Addis Ababa University

College of Natural Sciences

Cross-Lingual Linking Framework of Wikipedia Articles

Tehetena Alemu Nega

A Thesis Submitted to the Department of Computer Science in Partial Fulfillment for the Degree of Master of Science in Computer Science

Addis Ababa, Ethiopia

July 2020

Addis Ababa University
College of Natural Sciences

Tehetena Alemu Nega

Adviser: *Solomon Atnafu (PhD)*

This is to certify that the thesis prepared by *Tehetena Alemu Nega*, titled: *Cross-Lingual Linking Framework of Wikipedia Articles* and submitted in partial fulfillment of the requirements for the Degree of Master of Science in Computer Science complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the Examining Committee:

	Name	Signature	Date
Advisor:	Solomon Atnafu (PhD)		
Examiner:	Dida Midekso (PhD)		
Examiner:	Solomon Gizaw (PhD)		

Abstract

These days, unstructured and textual data on the web are rapidly increasing and can be find in different languages. We can structure this kind of data by using text analysis with Machine learning methods. The Semantic web uses the obtained structured information to facilitate information extraction on the web. Machine learning methods provide a different way of querying web data using statistical techniques. Among the existing machine learning methods, Latent Dirichlet Allocation (LDA) is the effective one and is a widely used method to analyze unstructured data and to perform allocation.However, LDA may not necessarily be efficient with short documents. Documents with fewer than 50 words are mostly found in tweet texts and Amharic Wikipedia articles. LDA model's depends on the language specific stop-words and on the stemming algorithm for finding the optimal topic. In this thesis, we study the application of machine learning approach that links similar Amharic Wikipedia article with its corresponding English Wikipedia article.

In this thesis, we designed a model which is based on LDA model. We evaluate our model by finding textual similarities between Amharic and English documents. Our model uses stability metric to find the topics in multilanguage texts. And we also propose a method to find conceptual similarity and aggregated it with the result of textual similarities. We use the aggregated result to link Amharic Wikipedia with English Wikipedia which becomes bi-lingual corpus. This allows to improve the Amharic information retrieval on the web. We show how accurate our model is in terms of precision and recall in order to perfectly align Amharic with English Wikipedia. We also showed that our topic stability metrics are related to the contents of the topics with proximity and topic coherence measure.

We analysed masses of text in Wikipedia using LDA, and observed wrong topic detection on Amharic Wikipedia Articles. Our model which is based on Wikipedia inter-language link and Word Embedding model improved the result of the LDA model. Our approach shows stability for unbalanced size of Wikipedia articles. Our experimental results demonstrate also that our proposed model improves the LDA based bilingual document similarity score and provides better overall accuracy and significantly outperforms previous studies.

Keywords: Semantic Similarity, Topic Models, LDA, Word Embeddings, Document similarity, Cross-lingual Linking

Dedication

To: My Family

To my beloved family: Parents, Sister and Brother, whose affection, love, encouragement and prays of day and night make me able to get such success and to my friends, who encouraged and supported me.

Acknowledgements

This thesis would not have been possible without the guidance and the help of lecturers and students of Computer Science Department of Addis Ababa University (AAU), whose insightful comments and encouragement inspired me throughout the work. Firstly, I would like to express my special appreciation and thanks to my advisor Professor Dr. Solomon Atnafu, you have been a tremendous mentor for me. I could not have imagined having a better advisor and mentor for my MSC study. Also, I would like to express my sincere gratitude to my co-advisor Melkamu Beyene for the continuous support, patience, motivation, and immense knowledge through my MSC. I would also like to extend my great appreciation to my friends emotionally throughout this study. My sincere thanks also go to my brother and sister, who facilitate me with materials such as good internet connection to download the required file for the research work. Without they precious support it would not be possible to conduct this research.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Acronyms

ALS	Alternating Least Squares
BoW	Bag-of-words model
BTM	Biterm Topic Model
CANDECOMP	Canonical Decomposition
CLLD	Cross Lingual Link Discovery
CDCR	Cross-Document co-reference resolution
CP	Canonical polyadic decomposition
HDP	Hierarchical Dirichlet Process
IE	Information Extraction
IR	Information Retrieval
KG	Knowledge graph
LD	Linked Data
LDA	Latten Discreet Allocation
LOD	Linked Open Data
LSA	Latent Semantic Analysis
LSI	Latent Semantic indexing
MLOD	Multilingual Linked Open Data
MT	Machine Translation
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
OWL	Web Ontology Language
PARAFAC	Parallel Factors Decomposition
pLSA	Probabilistic Latent Semantic Analysis
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attribute
RP	Random Projections
SPARQL	Semantic query language for databases.
STS	Semantic Textual Similarity

SVD	Singular value decomposition
W3C	World Wide Web consortium
WSD	Word Sense Disambiguation
YAGO	Yet Another Great Ontology

Chapter One: Introduction

1.1 Background

At the begging, the web is designed for communication between humans using shared knowledge. Later, the information on the Web is developed to be understandable by machine. However, search engines do not satisfy the information need of some users, because the information not connected to the knowledge base. The user cultural background including language may influence the required search result. Besides, the demand for information search varies among users. In many cases, users mostly will turn to English information because they might not get the required information on their native language. Unquestionably, traditional web search capacity limited to provide the optimal results for other languages English. Thus, such language bilingual users face a challenge to find a document in any language through queries of a single language. Let us assume a user who provides a query “King David with harp” “ንጉስ ዳዊት ና በገና”.

The worldwide web consortium (W3C) introduced the semantic web to provide answers for such sophisticated queries. The semantic web is a standard shift that brings structure to get meaningful content of the web [1]. The standard data publishing model in the semantic web is a resource description framework (RDF) [2]. RDF is a generic data representation model of a document in a uniform way. In fact, the conception of semantic web advance publishers is including mega web companies such as Google, Facebook, and Amazon to publish their data using semantic web data publishing formats such as RDF leading to the creation of the web of data.

Those early adopters produced a large independent data silo. This brought the need to integrate data silos from various sources. In order to give solutions for the above problems, in July 2006, Berners-Lee proposed a set of principles and technologies known as linked data (LD) [3].

It is widely known the vision of linked data is to make the Internet a global database. In fact, the semantic web is about making links between datasets understandable both by humans as well as machines. W3C initiated a community project to publish and Interlink the existing

open license datasets based on LD publishing principles. The extracted data set in this project formed the linked open data (LOD) cloud. To this end, DBpedia has been publicly available by extracting structured information from Wikipedia [1]. In fact, DBpedia interlinks the RDF triples with other knowledge bases such as YAGO[1] and freebase[2] in the LOD. The results from DBpedia are accessed by the SPARQL querying language.

The huge amount of information stored on Wikipedia has helped for the creation of the largest knowledge bases on the web of data. Wikipedia offers scale and multilingualism that models other knowledge bases and has the ability to evolve quickly and cover most subject disciplines [4]. Besides, knowledge bases created from Wikipedia have been applied in the application such as text categorization, indexing, clustering, and searching and other problem sets. Currently, there are 309 Wikipedia editions in different languages, with 299 active and 10 closed [5]. In a recent development, Wikipedia cross-language links exist between most parts of the edition of the language. For instance, the Italian Wikipedia knows more about Italian villages. With this in mind, the Amharic Wikipedia contains and knows exclusively about Amharic literature. But, some Amharic links miss inter-language links with other languages edition of Wikipedia including the English version. To demonstrate this idea, we can take the Amharic Wikipedia article "ጠፈር" [9] or "ዓዋ" [10] which is not connected to the English page "Universe" [11].

The main problem in linked data Internationalization is that some languages of the Wikipedia editions have few resources compared to other languages. For instance, the English Wikipedia has 6,010,615 articles pages currently whereas Amharic Wikipedia has articles count within 14,802 [5, 6]. However, by recommending Inter-language links between different Wikipedia pages, it is possible to increase the less-resourced Wikipedia articles.

We believe writing a query in a different language should return results in the language the user understands but that is not the case usually. For Instance, Google uses the location of the user to run a search. Following this, one of the factors to ensure the efficiency and effectiveness of multilingual web search results is by linking similar documents in a different

1 http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

2 http://www.freebase.com/

language. Significantly, cross-lingual entity linking is an essential technique for numerous search and NLP applications.

At present, Wikipedia language editions are linked to each other manually by each contributor via graphical user interfaces. However, sometimes Wikipedia contributors forget to indicate the corresponding page for all languages or even for one language. Sometimes, pages are incorrectly mapped. In addition, when the number of disconnected pages rises for a given language, the task of manually connecting those pages with other language edition becomes time consuming. For this reason, linking manually different Wikipedia editions is not adequate.

Currently, Wikipedia has more than 280 language editions [21, 22]. Each Wikipedia edition has its individual community and unique collection of content. In fact, seven of the language editions found in the portal page of Wikipedia, have millions of articles and have a higher growth rate [21]. Among these seven editions, the English Wikipedia is the largest site of them all. These seven languages available on the Wikipedia main portal pages [21] are the gateway to the other language release of Wikipedia. However, even now it is not promising for Wikipedia users to discover any data on their preferred language. Thus, cross-lingual link discovery (CLLD) is utilized to handle the issue of these language barriers in using knowledge based such as Wikipedia in LOD.

Cross-lingual entities linking on web give the opportunity for semantic search. One of the benefits is for search engine users to find direct answers to their information needs instead of only document. Therefore, the main goal of our study is to link Amharic with English knowledge base.

1.2 Motivation

Multilingual knowledge bases are important in knowledge sharing and communication in today's world. Most knowledge bases including DBpedia1, YAGO2, and Free Base3 are built from different Wikipedia editions. In the Meantime, knowledge bases from non-English languages are less than the English one. Above all, multilingual knowledge bases suffer from the data sparsity problem of a non-English knowledge base. For this reason, there is limited coverage of cross-lingual Wiki-based knowledge links on the web. In fact, multilingual knowledge bases help to break the language barrier on information sharing in the worldwide

web. In particular, no study, to our knowledge, has considered English-Amharic comparable corpus build from Wikipedia editions to increase non-English languages in linked data.

In a global society, users are looking for information in their native or non-native language. And the cross-language search helps users to query in their native language to acquire the required result. Though, such multilingual language searches need multilingual knowledge systems to share information in a standardized way between different languages on the web. Knowledge sharing communities such as Wikipedia with information extraction and machine translation has advanced the automated construction of machine-readable knowledge bases. Wikipedia based knowledge bases are built in several different languages, however, multi-lingual ontologies based on the wiki still face the limited coverage of cross-lingual knowledge links.

Aligning the different knowledge bases spread on various Wikipedia editions will enhance the construction of a coherent knowledge base and benefit many applications. When the two Wiki knowledge bases are in different languages, we call it the cross-lingual knowledge linking problem [4]. Since the cross-lingual alignment of articles by human labor Automatic is very costly and error-prone, cross-lingual knowledge linking is a highly desirable task. For this reason, we need a real-time system that automatically extracts from massive raw texts and updates the knowledge base.

Many information systems on the web that need access to structured information are based on open knowledge bases such as DBpedia, YAGO, and Freebase. Most of the time that knowledge base is made from semi-structured knowledge such as Wikipedia. Open data includes an ever-growing number of data silos using DBpedia as a central hub. Internet users usually search for information dispersed to different datasets. In linked data knowledge bases or knowledge graphs are connected for an integrated view of the data. Knowledge graph (KG) represents the collection of interlinks real-world entities to show how things relate to each other [5]. Knowledge graphs assist to provide a high-quality answer for search engines.

Google incorporates the knowledge graph to improve search results from different sources [6]. Google provides actual information about the subject of your query rather than just a list of links. Google uses a page rank to determine site importance using human-generated links [7]. It is assumed that web pages linked to many important pages are also important. This

conclusion hinges on cross-lingual links between pages' help search engines to accept queries in various languages and display the needed information regardless of language. Knowledge linking is a core problem in the knowledge base and represents a new research direction.

Basically, Cross-lingual knowledge linking globalizes multilingual knowledge linking. Similarly, Wikipedia Interlingua link used to link same concept across language for multilingual knowledge base. Wikipedia inter-language links can appear as links in the "Languages" section or as inline links in the text of the article. These two types of links are handled differently. Unfortunately, language link searching and then editing is demanding as well as time consuming. To the best of our knowledge, there is no previous work has been studied on linking English Wikipedia with Amharic Wikipedia.

In this research work, we plan to consider performing tagging and recommending task on real world articles that depends on Amharic and English Wikipedia Dump files, by using practical system that is in the process of development. Cross-language link discovery and recommendation is a useful work in improving interlinking of Wikipedia and will be used to facilitate the task of asking complex Amharic queries from Semantic Application. We use Wikipedia inter-language link along with machine learning technique such as topic modeling and word embedding to find inter-language link between two similar Wikipedia articles in different language.

1.3 Statement of the Problem

Wikipedia is a multilingual encyclopedia on the web. Wikipedia can connect its different language editions by using cross-lingual links. Since cross-lingual links are established by human contributors, all Wikipedia pages do not have a cross lingual links. In fact, more than 80% of Wikipedia languages editions have fewer than one hundred thousand articles [18]. For this reason, users of Wikipedia pages cannot have cross-lingual links in user's mother tongue.

Nowadays, the growth of LOD clouds creates a pressing issue of heterogeneity. Linked Open Data (LOD) is linked data that is also open data means its uses of open sources. One of the causes for semantic heterogeneity is the natural language variation in knowledge bases. Wikipedia faces the problem of language barriers. For instance, anchored links in Wikipedia are created within the same language. In addition, the absence of links between different language domains limits the knowledge sharing and the discovery in LOD cloud. Because

most language editions contain small fraction of information in Wikipedia, English-to-Amharic cross-language link discovery is supposed to use dictionary based translation from English articles. The other problems in the English-to-Amharic cross-language link discovery are the lack of multilingual semantic resource to map Amharic ontology with English ontology. For instance, other languages that use machine translation or dictionary or a community effort have similar problems.

But, by using the English-to-Amharic cross-lingual link recommendation, it is possible to fill the gap in article coverage between Amharic and English Wikipedia editions. Huayu Lu et al. in [19] have been able to link Chinese and English knowledge bases. The method of adapting different language for Cross-language link between articles in [19] has its own challenges and need further study.

NLP and machine learning can be used to extend Wikipedia RDF-triple. First machine learning technique can be applied to find candidate class and candidate defining axioms. Then, textual evidence is extracted from the list page and other Wikipedia attributes. The result of machine learning with textual evidence can be used to select super class in Wikipedia RDF-triple.

We also have to determine the similarity degree and the level so that we can map Amharic synonyms words such as "በላ and ተመገበ" in Amharic Wikipedia title to English Wikipedia title. The Wikipedia titles are retrieved from extracted RDF Triple of Amharic Wikipedia. A statistical translator or Google API can be used interchangeably to translate the Amharic Wikipedia title into English or French phrase. Then machine learning technique along with natural language processing technique used to preprocess the translating phrase. First we clean text to eliminate less useful part of the text by removing stopwords and remove some special characters. the second part of the preprocessing step is annotation means we apply structural markup and part-of-speech tagging. In the third part of the preprocessing step normalization of the text has been done using stemming and lemmazation. Finally, we apply statistical and machine learning techniques on preprocessed data using LDA topic modeling and word embedding used to mine meaningful topic for the document and make document similarity. After preprocessing, we need to generate an algorithm based on the ranking result of the

similarity to interlink concepts. We plan to use Wikipedia feature such as interlanguage Link in computing Similarity between documents.

There are different methods to extract linkage such as Named Entities or thesaurus concepts and also Machine learning. This motivates our works by identifying the appropriate alignment technique of the Amharic Wikipedia titles with Other language Wikipedia Pages using its list pages and attributes.

In recent years, a survey of literature on Similarity between Wikipedia articles is scarcely reported [20]. In addition, lack of evaluation benchmark that automatically measure cross-lingual similarity methods. Unfortunately, few studies have been carried out on language-independent similarity methods between Wikipedia articles written in different languages. Therefore, the aim of this thesis is to develop a language-independent approaches for measuring similarity in Wikipedia. Also our study considers linking Wikipedia knowledge base with DBpedia Ontology.

This thesis work seeks to address the following research questions:

- ✓ What are the characteristics of similar articles in different language edition of Wikipedia?
- ✓ Can language-independent approach be used to find similar documents in the web? And the following sub-questions are investigated
    - o How accurate is a language-independent method compared to previous linguistic resources a method that depends on machine translation systems or WordNet?
    - o How does data sparsity of a document affect the accuracy of the method?
    - o What kind of language-independent method works best for cross-lingual similarity for all Wikipedia editions?

1.4 Objectives

The general objective of this work is to design a generic language-independent Framework for inter-language linking of Wikipedia articles.

The specific objectives of the study are to:

- Review related work in cross-lingual ontology mapping.
- Identify the language specific characteristic of ontology mapping.
- Identify methods to extract Amharic and English Wikipedia dump files.
- Design a framework for mapping Wikipedia inter-language links.
- Test the designed framework on Amharic Wikipedia.
- Evaluate the performance of the system.

1.5 Methods

In order to conduct this research work, the methodologies mentioned below will be used to select and implement appropriate methods and techniques.

Literature Review

Different literatures that are relevant to the subject of this thesis work will be reviewed. Documents which can add value from a variety of sources, such as journals, proceedings of conferences, books, research works, and materials from the internet will be considered.

Data Collection Method

The pages of all English and Amharic Wikipedia articles and inter language link were extracted from a data dump of Wikipedia from December 2017 will be created [61]. This includes standard page titles and article content, page number and language id.

Tools

A variety of natural language processing tools are required to implement prototype and to evaluate the findings of the research. A tool which best fits for data analysis purpose is used.

Prototype Development

In order to show how our framework functions on cross-lingual Amharic Wikipedia with English Wikipedia article alignment, a prototype system will be developed and tested. The accuracy of the prototype will be evaluated.

Experimental Evaluation

An experimental research method will be used throughout this thesis. The research procedure involves investigation of the problem area, proposing solution and finally evaluation of the implementation of the framework. Best, two performance metrics will be used to evaluate the similarity degree. To check the performance of our approach, we plan to use the latest English and Amharic Wikipedia dumps as a test sample. We then check the linking result of implementation for correctness with an experts' view.

1.6 Scope and Limitations

We consider conducting experiments on the developed framework for mapping Wikipedia inter-language links. Most research on Cross lingual knowledge linking has been carried out in a specific language. But, they do not consider Amharic to English language knowledge mapping, which have its own challenge. The developed framework will be generic enough to work for all Wikipedia articles but experimental test will be conducted only for Amharic and English Wikipedia Article.

1.7 Application of Results

Most of the Wikipedia content is unstructured and only provides textual search with no capability to support sophisticated queries. This study helps Internet users to pose complex queries in their native language. Moreover, the Amharic Wikipedia can be a central interlinking hub where all Amharic linked datasets will be interlinked. It can also be a guideline on how Amharic linked data could be published and how Amharic characters can be handled in linked open data.

1.8 Organization of the Thesis

This chapter include chapter one with introduction. The remaining parts of this thesis report is organized as follows. In Chapter Two, we present a literature review in cross-lingual link discovery methods. Chapter Three covers related work to the study. Chapter Four is about data collection and preprocessing. Chapter Five presents the proposed framework with detail description. Implementation of the proposed framework with its evaluations and analyses are presented in Chapter six. Chapter seven presents the conclusion and provides the future research direction.

Chapter Two: Literature Review

2.1 Overview on Cross-lingual link discovery

Cross-lingual link discovery deals with the search of a potential link between two different Wikipedia language articles. Nowadays, the growth of Chinese document on the web directs most recent studies to focus on Cross-lingual link discovery on Chinese languages. For example, English to Chinese Cross-Lingual link discovery has been studied in [21, 22]. Cross-Lingual link discovery are among the most widely investigated types of Wikipedia link discovery. The main problem of Cross-lingual links in Wikipedia is when human editors sometimes do not establish a link for a given page. For this reason, not all Wikipedia pages have Cross-Lingual links. A study on Cross-lingual link discovery between English to traditional Chinese articles is presented by Liang-Pu et al. [21]. The works deal with the finding of the missing links in Wikipedia articles.

Automatic Cross-lingual links discovery methods have been proposed by different researchers [22, 23, 24] to automate link discovery. Most early studies as well as current work on cross-lingual link discovery are implemented either by matching ontology or by schemas. Several literature reviews of schema matching [25, 26, 27] and ontology matching [21] have also been published over the past decade. In fact, schema and ontology matching approaches are classified in two main groups named element level matching and structure level matching [21]. Element-level matching is computed by analyzing only the meaning and the content of element between two schemas or ontologies. While structure-level approaches use hierarchies and compute similarities by looking at the relationship that the single components and the component groups have.

Element-level matching approach is either a constraint-based or language based. Constraint-based approach considers data types values ranges, keys uniqueness of values. But, Language-based approach uses soundex edit distance and NLP. Element levet algorithms are used to decide potential matching structures. Thus, structural-level matching approach can be either taxonomy-based or model-based approach. Then, the result of the matching from both structure methods can be completed using techniques such as linguistic matching, constraint-based matching and so on.

The task of cross-lingual link discoveries (CLLD) involves connecting concepts in knowledge base with concepts in other articles. Respectively, the process of automatically distinguishing concept mentions in a text document and connecting it to a concept referent in a knowledge base (KB) such as Wikipedia is called Wikification. The task of Wikification involved first in disambiguating a mention in a document and defines the comparable Wikipedia title. Wikification within English language has been studied broadly in [29, 30, 31]. At the point when Wikification engaged with two different natural languages, it is named Cross-Lingual Wikification or Entity-Linking [32]. As indicated by [32], Text Analysis Conference (TAC), Knowledge Base Population (KBP), Entity Linking Tracks have examined connecting significant mentions in the English Wikipedia with the key concept of the text on Spanish and Chinese language [32].

The proposed system of Liang-Pu et al. in [22], enables users to automatically tag potential links for a given article. In addition, the system is able to automatically find cross-language link of the tags. The system uses both English and Chinese Wikipedia dmp files to find missing links of Wikipedia articles. The main goal of the paper is to find missing links and automatically tags articles. The paper utilizes two translation mechanisms to find the corresponding cross-lingual translation. These two translation mechanisms are called Pattern translation and Google translation practices by the study of Lucene software that deals with ambiguous phases in articles. On the other hand, to automatically tag articles the system is first engaged with candidate discovery. N-gram tokenizer and maximum matching algorithm are adapted to appropriately use the whole keyword as an anchor.

After finding potential terms and relevant pages in Chinese Wikipedia, Chinese terms are connected with the English candidate terms. While finding a term, the candidate could have the same title but with different meanings. There are also terms overlapping problems for the candidate. To select adequate or better candidate, the system first checks long term in Wikipedia on its anchor look-up table to find a true meaningful term and uses Wikipedia redirect pages to find disambiguation information and candidate terms. To evaluate the performance and the correctness of cross-lingual link discovery, randomly selected four bilingual news articles from Yahoo are used and tagged with Chinese articles with human experts. The experimental result proved that the proposed method that is the maximum match algorithm has a better performance and 8% higher coverage than the original Wikipedia

anchor link transformation technique. The authors in [22] argued that the proposed system has low precision for WSD (Word Sense Disambiguation). Thus, in order to gain a higher performance, they proposed machine learning.

A more general technique which is practical to all languages in Wikipedia has been implemented [32]. The technique proposed by [32] does not consider a machine translation technology and it is practical for low resourced languages. Principally, the task of Wikification is to deal with title ambiguity and with variability. Titles ambiguity and variability means that titles can be expressed in multiple forms including in synonym words. The issues of ambiguity and variability can be handled by calculating similarity between the mention of the documents and the candidate Wikipedia article titles. The other challenge of Wikification is the matching of words from one language article to a title of another language article. Such challenges of cross-lingual matching of titles have been handled in [32] using multilingual title and word embedding. Multilingual titles and word embedding refer to vector representation without language restriction of both distinct mentions and individual Wikipedia title in order to compute meaningful similarity.

The paper in [32] addresses the challenge of generating English title candidate from the given foreign mention by following the inter-language links in Wikipedia. However, this proposed system does not consider missing Wikipedia pages in another language and proposed to use a transliteration model. The authors in [32] argue that the proposed disambiguation method which identifies the most probable title is conceptually the same, so our method could generalize the same as this case.

The paper addresses the test of creating English title from Wikipedia articles. Nevertheless, the proposed framework in [32], does not consider missing Wikipedia pages in other languages. The authors in [32] oppose that the proposed disambiguation technique which recognizes the most acceptable title works for all languages. So, our strategy could sum up like this case.

2.2 Cross-lingual link Recommendation

Wikipedia is the largest online encyclopedia that covers many written languages in the world. The different language editions of Wikipedia are being connecting to each other using a Wikipedia cross-lingual link. However, cross-lingual links of Wikipedia articles are

established by human contributors. For this reason, not all Wikipedia pages have cross lingual links. Under those circumstances, not all Wikipedia pages have their corresponding pages on other language edition. The proposed system by authors in [33] reduces knowledge gaps in Wikipedia by recommending articles for creation. To identify missing articles, articles should be ranked according to their importance. On the other side, the system recommends important missing articles to editors based on their interests.

The first step of the recommendation system is the identification of the missing links. Then, the recommender system associated with the ranking of the missing articles is ordered according to the significance of the article. The last operation of the recommendation system is to suggest missing articles based on the interest of the editors.

The authors validated the proposed system by running large-scale controlled experiments that involved twelve thousand French Wikipedia articles. The authors in [32] stated that their results demonstrate that personalized article recommendations are the effective way of increasing the creation rate of important articles in Wikipedia.

Study of both English mono-lingual and English-to-Chinese documents linking has been covered by the recent research on link discovery in [33, 34]. However, linking other languages documents to the English language has its own problem's need that should be addressed. Chinese document linking with its English counter version is first addressed by paper [34].

Chinese Wikipedia is modified by different part of the world where users have different knowledge and language variation. Given that, it's important to segment texts in order to link Chinese documents to English documents. Based on the language's specific issues for document linking, the paper proposes in [34] a link discovery framework that recommends cross-lingual link in Chinese to target document in English. The proposed Chinese-to-English Cross-linking framework utilizes three main tasks. The first task is Chinese word segmentation, and then proceeds to the Chinese to English translation. The system uses Machine Translation Such as Google Translate API and triangulation which uses a page name matching. The triangulation translation approach mines in one language and identify target documents translated into the second language. The final task of the proposed Chinese English linking is link mining (ML). The methods used state of the art mono-lingual link discovery named Link mining (ML) in [36] and Page Name Matching (PNM) method in [35] used to

weight phrases of links based on different instances. Link mining method applied to mine links, existed in a single language version of Wikipedia. Link mining method constructs link table, T link, of mono-lingual anchor-to-target (a→d) pairs. Link table predict the likelihood succession of terms as an anchor.

Chinese to English link discovery is important to first build a table of relating English and Chinese document. Then, for a new Chinese document, the system identifies all substrings using n-gram from page title table list of the anchors, and then matches it with the corresponding English document. The recommended anchors are either article titles that readers might look up or ones that were previously linked by the human editors. The recommended anchors are the final step of link discovery approach. The experimental result demonstrates that link discovery framework has successfully recommend Chinese-to-English cross-lingual links. The system evaluation obligated to use 50 links per document. On the other hand, related to recommending English cross-lingual linking, the authors created 36 new Wikipedia topics to mine the remaining corpus created in the form of table. According to the experimental result, their link discovery framework can efficiently recommend Chinese-to-English cross-lingual links including recommending Wikipedia articles for further reading. The authors also claim that segmentation and machine translation are not important for link discovery. However, the authors in [37] suggest linkage factor graph model to a further improvement in performance.

2.3 Latent Dirichlet Allocation (LDA)

In this section, we will give a brief introduction to Latent LDA Latent Dirichlet Allocation (LDA) first introduced by Blei et al. [38]. LDA is the most popular topic modeling technique used for text mining. It can be used to model and finds the relevant topics along with its organization in unstructured corpus. In this particular case, the LDA model consumes a number of documents and treats those documents as probability distributions over latent topics. This suggests that a given document probability distribution of individual topics defines how likely each word is to appear in a given topic.

LDA assumes documents are produced from a mixture of topics, where a topic is a probability distribution over this set of terms. Those topics then generate words based on their probability

distribution. The generated words in each document are related to each other. In a simple way, a "document" is a "bag of words" with no structure beyond the topic and word statistics.

Figure 2.1 shows a concise way of visually representing the dependencies among the model parameter. The boxes are "plates" representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document. Only the shaded plate represents words that are observed. α and η are the parameters of the respective Dirichlet distributions. And the Plate represents repetition.

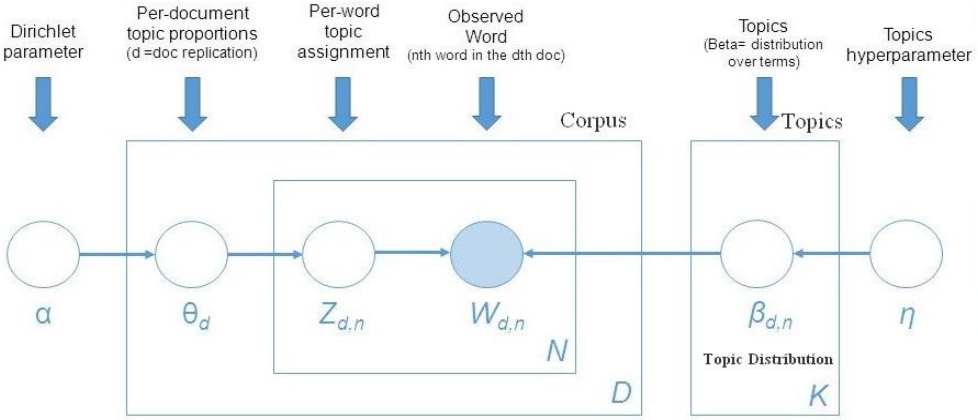

Figure 2.1: *A Graphical model representation of LDA.*

Given a corpus D consisting of M documents, with document d having N_d words ($d \in \{1...M\}$), the generative story of LDA for Document D consists of the following steps [38]:

LDA requires the definition of a set of hyperparameters α and β used in the inference. In above generative process, all known variables are words in documents while others are latent variables (φ and θ) and hyper parameters (α and β). In order to infer the latent variables and hyper parameters, the probability of observed data D is computed and maximized as follows:

According to the generative graphical model depicted in Figure 2.1, the join distribution of observed and hidden variables can be written as: -

$$\mathcal{P}(D|\alpha, \beta) = \prod^{M}_{d=1} \int \mathcal{P}(\theta_d|\alpha) \left(\sum_{n=1}^{Nd} \mathcal{P}(Z_{dn}|_d)\, \mathcal{P}(w_{dn}|\, Z_{dn,}\, \varphi)\, \mathcal{P}(\varphi|\, \beta)\right) d\theta_{\,d}\, d_p \qquad (1)\ [38]$$

With LDA, the terms in the collection of documents produce a vocabulary that is then used to generate the latent topics. Documents are treated as a mixture of topics, where a topic is a

probability distribution over this set of terms. Each document is then seen as a probability distribution over the set of topics. We can think of the data as coming from a generative process that is defined by the joint probability distribution over what is observed and what is hidden.

The words with highest probabilities in each topic usually give a good idea of what the topic is. LDA could be further extended with Variational Bayesian, expectation–maximization algorithm, and Gibbs sampling.

2.3.1 Biterm Topic Model (BTM)

LDA able to detect topics well for normal texts. However, low resourced language such as Amharic has short texts, usually have less than 50 characters. In fact, the sever data sparcity problem makes the topic modeling in short texts difficult and unreliable. To solve thos issue, many researchers have proposed various effective approach based on "Biterm Topic Model" [39, 40].

A Biterm Topic Model learns topics by directly modeling the generation of word-word co-occurrence patterns instead of single words. (e.g., Biterms) at corpus-level rather than at document-level. In contrast, LDA model word-document co-occurrences.

Unlike, LDA models that consist of the word occurrences, a biterm consists of two words co-occurring in the same context in a corpus. Biterm topic model (BTM) effectively models short text by capturing the rich global word co-occurrence information. In fact, a biterm is generated by drawn two words independently from a same topic. In other words, the distribution of a Biterm b=(wi,wj) is defined as:

$$P(b) = \{sum\_k\{P(wi|z)*P(wj|z)*P(z)\} \quad (2)\ [38]$$

With Gibbs sampling algorithm, we can learn topics by estimate P(w|k) and P(z).

Though, short-text has a sparse structure with many highly related words never co-occur together. Still, BTM may lose many possible coherent and word co-occurrence patterns in the corpus. Short text topic modeling using word embedding has been addressed by Ximing et al [42]. The author use word embedding to address the issues of data sparsity and lack of co-occurrence words in Wikipedia article of low resourced language.

2.3.2 Word Embedding

Word embedding is one of the models to represent document vocabulary in the vector space model. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. Word embedding allow us to capture similar context such as semantic and syntactic similarity between words.

All previous approaches to learn cross-lingual representations have been based on some form of language model or matrix factorization [41]. In contrast, Søgaard et al. [43] propose an approach that does without any of these methods, but instead relies on the structure of the multilingual knowledge base Wikipedia, which they exploit by inverted indexing. Their method is based on the intuition that similar words will be used to describe the same concepts across different languages.

In Wikipedia, articles in multiple languages deal with the same concept. We would typically represent every concept with the terms that are used to describe it across different languages.

2.3.3 The Bilingual Topic Modeling

In monolingual settings, the majority of text mining research using topic models is based on the [44, 38] models and its variants. pLSA and LDA have found applications in document clustering, text categorization and ad-hoc information retrieval, but are not suited for cross-lingual text-mining since they were designed to work with monolingual data.

In multilingual settings, knowledge is mined from text by relying on machine-readable multilingual dictionaries or by using multilingual data. Since machine-readable dictionaries are not available for all languages pairs, the latter approach is more flexible. Multilingual data either refers to parallel corpora or comparable corpora. A parallel corpus is a collection of documents in different languages, where each document has a direct translation in the other languages. Hence, a parallel corpus is data-aligned at the sentence level. Parallel corpora are high-quality multilingual data resources, but they are not widely available for all language pairs and they are limited to a few narrow domains (e.g., the parliamentary proceedings of the Europarl corpus [45]). Therefore, text mining from comparable corpora has gained interest over the last few years. A comparable corpus is a collection of documents with similar content which discusses similar themes in different languages, where documents in general are not

exact translations of each other and are not strictly aligned at the sentence level. Unlike parallel corpora, comparable corpora by default comprise both shared and non-shared content

2.4 Summary

Recently, the World Wide Web is enriched with information other than English. However, much of the data we found on the web are in unstructured form and multilingual coverage is unbalanced. Since, the size of the non-English contents on the web is increasing at a very high rate. Multi-lingual or cross-lingual searching has become the upcoming global knowledge sharing trends. Thus, we need an effective method to extract, organize and search meaningful information from sites in different language on the web. However, there is a challenge to extract meaningful information from those unstructured text. In fact, unstructured text written in different natural languages has name variations and entity ambiguity. This concern of unstructured data analysis demands the bridge of Web data with knowledge base and raises a research in entity linking.

Natural language processing (NLP) techniques resolve the challenge of extraction of important information from unstructured data and represent into a structured format. The structure format of the data can be used for further data and document analysis. An example of NLP and unsupervised machine learning algorithms is probabilistic topic modeling method for discovering the common theme of textual data. In fact, topic modeling is gaining the attention of researchers on different number of application such as cross-lingual entity linking because of the expansion of textual data [55, 56, 57]

Little research has been done on cross lingual linking of a low resourced language [58,59]. Previous works in this field focused primarily on bilingual dictionary data. Parallel corpus is one of the most invaluable resources for many multilingual natural language processing applications. Under-resourced language such as Amharic language lack parallel corpus such as bilingual dictionary data. As a result, we propose a method that computes similarity between equivalent Amharic and English Wikipedia articles. We first build Amharic-English Bilingual Corpus of Wikipedia's Articles to automatically align Amharic Wikipedia article with its corresponding English article using probabilistic topic models. The aims of the proposed Bilingual Corpus are to support future natural language processing research works such as information extraction and cross-lingual entity Wikification.

Chapter Three: Related Work

This chapter presents related works on Amharic to English cross lingual link discovery system. The chapters have two section Cross-lingual link discovery and Cross lingual link recommendation.

The language independent similarities between entities have been studied by Beyene Melkamu et al [51]. The aim of the research is to detect co-referent entities located in different Linked Open Data (LOD) sources in different natural languages. The author used two different approaches structural and textual similarities. The first approach learns the structural similarities of entities of RDF graphs. Multi-scale analysis of the RDF graph is used to find the structural similarities. This is implemented as a tensor decomposition of RDF graph. The second approach tries to find textual similarities of entities. To elaborate, for each entity mention a textual data coming from the web documents have been associated. Textual data has been preprocessed and each entity is represented in a high dimensional space (using tensor) with each dimension corresponding to terms. Lastly, a global similarity is computed from textual and structural evidence using aggregation mechanism linear opinion pool method. A global similarity vector corresponds to potential co-referent entities.

The experiment is conduct using RDF data sets taken from the 2014 French and English localized DBpedia editions. The structural analysis is conducted by entities and predicates of the person class entities of predicates starting with subjects and objects from the person class. Predicates occurring less than 100 times are omitted and common predicate has been identified through inverse functional properties and their transitive. The authors use 200 entities from the English DBpedia edition that are connected by owl:sameAs ontology to the French DBpedia so as to evaluate this approach. From the preprocessed datasets, a RESCAL factorization of the tensor is prepared and its cosine distance has been computed. The result cosine similarity is converted to vector probability as a result of the structural analysis. On the other hand, textual information is extracted from Wikipedia and other sources, then as stop word removal and stemming been conducted. Entities are represented as a weighted vector using if.idf weighting scheme. The proposed approach by the author have a high results of average precision of 95%. However, there is lack of bench mark data set on the research work. On the other side, we learned that structural evidence revealed performance irregularities with

different parameters such as the number and type of predicates used. Thus, combining the structural and textual information of entities provide a realistic result to use a web of document with the web of linked open data.

In 2016 Beyene Melkamu et al [60] identified an automatic entity co-reference resolution in a multilingual linked open data (MLOD). The authors use DEDICOM tensor factorization model and introduce relations weight into a tensor factorization. The author adapted three phase approach to enhance existing SRL techniques by reasoning techniques. In the first approach, statistical relational learning (SRL) with factorization of three-way tensor is used to compute structural similarity between entities. In the second approach, entities' textual data are projected into a cross-lingual topic space using latent Dirichlet allocation (LDA) to find the textual similarities between entities. Then, finally belief aggregation strategy is used to combine textual with structural similarity. Similar to the previous research by same author, there is a lack of standard benchmark datasets that can be representative to the amount of heterogeneity of the research work. As a consequence, the experiment of the study is conducted on dataset from French and English DBpedia sub graph by introducing SPARQL query.

The algorithm of the research study outperforms the state of the art approaches based on tensor decomposition for the task of entity co-reference resolution in a MLOD setting. The model is evaluated against RESCAL and RELATIONAL and result in a better performance improvement. The proposed model ECRMLOD's precision and recall for co-referent entities has been calculated. Also ECRMLOD have 5-7 percent 62-score performance increase by integrating unstructured textual information from unstructured sources. The result shows that ECRMLOD has a consistent performance among intra-language and inter-language data sources. The authors consider scalability and syntactical heterogeneity problem by introducing entities textual descriptions from unstructured sources as an additional source of knowledge. The algorithm is efficient and scalable to extract textual descriptions of LOD entities from the web of documents. We learned that by using bilingual LDA model to project textual description into a language independent latent space, the authors are able to get an effective result. In this research work, we learned that DEDICOM is better than RESCAL to the task of entity co-reference resolution in a MLOD setting.

Many documents similarity measures depend on language-dependent resources such as dictionaries or machine translation (MT) systems. Unfortunately, under-resourced languages lack the linguistic resources. Recently, an approach to measure cross-lingual similarity across Wikipedia has been studied by Monica [20]. The author developed language independent approaches to measure cross-lingual similarity in Wikipedia. The first key contribution of the research work is Wikipedia similarity corpus to learn the similarity characteristics of Wikipedia articles. The corpus facilitates the comparison of various approaches for measuring cross-lingual similarity. The corpus is created from eight language pairs including both well-resourced and under-resourced languages of Wikipedia articles. The second main contribution of the study is language-independent approaches to measure cross-lingual similarity in Wikipedia articles.

The author conducted four different experiments to investigate a lightweight" language-independent features. The first experiment uses Wikipedia links to align similar sentences. The second experiment studies the usefulness of content similarity features. The third experiment examines the use of structure similarity features (such as the ratio of section length, and similarity between the section headings). The final experiment considers a combination of these features in a classification and a regression approach. The performance of the study is evaluated against the human judgments in the similarity corpus. The strength of the study is that it can be applied to any languages written in Latin script and non-Latin script languages. However, non-Latin script languages to use this approach first the script need to be transliterated. The approaches achieved best performance with the combination of all features in a classification and a regression approach. The results of the study show that the random forest classifier was able to classify 81.38% document pairs correctly in a binary classification problem, 50.88% document pairs correctly in a 5-class classification problem, and RMSE of 0.73 in a regression approach. Thus, the result achieved by the study is higher compared to classifier algorithms using machine translation and cosine similarity of the tf-idf scores. This study also shows that language-independent approaches can be used to measure cross-lingual similarity between Wikipedia articles.

These studies have identified that there is a need to improve similarity in Wikipedia. This study focus on improve similarity in Wikipedia articles between Amharic and English

languages. To the best of our knowledge there is no previous work that link that Amharic and English Wikipedia language versions.

Summary

As we have already discussed, the author language independent approach [20] does not incorporate non-Latin script languages without transliterated. However, our method will address non-Latin script languages such as Amharic without any further processes. Moreover, we will develop a tool to automatically align English and Amharic Wikipedia Corpus.

Beyene Melkamu [51] uses the aggregate the structural and textual evidence to find document similarity. The linear average of the two evidences may negatively affect the global similarity result. Besides, the study needs more vocabulary coverage of the parallel/comparable corpus for efficient result. But, this also degrades the overall performance of the proposed system. This implies the model might not have a good result for short text such as Amharic Wikipedia. Thus, the alignments of Amharic Wikipedia with rich English Wikipedia need further investigation.

Chapter Four: Cross-Lingual Linking Framework of Wikipedia Articles

4.1 System Architecture

In this chapter, we present the proposed Cross-lingual knowledge linking framework. The first part of this chapter presents the details on proposed framework. The second part explains the proposed Bi-Lingual Latent Dirichlet Allocation (LDA) model and Wikipedia inter-language Link model and Cross-lingual Entity Linking implementation.

The system architecture shown in Figure 4.1 is the proposed framework for Cross-lingual linking of Wikipedia. The framework comprises three components: Bi-Lingual LDA Model, Wikipedia Inter-language Link Model and Cross-Lingual Entity Linking Model.

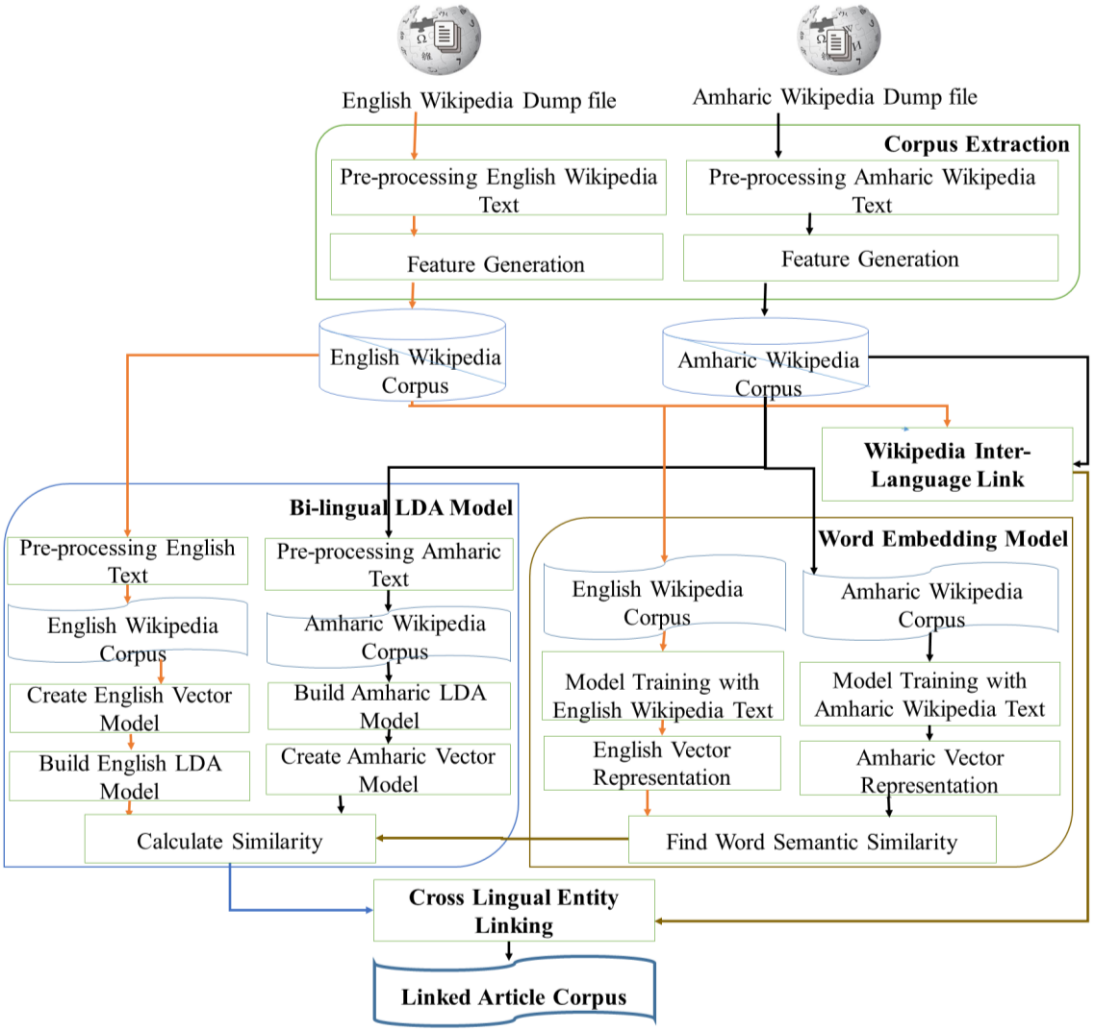

Figure 4.1: Cross-lingual Linking framework

4.1.1. English Wikipedia Corpus Extraction

We extract full content of English Wikipedia articles from Wikipedia dump downloaded from MediaWiki on December 2017 [61], which is available in XML format. This single Wikipedia dump file contain all articles with the respective content and meta-information (such as page id, creation date etc.). Corpus preprocessing based labeling performed with python and Jupyter notebook. The data storage in every step was by using txt files encoded in UTF-8.

Typically, the size of the English-language edition of Wikipedia is much higher than the Amharic-language edition on the same topics. The Wikipedia dump file XML format and the Wikimedia markup of the articles contain lots of information such as formatting that is irrelevant to the statistical language modeling, where we are concerned only with the words and how they form sentences. In fact, the data from Wikipedia xml dump is a semi-structural. Therefore, we need further preprocessing on those data, prior to applying any analysis techniques.

The first step of preprocessing and training data preparation is to remove the stop words from the articles because they do not contribute to the theme of the article's content. Similarly, stemming or lemmatization is an effective process in order to treat various inflected forms of words as a single word as they essentially mean the same.

Preprocessing

We follow a typical workflow of data preparation for natural language processing (NLP). Textual data is transformed into numerical feature vectors required as input for the LDA machine learning algorithm.

Check the Corpus

Large Wikipedia dump file is resulted in a very large corpus file. Given its enormous size, you may have difficulty reading the full file into memory at one time.

Raw Text Extraction

The textual data extracted from the Wikipedia XML dump file is then normalized by removing numbers, punctuation and other special characters and using lowercase. Raw text extracted from Wikipedia dump were tokenized splits the sentences into words (tokens) that are

separated by whitespace and made lowercase; URLs, and HTML tags, and common English" stop-words" were stripped; infrequent words (which occurred less than 1 time in the entire corpus) and extremely short text (with less than 50 words) were removed; and words were lemmatized. These operations are performed using the NLTK Python library's default POS tagger [62].

Stop Word Removal

We only want to analyze English and Amharic Wikipedia articles and have to identify the language since no such tag is available in our data set. A simple classification between English and Amharic as the primary language is achieved by comparing the fraction of stop words in the text. Stop words are the most common words of a given language such as "a", "of", "the", "and" in English.

Filtering out stop words and stemming

The last preprocessing steps are filtering out the English, as these common words presumably do not help in identifying meaningful topics, and stemming the words such that, for instance, "test", "tests", "tested", and "testing" are all reduced to their word stem "test".

After having trained an LDA model, we inspect the topics and identify additional stop words, which are filtered out for the subsequent model training. This procedure is repeated, as long as stop words appear in the lists of top words.

Feature generation

The feature vectors are then generated following a simple bag-of-words approach using gensim doc2bow. Each document is represented as a vector of counts, the length of which is given by the number of words in the vocabulary, which we set to 350, 4880, for Amharic and English Wikipedia Corpus. The CountVectorizer is an estimator that generates a model from which the tokenized documents are transformed into count vectors. Words have to appear at least in two different documents and at least four times in a document to be taken into account.

After the data was preprocessed and filtered, the Wikipedia English language Wikipedia Corpus was divided into an 80/20 training and testing data split.

Now we can present the algorithms that automatically generate our English datasets.

We extract a plain text from the English Wikipedia dump by using English language stop words with Unicode encode scheme by following the natural language preprocessing steps mentioned above.

The algorithm for extracting English corpus is presented in Algorithm 4.1.

Algorithm 4.1: Algorithm for extracting English Wikipedia corpus

Input : English Wikipedia Dump

Output : English Wikipedia Corpus

Split title, text , page in from English Wikipedia input Dump

Apply Preprocessing and transformed to Vector

Textual data extracted and normalized m

Apply Tokenization on raw text

Remove Stop words on raw text

Apply stemming on raw text

Return Normalized text

4.1.2. Amharic Wikipedia Corpus Extraction

The first step of preprocessing and training data preparation is to remove the stop words and stemming or lemmatization. Amharic Wikipedia extractor passes similar process as that of the English Wikipedia Corpus, except stop words and lemmatization treated for Amharic specific cases.

Preprocessing

Amharic textual data is converted into numerical feature vectors used as input for the LDA machine learning algorithm.

Check the Corpus

Large Amharic Wikipedia dump file converted to corpus which is difficult to read full content on memory at a time.

Extract Raw Text

The textual data extracted from the Amharic Wikipedia XML dump file is then normalized by removing numbers, punctuation and other special characters and using lowercase. Then, raw text extracted from Wikipedia dump converted into tokens. We remove URLs, and HTML tags, and common English" stop-words", infrequent words, extremely short texts less than 50 words. And words lemmatized using NLTK Python library's [62].

Stop Word Removal

We create our own Amharic Stop word lists. Stop words are the most common words of a given language such as "a", "of", "the", "and" in English. And the Amharic Stop words are "እዚህ", እዚያ, ነው", "", "". Lists of stop words for different languages are provided by NLTK.

Filtering out stop words and stemming

We filtering out those stop words from the dump file of Wikipedia. After having trained an LDA model, we filtered out the succeeding model training. This procedure is repeated until all stop-words removed on the text.

Feature generation

The feature vectors are then generated using bag-of-words approach using gensim doc2bow. After the data was preprocessed and filtered, the Wikipedia both Amharic language Wikipedia Corpus were divided into an 80/20 training and testing data split.

Now we can present the algorithms that automatically generate our Amharic datasets.

We are interested in the plain text of each Wikipedia article. But the content of Wikipedia article might contain elements such as list, tables, and images and excerpts with irregular indentation (e.g. outdent). So, we remove each article noisy element and convert to raw text for further data analysis.

Extract a plain text from the Amharic Wikipedia dump by using Amharic language stop words with Unicode encode scheme by following the natural language preprocessing steps mentioned above.

Algorithm 4.2: Algorithm for extracting Amharic Wikipedia corpus

Input : Amharic Wikipedia Dump

Output : Amharic Wikipedia Corpus

Split title, text , page in from Amharic Wikipedia input Dump

Apply Preprocessing and transformed to Vector

Textual data extracted and normalized m

Apply Tokenization on raw text

Remove Stop words on raw text

Apply stemming on raw text

Return Normalized text

4.1.3. Bi-lingual LDA Model

The Bi-LDA component of the framework is based on an unsupervised method LDA that enables to identify Amharic and English topics. We adapt LDA model for both English and Amharic Corpus separately. LDA parameters have been tuned to optimize training for recognizing the right topic in Amahic and English texts. The model for both are saved separetly in the disk. After wards, we project each langauge LDA model in different vector space. Finally, we implement a Hellinger distance similarity between Amahric and English texts LDA vectors. Hellinger distance provide good result useful for similarity between probability distributions such as LDA topics.

We adopted unsupervised machine learning algorithms LDA model, which extracts latent topic features of Wikipedia corpus, to improve the accuracy for cross lingual linking model. However, directly applying LDA has some limitation espcially on short texts. Such topics are fixed and uncorrelated topics and so on. The main reason is LDA implicitly capture document-

level word-cooccurance patterns to reveal the topics. So , topic modeling on short texts suffer from data sparcity. We adap BTM - Biterm Topic Model to mitigate the issues of sparcity. Beside LDA limitation, LDA is the leader in topic modeling field. At the same time,Word2vec gave a higer quality of word vectors than those obtained by LDA model. For this reason, experiment have been conducted with augmented LDA model with word Embedding to improve the accuracy of Cross lingual linking model. The first method was using LDA over the bi-gram and the second method includes training LDA model over word embedding and probabilistic topic model.

In this phase, we have used Genism and NLTK (Natural Language Toolkit) library for LDA topic modeling. In fact, NLTK is used have active discussion form. NLTK is in text processing libraries for clustering, classification, tokenization, stemming, wrappers for industrial strength NLP libraries. Also, we use word2vec pre-trained model for word embedding algorithm.

We combine the document similarity result of bilingual LDA model with the result of Wikipedia Inter-language link model to find those missing links from Wikipedia Inter-language Linking model. Also we use Sesnsetovec of Genisim to model sense distinctions of words for a better result.

4.1.3.1.LDA Model

We made an LDA model estimation and discovery topic distribution of unseen Amharic and English Wikipedia corpus. We separately train an LDA model using an English Wikipedia Corpus and English Wikipedia Corpus.And saved the model to the disk separtly for both languages. We reload the registered model for pre-tranining the model and to complete document similarity. In fact, we update the model by repeatdly tranining with the new English and Amhric Wikipedia Coprus.

This LDA model test and infer topic distribution and compare them to the rest of the corpus to find the most similar article. We use Gensim pakage for topic modeling. Gensim is highly optimised content and easy and felxible for implementtion. The algorithm of LDA is presented as follow.

Algorithm 4.3: Algorithm for LDA Model

Input : Amharic Wikipedia Corpus, English Wikipedia Corpus

Result: A Wikipedia corpus with links mapped from the English Wikipedia and Amharic Wikipedia

Split title, text , page in from Amharic Wikipedia input Dump

for page in AmCr do

    Remove Stop words on raw text English and raw text Amharic

    Apply preprocessing on raw text English and raw text Amharic

    Build LDA Model English on raw text English

    Build LDA Model English on raw text Amharic

    Create English Vector Model from English LDA Model

    Create Amharic Vector Model from Amharic LDA Model

Return Hellinger distance similarity English Vector Model and Amharic Vector Model

4.1.3.2. Word Embedding Model

We used LDA to represent the text as mixtures of topics at document level. However, it is important to be able to extract meaning from text at word, paragraph, document for having a good document similarity model. For this reason, we use word embedding using word2vec to find a vector space representation of the word. We use continuous bag of words (CBOW) word2vec approaches to create word vector representation. The model learns semantic and syntactic similarity between words in Wikipedia corpus.

We first tokenize each text to list of words and trained it to word2vec for model training. We then determine the computation complexity and the accuracy of the model by its dimension. Then, we built vocabulary from both English and Amharic Wikipedia corpus and trained all the words to the word2vec model. Finally, we visualize the similarity and relation between words into two dimensional spaces. The algorithm for Word2Vec training is presented as follow.

List 4.4: Algorithm for Word2Vec Training

```
Data : train_data
Result: A wikipedia corpus in vector representation
for sentence_group in train_data do
   text = tokenize(sentence_group)
        print(len(text))
        print(len(train_data))
        Result = text.split()
   end
end
return Result
```

4.1.4. Wikipedia Inter-language Linking

This model is responsible to align automatically Amharic and English Wikipeia corpus based on their interlanguage links. We use Wikimedia SQL dump of interlanguage to map English Wikipedia with Amhric Wikipedia [61].

The two Wikipeida corpra are mapped by using Wikipedia page id and language number of articles from the SQL dump. Then, for each Amharic title we mange to find the equivalent English title. Basically, Inter-language links show that the page Abiy Ahmed in English Wikipedia is the same entity with ዐቢይ አህመድ in AmharicWikipedia.

Finally, We build a text corpus from Wikipedia using inter-wiki links and able to align content at the document level. The created corpus is a straightforward example of a comparable corpus, since the aligned article pairs may range from being almost completely parallel to containing non-parallel sentences. In fact, Wikipedia provides mapping between any pair of languages. Thus, the results show that the model has a good fit for any type of langauge including non latin character set such as Amharic. However, the limitation of this model is not all Wikipedia articles have language links. For this reason, we propose unsupervised

model for learning textual similarity between Wikipdia articles named Bi-Lingual LDA model.

List 4.5: Algorithm for Wikipedia Inter-language Linking

Input : AmCr : Amharic Wikipedia Corpus **Input** : English Wikipedia Corpus **Input** : M: Mapping between page titles **Output** : distance between Vectors

Split title, text , page in from Amharic Wikipedia input Dump
for page in Linked Corpus do
if link in M then
G[page].addMap(M.to AmCr(link)
Return G

4.1.5. Cross-Lingual Entity Linking Model

Cross-Lingual entity link model is designed to find missing links by aligning corpus by Inter-language link model and use the result of Bi-Lingual LDA Model to align the rest of articles. By implementing both interlanguage link and LDA, we are able to find and align more articles than a single method.

List 4.6: Cross-Lingual Entity Linking Algorithm

Data: AmCr: Amharic Wikipedia Corpus Data: EngCr: English Wikipedia Corpus Data: aggregatedResult: English Wikipedia Corpus
Data: M: Mapping between page titles Result: A Wikipedia corpus with links mapped from the English Wikipedia and Amharic Wirkipedia

```
for page in AmCr do
   large page = M.toEngCr (page) links = aggregatedResult. aggregated(Amharic Wikipedia Corpus, English Wikipedia Corpus , LDAResult, InterlanguageResult)
        if link in M then
        G[page].addMap(M.toAmCr(link)
        end
end
return G
```

4.2 Summary

The main contribution of this research work is a novel approach that aligns web data with a knowledge base such as Wikipedia. We propose a multilingual topic model specifically considered to handle non-parallel data called BiLDA. In BiLDA, when two documents share their topics completely, they are called an aligned document pair. In this thesis work, we are able to show BiLDA model to align two languages pair Amharic and English knowledge base.

While the previous methods [56] are able to make effective use of parallel sentence and documents to learn cross-lingual word representations, they neglect the monolingual quality of the learned representations. Ultimately, we do not only want to embed languages into a shared embedding space, but also want the monolingual representations do well on the task at hand.

Word embedding is used to describe Wikipedia concepts with LDA model. We use word embedding to learn cross-lingual word representations. In this way, we are directly provided with cross-lingual representations of words without performing any optimization whatsoever. As a post-processing step, we can perform dimensionality reduction on the produced word representations using the LDA model. We combine the document similarity result of LDA with Wikipedia article language links with word embedding to acquire better approach.

We study Amharic LDA hyper-parameters α and β and their effect on Amharic topic identification. This study shows LDA is an efficient method with the right choice of hyper-parameters for the specific language. Another important finding of the study is the impact of using a language specific stopwords and the stemming algorithm on topic identification.

Chapter Five: Prototype and Evaluation

In this Chapter, we will describe the implementation of the proposed method which is Bi-lingual alignment. The development environment, the tools, and languages we used for development are briefly discussed.

5.1 Development Environment

The development environment and the tools we used for the development are presented below:

We used a Laptop Computer with a specification of an 8GB RAM, Intel Core if Processor that runs on a windows 10 Operating System, Intel core i7 processor with processing speed of 2.5GHz.

Tools and Languages

During the development process the tools and languages we used are listed below.

The Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

Gensim

Gensim (Rehurek 2008) is a free Python library for topic modelling related to automatic extraction of semantic topics from documents, document indexing and similarity retrieval with large corpora. There are several efficient multicore implementations of popular algorithms in Gensim, such as online Latent Semantic Analysis (LSA/LSI/SVD), Latent Dirichlet Allocation (LDA), Random Projections (RP), Hierarchical Dirichlet Process (HDP) or word2vec deep learning. Once the semantic topics are discovered, the plain text documents can be queried for topical similarity against other documents.

NLTK

The Natural Language Toolkit (NLTK) is a natural language toolkit used in statistical natural language processing in Python programs that work with human language data for applying in. It provides lexical resources such as WordNet, text processing NLP libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning, and wrappers.

Python

Python is a popular programming language from web development to data visualization. It is a versatile, powerful, and a community of supporter. Python is used for web development (server-side), software development, mathematics and system scripting.

5.2 Wikipedia Page Aligner and the Bilingual Corpus

The tools align English Wikipedia page id and title with Amharic Wikipedia page id and title. The Bilingual Corpus is built based on English Wikipedia and Amharic Wikipedia pages that have proper page and language link between articles.

We assess the effectiveness of Wikipedia aligner by means of human contributors. The systems have been able to correctly align all Amharic Wikipedia articles that have inter-language link with the corresponding English Wikipedia Articles.

5.3 Data Set

As we only consider English Wikipedia and Amharic Wikipedia pages that have language link with Amharic Wikipedia version, we removed those documents that don't have inter-language link between Amharic and English Wikipedia Articles. This dataset is made up by 20,764 and 14,000 unique English and Amharic pages respectively, all of them with their corresponding page id and language link id. We used Wikipedia xml dump data set from Wikidata on December 2017 found from Wikipedia foundation project.

5.4 Experiment

Experimental Data and preprocessed Result

The experiments were performed with Cisco UCS C220 M4 UCSC-C220-M4S E5-2623V3 3.0GHz 80 GB RAM 12G RAID 1U Server for experimentation. The main focus of the experiments was to calculate number of correctly mapped English and Amharic Wikipedia articles. We consider the stop word for both articles, however, we do not consider stemming or lemmatization techniques. After a series of experiments, it was found that the algorithm need high tech machine to process large size of the English Wikipedia articles.

Table 5.1 Experiment data and preprocessed results

Language	Articles	Size Compressed xml dump
Amharic	3,522	5.14 Mb
English	5,321,200	13.6 Gb

Bi-Lingual Corpus

In this research, we have selected Wikipedia as a data source for Cross-Lingual Knowledge Linking. However, there exists no publicly available and preprocessed aligned Bilingual Amharic with English Wikipedia Corpus for research. NIST Text Analysis Conference's Knowledge Base Population (TAC-KBP) queries and documents are not available in Amharic but only for ten languages including English. Essentially, the aim of TAC-KBP is to link a given named entity mention from a source document to an existing knowledge Base (KB) [26]. But, those data from TAC-KBP are not directly usable for cross-language knowledge linking purpose. For this reason, we decided to prepare our own dataset (Knowledge Base) by following a series of steps in the following sections.

An EL (Entity Linking) system is also required to cluster mentions for those NIL entities that do not have corresponding KB entries [63]. We will need to use two different data dumps to make English to Amharic multilingual aligned version of the Corpus (Knowledge Base). For this study, we analyzed the English and Amharic version of Wikipedia dumps collected from MediaWiki [63].

A parallel corpus of Amharic and English language of Wikipedia has been extracted from Amharic and English Wikipedia Corpus presented in Chapter four using interlanguage link. We use Wiki inter-language link recorded as latest-page links as sql dump and latest-pages-articles. Aligning of parallel corpora at sentence level is a prerequisite for many areas of linguistics research. During translation, sentences can be split, merged, deleted, inserted or reordered by the translator. This makes alignment the non-Trivial task.

Parallel corpora are not available for all Amharic languages except for Bibile Data. A statistical method in multilingual research domains require huge parallel/comparable corpora. We used different tools for creating comparable articles from Wikipedia for Amharic and languages based on inter-language links and preprocessing the created comparable corpus. We use Wikipedia Page-link and Language link sql dump to extract the Wiki inter-language link. Comparable corpus is used to match similar documents on Topic level.

Corpus preprocessing based labelling, Topic Modeling and Concept Modeling [Bi-Text and co-Reference (Relation Modeling)] phase was performed with python and Jupyter notebook. The data storage in every step was by using txt files encoded in UTF-8. The easiest way to access the full Wikipedia content is to download the Wikipedia dump, which is basically a single XML file containing all articles with the respective content and meta-information (like page id, creation date etc.)

Table 5.2: Sample English Corpus extracted from English XML based Wikipedia dump file

No	Page Id	Article Content
0	689s	Asia is Earths largest and most populous continent located primarily in the Eastern and Northern Hemispheres It shares the continental landmass of Eurasia with the continent of Europe and the continental landmass of AfroEurasia with both Europe and Africa Asia covers an area of about 30 of Earths total land area and 87 of the Earths total surface area The continent which has long been home to the majority of the human population was the site of many of the first civilizations Asia is notable for not only its overall large size and population but also dense and large settlements as well as vast barely populated regions Its billion people constitute roughly 60 of the worlds population In general terms Asia is bounded on the east by the Pacific Ocean on the south by the Indian

		Ocean and on the north by the Arctic Ocean The western boundary with Europe is a historical and cultural construct as there is no clear physical and geographical separation between them The most commonly accepted boundaries place Asia to the east of the Suez Canal the Ural River and the Ural Mountains and south of the Caucasus Mountains and the Caspian and Black SeasChina and India alternated in being the largest economies in the world from 1 to 1800 CE China was a major economic power and attracted many to the east and for many the legendary wealth and prosperity of the ancient culture of India personified Asia attracting European commerce exploration and colonialism
1	23033	Portugal Portuguese officially the Portuguese Republic is a sovereign state located mostly on the Iberian Peninsula in southwestern Europe It is the westernmost country of mainland Europe being bordered to the west and south by the Atlantic Ocean and to the north and east by Spain Its territory also includes the Atlantic archipelagos of the Azores and Madeira both autonomous regions with their own regional governments At 17 million km2 its Exclusive Economic Zone is the 3rd largest in the European Union and the 11th largest in the world Portugal is the oldest state on the Iberian Peninsula and one of the oldest European nation states its territory having been continuously settled invaded and fought over since prehistoric times The PreCelts Celts Carthaginians and Romans were followed by the invasions of the Visigoths and Suebi Germanic peoples Portugal as a country was established in the aftermath of the Christian Reconquista against the Moors who had invaded the Iberian Peninsula in 711 AD After the Battle of São Mamede where Portuguese forces led by Afonso Henriques defeated forces led by his mother Theresa of Portugal the County of Portugal affirmed its sovereignty and Afonso Henriques styled himself Prince of Portugal
2	21482	Nairobi Kenya capital city urban core city Nairobi National Park large game reserve known breeding endangered black rhinos home giraffes zebras lions
3	21175	Nitrogen chemical element symbol N atomic number first discovered isolated Scottish physician Daniel Rutherford

In addition to this, we also performed separate approach to identify similarity between Amharic and English Wikipedia article base on LDA Model.

Table 5.3: Sample Amharic Corpus extracted from English XML based Wikipedia dump file

No	Page Id	Article Content
0	9952	እስያ የአለም ትልቁ አህጉር ነው ስሙ እስያ የወጣ ከግሪክ Aσία አሲያ ሲሆን መጀመርያ በጽሕፈት የተገኘው በሄሮዶቶስ ታሪክ መጻሕፍት ውስጥ ነበር በሄሮዶቶስ ዘንድ እስያ ማለት አናቶሊያ ትንሹ እስያ ወይም ፋርስ መንግሥት ግዛት ነበረ ሄሮዶቶስ ስለ ስሙ መነሻ ግን እርግጥኛ አይደለም ግሪኮች እስያ ከፕሮሜጤዎስ ሚስት ሄሲዮኔ እንደ ተሰየመ ሲያስቡ ልድያውያን ግን ከኮቱስ ልጅ አሲያስ እንደ ሆነ አሰቡ ይለናል
1	8413	ፖርቱጋል ወይም ፖርቹጋል ፖርቱጊዝ Portugal ፑርቱጋል በኤውሮፓ ውስጥ የሚገኝ ሀገር ነው የፖርቹጋል መንግሥት በምዕራብና በደቡብ ከአትላንቲክ ውቅያኖስ በምሥራቅና በሰሜን ከኤስፓኝ መንግሥት ጋር ይዋሰናል ስም የፖርቱጋል ስም ከአንድ ሮማይስጥ ወደብ ስም ፖርቱስ ካሌ የካሌ ወደብ የአሁንም ፖርቱ መጥቷ በዚህ ሁሉ ምክንያት የፖርቱጋል መንግሥት ታላቅነት በሌላው ዓለም እንደ ታወቀ እንደዚሁም ፩ኛ ትልቅ መንግሥት መሆኑ ፪ኛ ክርስቲያን መሆኑ ኢትዮጵያ ባሉት የውጭ አገር ሰዎች አፍ እየተነገረ ዝናው በኢትዮጵያ ቤተ መንግሥት ዘንድ መሰማት ቻለ ዋቢ ምንጭ Weather forecast for Portugal Travel and Tourism office website
2	3696	Empty
3	12505	Empty

Table 5.4: Sample Wikipedia inter-language link Corpus generated form Wikipedia page link SQL and language link SQL files

	English Page Id	English Page Title	Amharic Page Id	Amharic Page Title
0	689	Asia	9952	እስያ
1	23033	Portugal	8413	ፖርቱጋል
2	21482	Nairobi	3696	ናይሮቢ
3	21175	Nitrogen	12505	ናይትሮጅን

Table 5.5: Sample Bi-Lingual Sentence-Aligned Corpus from Wikipedia

	Amharic Article Content	English Article Content
0	"'እስያ"' የአለም ትልቁ አህጉር ነው። ስሙ «እስያ» የወጣ ከግሪክ Aσία (/አሲያ/) ሲሆን መጀመርያ በጽሕፈት የተገኘው በሄሮዶቶስ ታሪክ መጻሕፍት ውስጥ ነበር።	Asia is Earths largest and most populous continent located primarily in the Eastern and Northern Hemispheres It shares the continental landmass of Eurasia with the continent of Europe….
1	እናት ጨቅላዋን ስታመግብ፣ ዛንዚባር"'ሴቶች"' የሰው ልጆች አንስቶች ናቸው። እናቶቻችን፣ እኅቶቻችን፣ ሴት ልጃቸን ናቸው።	A woman is a female human being The term woman is usually reserved for an adult with the term girl being the usual term for a female child or adolescent
2	ስንዴ Triticum ኢትዮጵያ ውስጥ እንዲሁም በመላ አለም የሚገኝ የምግብ እህል ተክል ወገን ነው	Wheat is a grass widely cultivated for its seed a cereal grain which is a worldwide staple food
3	Empty	Empty

This English Wikipedia dump file contains about 11 million entries and has an overall size of 44 GB (unzipped). However, the English Wikipedia comprises only 4.45 million articles (as of February 2013), as there are many additional pages listed in the Wikipedia dump which are no articles, like category pages, redirects, talk pages and other pages such as images and charts. In fact, Language Links Wikipedia dump is updated every six months. Hence, to get appropriate language links between Amharic and English document, we

download the English and Amharic Wikipedia xml dump within one-week time range difference.

We developed LDA topic model, to extract parallel sentences from the preprocessed data. We further propose bilingual alignment model based on LDA to produce a Bi-Lingual Corpus. Bi-Lingual Corpus is needed to meet the demand for the multi-lingual corpus for Cross-Lingual Wikification purpose. The proposed LDA assigns a score to each extracted Bi-Lingual Corpus pair, which measure which is a measure of the degree of parallelism between the two sentences in the pair. These scores allow us to select only those sentences having a certain degree of parallelism.

5.5 Evaluation and Discussion

Model Training and Evaluation

In this section, we conduct experiments on Amharic Wikipedia Corpus, to demonstrate the effectiveness of our proposed approach. We take two typical topic models as our baseline methods, namely LDA, and Word embedding. We utilized each of the trained models (LDA models) to topic model both Amharic and English Wikipedia Articles.

With LDA topic modeling, evaluation performance involves to identify logical subgroupings in our data without prior knowledge of the data and best grouping that can be found. We should also take in mind; evaluating LDA examination of homogeneity of the words comprising the documents in each grouping is often done.

The experiment is conducted on the proposed system using an English–Amharic Wikipedia Bi-Lingual corpus we built, to show that it significantly outperforms the standard Word embedding in terms of perplexity and performance in document alignment. We use document-aligned comparable corpus containing 3989 English and Amharic Wikipedia After the preprocessing the data we evaluate the accuracy of the LDA Topic model using Gensim. Gensim help to report on the perplexity of a topic model on held-out evaluation texts thus facilitates consequent fine tuning of LDA parameters (e.g. number of topics).

Topic models learn topics—typically represented as sets of important words—automatically from unlabeled documents in an unsupervised way. This is an attractive method to bring structure to otherwise unstructured text data, but Topics are not guaranteed to be well interpretable; therefore, coherence measures have been proposed to distinguish between good and bad topics.

Once you have your topics the next step is to determine the ideal model to use in our classifier that is BTM. In different circumstances we may train the model over again with different parameter. In fact, LDA model meet randomness in both training and inference step. Therefore, LDA model generate different topic every time we train on the same corpus.

Therefore, we stabilize the topic generation process by setting a fixed Random Number for LDA Model.

After data was collected and preprocessed, those models were trained in an effort to evaluate the optimal hyper parameter settings for each model. As mentioned above, we use bilingual Corpus we extracted from both English and Amharic Wikipedia dump using inter language link between them. And to report the evaluation results, we use three metrics: Recall (equation 4), Precision (Equation 5) and F-measure.

The most single important step in setting up a good topic modeling system is good preprocessing. We have done preprocessing in the following steps: -

1 ***Stopword removal*** using NLTK's English stopwords dataset.
2 ***Bigram and trigram collocation detection*** (frequently co-occurring tokens) using gensim's phrases. This is our first attempt to find some hidden structure in the corpus.
3 ***Lemmatization*** (using Gensim's lemmatize) to only keep the nouns. Lemmatization is generally better than stemming in the case of topic modeling since the words after lemmatization still remain unstable. However, generally stemming might be preferred if the data is being fed into a vectorizer and isn't intended to be viewed.

Hyperparametr Training

The main hyperparameter that we have to correctly in LDA models set is the number of topics. We select the optimal number of topics using some pyldavis. We give the number of topics, LDA starts shuffling the topic distribution in each document and the word distribution in each topic until the final results shows a high segregation of topics. To perform LDA, we'll use Gensim.

We evaluate the Topic model by proximity and coherence of their topics. In our study, we use popular and effective topic coherence the average point-wise mutual information between pairs of top words in a topic. The top words in each topic should have high

similarity according to a good word embedding model that is a topic model using word vectors.

Compute Model Perplexity and Coherence Score

We use Model perplexity and in particular, topic coherence measure to judge how good a given topic model is. The preprocessing of the data and the LDA and Word2Vec model will be explained and applied. We use NLTK (Natural Language Toolkit) for text processing and Visualization such as easy to use text preprocessors, and text-based model techniques. And the next step is to read the text files and done some text cleanup such as stop-word removal, such as removing newline characters, quotes, and extra spaces. Then from the cleaned texts then we created tokens for every sentence. Once we have the tokens, we need to stem them to their root form.

The objective of this thesis work topic modeling to find similar English Wikipedia article for Amharic Article. We apply Stemming and Lemmatized to reducing the derived words to their root form. Root form of a word increases the strength of the words in the document. After finding root words generate a document term frequency matrix to create an LDA model.

The document term frequency matrix counts how frequently a term appears in a document. To generate a document term frequency matrix, we create a dictionary of documents using the both English and Amharic Wikipedia corpora. The dictionary method, dictionary method traverses each document and assigns a unique id to each unique token along with their counts. The dictionary then converted into a bag-of-words using the doc2bow method to convert the data to require by the LDA Model. The result of the LDA Model into a bag-of-words is a list of vectors equal to the number of documents. Each document vector has a series of tuples with token id and token frequency pair. We can invoke the LDA Model from the Gensim package with our data and other required parameters.

After the preprocessing the data we evaluate the accuracy of the LDA Topic model using Gensim. The most important tuning parameter for LDA models is number of topics and

passes. As we mention above, the size of Amharic corpus is short while English Wikipedia corpus is large. Therefore, we experiment the optimal LDA parameter for Amharic and English Wikipedia corpus by LDAvis. Therefore, GridSearch performed to determine the best LDA model. We evaluate accuracy of the model by calculating the perplexity metric. It measures he goodness of the model. The lower the value better is the model.

It is difficult to assess LDA subjects, Because of many parameters and each run gives slightly different result. Thus to compare topic models we propose Data-oriented evaluation and Bilingual Corpus alignment evaluation. First we use a Data-oriented Topic Evaluation Methods that is a perplexity to measure the model fit between dataset. The model with higher log-likelihood and lower perplexity (exp (-1. * log-likelihood per word)) is considered to be good. (Log Likelihood: Higher the better, Perplexity: Lower the better). The Result we get from the perplexity Evaluation is high for Amharic corpus than for English Corpus.

We Evaluate of the Alignment using the aggregated Similarity Precision and Recall to measure whether a document is correctly alignment or not.

Coherence Scores

On a different note, perplexity might not be the best measure to evaluate topic models because it doesn't consider the context and semantic associations between words. Probability-weighted word lists produced by the topic model allow correspondence of semantic topics to semantic concepts representable by a topic model. We use Coherence evaluation to rate coherence or interpretability of a set of words by a topic model. However, when evaluating topic model, its metrics usually do not agree.

Bigram Cross Validation

The first method was using LDA model directly over the unigrams and the second method involves training LDA over Word embedding for concept modeling and document Similarity.

Precision and Recall

We directly evaluate document-aligned comparable corpus of English and Amharic Wikipedia to measure the similarity between Amharic and English Wikipedia Article using precision and recall. We used 10 samples labeled Amharic-English Wikipedia articles for training. The LDA Model perform well on finding textual similarity task.

In this paper, we use Skit learn library to measure the precision, Recall and F-Score that result in good overall accuracy more than 97%. Thus we can see result is promising and outperforms all the baseline methods.

All the top informative Word Embedding model features seem to perform well, except the model find analogy for some words. For instance, the system finds Man analog to woman but wrongly analogy for Doctor which is Pizza.

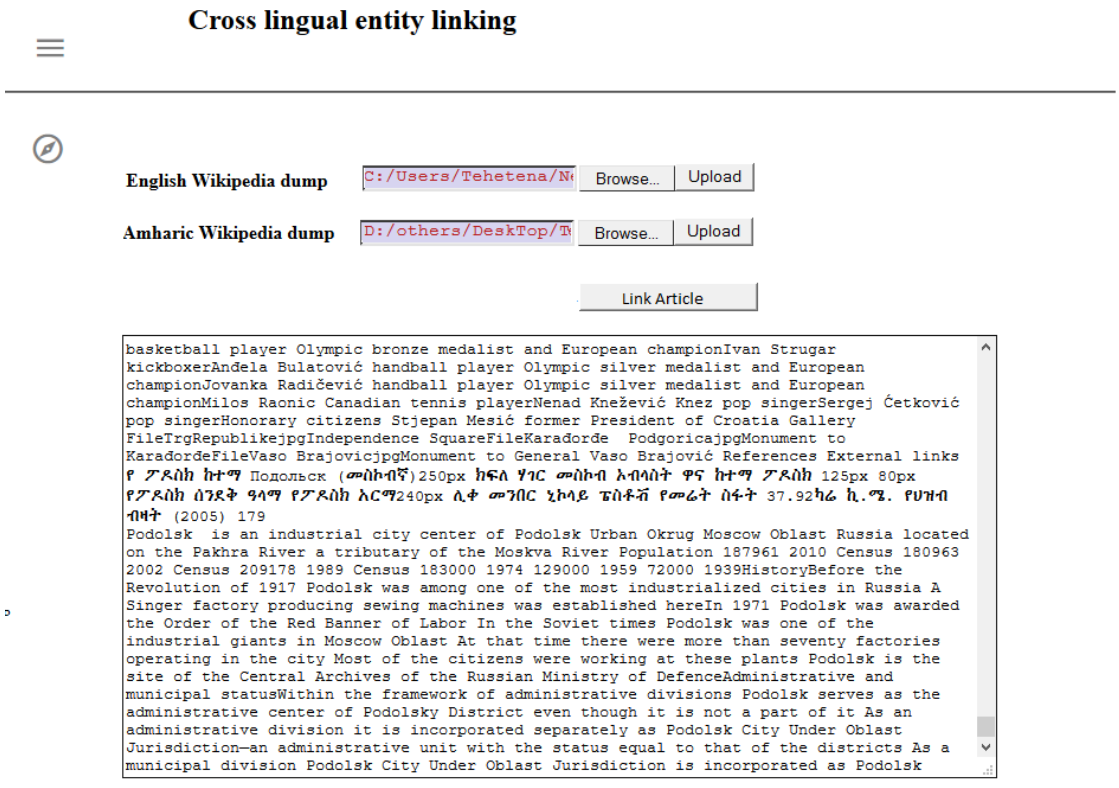

Figure 5.1: User Interface of the proposed Cross-lingual Linking framework

On the other hand, LDA topic model have difficulties in finding short text topics but provide a better result when the size of the text increases. We believe our Bilingual LDA model have quite similar performance to the result achieved by Melkamu Beyene et al [56]. We use Bilingual LDA model with Word Embedding model to find Semantic Similarity of English documents. Lingual Entity Model uses both Bilingual LDA model result with Word Embedding model to link Similar Amharic and English Wikipedia Articles.

Chapter Six: Conclusion and Future Work

6.1 Conclusion

In this thesis work, we have contributed to a new way of study on linking Amharic Wikipedia with English Wikipedia through the use of Topic Modeling. The aim of the study is to find a method that measure cross-lingual similarity between Amharic and English interlanguage-linked Wikipedia articles. The basic methodology of our approach is to facilitate production of bilingual corpus of English and Amharic Wikipedia. Our study base on the previous work done by Melkamu Beyene et al. linking of French Wikipedia with English Wikipedia using LDA.

In evaluation section observations were gathered on both topic model clustering quality and topic coherence. This provides insight into how general topic models perform when dealing with short bilingual documents. Broadly translated our findings indicate that the result of the language-independent model depends on document characteristics such as the length of document length and word frequency. While, using Wikipedia interlanguage-link we can find 98 percent of similar articles between English and Amharic language.

In our experiment, we use a bilingual corpus extracted from English and Amharic Wikipedia database dump. Our experimental data is not a parallel corpus. Thus, we extracted bi-lingual texts from the collected English and Amharic articles based on Wikipedia inter-language links. And we try to improve topic extraction from the Amharic Wikipedia by using Wikipedia language links.

Finally, we evaluated the proposed bilingual topic models in terms of perplexity and accuracy of bilingual pair of text extracted from both English and Amharic Wikipedia article. Also we had calculated the precision and recall of the proposed framework and achieve better overall accuracy than previous methods.

6.2 Contribution

The contributions of this thesis are:

- We create a bi-lingual Corpus from Amharic and English Wikipedia Dump files using Wikipedia inter-language links.

- We develop a language-independent method that can be used to measure the similarity between Amharic Wikipedia and English Wikipedia articles based on LDA and Word Embeeding. We identify cross-lingual similarity between Amharic and English Wikipedia articles. We calculate textual similarity based on LDA and word Embedding models: We use LDA Model to infer document-topic distributions from a set of Wikipedia corpus and Word embedding to infer word-topic distributions from a set of Wikipedia corpus. The document-topic distribution provides an overview of what topics are found within each of the Wikipedia article, and in what proportion.
- We compare similarity between-Amharic and English Wikipedia articles using domain expert to check the accuracy of the proposed model.

6.3 Future Work

In this thesis work, we proposed a bilingual topic model based on word embedding to mine similar bilingual Wikipedia topics to address the cross-lingual entity linking issue. A tremendous amount of research and development focused on English Wikipedia. However, English Wikipedia is not representative of all Wikipedia languages. Future studies should consider other edition of Wikipedia articles. For instance, incorporate Amharic morphological analysis on our Amharic corpus extraction model. Also, in future we are planning to include structural similarity of document to enhance the quality of the cross-lingual linking algorithms. Other possible enhancement is to develop knowledge graph from the corpus, so that it can be useful for many NLP applications. Also, it is possible to include integrating trilingual entity linking and discovery task which requires to take three raw texts from three languages for instance English, Amharic and French as input. And it automatically extracts mentions and link them to English knowledge bases and cluster entity mentions across documents in different languages. Finally, we recommend the use our model in order to create an Amharic and English knowledge base, based on TAC Knowledge Base Population.

References

[1] Scientific American, “The Semantic Web,” 2001.

[2] Wikipedia, 6 March 2016. [online] Available: https://en.wikipedia.org/wiki/Semantic_Web.

[3] Tim Berners-Lee, Retrieved from W3C Design Issues, 27 07 2006. [online]Available: https://www.w3.org/DesignIssues/LinkedData.html.

[4] Lih Andrew, “resource, Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news,” vol. 3, No. 1, pp. 1-31, 2004.

[5] Wikipedia, the free encyclopedia, “List of Wikipedia,” Feb 2020. [online]Available: https://en.wikipedia.org/wiki/List_of_Wikipedias.

[6] Wikipedia, the free encyclopedia, “Size of Wikipedia,” May 2016. [online]Available: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[7] Wikipedia, the free encyclopedia, “አዲስ አበባ,” June 2016. [online]Available: https://am.wikipedia.org/wiki/አዲስ_አበባ

[8] Wikipedia, the free encyclopedia, “Addis Ababa,” June 2016. [online]Available: https://en.wikipedia.org/wiki/Addis_Ababa

[9] From Wikipedia, the free encyclopedia, “ጦፈር,” July 2016. [online]Available: https://am.wikipedia.org/wiki/ጦፈር.

[10] From Wikipedia, the free encyclopedia, “ጎዋ,” July 2016. [online]Available: https://am.wikipedia.org/wiki/ጎዋ.

[11] From Wikipedia, the free encyclopedia, “Universe,” 2016. [online]Available: https://en.wikipedia.org/wiki/Universe.

[12] From Wikipedia, the free encyclopedia, “List of cities and towns in Ethiopia,” July 2016. [online] Available: https://en.wikipedia.org/wiki/List_of_cities_and_towns_in_Ethiopia.

[13] From Wikipedia, the free encyclopedia, “Liste des villes d'Éthiopie,” July 2016. [online]Available: https://fr.wikipedia.org/wiki/Liste_des_villes_d'Éthiopie.

[14] From Wikipedia, the free encyclopedia, “መደብ:የኢትዮጵያ ከተሞች,” July 2016. [online]Available: https://am.wikipedia.org/wiki/መደብ: የኢትዮጵያ_ከተሞች.

[15] From Wikimedia, the free encyclopedia, “Brucointestino,” Feb 2020. [online]Available: https://meta.m.wikimedia.org/wiki/User:Brucointestino

[16] Göbölös-Szabó J, Prytkova N, Spaniol M, Weikum G., “Cross-Lingual Data Quality for Knowledge Base Acceleration across Wikipedia Editions,” In Proceedings of the 10th International Workshop on Quality in Databases (QDB 2012), 2012.

[17] Adida, Ben, and Mark Birbeck, “RDFa Primer, Bridging the Human and Data Webs. W3C Working Group Note, W3C (2008).,” In W3C , 2008.

[18] From Wikipedia, the free encyclopedia, “List of Wikipedias,” May 2016. [online]Available: http://meta.wikimedia.org/wiki/List_of_Wikipedias.

[19] Miao, Qingliang, Huayu Lu, Shu Zhang, and Yao Meng, “Cross-lingual link discovery between Chinese and English wiki knowledge bases,” In Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27), 2013.

[20] Paramita, Monica Lestari. "Language-Independent Methods for Identifying, "Cross-Lingual Similarity in Wikipedia." Doctoral dissertation., University of Sheffield, 2019.

[21] Otero-Cerdeira, Lorena, Francisco J. Rodríguez-Martínez, and Alma Gómez-Rodríguez, “Ontology matching: A literature review,” Expert Systems with Applications, vol. 42, No. 2, pp. 949-971, 2015.

[22] From Wikipedia, the free encyclopedia, “Wikipedia:Portal,” [online]Available: https://en.wikipedia.org/wiki/Wikipedia:Portal. [In Proc 18 September 2019].

[23] Sorg, Philipp, and Philipp Cimiano, “Enriching the Cross- lingual Link Structure of Wikipedia – A Classification-Based Approach,” In Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence, 2008.

[24] Ji, Heng, Joel Nothman, Ben Hachey, and Radu Florian., “Overview of TAC-KBP2015 Tri-lingual Entity Discovery and Linking,” In TAC, 2015.

[25] Štajner, Tadej, and Dunja Mladenić., “Cross-lingual document similarity estimation and dictionary generation with comparable corpora,” Knowledge and Information Systems 58, vol. 58, No. 3, pp. 729-743, 2019.

[26] Gracia, Jorge, Elena Montiel-Ponsoda, and Asunción Gómez-Pérez, Cross-lingual Linking on the Multilingual Web of Data, vol. 936, Boston (USA): ISWC 2012, 2012.

[27] HOSSAIN, JAFREEN, NOR FAZLIDA MOHD SANI, LILLY SURIANI AFFENDEY, ISKANDAR ISHAK, and KHAIRUL AZHAR KASMIRAN, “SEMANTIC SCHEMA MATCHING APPROACHES: A REVIEW,” Journal of Theoretical & Applied Information Technology, vol. 62, No. 1, 2014.

[28] Abubakar, Mansir, Hazlina Hamdan, Norwati Mustapha, and Teh Noranis Mohd Aris, “Instance-based ontology matching: a literature review,” In International Conference on Soft Computing and Data Mining, Johor, Malaysia, 2018.

[29] Yamada, Ikuya, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji, “Using encyclopedic knowledge for named entity disambiguation,” In11th conference of the European Chapter of the Association for Computational Linguistics, 2006.

[30] Mihalcea, Rada, and Andras Csomai.., “Wikify!: linking documents to encyclopedic knowledge,” In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM’07, 2007.

[31] Ratinov, Lev, Dan Roth, Doug Downey, and Mike Anderson, “Local and global algorithms for disambiguation to wikipedia,” InProc. of the Annual Meeting of the Association of Computational Linguistics (ACL), 2011.

[32] Tsai, Chen-Tse, and Dan Roth, “Cross-lingual Wikification Using Multilingual Embeddings,” In Conference: Proceedings of the 2016 Conference of the North

American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[33] Huang, Wei Che Darren, Shlomo Geva, and Andrew Trotman., "Overview of the INEX 2009 Link the Wiki Track," In Proceedings of INEX 2009, 2010.

[33] Tang, Ling-Xiang, Shlomo Geva, Andrew Trotman, Yue Xu, and Kelly Y. Itakura., "Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery," In Proceedings of NTCIR-9, 2011.

[34] Chen, Liang-Pu, Yu-Lun Shih, Chien-Ting Chen, Tsun Ku, Wen-Tai Hsieh, Hung-Sheng Chiu, and Ren-Dar Yang, "English-to-Traditional Chinese Cross-lingual Link Discovery in Articles with Wikipedia Corpus," In Proceedings of the Twenty-Fourth Conference on Computational Linguistics and Speech Processing, 2012.

[35] Geva, Shlomo. "Gpx: Ad-hoc queries and automated link discovery in the wikipedia." In International Workshop of the Initiative for the Evaluation of XML Retrieval, pp. 404-416. Springer, Berlin, Heidelberg, 2007.

[36] Geva, Shlomo., "Ad-Hoc Queries and Automated Link Discovery in the Wikipedia," In Proceedings of INEX 2007, 2008.

[37] Tang, Ling-Xiang, Andrew Trotman, Shlomo Geva, and Yue Xu., "Cross-Lingual Knowledge Discovery: Chinese-to English Article Linking in Wikipedia," In Information Retrieval Technology, 2012.

[38] Blei, David M., Andrew Y. Ng, and Michael I. Jordan., "Latent dirichlet allocation," Journal of machine learning research, pp. 993-1022, 3(Jan) 2003.

[39] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng " A biterm topic model for short texts," In Proceedings of the 22nd international conference on World Wide Web (pp. 1445-1456). ACM., May 2013.

[40] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng, "Btm: Topic modeling over short texts," IEEE Transactions on Knowledge & Data Engineering, 2014.

[41] Shi, Tianze, Zhiyuan Liu, Yang Liu, and Maosong Sun, “Learning Cross-lingual Word Embeddings via Matrix Co-factorization,” In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, January 2015.

[42] Ximing Li, Ang Zhang, Changchun Li, Lantian Guo,” Relational Biterm Topic Model: Short-Text Topic Modeling using Word Embeddings”, March 2019

[43] Søgaard, Anders, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen, “Inverted indexing for cross-lingual NLP,” In The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP 2015), 2015.

[44] Hofmann, Thomas., “Probabilistic Latent Semantic Analysis,” Jan 2013.

[45] Koehn, Philipp., “Europarl: A parallel corpus for statistical machine translation,” In MT summit (Vol. 5, pp. 79-86), 2005.

[46] Kolda, Tamara G., and Brett W. Bader., “Tensor decompositions and applications,” SIAM review, vol. 51, No. 3, pp. 455-500, 2009.

[47] Gauvin, Laetitia, André Panisson, and Ciro Cattuto, “Detecting the community structure and activity patterns of temporal networks: a non-negative tensor factorization approach,” PloS one, vol. 9, No. 1, 2014.

[48] Sapienza, Anna, André Panisson, J. T. K. Wu, Laetitia Gauvin, and Ciro Cattuto, “Anomaly Detection in Temporal Graph Data: An Iterative Tensor Decomposition and Masking Approach,” InAALTD@ PKDD/ECML, September, 2015.

[49] Zhang, Lefei, Liangchen Song, Bo Du, and Yipeng Zhang., “Nonlocal Low-Rank Tensor Completion for Visual Data,” In IEEE transactions on cybernetics, 2019.

[50] Parag Singla , Pedro Domingos, “Entity resolution with markov logic In Data Mining,” In ICDM'06. Sixth International Conference, December, 2006.

[51] Beyene, Melkamu, Pierre-Edouard Portier, Solomon Atnafu, and Sylvie Calabretto., "Tensor factorization for cross lingual entity co-reference resolution in the linked open data," In Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems (pp. 17-23). ACM., October, 2015.

[52] Nickel, Maximilian, Volker Tresp, and Hans-Peter Kriegel, "Factorizing yago: scalable machine learning for linked data," In Proceedings of the 21st international conference on World Wide Web ACM, April, 2012.

[53] Nikolaus, Ralf, "Learning the parts of objects by non-negative matrix factorization," 1999.

[54] Evangelos E. Papalexakis Christos Faloutsos Nicholas D. Sidiropoulos, "Parcube: Sparse parallelizable tensor decompositions," In Joint European Conference on Machine learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, September, 2012.

[55] Zhang, Tao, Kang Liu, and Jun Zhao, "Cross lingual entity linking with bilingual topic model," In Twenty-Third International Joint Conference on Artificial Intelligence, 2013, June.

[56] Balikas, Georgios, "Mining and learning from multilingual text collections using topic models and word embeddings," Doctoral dissertation, 2017.

[57] Heyman, Geert, Ivan Vulić, and Marie-Francine Moens, "C-BiLDA extracting cross-lingual topics from non-parallel texts by distinguishing shared from unshared content," Data Mining and Knowledge Discovery, vol. 30, No. 5, pp. 1299-1323, 2016.

[58] Wintrode, Jonathan, and Sanjeev Khudanpur, "Limited resource term detection for effective topic identification of speech," In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2014.

[59] Liu, Chunxi, Matthew Wiesner, Shinji Watanabe, Craig Harman, Jan Trmal, Najim Dehak, and Sanjeev Khudanpur, "Low-Resource Contextual Topic Identification on

Speech," In2018 IEEE Spoken Language Technology Workshop (SLT), IEEEx, December 2018.

[60] Melkamu Beyene, Pierre-Edouard Portier, Solomon Atnafu, Sylvie Calabretto. (2016). "Dataset linking in a multilingual linked open data context", In Proc. 8th International Conference on Management of Digital EcoSystems.

[61] Wikimedia, "Wikimedia Downloads". 6 March 2007. [online] Available: https://dumps.wikimedia.org /.

[62] From NLTK, "Natural Language Toolkit," Feb 2020. [online]Available: https://www.nltk.org/

[63] "MediaWiki," Wikimedia Foundation, Augest 2019. [online]Available: https://www.mediawiki.org › wiki › MediaWiki.

Annexes

Annex A: Source code of the Word Cloud

A wordcloud is used to illustrate the main topics on Bi-Lingual Corpus.

In [1]:

```
# modules for generating the word cloud

from os import path, getcwd

from PIL import Image

import numpy as np

import matplotlib.pyplot as plt

from wordcloud import WordCloud, ImageColorGenerator
```

In [2]:

```
# modules for generating the word cloud

from os import path, getcwd

from PIL import Image

import numpy as np

import matplotlib.pyplot as plt

from wordcloud import WordCloud, ImageColorGenerator
```

In [3]:

```
# Read the whole text

text = open(path.join(d, 'constitution.txt')).read()

wordcloud = WordCloud(background_color="white",width=1600, height=800).generate(text)
```

```
# Open a plot of the generated image.

plt.figure( figsize=(20,10), facecolor='k')

plt.imshow(wordcloud)

plt.axis("off")

#plt.tight_layout(pad=0)

#plt.show()

plt.savefig('wordcloud.png', facecolor='k', bbox_inches='tight')
```

DECLARATION

I, the undersigned, declare that this thesis is my original work and has not been presented for a degree in any other university, and that all source of materials used for the thesis have been duly acknowledged.

Declared by:

Name: _Tehetena Alemu Nega_____

Signature: _____

Date: _____

Confirmed by advisor:

Name: _Solomon Atnafu (PhD)_____

Signature: _____

Date: _____