|}{|C_k|}$. Unification for other component recalls and precision can be done similarly. So the final definition of BLANC can be succinctly stated as:

$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|}$ . . . . . (7)

$R_n = \frac{|N_k \cap N_r|}{|N_k|}, \quad P_n = \frac{|N_k \cap N_r|}{|N_r|}$ . . . . . (8)

$F_c = \frac{2|C_k \cap C_r|}{|C_r| + |C_k|}, \quad F_n = \frac{2|N_k \cap N_r|}{|N_r| + N_k|}$ . . . . . (9)

BLANC = $\frac{F_n + F_c}{2}$

- $C_k \backslash T_r$ are key coreference links missing in the response;
- $N_k \backslash T_r$ are key non-coreference links missing in the response;
- $C_r \backslash T_k$ are response coreference links missing in the key;
- $N_r \backslash T_k$ are response non-coreference links missing in the key.

Considering the example given above, we have

C_k =18

N_k =24

C_r =16

N_r = 96, and

$C_k \cap C_r$=11

$N_k \cap N_r$=24

$P_c = \frac{11}{16} = 0.6875$

$R_c = \frac{11}{18} = 0.6111$

$F_c = \frac{2 * P_c R_c}{P_c + R_c} = 0.6470$

$P_n = \frac{24}{96} = 0.25$

$R_n = \frac{24}{31} = 0.7742$

$F_n = \frac{2 * P_n R_n}{P_n + R_n} = 0.3779$

BLANC = $\frac{F_n + F_c}{2}$

$\frac{0.3779+0.6470}{2}$=0.5125

The task is ambiguous to calculate manually, So that we must calculate Precision and recall for each cluster of the test set document, and we take average values. The evaluation is performed automatically and get a precision of 0.7756 and recall of 0.5476.

CEAF - This metric proposed by [47] is an improved version of B3 which had a pitfall that entities could be used more than once during evaluation. As a result, chains with same entity and multiple entity mention chains are not taken into account. Hence, to overcome this CEAF outputs the best one-to-one mapping between gold and system predicted entities, it gives the count of common mentions that pertain to both system and gold labels. This entity based similarity is referred to as $\phi(S_i, S_j)$ or Gold G_i and system S_i. The best similarity measure is denoted as $\phi(g^*)$. Thereby Precision P and Recall R are calculated as follows:

P= $\frac{\phi(g^*)}{\sigma_i \phi(S_i,S_j)}$ . . . . . (1)

R= $\frac{\phi(g^*)}{\sigma_i \phi(G_i,G_j)}$ . . . . . (2)

F= $\frac{2*PR}{P+R}$ . . . . . (3)

Using Ceaf$_e$ evaluation metric our system performs a precision of 35.76% and a recall of 77.47%. In general the performance of the developed system summarized in table as follow using the stated metrics;

ID	Bcube			CEAFe			BLANC		
	P	R	F	P	R	F	P	R	F
	0.9574	0.4403	0.6032	0.3576	0.7747	0.4893	0.7756	0.5476	0.5375

Table 19: Coreference resolution Bcube, Ceafe and BLANC results on the evaluation set

MUC [44] is a link-based metric. Given a document d,recall is computed as the number of common links between the key chains and the system chains in d divided by the number of links in the key chains. Precision is computed as the number of common links divided by the number of links in the system chains. Below we show how to compute (1) the number of common links, (2) the number of key links, and (3) the number of system links. To compute the number of common links, a partition P(S_j) is created for each system chain S_j using the key chains. Specifically, P(S_j) = {C_j^i: i = 1, 2, . . . , |K(d)|} . . . . . .(1)

Each subset C_j^i in P(S_j) is formed by intersecting S_j with K_i. Note that $|C_j^i| = 0$ if S_j and K_i have no mentions in common. Since there are |K(d)|∗|S(d)| subsets in total, the number of common links is c(K(d),S(d)) = $\sum_{j=1}^{|S(d)|} \sum_{i=1}^{|K(d)|} w_c(C_j^i)$, where,

$w_c(C_j^i)$ = $\begin{cases} 0 \; if |C_j^i|=0 \\ |C_j^i|-1 \; if |C_j^i|>0 \end{cases}$

Intuitively, $w_c(C_j^i)$ can be interpreted as the weight of C_j^i. In MUC, the weight of a cluster is defined as the minimum number of links needed to create the cluster, so $wc(C_j^i) = |C_j^i| - 1$ if$|C_j^i| > 0$. The number of links in the key chains, K(d), is calculated as:

k(K(d)) = $\sum_{i=1}^{|k(d))|} w_k(K_i)$ . . . . . (3) where $w_k(K_i) = |K_i| - 1$. The number of links in the system chains, s(S(d)), is calculated as:

s(S(d))=$\sum_{j=1}^{|S(d)|} w_s(S_j)$ . . . . . .(4), where $w_s(S_j) = |S_j| - 1$.

ID	MUC		
	P	R	F
	0.3977	0.5000	0.4430

Table 20: Coreference resolution results MUC on the evaluation set

6.2.2 Performance of Entity Linking

For entity linking, entity labels are predetermined by collecting characters that appear in all three sets;characters that do not appear in any of the three sets are put together and labeled as Unknown. This is reasonable because it is not possible for a deep learning model to learn about characters that do not appear in the training set. Likewise, characters that appear in the training set but not in the other sets cannot be evaluated. A total of 18 labels are used for entity linking that consist of the top 17 most frequently appeared characters across all sets.

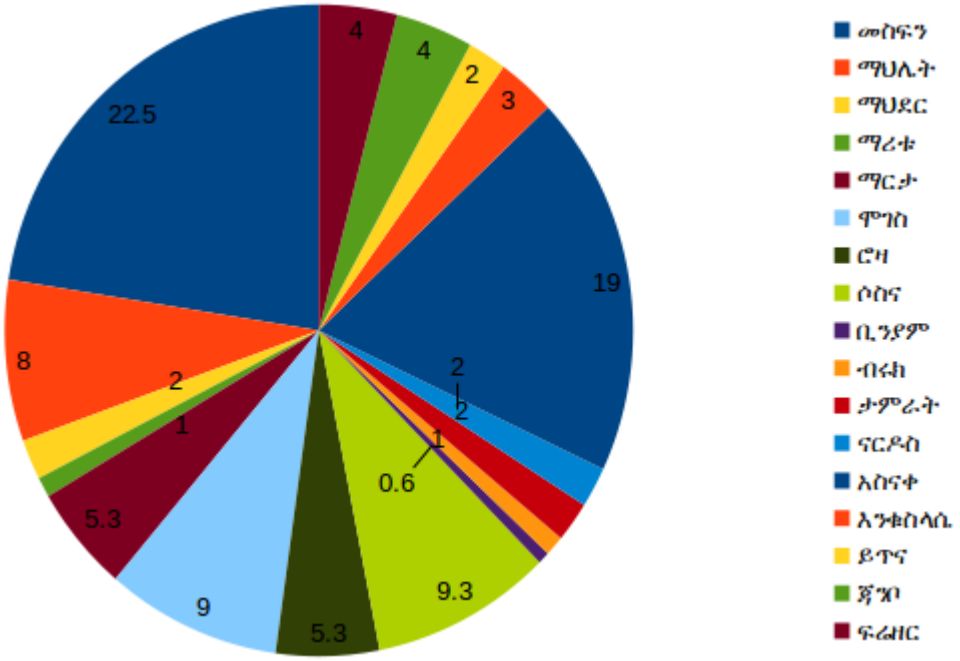

Figure 24: Character labels used for entity linking.

Two metrics are used to evaluate the entity linking models. One is the micro-average F1 score whose precision (P) and recall (R) are measured by taking (D: as a set of documents, N: as the total number of mentions in D, $C_m^{s/o}$: the cluster from the system (s) or the oracle (o) that the mention m belongs to):

$$P = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{|C_m^s \cap C_m^o|}{|C_m^s|} \qquad R = \frac{1}{N}\sum_{d\in D}\sum_{m\in d}\frac{|C_m^s \cap C_m^o|}{|C_m^o|}$$

The micro-average F1 tends to weigh more on frequently occurring entities so it is useful if you need to know the raw prediction power of our model. The other is macro-average F1 score that measures the micro-average F1 for each entity e, say F_1^e, and takes the average, that is $1/|\mathrm{E}| \sum_{e\in E} F_1^e$ where E is the set of all entities.

ID	Singular			Plural			All		
	P	R	F	P	R	F	P	R	F
1	0.3835	0.3835	0.3835	0.3978	0.3893	0.3935	0.3832	0.3697	0.3763

Table 21: The Performance of the Entity linking system.

Since we take only singular mentions, a pseudo-singular dataset is created where exactly one entity is chosen for each plural mention based on the closest matching previous speaker or if there is none, chosen randomly. Thus, the models trained on this pseudo-singular dataset always predicts one entity per mention. The label accuracy considering only 5 entities, that are the 4 main characters (**መስፍን**, **ማህሌት**, **ሶስና**, **አስናቀ**) character who appears in the test set, training set and development set as common, and all the others as one entity is 80.65%. The macro average between the F1 scores of the 5 entities are 77.2%.

Following [15], the labeling accuracy (Acc) and the macro-average F1 score (F1) are used for the evaluation (C: the total number of characters, $\mathrm{F1}_i$: the F1-score for the i'th character):

$$\mathrm{Acc} = \frac{number\ of\ correctly\ Identified\ Mentions}{number\ of\ All\ Mentions}$$

$$\mathrm{F1} = \frac{1}{C}\sum_{i=1}^{c} F1_i$$

The test set data contains total 274 mentions, from this we have 266 mentions are identified, and 45 of them are not referents of the actual characters. And there are 8 mentions which refers actual characters but identified as non character referents(False Negatives).

	Actual Mentions	Actual Non Mentions
Pred. Mentions	221	45
Pred. Non Mentions	8	2501

Table 22: Count break down of mentions in our corpus after disambiguation.

Acc=$\frac{221}{274}$ =80.65

The macro F1 measure of the system are calculated by averaging the f1- score each main characters.

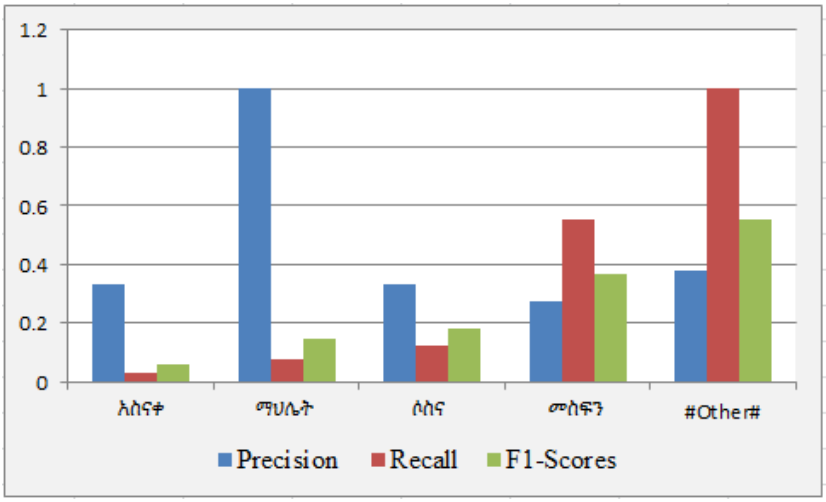

Figure 25: Character identification accuracy

Table 23 shows the overall accuracy of the proposed system in terms of F- measure. Two types of evaluations are performed for this task. The first one is based on the 5 characters 4 of them are the main characters and every others are grouped together as one entity called others(main + other). the second type of evaluation is based on 51 characters comprising all characters appeared in the dataset, except for the ones appear either in the training or the evaluation set but not both, which is grouped to the Others (ALL).

Season	Main + Other		All entities	
	Acc	F1	Acc	F1
1	80.65%	77.2%	85.5%	56.71%

Table 23: The Overall labeling scores of the proposed systems.

For the overall performance, the proposed system outperforms for Main + Others with the labeling accuracy of 80.65% and the macro-average F1 score of 77.2%. Using all entities the model proposed in this work is able to achieve F1 scores of 56.71% in B^3 for scene-delimiter documents. From the results of the systems, we are able to deduce the pros and cons for each system.

6.3 Result and Discussion

Based on the evaluation results, several interesting observations can be made to show how different system architectures affect model performance on this task. The identification of mentions was evaluated according to standard measures of precision, recall and F1, with respect to the gold standard provided by the metrics. The other metrics used in this task are the ones commonly used in previous evaluation campaigns. MUC - based on links, measures precision and recall by comparing gold and system predicted pairs of mentions; BCUBE - based on mentions, computes recall and precision separately for each mention and averages the results; CEAF - associates each predicted entity to the most similar gold entity, using one of two possible variants of similarity, i.e. based on entities and BLANC - uses a variation of Rand index [82] for evaluating coreference resolution.

To the best of our knowledge, this paper provides the first extensive framework for resolving referents for personal mentions, which is a critical task in any resolution task. Omitting twinless system mentions from the training data while keeping the number of correct mentions constant should improve the coreference resolution performance, because a more precise coreference resolution model is obtained.

Chapter Seven

7 Conclusion, Contribution and Recommendation

In this chapter we try to address the brief summary of the research work including the main contribution and future works that could be extended from this work.

7.1 Conclusion

In this research work, we propose a novel NLP task called character identification that aims to find the global entities for all personal mentions, representing individual characters in the contexts of multiparty dialogue. It is shown that the approach for solving this problem has been to model the task as co-reference resolution followed by entity linking for assigning character labels to clusters of named entity mentions. In this work, we have built a character identification system for multi-party dialogues. We have developed a neural approach to coreference resolution using agglomerative CNN which aggregates the feature groups into mention, mention pair representations, cluster and mention-cluster embeddings.

The work presents a deep learning approach to character identification in multiparty dialogues relying only on transcription TV series data. We have used a pretrained word embedding of Amharic Word2vec found from facebook fasttext corpus and have calculated the cluster purity scores for scene delimited documents on the Sewlesew transcripts. Our annotation scheme allows the development of large Dialogue data set with the personal mentions and their referent characters with the help of linguistics expert. Hence, the work provides baseline approaches and results using deep leaning models in order to tackle the task of character identification.

We further disambiguate our corpus and introduce generic groupings of mentions with abstract referent entities. The nature of this corpus is analyzed with potential challenges and ambiguities identified for future investigation. Hence, this work provides baseline approaches and results using existing coreference resolution systems. We also propose a CNN mention-to mention ranking model that provides better mention and mention-pair representations learned from feature groupings of dialogue-specific features.

In this work we develop a generic model for character identification system using different sub components, which contains major components such as preprocessing, corpus annotation and

character identification including coreference resolution, entity linking. Each component is comprised of their own sub-components and algorithms. We used a python Programming language as a developmental tool and other Libraries such as publicly available ngram part-of-speech tagger named TnT Tagger, python based morphological analyzer for Amharic named HornMorpho. When evaluating the system, special emphasis is placed on accuracy and F-1 measures on the main characters as the system will be examined across all entities, as well as across the main characters specifically. We found out that our model best worked on scene delimited documents with an F1 score of 77.2%. Also, we got a character identification accuracy of 80.65% on the 4 main characters of the annotated corpus.

7.2 Contribution

The purpose of this research work is to contribute for the development of character identification system on Amharic multiparty dialogues. Character Identification is one crucial block in natural language understanding because it allows the model to link entities to all their different mentions. Character identification is steppingstone to facilitate and provide entity specific knowledge for systems like question answering and dialogue generation.

7.3 Recommendation

This research work requires us to build a system which can identify different mentions in multiparty dialogues as corresponding characters in the show. As a future work, increase the size of the corpus with high-quality with disambiguate annotation and Global or External features integration are recommended to enhance the performance of the system designed. Augmenting and Enlarge existing corpus to tackle plural and collective mentions. Another recommendation is improving the work for character mining and emotion detection.

8 References

[1] Kevin Clark and Christopher D. Manning. 2015. Entity-Centric Coreference Resolution with Model Stacking. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL'15, pages 1405–1415.

[2] Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics, ACL'15, pages 1416–1426.

[3] Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015.A Joint Framework for Coreference Resolution andMention Head Detection. In Proceedings of the 9th Conference on Computational Natural LanguageLearning, CoNLL'15, pages 12–21.

[4] HengJi, Joel Nothman, Ben Hachey, and RaduFlorian.Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking. In Proceedings of Text Analysis Conference, TAC 15, 2015.

[5] Kevin Clark. Neural coreference resolution." (2015).

[6] Temesgen Dawit, 2014. Design of Amharic Anaphora Resolution Model, Master Degree thesis AAU.

[7] Mahboob Alam Khalid, ValentinJijkoun, and Maarten De Rijke.The impact of named entity normalization on information retrieval for question answering. In European Conference on Information Retrieval, pages 705–710. Springer, 2008.

[8] Ben Hachey, Will Radford, Joel Nothman, Matthew Honnibal, and James R. Curran. Evaluating entity linking with wikipedia.Artificial Intelligence, 194:130 – 150, 2013. URL http://www.sciencedirect.com/science/article/pii/S0004370212000446.

[9] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw.MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP'13, pages 193–203, 2013.

[10] Yu-Hsin Chen and Jinho D. Choi. Character identification on multiparty conversation: Identifying mentions of characters in tv shows. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 90–100, Los Angeles, September 2016. Association for Computational Linguistics.
URL http://www.aclweb.org/anthology/W16-3612.

[11] Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 2011a. Local and Global Algorithms for Disambiguation to Wikipedia. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. ACL'11, pages 1375–1384.

[12] van Deemter, K., Kibble, R.: On coreferring: coreference in MUC and related annotation schemes. Comput. Linguist. 26(4), 629–637 (2000).

[13] Rada Mihalcea and Andras Csomai. 2007a. Wikify!: Linking Documents to Encyclopedic Knowledge.InProceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management. CIKM'07, pages 233–242.

[14] Alex Olieman, JaapKamps, Maarten Marx, and ArjanNusselder.A Hybrid Approach to Domain-Specific Entity Linking. In Proceedings of 11th International Conference on Semantic Systems, SEMANTiCS'15, 2015.

[15] Henry Y. Chen, Ethan Zhou, Jinho D. Choi,Robust Coreference Resolution and Entity Linking on Dialogues: Character Identification on TV Show Transcripts. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 216–225,Vancouver, Canada, August 3 - August 4, 2017. c 2017 Association for Computational Linguistics.

[16] Sameer Pradhan, Alessandro Moschitti, NianwenXue, Olga Uryupina, and Yuchen Zhang. CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes. In Proceedings of the Sixteenth Conference on Computational Natural Language Learning: Shared Task, CoNLL'12, pages 1–40, 2012.

[17] Amitava Kundu, Dipankar Das, and SivajiBandyopadhyay. Speaker identificationfrom film dialogues. In Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on, pages 1–4. IEEE, 2012.

[18] Marco Rocha. Coreference Resolution in Dialogues in English and Portuguese. InProceedings of the Workshop on Coreference and Its Applications, CorefApp'99, pages53–60, 1999.

[19] Daniel Jurafsky and James H. Martin, 2009, Speech and language processing, An Introduction to Natural Language Processing, Computational Linguistic, and Speech Recognition, 2nd Edition, Pearson Prentice Hall, New Jersey.

[20] Girma A. Demeke, "Manual Annotation of Amharic News Items with Part-of-Speech Tags and its Challenges", ELRC Working Papers Vol. 2; number 1: March, 2006.

[21] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4):115–133, 1943.

[22] Paul Werbos. Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. Harvard University, 1974.

[23] Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. (2004). The Automatic Content Extraction (ACE) Program–Tasks, Data, and Evaluation. In Proceedings of Language Resources and Evaluation Conference, (LREC 2004), pages 837–840.

[24] Seokhwan Kim, Rafael E. Banchs, and Haizhou Li. Towards Improving Dialogue Topic Tracking Performances with Wikification of Concept Mentions. In Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGDIAL'15, pages 124–128, 2015.

[25] Iulian V. Serban and Joelle Pineau. 2015. Text-Based Speaker Identification For Multi-Participant Open Domain Dialogue System. Department of Computer Science and Operations Research, Universite ´de Montreal ´ .

[26] Thien Huu Nguyeny, Avirup Sil, Georgiana Dinu and Radu Florian. Toward Mention Detection Robustness with Recurrent Neural Networks. IBM T.J. Watson Research Center, Yorktown Heights, New York, USA.

[27] Vaclav Nemcik (2006). "Anaphora Resolution", Master's Thesis, Masaryk University, Brno.

[28] Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In Empirical Methods in Natural Language Processing (EMNLP), pages 1971–1982.

[29] Kevin Clark and Christopher D. Manning. 2016. Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).

[30] Kamaldeep kaur and Vishal Gupta (2012, June). Name Entity Recognition for Punjabi Language. IRACST– International Journal of Computer Science and Information Technology & Security (IJCSITS),vol. 2, pp. 561-567.

[31] Srihari, R., Niu, C., & Li, W. (2000, April). A hybrid approach for named entity and sub-type tagging. In Proceedings of the sixth conference on Applied natural language processing (pp. 247-254). Association for Computational Linguistics.

[32] Weiqian Yan and Kanchan Khurad.2017. Entity linking with people entities on Wikipedia. Department of Math and Computer Science Emory University.

[33] Appelt, D. and Israel, D. (1999) Introduction to Information Extraction Technology: IJCAI-99 tutorial <http://www.ai.sri.com/ appelt/ie-tutorial/IJCAI99.pdf> (Accessed on 05/09/10).

[34] Feldman, R. and Sanger, J. (2007) The Text Mining Handbook: Advanced Approaches In Analyzing Unstructured Data. New York: Cambridge University Press.

[35] Cowie, J. and Lehnert, W. (1996) "Information Extraction", Communication of the ACM, 39(1), pp. 80-91.

[36] Daniel Khashabi, et.al. CogCompNLP: Your Swiss Army Knife for NLP,11th Language Resources and Evaluation Conference.(2018).

[37] K. Raghunathan, H. Lee, S. Rangarajan, N. Chambers, M. Surdeanu, D. Jurafsky, and Ch. Manning. 2010. A multi-pass sieve for coreference resolution. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 492–501.

[38] M. E. Calif , R. J. Mooney. Relational Learning of Pattern Match Rules for Information Extraction. Proceedings of the ACL Workshop on Natural Language Learning Spain July 1997.

[39] Line Eikvil, Information extraction from World Wide Web, a survey July 1999.

[40] Martha Yifiru Tachbelie and Wolfgang Menzel.(2007). Amharic Part-of-Speech Tagger for Factored Language Modeling. International Conference RANLP 2009 - Borovets, Bulgaria, pages 428–433.

[41] 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, pages 47–54, 2005.

[42] M. Getachew. Automatic part of speech tagging for Amharic language: An experiment using stochastic hmm. Master's thesis, Addis Ababa University, 2000.

[43] Anbessa Teferra& Grover Hudson. Essentials of Amharic,2007.

[44] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model theoretic coreference scoring scheme. In Proceedings of the 6th conference on Message understanding, pages 45–52. Association for Computational Linguistics.

[45] Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In The first international conference on language resources and evaluation workshop on linguistics coreference, volume 1, pages 563–566. Citeseer.

[46] Ladefoged, Peter. 2001. A Course in Phonetics. Fort Worth; Harcourt.

[47] Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 25–32. Association for Computational Linguistics.

[48] Michael Gasser. 2011.Horn Morpho: a System for Morphological Processing of Amharic, Oromo, and Tigrinya. Conference on Human Language Technology for Development Alexandria.

[49] Hobbs, J.: Resolving pronoun references. Lingua 44, 311–338 (1978)

[50] Kameyama, M.: Recognizing referential links: an information extraction perspective. In: ACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts (1997)

[51] Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., Jurafsky, D.: Stanford's multipass sieve coreference resolution system at the CoNLL-2011 shared task. In: Proceedings of the CoNLL 2011 Shared Task, Portland (2011)

[52] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., Jurafsky, D.: Deterministic coreference resolution based on entity-centric, precision-ranked rules. Comput. Linguist. 39(4), 885–916 (2013)

[53] A. Clark. Pre-processing very noisy text. In Proc. of Workshop on Shallow Processing of Large Corpora, pages 12–22, 2003.

[54] Uryupina, O.: Corry: a system for coreference resolution. In: Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval'10), Uppsala (2010).

[55] Bikel, D.M., Miller, S., Schwartz, R., Weischedel, R.: An algorithm that learns what's in a name. Mach. Learn. 34(1), 211–231 (1999).

[56] Borthwick, A.: A maximum entropy approach to named entity recognition. Ph.D. thesis, New York University (1999).

[57] Mulugeta, W and Gasser, M. (2012), "Learning morphological rules for Amharic verbs using inductive logic programming", SALTMIL-AfLaT Workshop on Language Technology for Normalisation of Less-Resourced Languages, Istanbul.

[58] McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of Conference on Computational Natural Language Learning, Edmonton, pp. 188–191 (2003).

[59] Fabio Ciravegna, Claudio Giuliano,Nicholas Kushmerick,Alberto Lavelli and Ion Muslea, Adaptive Text Extraction and Mining (ATEM 2006),11th Conference of the European Chapter of the Association for Computational Linguistics Proceedings of the Workshop on April 4, 2006.

[60] Cunningham, Hamish. GATE, a General Architecture for Text Engineering. Computers and the Humanities. 36. 223-254. (2002).

[61] Yamada, H., Kudoh, T., Matsumoto, Y.: Japanese named entity extraction using support vector machines. Information Processing Society of Japan, SIG Notes NL 142-17 (2001).

[62] Daume III, H., Marcu, D.: A large-scale exploration of effective global featuresn for a joint entity detection and tracking model. In: Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing, Vancouver (2005).

[63] Denis, P., Baldridge, J.: Global joint models for coreference resolution and named entity classification. In: Procesamiento del Lenguaje Natural 42. SEPLN, Barcelona (2009).

[64] Bjkrn Gamback and Utpal Kumar Sikdar.2017.Named Entity Recognition for Amharic Using Deep Learning: IST-Africa 2017 Conference Proceedings. IIMC International Information Management Corporation.

[65] ባየ ይማም ፣ጰጰጸጷ የአማርኛ ሰዋሰው ፤ ትሮመሮማሮማሮድ.

[66] Grover Hudson. 1999. Linguistic analysis of the 1994 Ethiopian census. Northeast African Studies, 6:89–107.

[67] Summary and statistical report of the 2007 population and housing census, URL *www.csa.gov.et/docs/cen2007*$_f$*irstdraft.pdf*., accessed on November 2, 2010.

[68] "Amharic Language", URl: http://www.lonweb.org/link-amharic.htm, last accessed April 02,2018.

[69] Daniels, Peter T. 1992. Contacts between Semitic and lndic scripts, Contacts between cultures: selected papers from the 33rd international Congress of Asian and North African Studies, Vol. 1, West Asia and North Africa, Amir Harrak, ed., 146-152. Edwin Mellen: Lewiston, N.Y.

[70] Daniels, Peter T. 1996. The first civilizations, The World's Writing Systems, Peter T. Daniels and William Bright, eds., 21-32. New York: Oxford University Press.

[71] Getatchew Haile. 1996. Ethiopic writing. The World's Writing Systems, Peter T. Daniels and William Bright, eds., 569-576. New York: Oxford University Press.

[72] Typical CNN architecture is under license CC BY-SA 4.0
URL *https* : *//commons.wikimedia.org/wiki/File* : *Typical*$_c$*nn.png*

[73] Hudson, Grover. 2001. Aspects ofthe history ofEthiopic writing, Bulletin ofthe Institute Ethiopian Studies 25. 1-12.

[74] Kaixin Ma, Catherine Xiao and Jinho D. Choi. 2017.Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics- Student Research Workshop, pages 49–55.

[75] Armbruster, Fishman(1986). The high central vowel in Amharic:New approaches to an old problem.

[76] PoDoLsKy, B. (1991).Historical phonetics of Amharic. University of Tel-Aviv.

[77] Kane, Thomas L. 1975. Ethiopian Literature in Amharic. Wiesbaden: Otto Harrassowitz.

[78] Grave, Edouard and Bojanowski*,Learning Word Vectors for 157 Languages.Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018).

[79] Cowie, J., & Lehnert, W. Information extraction. In Special naturallanguage processing issue of the communications of the ACM (Vol. 39, pp.80–91). New York, NY, USA, 1996.

[80] Hayward, Katrina, and Richard Hayward. 1999. Amharic, Handbook ofthe International Phonetic Association, page 41-49. Cambridge: Cambridge University Press.

[81] Ng, V.: Machine Learning for Coreference Resolution: From Local Classification to Global Ranking. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL), Ann Arbor, MI, pp. 157–164 (2005)

[82] Recasens, Marta & Eduard Hovy (2010). BLANC: Implementing the Rand index for coreference evaluation. Submitted.

[83] Ma, Kaixin & Xiao, Catherine & Choi, Jinho. (2017). Text-based Speaker Identification on Multiparty Dialogues Using Multi-document Convolutional Neural Networks. 49-55. 10.18653/v1/P17-3009.

[84] Francois Mairesse, Marilyn A. Walker, Matthias R. Mehl, and Roger K. Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research pages 475–500.

[85] Jorg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Proceedings of LREC pages 2214–2218.

[86] Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Rmsprop: Divide the gradient by a running average of its recent magnitude. Neural networks for machine learning, Coursera lecture 6e.

[87] Faber, Alice. 1997. Genetic subgrouping of the Semitic languages. The Semitic Languages,·Robert Hetzron, ed., 3-15. London: Routledge.

Appendices

Appendix A: List of Stop Words

ሁሉ
ሁሉም
ኋላ
ሁኔታ
ሆነ
ሆኑ
ሆኖም
ሁል
ሁሉንም
ላይ
ሌላ
ሌሎች
ልዩ
መሆኑ
ማለት
ማለቱ
መካከል
የሚገኙ
የሚገኝ
ማድረግ
ማን
ማንም
ሰሞኑን
ሲሆን
ሲል
ሲሉ
ስለ
ቢቢሲ
ቢሆን
ብለዋል
ብቻ
ብዛት
ብዙ
ቦታ
በርካታ
በሰሞኑ
በታች
በኋላ
በኩል
በውስጥ
በጣም
ብቻ
በተለይ
በተመለከተ
በተመሳሳይ
የተለያየ
የተለያዩ
ተባለ
ተገለጸ
ተገልጿል
ተጨማሪ
ተከናውኗል
ችግር
ታች
ትናንት
ነበረች
ነበሩ
ነበረ
ነው
ነይ
ነገር
ነገሮች
ናት
ናቸው
አሁን
አለ
አስታወቀ
አስታውቀዋል
አስታውሰዋል
አስካሁን
አሳሰበ
አሳስበዋል
አስፈላጊ
አስገነዘቡ
አስገንዝበዋል
አብራርተዋል
አበራርተው
አስረድተዋል
እስከ
አባከህ
አባከሽ
አባከዎ
አንድ
አንጻር
እስኪደርስ
እንኳ
እስከ
እዚሁ
እና
እንደ
እንደገለጹት
እንደተገለጸው
እንደተናገሩት
እንደአስረዱት
እንደገና
ወቅት
እንዲሁም
እንጂ
እዚህ
እዚያ
እያንዳንዱ
እያንዳንዳቸው
እያንዳንዷ
ከ
ከኋላ
ከላይ
ከመካከል
ከሰሞኑ
ከታች
ከውስጥ
ከጋራ
ከፊት
ወዘተ
ወይም
ወደ
ዋና
ወደፊት
ውስጥ
ውጪ
ያለ
ያሉ
ይገባል
የኋላ
የሰሞኑ
የታች
የውስጥ
የጋራ
ያ
ይታወሳል
ይህ
ደግሞ
ድረስ
ጋራ
ግን
ገልጿል
ገልጸዋል
ግዜ
ጥቂት
ፊት
ደግሞ
ዛሬ
ጋር
ተናግረዋል
የገለጹት
ይገልጻል
ሲሉ
ብለዋል
ስለሆነ
አቶ
ሆኖም
መግለጹን
አመልክተዋል
ይናገራሉ

Appendix B: List of Titles

አቶ
ወ/ሮ
ወ/ሪት
ዶ/ር
ሸህ
ቄስ
ከቡር
ክብርት
ሻምበል
ኮሎኔል
አስር አለቃ
አለቃ
ብላታ
ሀኪም
ነጋድራስ
ሀጂ
አርቲስት
አፈ-ጉባኤ
የተከበሩ
አምባሳደር
ኮማንደር
ብርጋድየር ጀኔራል
ሌተናል ኮሌኔል
ሼም
ፕ/ር
አፄ
መቶ አለቃ
ሚስተር
ጠ/ሚ
ሚኒስትር ድኤታ
ብፀአ

አምሳ አለቃ
ሻለቃ
ጀኔራል
ጀነራል
ፕሮፌሰር
ወታደር
ኢንጅነር
ድያቆን
ባላምበራስ
ብላቴን ጌታ
ፊታውራሪ

ዶክተር
ተመራማሪ
ከንቲባ
ሊቀመንበር
ምክትል
ሳጅን
አ/አለቃ
ከንቲባ
ከቡር
ሎሬት
ሀምሳ አለቃ
አሰልጣኝ
አምበል
ኡስታዝ
ኢንስትራክተር
ኢንጂነር
ሰአሊ
ፒያኒስት
ሚ/ር
ጠ/ሚኒስትር

ብላታ
አባ
ደጃዝማች
ኮሎኔል
ሜጀር
ጀነራል
በጅሮንድ
መምህር
ግራዝማች
ብላቴን ጌታ
ባላምባራስ

ጠ/ ሚኒስትር
ፕሬዝዳንት
ካፒቴን
ፓትሪያርክ
ፕ/ት
አፈ ጉባኤ
ማእድንና ኢነርጂ
ሚኒስትር
ወይዘሮ
ጠቅላይ ሚኒስትር
ሸክ
ዋና ዳይሬክተር
ዳይሬክተር
ኢንስፔክተር

ሊቀ ጠበብት
ዶክተር
ሻንበል
ነጋድራስ
ኮሎኔል
ልኡል
ራስ
አቡነ
መምህር